AGARD-LS-185

AGARD-LS-185

AGARD

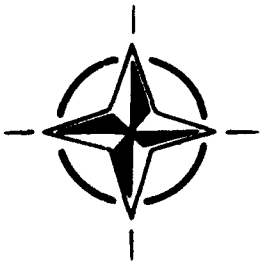
ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT
7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AGARD LECTURE SERIES 185

Machine Perception

(La Perception de l'Environnement par Senseurs Automatiques)

This material in this publication was assembled to support a Lecture Series under the sponsorship of the Guidance and Control Panel of AGARD and the Consultant and Exchange Programme of AGARD presented on 3rd—4th September 1992 in Hampton, VA, United States, 14th—15th September 1992 in Neubiberg, Germany and 17th—18th September 1992 in Madrid, Spain.



NORTH ATLANTIC TREATY ORGANIZATION

Published August 1992

Distribution and Availability on Back Cover

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AGARD LECTURE SERIES 185

Machine Perception

(La Perception de l'Environnement par Senseurs Automatiques)

A1

This material in this publication was assembled to support a Lecture Series under the sponsorship of the Guidance and Control Panel of AGARD and the Consultant and Exchange Programme of AGARD presented on 3rd—4th September 1992 in Hampton, VA, United States, 14th—15th September 1992 in Neubiberg, Germany and 17th—18th September 1992 in Madrid, Spain.

North Atlantic Treaty Organization
Organisation du Traité de l'Atlantique Nord

The Mission of AGARD

According to its Charter, the mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community;
- Providing scientific and technical advice and assistance to the Military Committee in the field of aerospace research and development (with particular regard to its military application);
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Exchange of scientific and technical information;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

Published August 1992

ISBN 92-835-0684-7

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

INTRODUCTION

Reiner Onken

Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
8014 Neubiberg
Germany

The objective of this Lecture Series is to present both the basic ideas and approaches of machine perception, here for vision and speech understanding, and a number of related applications, in particular for guidance and control.

Machine perception has become a topic of increased interest to the guidance and control community since the capability of autonomous process management and control is in reach in many fields including aerospace guidance and control. A great number of demonstration programs have been conducted worldwide and many new ones are underway, encouraged by the advent of more and more powerful computational architectures and performance.

This can be viewed as one of the major technology push impacts to guidance and control. With increased awareness of the potentials of these techniques exploitation in applications is demanded which will trigger the requirement pull process with the effect of intensifying the application-oriented research and development on this field.

To a great extent, the basic approach to machine perception in vision and speech recognition and understanding is developed upon what is known from animals and human perceptual mechanisms. Although the human perceptual capabilities are by far not reached at the time being, the pace of progress is amazing and there are even aspects in machine perception where the human capabilities are surpassed by the machine.

The task of vision, for example, whether for brains or for machines, is to extract useful information from light in a way to infer relevant properties of visible objects, i.e. their light reflectances, the individual or the machine needs to interact with in the world about it. One has identified in the brains of various creatures structures specialised for this kind of goal-oriented job.

There is the understanding in process control that pursuing certain preestablished goals requires situational knowledge, possibly the generation of a goal-oriented plan and certainly its execution. This, in turn, cannot be achieved satisfyingly without perception, including a structure of anticipation. This knowledge structure of the socalled perception-action cycle , where the gained information is to be embedded and represented, is often referred to as 'situation representation'. For all systems known so far, including the human brain, the situation representation has to comply with requirements for computational efficiency. Information compression and condensation has to be achieved for efficient handling of the knowledge (like content adressability), and the information being kept ready should be as complete and detailed as possible with secure information retrieval capability.

The brain structures are representing more or less only one common design decision in terms of a kind of trade off solution under the given biological constraints. As the machine can be diversified in architecture, complying to the different application requirements, the machine might be more flexible through the combination of complementary, dissimilar solutions serving the different performance aspects. This kind of representation in machine perception could, in principal, be more complete and more detailed, and could therefore avoid mismatches and illusionary effects, for instance, humans are suffering from. This can be taken as a promising perspective, although, since a comprehensive representation would be much more complex, less easily manageable and considerably larger in size, computational limitations still are prevailing.

Airborne missions have become more complex and stressful to the pilot. Scenarios now require threat avoidance, rapid replanning and reconfiguration of navigation modes in the presence of electronic warfare like jamming of navigation aids such as GPS, management of electromagnetic energy emissions in heavily defended areas, and continuous monitoring of avionics system status in terms of fault detection and isolation and fault tolerant reconfiguration. That is the scene, activating the requirement pull process, looking for diagnostic and decision-making functions being performed autonomously.

In airborne guidance and control both completely autonomous process control and autonomous knowledge-based assistance for the pilot in process control are of prime interest, including autonomous situation assessment, planning with decision-making and problem solving and execution services.

The lectures start on the first day with machine perception of speech, its recognition and understanding (Mangold). This perceptual task is very essential for operator (crew) assistance in order to offer natural communication means human individuals are used to. The source of information to be perceived is the human being himself. Speech production is based on the specific sound generation which is possible using the articulatory organs. Man has developed very special decoding and understanding mechanisms to extract from the speech signal all the information.

The remaining part of the lectures are exclusively devoted to vision, starting with approaches for sensing and interpretation of 3D shape and motion (Kanade) and elementary functions to be implemented on an electronic retina (Zavidovique).

The capabilities and performance of vision systems using monocular stereo, and image sequence analysis with pixel and feature processing will be discussed in the third lecture (Baker), as will their respective utilities to vision-based autonomous guidance. The principal focus will be on the relationship between optic flow technique for image pair analysis of motion and depth and spatio-temporal manifold analysis.

The second day is more application-oriented. It starts with a lecture on 3D vision application for navigation and control of mobile robots (Garibotto). This contribution describes a binocular stereo vision module for obstacle detection with no precise calibration at fast rate, a trinocular stereo vision based on segment primitives for the reconstruction of free space for navigation, and landmark detection for self-positioning and orientation of the mobile vehicle.

The following contribution adresses image sequence understanding with application examples like road vehicle guidance with obstacle avoidance, vehicle docking and aircraft landing approach guidance (Dickmanns). High-level spatio-temporal models of the processes of interest in the real world are exploited for automatic feature tracking. Other properties like feature grouping through 'Gestalt'idea, fixation-type vision, feature adaptation to the actual shape and feature selection in a situation context are incorporated in this approach.

The last lecture considers two scenarios of the application of 3D computer vision using passive imaging sensors (Evans). First, a general scene is analysed without any prior information concerning its structure. This would be the case when wishing to control, for example, a vehicle moving off-road across unknown terrain. Secondly, in the converse case the motion is analysed of a well defined object, for example when tracking a known aircraft. A review of techniques used will be presented followed by further description of particular systems.

The lecturers come from several of the participating AGARD countries, specifically France, Germany, Italy, the United Kingdom and the United States. There are seven lectures followed by a round table discussion at the end of the second day.

Abstract

Human perceptual capabilities involve the extraction of task-oriented information from environmental stimuli through physical sensing and the use of background knowledge.

There are many activities underway aimed at providing similar capabilities of artificial machine perception. Some success is achieved by exploiting what is known of corresponding human cognitive processes and by making use of the increasing power of information processing techniques. For this purpose, the recognition of sharply contrasted as well as fuzzy patterns (stationary or dynamically changing) plays an important role along with other aspects of processing of complex information structures.

These techniques are beginning to be applied in guidance and control, in particular with regard to artificial visual perception and speech understanding. This application promises major benefits with the advent of autonomous vehicle and mission control, and of intelligent systems for situation awareness support of human operators.

This Lecture Series covers the following subjects:

- Pattern recognition techniques
- Real time visual machine perception, principles and applications in G&C
- Real time speech recognition and understanding in the G&C domain.

This Lecture Series, sponsored by the Guidance and Control Panel of AGARD, has been implemented by the Consultant and Exchange Programme.

Abrégé

Les capacités de perception humaines permettent l'extraction de données orientées-tâches des stimuli du milieu environnant par le biais de la détection physique et par l'application de connaissances préalables.

Un grand nombre d'activités sont entreprises à l'heure actuelle, dans le but de créer des capacités similaires de perception artificielle machine. Un certain progrès est réalisable en exploitant les processus cognitifs humains connus et en se servant de la puissance de calcul grandissante des techniques de traitement des données. Dans ce contexte, la reconnaissance d'images à contrast marqué, ainsi que de motifs flous (stationnaires ou en évolution dynamique) joue un rôle important, conjointement avec d'autres aspects du traitement des structures de données complexes.

Ces techniques commencent à trouver des applications dans le domaine du guidance et du pilotage, en particulier en ce qui concerne la perception visuelle et la reconnaissance de la parole. Cette dernière application doit donner de bons résultats avec l'arrivée du contrôle autonome des véhicules et des missions et de systèmes intelligents d'aide à la perception de la situation.

Ce cycle de conférences portera sur les sujets suivants:

- les techniques de reconnaissance de motifs
- la perception visuelle machine en temps réel, principes et applications dans le domaine du guidage et du pilotage
- la reconnaissance et la comprehension de la parole, aspects guidage et pilotage.

Ce cycle de conférences est présenté par le Panel AGARD de Guidage et de Pilotage; et organisé dans le cadre du programme des Consultants et des Echanges.

List of Authors/Speakers

Lecture Series Director: Prof. Dr R. Onken
Universität der Bundeswehr München
Fakultät für Luft- und
Raumfahrttechnik
Institut für Systemdynamik und
Flugmechanik
8014 Neubiberg
Germany

AUTHORS/SPEAKERS

Dr H.H. Baker
SRI International, EK 233
333 Ravenswood Ave.
Menlo Park, CA 94025
United States

Mr T. Bernard
Paris-Sud University
Institut d'Electronique Fondamentale
91405 Orsay
France

Prof. Dr E.D. Dickmanns
Universität der Bw München
Werner-Heisenberg-Weg 39
8014 Neubiberg
Germany

Mr R. Evans
Roke Manor Research Ltd
Roke Manor, Romsey
Hampshire SO51 0ZN
United Kingdom

Dr G. Garibotto
ELSAG Central Research Dept
Via Puccini, 2
16154 Genova
Italy

Prof. T. Kanade
Computer Sciences Dept
Carnegie-Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
United States

Mr H. Mangold
Daimler-Benz AG, Forschungsinstitut
Wilhelm-Runge-Str. 11
7900 Ulm
Germany

Dr S. Masciangelo
ELSAG Central Research Dept
Via Puccini, 2
16154 Genova
Italy

Prof. B. Zavidovique
Paris-Sud University
Institut d'Electronique Fondamentale
Bât 220
91405 Orsay
France

Contents

Perception Based and Machine-Oriented Signal Processing Within Speech Understanding Systems

Helmut Mangold
Daimler-Benz, Research Center Ulm
Institute for Information Technology
7900 Ulm, Germany

Summary

Automatic recognition and understanding of speech signals is one of the key issues of advanced information technology. Language and speech are the relevant topics of cognition and therefore to understand spoken and written language offers basic capabilities for universal processing of information.

Speech is man's generic communication medium. Information transfer is widely done by speech communication between humans. There is a basic commonality of understanding each other's spoken messages. This common understanding must be the basic of machine understanding too.

Automatic recognition and understanding of spoken language is done in a multistep approach, which starts with the low level signal processing. The output of the recognition step is word recognition. Many possible words, the so called word hypotheses are the basis for intensive linguistic processing.

Linguistic processing cares for syntactic analysis and semantic analysis. The semantic analysis needs again many additional parameters from spoken language, like intonation and prosody to derive the meaning of a spoken phrase.

All the processing of natural speech is narrowly related to human information processing. It is therefore possible to learn much from our human processing or from models of this processing. On the other side statistical methods of information processing offer rather systematic and in many cases advanced methods for handling much of the information contained in speech using purely statistic approaches. To estimate the advantages of the more statistical approaches or more rule based approaches will be a great challenge for future research. Human perception will always be a guide how to process speech with machines.

1. Speech - Man's Tool for Communication

Speech as man's generic communication medium is fully adapted to the capabilities of the human individual. Speech production is based on the specific method of sound generation which is possible using the articulatory organs and, on the other side, perception is based on very special methods to extract all the relevant information from the speech signal, which is encoded through the time- and frequency characteristics of this signal.

But this level of signal processing is only a very small part of the human processes which are involved if we produce and perceive speech. It has become rather common to call the speech signal as spoken language. This terminology shows clearer that many scientific areas are contributing to these processes and have therefore to be addressed if we want to compare human speech perception and machine perception of spoken language. It is quite clear that due to the inherent adaptation between speech production and speech perception a good understanding of the generative processes necessary to produce speech signals may be helpful for designing and understanding all the methods which are relevant for machine perception of speech, and that of course a deep understanding of human speech perception may be helpful too.

This multilevel process of speech perception and understanding ranges from low-level signal processing up to high level cognitive processes. Speech signals are our natural tool for human information transfer and, far beyond this, speech and language are the basis of nearly all our cognitive processes. We shall therefore have to care about signal processing, parameter extraction, phonetic coding, linguistic structuring and analyzing, and finally about all the cognitive processes which we include in realizing natural language dialogues.

2. The Speech Signal

2.1 Signal Characteristics Based on the Natural Production Process

In a communication theoretic based view of the speech signal we may interpret it as a complex coded signal which includes different sorts of information that are coded in very specific manners. This may be easily understood if we look at the natural speech production process.

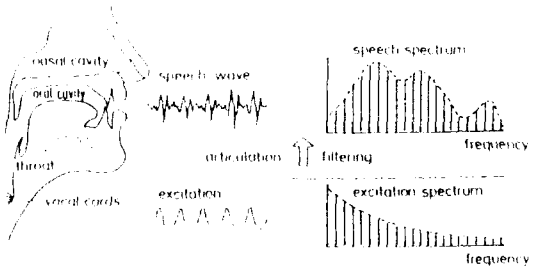


Fig.2.1: Principle of natural speech production (voiced sounds).

From Fig.2.1 we may see that the natural articulation system first produces an excitation signal resulting from the larynx for voiced sounds like vowels, and a noise signal for unvoiced sounds like the fricatives. This excitation signal covers a broad

spectral range. It consists of a collection of many harmonic frequencies in the case of the voiced excitation signal and of a noise spectrum in the unvoiced case. The basic pitch frequency distinguishes male and female voices and gives a good deal of the information which is relevant for natural intonation and for the prosodic part of the speech signal. For male voices this basic frequency is centered at around 100 Hz, for female voices it is about twice this value at around 200 Hz.

The actual sound information is modulated on this basic excitation spectrum. The envelope of the speech spectrum carries through its spectral resonance characteristics, the formants, the information about different sounds. So, we have mainly two parts in every speech signal, the excitation, which carries much of the prosodic information and the short term spectral envelope, which is representing the phonemic quality.

This short term spectral envelope is permanently changed through the process of articulation. This has led to a vivid optical representation of speech signals as three-dimensional spectrograms, called sonagrams. Such a sonagram of the German word "lesen" is shown in Fig.2.2.

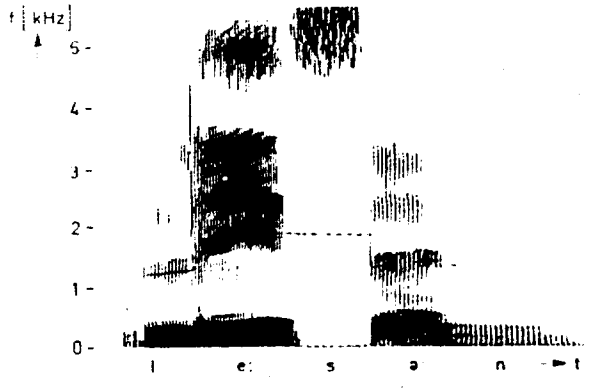


Fig.2.2: Sonagram of the German word "lesen" with indication of the second formant.

The horizontal axis represents the time scale, the vertical axis the frequency scale. The energy of the different frequencies is represented through the darkness. The darkest areas represent the formants, which are the resonances of the vocal tract and which represent different sounds. This means that the most important information is represented by these formants.

The course of the second formant is manually drawn into the sonagram. The position of this formant is continuously changing as the sounds change during the articulation. Such a sonagram seems to be rather easily readable and some attemps have been undertaken to use spectrograms as another representation of speech, e.g. for deaf people, but in practice spectrogram reading needs extensive training and even then it is not possible to do it in realtime. This means finally that optical perception of relevant speech information is practically not possible. But our natural speech perception system is based on spectral analysis and higher level parametrical analysis of a similar manner.

2.2 Natural Decoding of Speech Signal Information

The decoding of the information contained in the speech signal is done in a multilevel process. The primary processing is done within the different parts of our external and internal ear. The sensitivity range of the ear is extremely high. Its lower limit is given by the noise produced through hydrogen molecules in the air. The whole range reaches up to 120 dB. This huge range is necessary to guarantee that the ear can perceive every sound or noise which is practically possible.

Fig.2.3 gives a schematic overview about the primary organ. The middle ear is mainly responsible for a resistance adaptation of the resistance of the air to the resistance of the liquid within the inner ear. This inner part of the ear consists of a spiral tube which is separated into two parts through the basilar membrane. This carries around tenthousand sensors to measure the movement of this membrane. The membrane itself realizes a sort of mechanical short-time frequency analysis, producing nothing else than a spectral pattern like that in Fig.2.2.

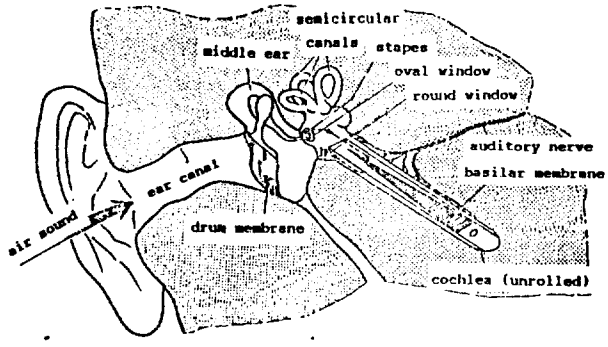


Fig.2.3: Schematic drawing of the structure of the inner ear with the cochlear tube stretched from spiral form to a linear form for clearness.

The endings of the auditory nerve are directly processing the signal from the basilar sensors. The auditory nerves do not only transmit the pulse frequency coded signal, but through intensive interaction of neighbouring nerves many enhancements of the spectral resolution are realized. In physics we have the basic principle that the product of spectral and time resolution in spectral analysis is constant. This means that always a better spectral resolution requires worse time resolution and vice versa. The mechanical spectral analyzer of the basilar membrane underlies of course the same rules. Only the very specific processing afterwards cares for a much better spectral and time resolution than might be possible through the mechanical analysis alone.

We have already seen that the dynamic range of our hearing covers around 120 dB in signal energy. This loudness sensitivity is nearly logarithmic, i.e. already the hearing cells on the basilar membrane have such an inherent logarithmic sensitivity. The spectral sensitivity is not uniform over the whole hearing range from around 16 Hz up to near to 20 kHz. Fig.2.4 shows the frequency dependent amplitude sensitivity of the ear which peaks in the 1 to 2 kHz range. Especially in this frequency range there is normally the important second formant of the different sounds, which is responsible for distinguishing many sounds from each other. Already a long time ago psychoacoustic experiments have shown that the transmission of the frequency range between around 800 Hz and 2 kHz is sufficient for getting a certain basic intelligibility (Zwi67).

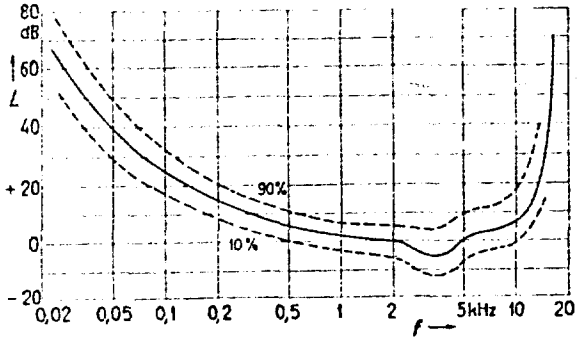


Fig.2.4: Frequency dependent amplitude sensitivity of human hearing.

A very important aspect of differentiating one spectral pattern from another one is frequency selectivity. This is usually measured by psychoacoustic experiments asking test listeners to detect small changes in the frequency of test tones. This leads to a perceptual frequency scale, which is constant over the first few hundred Hertz and which then decreases with increasing frequency. This degradation of the frequency resolution at higher frequencies is combined with improvement on temporal resolution at these higher frequencies. This fact is well adapted to the characteristics of the speech sounds themselves. The higher formants have usually higher bandwidth and it is therefore not necessary to analyse their mid frequencies as precise as for the lower formants. On the other side for sounds where the spectral energy is concentrated on higher frequencies like voiceless plosives, spectral changes are happening much faster than e.g. for vowels. Voiced sounds require therefore good spectral resolution, while voiceless sounds need good time resolution.

Combined with this varying spectral resolution is the spectral discrimination of neighbouring frequencies. It is highly amplitude dependent. This means that a frequency near to another one cannot be discriminated from the first if it does not reach a certain amplitude. Our hearing capabilities have a sort of band structure, where all frequencies which are near to each other are weighted with a bandfilter characteristic defined through the maximal frequency energy within this band. Fig.2.5 shows these bandfilter characteristics which are based on the one side on the nonlinear frequency sensitivity along the mel-scale and on the other side on the spectral masking which is done in the low level nervous processing (Pie85).

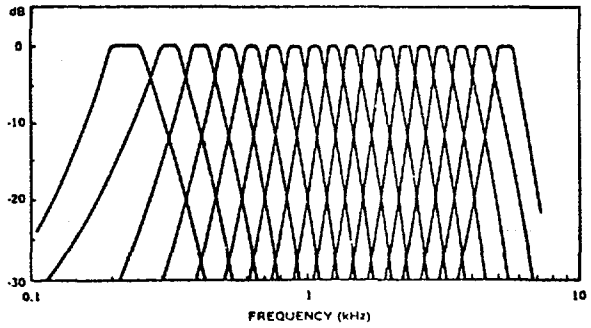


Fig.2.5: Frequency characteristic of 18 channels of a mel-scale based filter system as used for automatic speech recognition (similar to the filtering in the human auditory system).

The whole frequency scale is covered by 24 such frequency bands. Their bandwidths are highly different depending on the mel-scale. As we can see from the figure, where the frequency scale is logarithmic, such frequency masking works mainly upwards to higher frequencies.

Besides this spectral masking, we can also experience a time-dependent temporal masking. Such forward or backward masking is produced by stronger components coming before or after a weaker component.

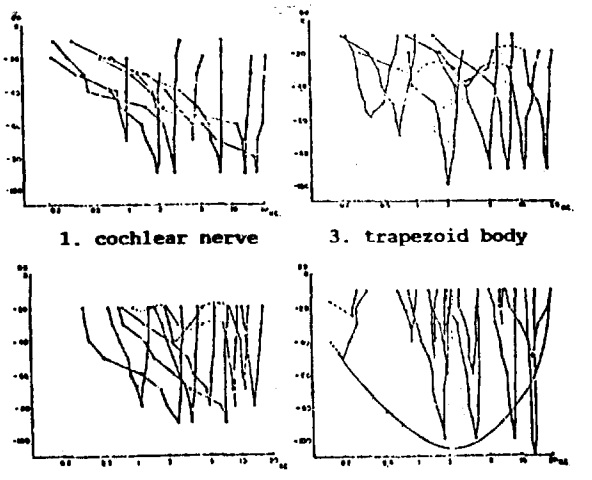


Fig.2.6: Enhancement of spectral selectivity on different positions of the auditory nerve apart from the basilar membrane.

The general idea of all these effects is to strengthen the strong components in the signal. This again is necessary to care for a good robustness of our human speech recognition process. Measurements in the lower level auditory nerves have shown this too, where the formants are systematically enhanced in the run of the nerve from the auditory cells. Fig.2.6 shows some spectral characteristics measured on auditory nerves on different positions from the auditory cells. the top left image shows spectral sensitivity of the cochlea itself for some few tones. The second image and the further images stem from nerves in the lower level of the brain, measured within the acoustic nerve. We can very clearly see, that the spectral sensitivity is more and more enhanced.

2.3 Robustness of the Decoding Process

Of course all the speech decoding done in the human perception process is not only based on the signal processing described. It includes much higher level processing, but many of the processing steps are already responsible for the high level of robustness which is possible in the human decoding process. We shall later see, that this robustness is by far better than the robustness we can today realize with machine recognition of speech.

Robustness concerns many aspects of speech perception, like

* wide dynamic range,
* tolerance against background noise,
* recognition of a large variety of different voices, dialects etc.,
* tolerance against spectral changes,
* high recognition rate even with badly articulated speech signals,
* resistance against nonlinear distortion.

Fig.2.7 gives an example for such a parameter dependency. Here the intelligibility for meaningless syllables is shown depending from the boarder frequency of a highpass and a lowpass filter for different speech levels. We can see that even with very small bandwidth there is still a good intelligibility of such meaningless syllables possible.

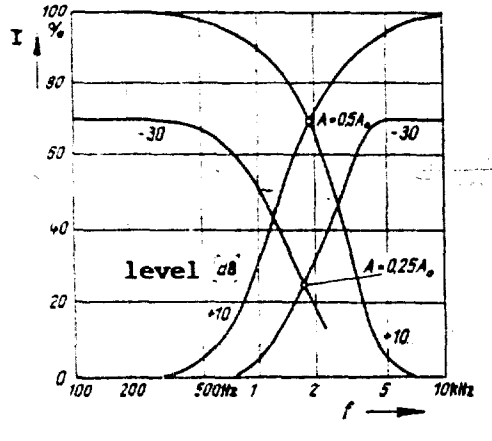


Fig.2.7: Intelligibility of meaningless syllables (logatomes) depending on the boarder frequency of a lowpass and a highpass filter.

Very interesting again is the fact that both curves have their crossover at around 2 kHz, the frequency where already in Fig.2.4 we have seen the highest auditory sensitivity.

3. Machine Recognition of Speech - Pattern Recognition

3.1 Structure of Word Recognition

Most today available speech recognizers are word recognizers, which are based on pattern recognition of spectral patterns like that in Fig.2.2. The basic structure of such a word recognizer is shown in Fig.3.1.

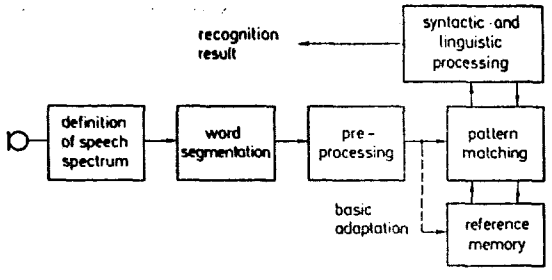


Fig.3.1: Basic structure of a word recognition system.

First the speech spectrum is continuously measured. Besides the static spectrum dynamic parameters like changes in the spectrum are measured too. In the last few years the usage of a mel-spectrum based analysis has proven to deliver optimal recognition results. Besides this approach there are still adaptive spectral filtering procedures used, where the spectral envelope is approximated through least squares approximation. This technique which is called linear predictive coding LPC gives a rather good approximation too (Ma76). Like the perception based approach this offers the possibility to make a detailed analysis of the spectral characteristics in a flexible manner. Fig.3.2 shows such an LPC-based spectral approximation for different degrees of the approximating filter.

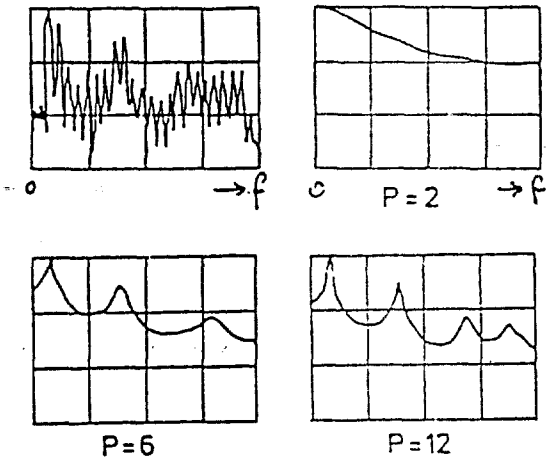


Fig.3.2: LPC-Analysis of a speech spectrum using different degrees of filtering. Upper left: speech spectrum

Using such a method for spectral estimation we get a spectral pattern for further processing like that in Fig.3.3, where we have shown a spectral pattern for the spoken word "They". Here we can clearly see, how the changing formants of the speech spectrum are modelled.

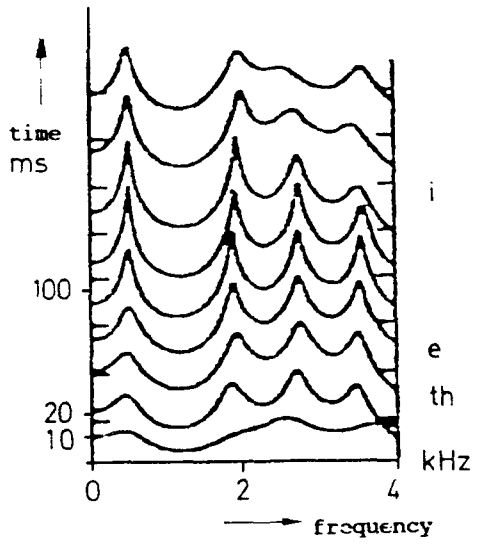


Fig.3.3: LPC-Spectrum of the word "They".

3.2 The Pattern Recognition Process

After the primary parameter definition some normalization stages are usually important for temporal and energy normalization. Through this processing it is possible to widen the dynamic range of the system. But it is of course possible too to include here some normalization which goes far beyond such rather simple procedures. This concerns mainly the normalization of different speakers' voices, to get a true speaker independent recognition.

Such a speaker adaptation is first done for the spectral parameters which define the specific voice sound of different speakers. One approximation may be used to adapt female and male voices to each other. But it is not yet possible to adapt all the dynamic variations of different speakers to each other. This will still be a topic for basic research. Some primitive approximations to this problem are already included in some existing word recognizers using a linear or a nonlinear time normalization of the varying speed of articulation.

Another important aspect of preprocessing is the enhancement of noise robustness. Due to many levels of perception our human perception of speech is highly robust against environmental noise. Fig.3.4 compares the capabilities of human perception and todays existing speech recognizers. We can see that existing word recognizers are still at least 10 dB away from the SNR which people can tolerate.

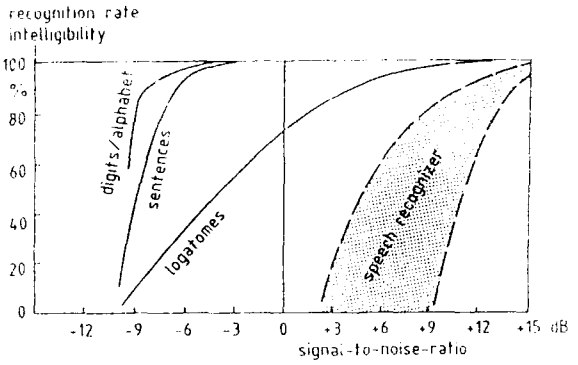


Fig.3.4: Human and machine recognition of speech under noisy conditions.

Especially the recognition of sentences uses a high degree of redundancy, while the good results of human digit recognition comes from the few numbers of possibilities to be distinguished.

The classification stage itself makes a more or less sophisticated comparison of a sort of reference pattern and the new pattern to be classified. The reference pattern is usually defined during the training process. For this training a user or many users have to utter every word to be recognized or at least some representative words for the vocabulary to be recognized. The system then stores this word patterns or special representations of the information contained within these patterns.

As shown in Fig.3.5 every classification makes a measurement of distances between a reference pattern and the new pattern.

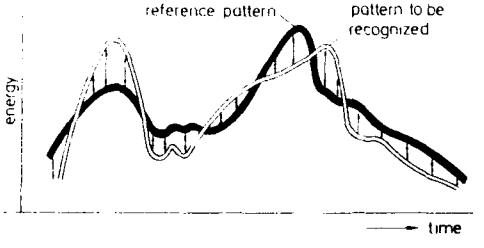


Fig.3.5: Pattern classfication through distance measurement.

Often the distance measurement includes some normalization procedures like in the dynamic time warp approach. The principle of this approach is shown in Fig.3.6.

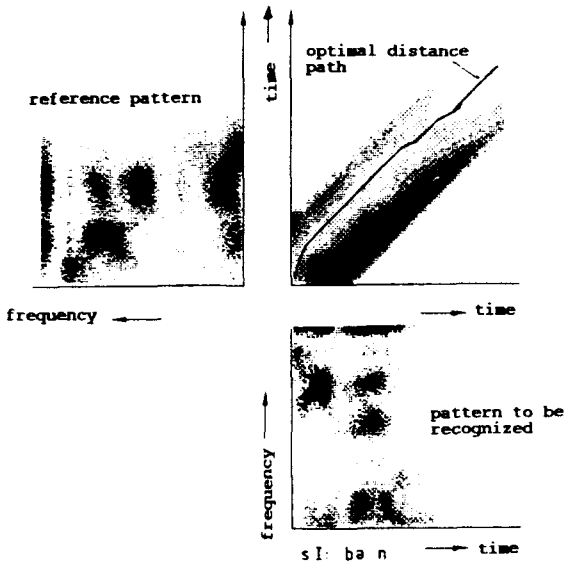


Fig.3.6: Principle of dynamic time warping DTW

DTW makes first a local comparison of all short time spectra (10 ms-spectra) of the reference pattern and the new pattern to be recognized. In a second step the best path through the resulting distance matrix is computed. This optimal distance path then is a measure on the double time scale how both spectral patterns may be optimally adapted to each other through dynamic adaptation of the time scales. If we may assume that the spectral deviations of both patterns are to be ignored - which is only allowed for speaker dependent recognition - then the deviation from the linear path is a good measure of similarity between both patterns.

Word recognizers based on this principle have brought the first breakthrough for practical applicability of word recognition due to their good recognition results in speaker adaptive word recognition (Cla92).

Another method of whole word based pattern recognition is done with artificial neural networks. Here again some assumptions about the physiological perception of speech are the basis for the technical approach. A neuron as the basic element of physiological processing consists of the cell corpus which has many dendrites arising from it. These dendrites are ending on other cells making contacts on their surface, the synapses. So they form a network for exchange of information. Fig 3.7 shows a schema of a physiological neuron and its electrical equivalent, the neural network basic element.

Through combination of many such neurons we can build a neural network which is able to

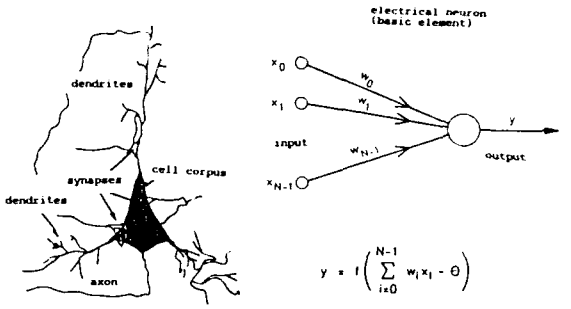


Fig.3.7: Schematic draw of a neuron and its electrical model.

make distance measurements between two-dimensional patterns. A schematic draw of such a network is shown in Fig.3.8. There are at least three signal layers necessary.

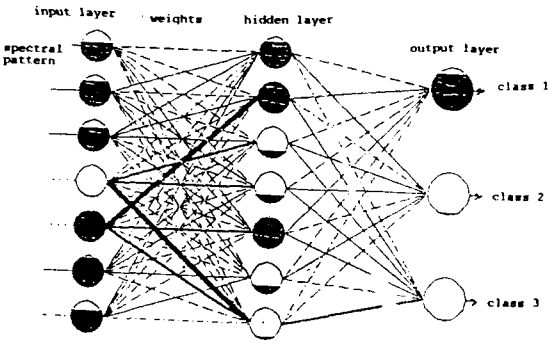


Fig.3.8: Schematic drawing of a neural network.

The first one is the input layer where we are inputting the result of the preprocessing, e.g. the spectral pattern of the word to be recognized. Following is the network of artificial neurons including the weighting factors w_i from Fig.3.7. The hidden layer combines the information from the training procedure. This means that we can interpret its function as a sort of reference pattern. The output layer finally combines the input from the input layer weighted with the information from the hidden layer to a measure of class membership. The darkness of the neurons within the layers gives first the spectral energy and finally the membership. Neural networks are nothing else than a distance measure scheme which usually includes some nonlinearity in the behaviour of the weighting factors. It is of course possible to include more than one hidden layer. But then the amount of training samples becomes very large. The advance of neural network speech recognizers lies in the fact that this technique concentrates on the discriminative aspects of the different spectral parameters.

Through intensive training the network is therefore able to learn even rather small distances between different word classes, e.g. to differentiate between phonetically rather similar words. The main drawback is still that the amount of training to make such differentiations is often not tolerable and so presently there is not yet any specific advantage of word recognizers based on neural networks compared to conventional statistic methods.

3.3 Capabilities and Limitations of Whole-Word Recognizers

The recognizers thus far described are based on purely whole word patterns. There is no knowledge included about the structure of speech or words, which consist of single sounds to be articulated in concatenation. The recognition process takes the word as the basic element with all the problems which are arising from the fact that e.g. normalization of rhythmic differences in the articulation of a word is not so easy. DTW has found a nice technique for this, but it has on the other side problems with adaptation of spectral changes for speaker independent or speaker adaptive recognition.

Another problem is the recognition of connected words with the methods mentioned. Here usually some parts of the words are coarticulated, such that the single words are no more articulated in the same manner as if they would have been spoken in isolation.

A more detailed adaptation to the structure of the language itself would therefore offer more possibilities to widen the scope of speech recognition to better word recognizers and on the other side to recognition of continuous speech and thus to real speech understanding systems.

4. The Phonologic Structure of Speech

4.1 Sounds and Phonemes

Historically the first approaches to automatic speech recognition started with attempting to recognize single sounds, or still more easier to recognize single letters to make an automatic typewriter. But all these attempts have not been very successful and so the practical solution was to make whole word pattern recognition for command applications. This is mainly due to the fact, that the word is the smallest unit which can easily be produced in isolation.

On the other side the smallest unit presently used in spectral pattern matching is the 10 ms-spectrum. The usual speaking rate of human speaking is around 20 sounds per second for even a fast speaker. If the spectrum of a word is calculated every 10ms then it is possible to describe every sound with around 5 spectral patterns. So, also rather short sounds like plosive bursts are at least described by one spectrum. This 10ms unit is a rather artificial unit which is only roughly oriented at the structure of the speech signal.

Much better units are phonologically based on distinctive parts of the continuous signal. Such units should fulfil at least the following criteria:

* They should have phonological meaning.
* They should be easily separable out of the continuous speech signal.
* They should not change too much if they are coarticulated with other units.
* Coarticulation of such units should not be possible too much.

We can at least identify two such units, the speech sound with its abstract representation the phoneme and the syllable, which is mainly a unit used in written representation of language but which has simultaneously an important aspect in spoken language.

The advantage of the phoneme as basic unit is the limited number of them. The usual large languages can be described by around 40 phonemes. But the number of syllables is between 100 and 1000 times larger, from which many are rather seldom. The phoneme seems to be a rather recommendable basis for a description of the language. A still pertinent problem is of course that there is no direct and reversible transform between phonemes in a word, its sound structure and the typing of the word. There are rule based systems to do this, but these sometimes miss the correct spelling. To use lexica needs on the other side extensive human work and never will be complete.

The question for the selection of the best units can perhaps be answered if we ask for our human perception. Here the answer is rather simple: It is surely not only a pure phonemic decoding. We experience this fact clearly if we want to recognize meaningless words. Even to recognize such meaningless syllables is complicated. On the other side long experience from optical spectrogram reading has shown that trained users are able to attain a correct phonetic decoding of between 80 and 90 percent.

4.2 Speech Structure and Perception Models

Our daily experience shows rather clearly that our speech perception process includes a huge amount of knowledge. The basic question will be if, and how this knowledge is practically combined with the existing structure of the speech signal itself. Is there e.g. a substantial amount of phonologic knowledge directly influencing the perception on a sound or word level?

Cole et.al. have described a basic collection of rules for such a perception model. These are(Co80):

* Words are recognized through the interaction of sound and knowledge.
* Speech is processed sequentially word by word. Each word's recognition locates the onset of the immediately following word and provides syntactic and semantic constraints to recognize the immediately following word.
* Words are accessed from the sounds which begin them.
* A word is recognized when the sequential analysis of its acoustic structure eliminates all candidates but one.

In this terminology the phonologic structure of the speech plays an important role.

Even if the definition does not include any intermediate structures like syllables, these may be included in the recognition of word structures. The composition of words from syllables and the relevance of syllable perception is shown very clearly in perception experiments. We have no problem to reconstruct missing sounds in a word, but we have much more problems to reconstruct missing syllables. Syllables may already have a certain semantic role, if we look at prefixes which may change totally the semantics of a word.

The stratification model of speech perception and speech structure in Fig.4.1 shows this fact. The linear structure of the phonemic chain is changed into a netstructure at the higher levels (Win83).

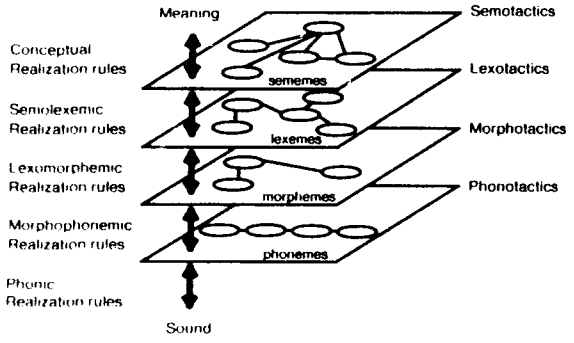


Fig.4.1: The stratification model of speech (from Win83).

5. The Role of Words and Sentences

5.1 The Word as a Semantic Unit

The bottom-up approach of speech perception which has been reflected in the existing work in automatic speech understanding has stressed the importance of all these small units, starting from a 10 ms feature vector over the phoneme, syllable up to the word. Other investigators, motivated chiefly by developments in generative linguistics, have proposed much larger units for perception like clauses or sentences(Pis75). The word plays here an intermediate role, as we already may see in the stratificational model from Fig. 4.1.

It is of course in the meantime clear that there is now sufficient psychological evidence that all these layers of analysis are available simultaneously. Many models of brain functions favour a layered model for the processes done in the brain, and of course these layers are permanently active during the process of perception. It has become clear from brain physiological studies that only if all layers are active a perception of speech is possible. Of course the problem is still under discussion how far speech based semantic processes need speech perception as a basic. Finally this means that cognitive processes are ultimately based on a language and speech processing procedure.

The word fulfills many of these requirements. It has a semantic meaning. As we know from some conversations, especially in foreign languages it is widely possible to arrange a fully word based conversation, leaving out all the rest of the sentence.

5.2 Syntactic and Semantic Structures

Words presented in a sentence context are more intelligible than presented in isolation. The same is true if we present words in a nonsense environment. Then the recognition of the word may be worsened. Some traditional assumptions about the contribution of syntax and semantics in the perception process underestimated the relevance of the cooperation of all the levels. This view gave them only the role to restrict the multitude of possible alternatives. The process of speech perception was in this model based on a strict serial organization, where the phonemic characteristics of the speech signal are more or less directly extracted from the acoustic properties of the signal.

Phonetic experiments in transcription of spoken language have shown in the meantime, that it is nearly impossible to decode the correct phonemic representation of an utterance without higher level lexical and syntactical information.

Finally it is important not to forget the prosodic information which exists on a rather low word level, but which is mostly relevant on the sentence or phrase level. Only in the last few years the importance of prosody for human perception is investigated deeper and this understanding then offers new chances for machine perception of speech.

5.3 Spoken Language and Information Processing

Communication and information processing are two very intensively connected topics. There is no information processing possible without any communication and we know that this communication process does not only cover the internal process of communication within the brain of a human but that the interpersonal communication is more or less the basic force for every advance in cognition. Spoken language communication is one of our basic communication media, it is at least the most spontaneous medium. Compared to written communication it offers so many additional parameters like intonation, prosody, stress to underline certain semantic facts and to give a much wider scope of information than it ever is possible through written language.

There is some psychophysical evidence that written and spoken language use the same phonetic code which is derived in a similar way from written or spoken information. This phonetic code could then be the basis for most of our language based information processing steps.

6. Machine Speech Understanding

6.1 Structures of Speech Understanding Systems

After these views into the structure of our human information processing, especially related to speech perception, it will now be interesting to look back again at the state of machine perception of speech. If we try to make a true analogy to our models of human speech perception we can have in principle two approaches, the strict serial system and the blackboard approach where every part of processing can permanently access to all the steps. Fig.6.1 shows the schematic structure of a serial speech understanding system.

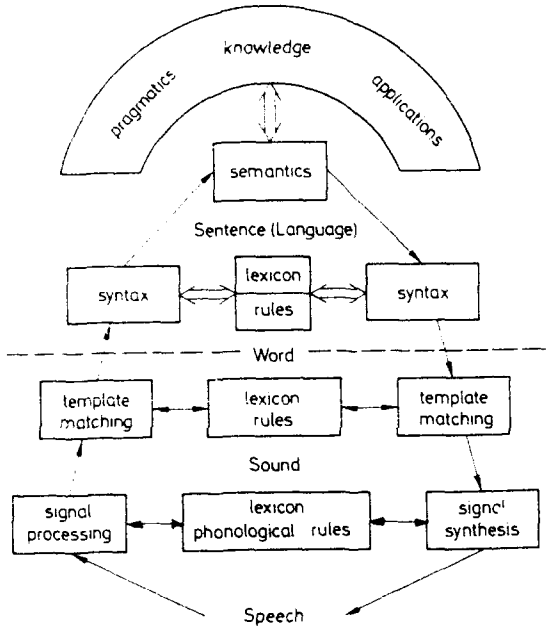


Fig.6.1: Steps in a serial speech understanding and dialog system.

Such a system includes not only the understanding stage up to the analysis of semantics but it must have additionally the reverse information channel for outputting of the answer.

All the steps which have to be treated start and end with the acoustic signal and they end with the semantic representation of the content of the spoken signal. The first steps in the analysis part are rather similar to a word recognizer, as was already described. Such speech understanding systems usually have to understand continuous speech and therefore it is never very helpful to consider the words as isolated events but it will be much better to represent every word by a collection of much smaller units, usually the phonemes. We shall see in the following chapter, which methods are today existing to recognize words on the basis of phonemes and how it is possible to care for different alternatives of every word and simultaneously to provide the following linguistic processing with a sufficient number of possible words for the sentence to be analyzed.

After the signal processing the linguistic processing is following which is based mainly on syntactic and semantic analysis. Of course the top level processing is depending on all the pragmatics based knowledge, which controls the dialogue and the internal knowledge processing. The output channel is doing rather similar things in a reverse manner. This means that from semantic concepts via syntactic design a text is created which then is transferred into an acoustic signal through phonologic steps and signal synthesis.

This linear approach to speech understanding gives good insight into the single steps and offers good possibilities for control of the different processing levels. A totally different approach is the blackboard based approach, where basically a simultaneous acces to all levels of signal processing is possible, from low level acoustic signals up to semantic and pragmatic processing. This approach offers the principal capability to make easy requests between all these domains, but the main problem is still, to decide, how all these domains are to be coordinated. Fig.6.2 gives a rough schema of such a blackboard based approach.

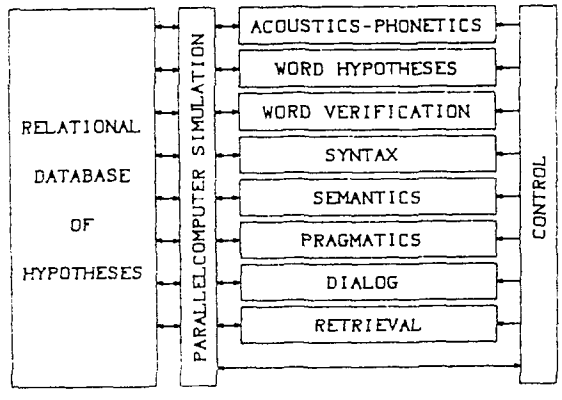


Fig.6.2: Blackboard approach for speech understanding.

The important part in every blackboard approach is the database where all the hypotheses about the results of the different parts are represented. It must of course include a measure for the vagueness of the special results which again could be the basis for interactions between the domains.

Of course the basic question is and will be, which of both concepts offers the best and on a long term basis the most possibilities for inclusion of much phonologic and linguistic knowledge and has simultaneously good capabilities for getting enough insight into the behaviour of the models. As we have already seen, psychoacoustics and psycholinguistics offer some ideas about this question, but it seems that our human information processing scheme does something serially and some other things are

done in parallel. At least the higher levels seem to have much parallelism using a sort of blackboard approach, while the very low level parameter extraction is done serially. Technical solutions of course prefer systems where most of the steps can be designed separately. This is the case in both examples, but the interaction in the serial system is much simpler. Therefore in most technically realized cases the serial approach is used and up to now is surely more advanced, even if in a long term sight this approach will be replaced through more and more parallelism.

6.2 Word Recognition in Speech Understanding systems

As we have already seen the most flexible way to describe continuous speech is on a basis of the phonemes or the sounds which describe the realization of the phonemes. Every word to be recognized can be modelled using such a phoneme chain. The single phoneme again can be modelled on the basis of spectral patterns or special features of such spectral patterns, like positions of formants, voiced/unvoiced characteristics or spectral energy distribution. Such a systematic model based approach is based on the theory of Markov Models, which had first been used to describe the statistical characteristics of written language. Fig.6.3 shows the results of a Markov Model for German written text, where statistical relations up to the degree 3 are used. The statistical degree r=0 uses only the distribution of letters and blanks in German texts, while r=3 includes the statistics of the distributions of the three following letters.

r = 0: aiobnin*tarsfneonlpiitdregedcoa*ds*e*dbieastn dnurlarsls*omn*keu**svdleeoieei*...

r = 1: er*ageplepfteiningeil*gerelen*re*unk*ves*mfc nzerurbom*...

r = 2: billunten*zugen*die*hin*se*sch*wel*war*gen nicheleblant*diertunderstim*...

r = 3: eist*des*nich*in*den*plassen*kann*tragen*wa zufahr*...

Fig.6.3: Markov Chains based on statistics of German texts.

Already with r=2 there are some short meaningful words received and this becomes better and better with rising r.

On the basis of Markov chains for spectral patterns we then model in a similar way the signal characteristics of spoken language up to the word level. Of course, as Markov himself has done, such a statistical modelling ist still possible beyond the word level. It is principally possible to model whole sentences, even the characteristics of texts can be included in a statistical model.

To recognize words it is then possible to use Hidden Markov Models HMM for every word and for every phoneme to be recognized, which can be trained through spoken speech and thus become more and more representative for the word to be recognized itself.

The basic structure which can be described by a Hidden Markov Model is shown in Fig.6.4.

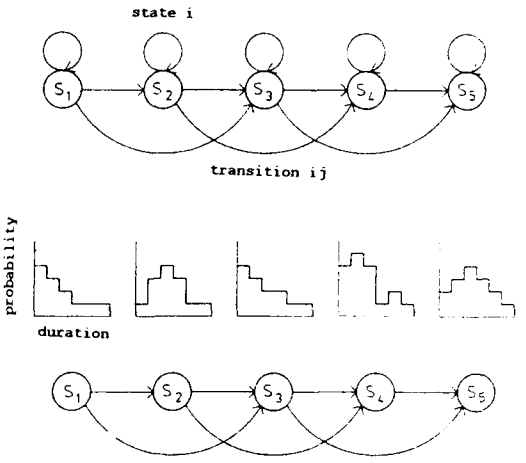


Fig.6.4: Basic structure of a Hidden Markov Model.

There are states and transitions, both with probabilities for them. These states Sn can be followed by another state but also by themselves. The structure of the model defines, which transitions are principally possible. Of course the most general model offers possibilities for every transition, but such models are practically not calculable due to restrictions in the statistical representation in a limited training material. So, experience is requested about the best structure for such models. Every state of a word model is again based on a smaller sound model, which usually has at least three states which model the onset, the stationary part and the final part of such a sound. The statistical model has to include not only durational models for every state but it must also have information about the probability of a selected spectral pattern being in the position of any state. This is necessary because the spectral variations in the articulation of different words are rather high. This can be seen in formant maps, where the position of the first two formants for the vowels have been analyzed. Such a map is shown in Fig.6.5.

If we look at such a plot, we can see, that there is much overlap of the different vowel spectra. This means that it is not possible to differentiate them clearly. This becomes much more complex with more dynamic sounds, which consist mainly of changing parameters. Therefore the characteristics of the different states in the HMM must be described by their probable distribution within the set of parameters, e.g. the spectrum. It has become usual to do this on a soft decision basis, meaning that the

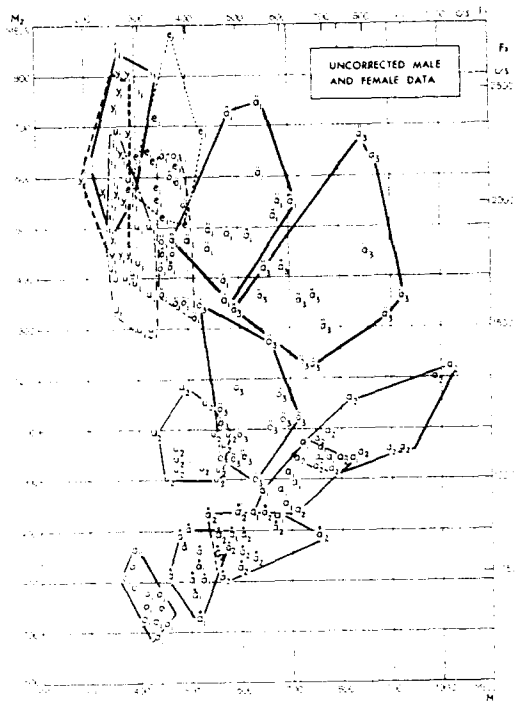


Fig.6.5: F2/F1-Plot of the Swedish vowels. (Fan59)

membership to one parameterset, e.g. in the F1/F2-area is described by the probability that a certain vowel has a certain F1/F2-parameter. Of course this needs immense statistic work with different voices and different examples of speech, but finally this leads to a chance to characterize the sounds even in a speaker independent way, if the statistical distribution of all the parameters is measured over many speakers.

It is highly astonishing how we human recognize speech in a widely speaker independent way. There seems to be not a long adaptation procedure necessary to recognize totally different voices, e.g. during a conversation with very differnt people. It is up to the moment not yet clear which sort of spectral and phonologic adaptation we can make to have a practically unlimited capability to recognize nearly every speaker. It seems obvious that mainly higher level processes are responsible for such a capability because there is no signal processing known which could do this. Since many years speech research has looked for the so called "distinctive features" in speech. These are parameters which could be independent of the special speaker and of the word where a special sound has been spoken. But there has nothing been found which fulfils all the expectations. For the moment therefore the solution is to adapt a word recognizer in a short training phase to a new speaker's voice. This is done with a spectral transformation. Fig.6.6 shows the principle of such a transformation. In a bilateral transformation the parameters (normally the spectral pattern) of the new speaker and of a well defined reference speaker are transformed into a new parameter area in such a way that the difference between both speakers becomes minimal. Through this transformation better results are possible than through a single sided transformation of the new speaker into a reference speaker.

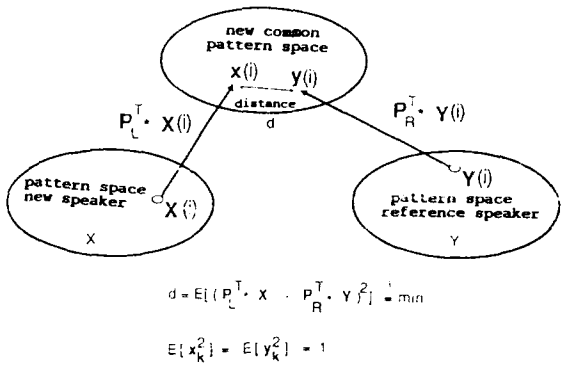


Fig.6.6: Principle of a two-sided transformation of speaker parameters.

If we look again on our human technique of adaptation such spectral adaptation is surely of minor importance, much more important seems to be an adaptation to the dynamic articulation.

After all these pattern oriented processing the word recognizer itself has again to identify the spoken word correctly. Using the Hidden Markov Technique it is again important to measure distances between the trained model and the chain of spectral states of the word to be recognized. Usually we get many word hypotheses. Especially in the case of continuous speech these hypotheses are defining a network of words which may all be possible at different time slots. Fig.6.7 shows the principle.

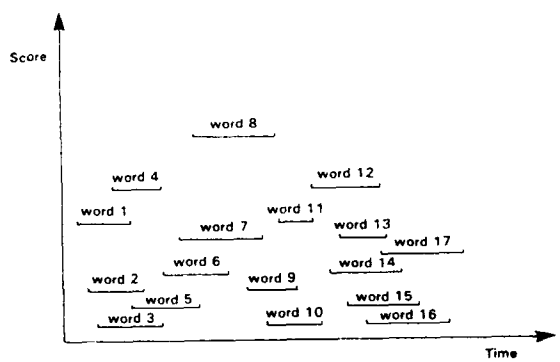


Fig.6.7: Word net as the result of the word recognition.

In a serial understanding system it will now be the task of the linguistic processing to define first the correct word chain and in the following stage to analyze all the contents of the phrase which had been spoken.

6.3 Language Models and Parsers

Similar to the definition of the most probable word, it is possible again to define the most probable chain of words using again statistical analysis of a huge collection of texts, which should be as far as possible representative for the texts to be analyzed. Then alone statistics may help to define from the word network the most probable sentence, based on the statistics of the most probable chain of words. We call such a method a language model, even if we know that every language model is rather restricted to the texts that had been the basis for the training of the model. So, if for example a speech understanding system should be able to write special letters for patent counselors, the training material should come from many such letters.

Such a statistics based approach has the advantage that there are no rules and it can be easily adapted to other applications if the training material is changed. The important drawback lies in the fact, that the language model may fail totally if the application domain is changed without new training. In some cases the result of such a recognizer may be worse than without any language model.

Therefore a systematic, rule based approach is an alternative which often gives better results on average texts, but of course it may totally fail on syntactic constructions for which there is no rule based model foreseen. Especially in the case of spontaneous speech understanding there are often phrases used which are not following any grammatical rule.

The approach of transformational grammar had seemed to offer a rather easy capability to derive very different grammatical structures from some basic principles. Fig.6.8 gives an example from (Win83).

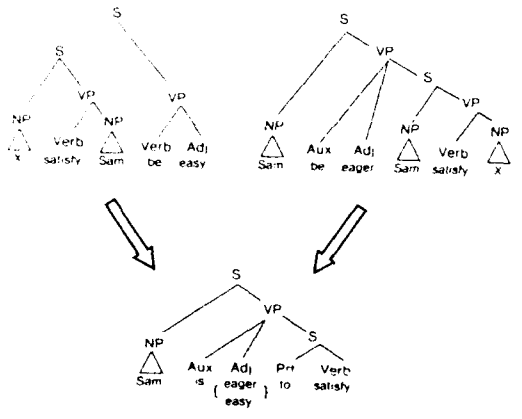


Fig.6.8: Sentences with different deep structure transformed into the same surface structure.

The deep structure of a sentence is related to the semantic content, while the surface structure is describing all the syntactic relations within this sentence. If there is a sentence with the same deep structure as another sentence it may be possible that they have different surface structures and vice versa. If we start with a syntactic analysis for the processing of the sentence we may see very similar surface structures for two sentences but the semantic content, represented by the deep structure is different.

Fig.6.9 shows a model of linguistic competence of the adult. This means that the main language capabilities are in a mature state and the actual usage is dominating over the acquisition of language capabilities.

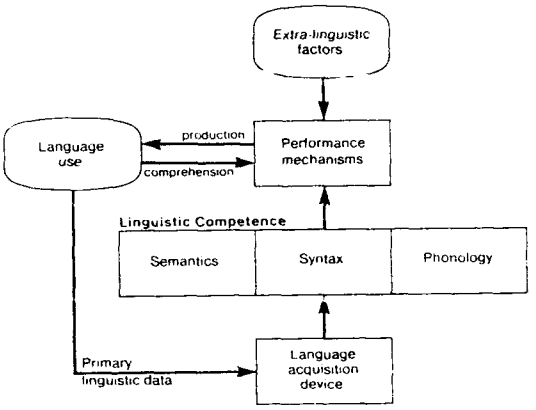


Fig.6.9: Model of the basic human language capability. From (Win83).

This model has three main components, the central linguistic competence, the language acquisition device and the performance mechanism. Linguistic competence is the source of our intuitions about grammatical structure. The language acquisition device is permanently bringing new information about deep and surface structures and is permanently widening the linguistic competence. Of course as already mentioned in the adult user this is no more as active as in the case of a child acquiring most of the linguistic competence. The model has the three main factors, semantics, syntactics and phonology in parallel as we have already seen in the blackboard model.

Another rather important relation happens within this model between the boxes for language use and the performance mechanism. The permanent interaction between the speech production mechanism and the perception mechanism has been stated many years ago already in the Motor Theory of Speech Perception. This theory says that every perception process is in parallel connected to an internal production process within the brain of the human perceiving the speech signal. All these theories very definitely state that there is an intensive interaction between both sides and that it is nearly impossible to perceive speech if the internal production capability is distorted. Of course it is clear that this does not concern the external mechanisms of speech production.

If we look at Fig.6.1 the syntactic processing stage refers to the word lexicon which is always the one basis of its analysis. The other thing are the necessary rules which identify the relations of words within a phrase or sentence. We can therefore state that the basic elements of a syntax are:

* a lexicon of allowed types of words,
* a collection of allowed types of sentences and
* a rule system combining both.

As an example for the problems with syntactic analysis we can look at two different syntax types. But the contents of the sentences are in this case totally similar.

Example sentences (1):
"Are there new papers from Maier?"
"Do you have five recently published reports from Mr. Miller?"
"Existed there a new report from the ministry?"

Equivalent syntactic description:
[presence][number][date][paper][author]

Example sentences (2):
"Has Mr. Maier recently written some new papers?"
"Has Mr. Miller newly published five new reports?"
"Has the ministry presently published a new paper?"

Equivalent syntactic description:
[auxiliary verb][author][date][verb][number][paper]

These two small examples may show that there are very many possible descriptions of the same fact. It is without any large amount of effort possible to create some thousand different versions of grammar describing the same content, but there are the same amount of versions which lead to misunderstanding.

Within today existing speech understanding systems the number of sentences allowed is rather restricted, being a basic problem how this can be permanently adapted to the actual versions of speaking habits. Every living language is permanently changing its habits and this means that even the syntactic constructions allowed are changing permanently. Every syntactic rule system should therefore have the capability to adapt itself to new speaking habits.

There are mainly two ways to realize adaptive grammar systems in understanding, to include elements of generative grammar or to do it in a sort of interactive learning through dialogue, which is in principle possible within a man-machine system.

6.4 The role of semantics and pragmatics

We know from our everyday experience that we do not only rely on our language knowledge if we try to understand the meaning of sentences spoken through a human partner, but we include much unconscious knowledge. These are elements which we call world knowledge or more general pragmatic knowledge. That is everything we know from the special application on which we make our discussions but far beyond this all the knowledge from our life. Therefore often understanding via a telephone call is less easier than a direct conversation, where we can include behaviour of our partners too.

The model of Fig.6.9 covers therefore only the limited and narrow speech model. It has for practical reasons to be widened with a special channel providing the non-speech experience and a knowledge base for all these non-speech experiences.

In the schema of a linear speech understanding system of Fig.6.1 this pragmatic and application oriented processing and database forms the top level processing part of the whole system. In our human processing this knowledge is surely distributed over the whole cognitive processes of the brain.

For a limited technical application of speech understanding there are some chances to include such knowledge in a practical accessible manner. It will then be intermixed with the semantic analysis part.

Semantic analysis may rely on many different aspects of the speech structure. The most important of them are represented through the following parameters:

* Syntactic structure
  The order of words within a phrase defines widely the semantic content of a sentence. The main problem is that there are extreme possibilities for ambiguities which may not be resolved through a syntactic analysis alone, but which need additional knowledge.

* Vocabulary
  The vocabulary can within technical systems be restricted to a rather limited amount of words. If a user is able to handle such a limited amount of words and he can express all his ideas with this lexicon, than it is possible to define the semantics of the words used in a rather consistent way, such that possible misunderstandings are rather limited.

* Prosody
  This parameter characterizes all the relevant aspects of extra-linguistic but speech oriented behaviour of a human. Examples are intonation, stress for words or sentences, rhythm of speaking, up to hesitations. A detailed analysis of such parameters is presently not yet possible in automatic systems, but there are many scientific approaches to use much more of these parameters for semantic analysis.

* Phonology and Articulation
  How sounds are spoken and how they are combined to words characterizes partly intonation and partly some special knowledge about the speaker himself. We can detect from this information something about things which are directly relevant on the background on which the speech to be understood is articulated. Here non-speech articulations, like ah's and hm's etc. are relevant too.

* Acoustics
  External noise, distortions, limited bandwidth give us some semantic in-

formation about the speech signal and its location of production and therefore about the speaker's present situation.

* Discourse structure
Every dialogue has a certain structure which depends on many factors, like speaker habit, dialogue content, dialogue stress, relevance of content etc. It is even for human auditors not easy to assess all these different aspects from the speech signal alone. For machine speech understanding it is presently nearly impossible to rely on such an analysis. Here still much research is necessary, which must include ergonomic aspects as well as application oriented and phonologic details.

* Dialogue stile
People are used to adapt themselve under different conditions to different stiles of dialogue. This aspect is narrowly related to the problems of analysis of discourse structure. It is more or less the top level aspect of the dialogue scenario.

These aspects which should be included in the semantic analysis task are widely intermixed with each other such that it is not so easy to separate them definitely and to describe their influence under semantic aspects in a very definite manner. Additionally some parameters are often only occasionally changed and give some unconscious information, but often do not reflect the conscious intention of the speaker. Often they reflect the special habits any special speaker has, and so they characterize more the speaker and not so much the semantics of the speech itself.

The basic tasks of semantic analysis are then:

* to create a logic description of the content of a sentence,

* to describe within this logic description relations with a world model, and

* describe possible semantic alternatives as a source for the future dialoge.

Practically this task needs very powerful tools for describing all the possibilities and relations efficiently and in such a way that definite semantics are coming out and not ambiguity.

Within the examples given for syntactic analysis we can see where some difficulties are. For example semantic rules may be:

* Make a list of all words which have been attributed to [author].
* If the word for the date is defined as a year, then check if this is a meaningful year (which should be between 1900 and 1992).

Within the first example, the list of authors is not easy to be implemented, because authors are here not only people with names but they can be an official agency, a confederation etc. All these can be the source of documents, and in the definition of our syntax they can be authors.

A good prosodic analysis will resolve some ambiguities, but most must be done by inclusion of pragmatic knowledge coming direct from the specific application and additional usage of world knowledge. Both can be gained within the understanding system through the dialogue.

6.5 Dialogue with Speech Understanding Systems

In our human communication we do many clarification by dialogue. Within these dialogues we learn not only the actual situation of information processing but the long term knowledge. That is the knowledge which we build up during our life and which we therefore call world knowledge. There are nice examples for computer-man dialogues from the first start of artificial intelligence research and the first investigation into man-machine dialogue.

These first examples have been realized with the system called Eliza at the MIT by Joseph Weizenbaum. It had been realized already in the sixties, when computers had become enough powerful to handle nearly in realtime such linguistic problems (Wei66). Fig.6.10 gives an example of such a dialogue.

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says i'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

I need some help, that much seems certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother

TELL ME MORE ABOUT YOUR FAMILY

Fig.6.10: Example of a man-machine dialogue with Eliza. Written language.

The analysis of meaning was not really realized in Eliza, it had been more or less only a pattern identification and we feel spontaneously that the dialogue is rather schematic and it turns around itself. Only an analysis of the deep structure may have overcome all these problems.

But of course the main problem was that no real pragmatics was implemented. So the dialogue itself was senseless and therefore it looks like a typical party dialogue, where people who have nothing to say to each other are speaking and have a nice entertainment.

A real pragmatik and semantic analysis, which includes knowledge must be based on extensive databases and the correct inclusion of all the knowledge stored in these databases. It is clear that this problem is again a language analysis problem because much of the knowledge in these databases will again be stored using language as the adequate medium.

7. Speech recognition and understanding and their applications

7.1 Technical state of speech recognition

Speech recognition systems today available are concentrating on very special tasks. In Fig.7.1 we have shown the available systems on a three dimensional specification map.

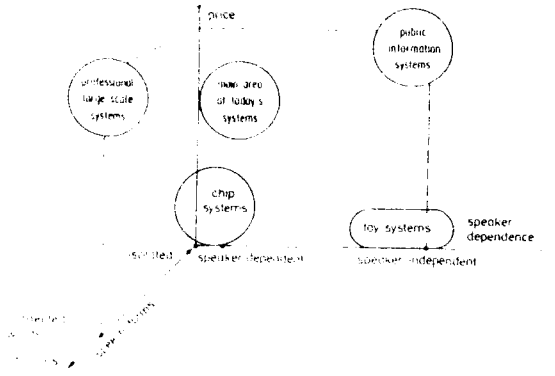


Fig.7.1: Three dimensional representation of the major aspects of speech recognition systems.

The relevant parameters used for this classification are:

* The system prize, which usually represents the technical capabilities of a system, ie. a good recognizer for isolated words with high recognition rate is usually more expensive than one with a limited recognition rate.
* The sort of speaking required, isolated or connected or totally continuous.
* The degree of speaker dependence, adaptation or totally speaker independence.

The main areas of practical systems concern the recognition of isolated words for command applications. These applications often require speaker independent recognition if they are used over the telephone in public applications. Another class of recognizers addresses the problem of connected words. Speaker independence is here still a problem because the coarticulation problems of different speakers are not so easy to be predicted and modelled. Another aspect, which could only be described in terms of prize is robustness against background noise, speaker variations, limited bandwidth etc. Finally we have not included in the presentation the vocabulary size, which can vary from very few words (10 to 20) for limited command input into machines up to many thousand words, when one wants to realize a dictation machine.

The recognition rates today possible differ very high, depending on the difficulty of the recognition task. It can be near to 100% for good quality speech, a limited vocabulary with trained speakers, but it can be 20% worse for untrained speakers in the same application task and it can even be as low as some ten percent for larger vocabulary under noisy conditions. Therefore it does not make much sense to give here figures. Every application task must be carefully investigated, user behaviour must be modelled and the man-machine dialogue must be designed as carefully.

The application of speech understanding is still not yet possible because practical and applicable speech understanding systems which can understand continuous speech input with naturally spoken sentences are not yet on the market. There are speech dialogue systems available with word recognition as input and with a continuous speech output. For most practical applications such systems fulfill the need of the user, if the user himself cares for a careful isolated or connected spoken input.

7.2 Forms of Dialogues

Fig.7.2 shows schematically how speech input and output may bring a human and a system together.

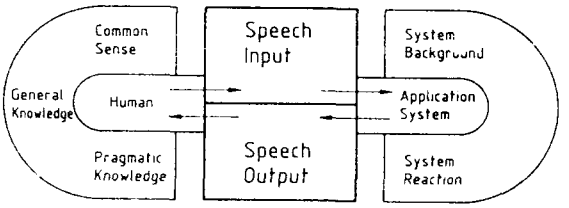


Fig.7.2: Functional relations in speech controlled systems.

On the one side we find the human operator with its knowledge, based on very different sources. On the other side there is the application system, which is containing different forms of information and which will show very specific reactions.

We have roughly two different forms of users, the occasional user and the professional user. The occasional user uses speech communication with machines only for very specific applications and rather rarely. He is not trained to usage of speech systems and handles them as if he would speak to a human. The professional user on the other side is a daily user and is trained to do the right things, ie. speak in the manner required and knowing the vocabulary allowed.

We can distinguish two forms of dialogues, the action dialogue and the information dialogue.

Fig.7.3 shows the essential elements of an action dialogue, where the user wants to get rather simple precise actions. The goals of this activity are rather clear, the user has to command his request and gets then hopefully the correct system reaction. here syntactic and pragmatic processing steps are mostly included covering very restricted and specific pragmatic aspects. Simple examples of such dialogues are speech based machine control.

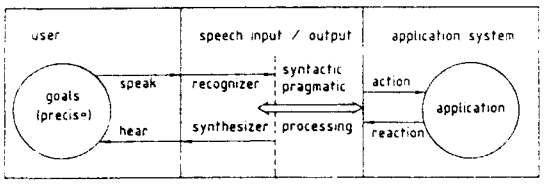


Fig.7.3: Structure of an action dialogue.

In Fig.7.4 the basic elements of an information dialogue are presented. Here the user does not want to produce direct actions but he wants to get information in a more or less natural dialogue. The primary goal of such a dialogue is to make a real information exchange.

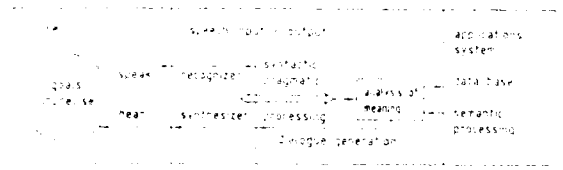


Fig.7.4: Structure of an information dialogue.

Usually here the level of information exchange goes much deeper than in the action dialogue. Therefore the analysis of meaning is the additional component characterizing such a dialogue. Examples of such dialoges are information systems, e.g. for flight time tables or for general public information like weather forecast. Such systems will become more and more important already in the near future and they will then need good speech understanding.

8. Future Developments

Machine perception of spoken and written language is surely one of the most advanced challenges of information technology. Speech is the basis of most of our cognitive processes. If we can get a deeper and deeper understanding of all the processes related to speech production and speech understanding we will get access to much better understanding of the understanding process itself. It is clear from the laborious research in speech understanding in the past that we are presently only in the begin-ning to understand speech and all the structure behind it better and that there is still a long way to go.

Presently available systems which can be useful tools for man-machine communication have in many areas profited from models of our human speech processing. Such models will in the future help to understand all the important processing steps better. A system approach to integrate the different steps into a more synergetic concept may be better than the purely linear step-by-step approach.

Deeper insight into the mechanisms of speech will help us not only in systems for easy information processing, it will help us in speech translation and in cooperative knowledge processing.

Speech interactive systems will offer us a true human access to machine information and they will in such a way widen the scope of practical applications of information technology in the same way as the basic insight into it.

Literature

(Cla92)
Class F, Katterfeldt H, Regel P: Methoden und Algorithmen der Worterkennung, in Mangold H: Sprachliche Mensch-Maschine-Kommunikation, Oldenbourg, München 1992.

(Co80):
Cole R A, Jakimik J: A Model of Speech Perception, in Perception and Production of Fluent Speech, Erlbaum, Hillsdale 1980.

(Fan59):
Fant G: Acoustic analysis and synthesis of Speech with application to Swedish. Ericsson Technics 1,1 (1959)

(Ma76):
Markel J.D, Gray A.H: Linear Prediction of Speech, Springer, Berlin 1976.

(Pie85):
Pieraccini R, Rainieri F, Giordana A, Laface P, Kaltenmaier A, Mangold H: Algorithms for Speech Data Reduction and Recognition, ESPRIT 85, Elsevier Science 85

(Pis75)
Some Stages of Processing in Speech Perception, in Structure and Process in Speech Perception, Springer, Berlin, 1975.

(Wei66):
Weizenbaum J:ELIZA, CACM 981966), 36-45.

(Win83)
Winograd T.: Language as a Cognitive Process, Addison-Wesley, 1983, Reading Mass.

(Zwi67):
Zwicker E, Feldtkeller R: Das Ohr als Nachrichtenempfänger, Stuttgart 1967

Sensing and Interpretation of 3D Shape and Motion

Takeo Kanade
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890, USA

Abstract

Robotics is where artificial intelligence meets the physical world. Computer vision provides robots with the perceptual capabilities which are especially critical for robots which operate in an unconstrained natural environment.

In computer vision, recovery of 3D shape and motion is the key to understanding scenes. Thus, the problem has attracted much of the attention of vision researchers over the last decade, and many sophisticated algorithms have been developed. I am going to talk about three recently developed methods for sensing and interpreting 3D shape and motion:

- **The factorization method for image sequence analysis**
- **Very fast range imaging by analog VLSI smart chip**
- **The multi-baseline stereo method.**

It is interesting to note that while the performance of these methods has exceeded that of previous methods, the algorithms themselves are simpler and more straightforward. In addition to enhanced performance, these algorithms are suitable for real-time parallel implementation by special hardware or VLSI.

The following three parts provide detailed descriptions of these methods.

The Factorization Method for Shape and Motion Recovery from Image Streams [1]

Inferring scene geometry and camera motion from a stream of images is possible in principle, but is an ill-conditioned problem when the objects are distant with respect to their size. We have developed a factorization method *that can overcome this difficulty by recovering shape and motion without computing depth as an intermediate step.*

An image stream can be represented by the $2F \times P$ measurement matrix of the image coordinates of P points tracked through F frames. We show that under orthographic projection this matrix is of rank 3.

Using this observation, the factorization method uses the singular value decomposition technique to factor the measurement matrix into two matrices which represent object shape and camera motion respectively. The method can also handle and obtain a full solution from a partially filled-in measurement matrix, which occurs when features appear and disappear in the image sequence due to occlusions or tracking failures.

The method gives accurate results, and does not introduce smoothing in either shape or motion. We demonstrate this with a series of experiments on laboratory and outdoor image streams, with and without occlusions.

1 Introduction

The structure from motion problem – recovering scene geometry and camera motion from a sequence of images – has attracted much of the attention of the vision community over the last decade. Yet it is common knowledge that existing solutions work well for perfect images, but are very sensitive to noise. We present a new method called the factorization method which can robustly recover shape and motion from a sequence of images without assuming a model of motion, such as constant translation or rotation.

More specifically, an image sequence can be represented as a $2F \times P$ measurement matrix W, which is made up of the horizontal and vertical coordinates of P points tracked through F frames. If image coordinates are measured with respect to their centroid, we prove the *rank theorem*: under orthography, the measurement matrix is of rank 3. As a consequence of this theorem, we show that the measurement matrix can be factored into the product of two matrices R and S. Here, R is a $2F \times 3$ matrix that represents camera rotation, and S is a $3 \times P$ matrix which represents shape in a coordinate system attached to the object centroid. The two components of the camera translation along the image plane are computed as averages of the rows of W. When features appear and disappear in the image sequence due to occlu-

[1] This research was performed by Carlo Tomasi and Takeo kanade, and was supported by the Defense Advanced Research Projects Agency (DOD) and monitored by the Avionics Laboratory, Air Force Wright Aeronautical Laboratories, Aeronautical Systems Division (AFSC), Wright-Patterson AFB, Ohio 45433-6543 under Contract F33615-87-C-1499, ARPA Order No. 4976, Amendment 20. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of DARPA or the U.S. government.

sions or tracking failures, the resultant measurement matrix W is only partially filled-in. The factorization method can handle this situation by growing a partial solution obtained from an initial full submatrix into a full solution with an iterative procedure.

The rank theorem precisely captures the nature of the redundancy that exists in an image sequence, and permits a large number of points and frames to be processed in a conceptually simple and computationally efficient way to reduce the effects of noise. The resulting algorithm is based on the singular value decomposition, which is numerically well-behaved and stable. The robustness of the recovery algorithm in turn enables us to use an image sequence with a very short interval between frames (an *image stream*), which makes feature tracking relatively easy.

We have demonstrated the accuracy and robustness of the factorization method in a series of experiments on laboratory and outdoor sequences, with and without occlusions.

2 Relation to Previous Work

In Ullman's original proof of existence of a solution [Ull79] for the structure from motion problem under orthography, as well as in the perspective formulation in [RA79], the coordinates of feature points in the world are expressed in a world-centered system of reference. Since then, however, this choice has been replaced by most computer vision researchers with that of a camera-centered representation of shape [Pra80], [BH83], [TH84], [Adi85], [WW85], [BBM87], [HHN88], [HJ89], [Hee89], [MKS89], [SA89], [BCC90]. With this representation, the position of feature points is specified by their image coordinates and by their depths, defined as the distances between the camera center and the feature points, measured along the optical axis. Unfortunately, although a camera-centered representation simplifies the equations for perspective projection, it makes shape estimation difficult, unstable, and noise sensitive.

There are two fundamental reasons for this. First, when camera motion is small, effects of camera rotation and translation can be confused with each other: for example, small rotation about the vertical axis and small translation along the horizontal axis both generate a very similar change in an image. Any attempt to recover or differentiate between these two motions, though doable mathematically, is naturally noise sensitive. Second, the computation of shape as relative depth, for example, the height of a building as the difference of depths between the top and the bottom, is very sensitive to noise, since it is a small difference between large values. These difficulties are especially magnified when the objects are distant from the camera relative to their sizes, which is usually the case for interesting applications such as site modeling.

The factorization method we present in this paper takes advantage of the fact that both difficulties disappear when the problem is reformulated in world-centered coordinates, unlike the conventional camera-centered formulation. This new (old – in a sense) formulation links object-centered shape to image motion directly, without using retinotopic depth as an intermediate quantity, and leads to a simple and well-behaved solution. Furthermore, the mutual independence of shape and motion in world-centered coordinates makes it possible to cast the structure-from-motion problem as a factorization problem, in which a matrix representing image measurements is decomposed directly into camera motion and object shape.

We first introduced this factorization method in [TK90a, TK90b], where we treated the case of single-scanline images in a flat, two-dimensional world. In [TK91] we presented the theory for the case of arbitrary camera motion in three dimensions and full two-dimensional images. This paper extends the factorization method for dealing with feature occlusions as well as presenting more experimental results with real-world images. Debrunner and Ahuja have pursued an approach related to ours, but using a different formalism [DA90, DA91]. Assuming that motion is constant over a period, they provide both closed-form expressions for shape and motion and an incremental solution (one image at a time) for multiple motions by taking advantage of the redundancy of measurements. Boult and Brown have investigated the factorization method for multiple motions [BB91], in which they count and segment separate motions in the field of view of the camera.

3 The Factorization Method

Given an image stream, suppose that we have tracked P feature points over F frames. We then obtain trajectories of image coordinates $\{(u_{fp}, v_{fp}) \mid f = 1, \ldots, F, p = 1, \ldots, P\}$. We write the horizontal feature coordinates u_{fp} into an $F \times P$ matrix U: we use one row per frame, and one column per feature point. Similarly, an $F \times P$ matrix V is built from the vertical coordinates v_{fp}. The combined matrix of size $2F \times P$

$$W = \begin{bmatrix} U \\ \hline V \end{bmatrix}$$

is called the *measurement matrix*. The rows of the matrices U and V are then registered by subtracting from each entry the mean of the entries in the same row.

$$\begin{aligned} \tilde{u}_{fp} &= u_{fp} - a_f \\ \tilde{v}_{fp} &= v_{fp} - b_f , \end{aligned} \tag{1}$$

where

$$a_f = \frac{1}{P} \sum_{p=1}^{P} u_{fp}$$

$$b_f = \frac{1}{P} \sum_{p=1}^{P} v_{fp} .$$

This produces two new $F \times P$ matrices $\tilde{U} = [\tilde{u}_{fp}]$ and $\tilde{V} = [\tilde{v}_{fp}]$. The matrix

$$\widetilde{W} = \begin{bmatrix} \tilde{U} \\ \hline \tilde{V} \end{bmatrix}$$

is called the *registered measurement matrix*. This is the input to our factorization method.

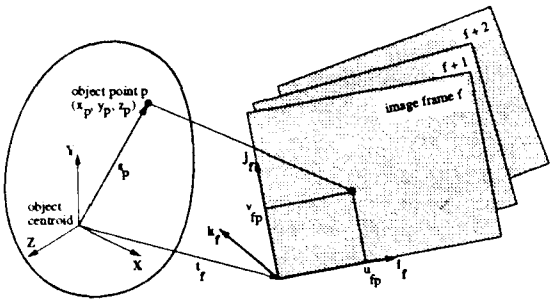


Figure 1: The systems of reference used in our problem formulation.

3.1 The Rank Theorem

We now analyze the relation between camera motion, shape, and the entries of the registered measurement matrix $\widetilde{W}$. This analysis leads to the key result that $\widetilde{W}$ is highly rank-deficient.

Referring to Figure 1, suppose we place the origin of the world reference system $x - y - z$ at the centroid of the P points $\mathbf{s}_p = (x_p, y_p, z_p)^T, p = 1, \ldots, P\}$, in space which correspond to the P feature points tracked in the image stream. The orientation of the camera reference system corresponding to frame number f is determined by a pair of unit vectors, $\mathbf{i}_f$ and $\mathbf{j}_f$, pointing along the scanlines and the columns of the image respectively, and defined with respect to the world reference system. Under orthography, all projection rays are then parallel to the cross product of $\mathbf{i}_f$ and $\mathbf{j}_f$:

$$\mathbf{k}_f = \mathbf{i}_f \times \mathbf{j}_f \, .$$

From Figure 1 we see that the projection (u_{fp}, v_{fp}), *i.e.*, the image feature position, of point $\mathbf{s}_p = (x_p, y_p, z_p)^T$ onto frame f is given by the equations

$$\begin{aligned} u_{fp} &= \mathbf{i}_f{}^T(\mathbf{s}_p - \mathbf{t}_f) \\ v_{fp} &= \mathbf{j}_f{}^T(\mathbf{s}_p - \mathbf{t}_f) \, , \end{aligned}$$

where $\mathbf{t}_f = (a_f, b_f, c_f)^T$ is the vector from the world origin to the origin of image frame f. Here note that since the origin of the world coordinates is placed at the centroid of object points,

$$\frac{1}{P}\sum_{p=1}^{P} \mathbf{s}_p = 0 \, .$$

We can now write expressions for the entries $\widetilde{u}_{fp}$ and $\widetilde{v}_{fp}$ defined in (1) of the registered measurement matrix. For the the registered horizontal image projection we have

$$\begin{aligned} \widetilde{u}_{fp} &= u_{fp} - a_f \\ &= \mathbf{i}_f{}^T(\mathbf{s}_p - \mathbf{t}_f) - \frac{1}{P}\sum_{q=1}^{P} \mathbf{i}_f{}^T(\mathbf{s}_q - \mathbf{t}_f) \\ &= \mathbf{i}_f{}^T \left(\mathbf{s}_p - \frac{1}{P}\sum_{q=1}^{P} \mathbf{s}_q \right) \\ &= \mathbf{i}_f{}^T \mathbf{s}_p \, . \end{aligned} \tag{2}$$

We can write a similar equation for $\widetilde{v}_{fp}$. To summarize,

$$\begin{aligned} \widetilde{u}_{fp} &= \mathbf{i}_f{}^T \mathbf{s}_p \\ \widetilde{v}_{fp} &= \mathbf{j}_f{}^T \mathbf{s}_p \, . \end{aligned} \tag{3}$$

Because of the two sets of $F \times P$ equations (3), the registered measurement matrix $\widetilde{W}$ can be expressed in a matrix form:

$$\widetilde{W} = RS \tag{4}$$

where

$$R = \begin{bmatrix} \mathbf{i}_1^T \\ \vdots \\ \mathbf{i}_F^T \\ \mathbf{j}_1^T \\ \vdots \\ \mathbf{j}_F^T \end{bmatrix} \tag{5}$$

represents the camera rotation, and

$$S = \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_P \end{bmatrix} \tag{6}$$

is the shape matrix. In fact, the rows of R represent the orientations of the horizontal and vertical camera reference axes throughout the stream, while the columns of S are the coordinates of the P feature points with respect to their centroid.

Since R is $2F \times 3$ and S is $3 \times P$, the equation (4) implies the following.

> **Rank Theorem**: *Without noise, the registered measurement matrix $\widetilde{W}$ is at most of rank three.*

The rank theorem expresses the fact that the $2F \times P$ image measurements are highly redundant. Indeed, they could all be described concisely by giving F frame reference systems and P point coordinate vectors, if only these were known.

From the first and the last line of equation (2), the original unregistered matrix W can be written as

$$W = RS + \mathbf{t}\mathbf{e}_P^T \, , \tag{7}$$

where $\mathbf{t} = (a_1, \ldots, a_F, b_1, \ldots, b_F)^T$ is a $2F$-dimensional vector that collects the projections of camera translation along the image plane (see equation (2)), and $\mathbf{e}_P^T = (1, \ldots, 1)$ is a vector of P ones. In scalar form,

$$\begin{aligned} u_{fp} &= \mathbf{i}_f^T \mathbf{s}_p + a_f \\ v_{fp} &= \mathbf{j}_f^T \mathbf{s}_p + b_f \, . \end{aligned} \tag{8}$$

Comparing with equations (1), we see that the two components of camera translation along the image plane are simply the averages of the rows of W.

In the equations above, $\mathbf{i}_f$ and $\mathbf{j}_f$ are mutually orthogonal unit vectors, so they must satisfy the constraints

$$|\mathbf{i}_f| = |\mathbf{j}_f| = 1 \quad \text{and} \quad \mathbf{i}_f^T \mathbf{j}_f = 0 \, . \tag{9}$$

Also, the rotation matrix R is unique if the system of reference for the solution is aligned, say, with that of the first camera position, so that:

$$\mathbf{i}_1 = (1,0,0)^T \quad \text{and} \quad \mathbf{j}_1 = (0,1,0)^T \, . \tag{10}$$

The registered measurement matrix $\widetilde{W}$ must be at most of rank three without noise. When noise corrupts the images, however, $\widetilde{W}$ will not be exactly of rank 3. However, the rank theorem can be extended to the case of noisy measurements in a well-defined manner. The next subsectionintroduces the notion of approximate rank, using the concept of singular value decomposition [GR71].

3.2 Approximate Rank

Assuming[2] that $2F \geq P$, the matrix $\widetilde{W}$ can be decomposed [GR71] into a $2F \times P$ matrix O_1, a diagonal $P \times P$ matrix Σ, and a $P \times P$ matrix O_2,

$$\widetilde{W} = O_1 \Sigma O_2 \,, \tag{11}$$

such that $O_1^T O_1 = O_2^T O_2 = O_2 O_2^T = \mathcal{I}$, where $\mathcal{I}$ is the $P \times P$ identity matrix. Σ is a diagonal matrix whose diagonal entries are the *singular values* $\sigma_1 \geq \ldots \geq \sigma_P$ sorted in non-decreasing order. This is the *Singular Value Decomposition* (SVD) of the matrix $\widetilde{W}$.

Suppose that we pay attention only to the first three columns of O_1, the first 3×3 submatrix of Σ and the first three rows of O_2. If we partition the matrices O_1, Σ, and O_2 as follows:

$$\begin{aligned} O_1 &= \underbrace{\left[\; O_1' \;\middle|\; O_1'' \;\right]}_{3 \quad P-3} \Big\}2F \\ \Sigma &= \underbrace{\left[\begin{array}{c|c} \Sigma' & 0 \\ \hline 0 & \Sigma'' \end{array}\right]}_{3 \quad P-3} \begin{array}{l} \}3 \\ \}P-3 \end{array} \\ O_2 &= \underbrace{\left[\begin{array}{c} O_2' \\ \hline O_2'' \end{array}\right]}_{P} \begin{array}{l} \}3 \\ \}P-3 \end{array} \,, \end{aligned} \tag{12}$$

we have

$$O_1 \Sigma O_2 = O_1' \Sigma' O_2' + O_1'' \Sigma'' O_2'' \,.$$

Let $\widetilde{W}^*$ be the ideal registered measurement matrix, that is, the matrix we would obtain in the absence of noise. Because of the rank theorem, $\widetilde{W}^*$ has at most three non-zero singular values. Since the singular values in Σ are sorted in non-increasing order, Σ' must contain all the singular values of $\widetilde{W}^*$ that exceed the noise level. As a consequence, the term $O_1'' \Sigma'' O_2''$ must be due entirely to noise, and the best possible rank-3 approximation to the ideal registered measurement matrix $\widetilde{W}^*$ is the product:

$$W = O_1' \Sigma' O_2'$$

We can now restate our rank theorem for the case of noisy measurements.

[2]This assumption is not crucial: if $2F < P$, everything can be repeated for the transpose of $\widetilde{W}$.

Rank Theorem for Noisy Measurements: *All the shape and rotation information in $\widetilde{W}$ is contained in its three greatest singular values, together with the corresponding left and right eigenvectors.*

Now if we define

$$\begin{aligned} R &= O_1' [\Sigma']^{1/2} \\ S &= [\Sigma']^{1/2} O_2' \,, \end{aligned}$$

we can write

$$W = RS \,. \tag{13}$$

The two matrices R and S are of the same size as the desired rotation and shape matrices R and S: R is $2F \times 3$, and S is $3 \times P$. However, the decomposition (13) is not unique. In fact, if Q is *any* invertible 3×3 matrix, the matrices RQ and $Q^{-1}S$ are also a valid decomposition of W, since

$$(RQ)(Q^{-1}S) = R(QQ^{-1})S = RS = W \,.$$

Thus, R and S are in general different from R and S. A striking fact, however, is that except for noise the matrix R is a linear transformation of the true rotation matrix R, and the matrix S is a linear transformation of the true shape matrix S. Indeed, in the absence of noise, R and R both span the column space of the registered measurement matrix $\widetilde{W} = \widetilde{W}^* = W$. Since that column space is three-dimensional because of the rank theorem, R and R are different bases for the same space, and there must be a linear transformation between them.

Whether the noise level is low enough that it can be ignored at this juncture depends also on the camera motion and on shape. Notice, however, that the singular value decomposition yields sufficient information to make this decision: the requirement is that the ratio between the third and the fourth largest singular values of $\widetilde{W}$ be sufficiently large.

3.3 The Metric Constraints

We have found that the matrix R is a linear transformation of the true rotation matrix R. Likewise, S is a linear transformation of the true shape matrix S. More specifically, there exists a 3×3 matrix Q such that

$$\begin{aligned} R &= RQ \\ S &= Q^{-1} S \,. \end{aligned} \tag{14}$$

In order to find Q we observe that the rows of the true rotation matrix R are unit vectors and the first F are orthogonal to corresponding F in the second half of R. These *metric constraints* yield the over-constrained, quadratic system

$$\begin{aligned} \mathbf{i}_f{}^T Q Q^T \mathbf{i}_f &= 1 \\ \mathbf{j}_f{}^T Q Q^T \mathbf{j}_f &= 1 \\ \mathbf{i}_f{}^T Q Q^T \mathbf{j}_f &= 0 \end{aligned} \tag{15}$$

in the entries of Q. This is a simple data fitting problem which, though nonlinear, can be solved efficiently and reliably. Its solution is determined up to a rotation of the

whole reference system, since the orientation of the world reference system was arbitrary. This arbitrariness can be removed by enforcing the constraints (10), that is, selecting the $x - y$ axes of the world reference system to be parallel with those of the first frame.

3.4 Outline of the Complete Algorithm

Based on the development in the previous sections, we now have a complete algorithm for the factorization of the registered measurement matrix $\widetilde{W}$ derived from a stream of images into shape S and rotation R as defined in equations (4) - (6).

1. Compute the singular-value decomposition $\widetilde{W} = O_1 \Sigma O_2$.
2. Define $R = O_1'(\Sigma')^{1/2}$ and $S = (\Sigma')^{1/2} O_2'$, where the primes refer to the block partitioning defined in (12).
3. Compute the matrix Q in equations (14) by imposing the metric constraints (equations (15)).
4. Compute the rotation matrix R and the shape matrix S as $R = RQ$ and $S = Q^{-1}S$.
5. If desired, align the first camera reference system with the world reference system by forming the products RR_0 and $R_0^T S$, where the orthonormal matrix $R_0 = [\mathbf{i}_1\ \mathbf{j}_1\ \mathbf{k}_1]$ rotates the first camera reference system into the identity matrix.

4 Experiment

We test the factorization method with two real streams of images: one taken in a controlled laboratory environment with ground-truth motion data, and the other in an outdoor environment with a hand-held camcorder.

4.1 "Hotel" Image Stream in a Laboratory

Some frames in this stream are shown in figure 3. The images depict a small plastic model of a building. The camera is a Sony CCD camera with a 200 mm lens, and is moved by means of a high-precision positioning platform. Camera pitch, yaw, and roll around the model are all varied as shown by the dashed curves in figure 4. The translation of the camera is such as to keep the building within the field of view of the camera.

For feature tracking, we extended the Lucas-Kanade method described in [LK81] to allow also for the automatic selection of image features. The Lucas-Kanade method of tracking obtains the displacement vector of the window around a feature as the solution of a linear 2×2 equation system. As good image features we select those points for which the above equation systems are stable. The details are presented in [Tom91, TK92].

The entire set of 430 features thus selected is displayed in figure 5, overlaid on the first frame of the stream. Of these features, 42 were abandoned during tracking because their appearance changed too much. The trajectories of the remaining 388 features are used as the measurement matrix for the computation of shape and motion.

The motion recovery is precise. The plots in figure 4 compare the rotation components computed by the factorization method (solid curves) with the values measured mechanically from the mobile platform (dashed curves). The differences are magnified in figure 6. The errors are everywhere less than 0.4 degrees and on average 0.2 degrees. The computed motion follows closely also rotations with curved profiles, such as the roll profile between frames 1 and 20 (second plot in figure 4), and faithfully preserves all discontinuities in the rotational velocities: the factorization method does not smooth the results.

Between frames 60 and 80, yaw and pitch are nearly constant, and the camera merely rotates about its optical axis. That is, the motion is actually degenerate during this period, but still it has been correctly recovered. This demonstrates that the factorization method can deal without difficulty with streams that contain degenerate substreams, because the information in the stream is used *as a whole* in the method.

The shape results are evaluated qualitatively in figure 7, which shows the computed shape viewed from above. The view in figure 7 is similar to that in figure 8, included for visual comparison. Notice that the walls, the windows on the roof, and the chimneys are recovered in their correct positions.

To evaluate the shape performance quantitatively, we measured some distances on the actual house model with a ruler and compared them with the distances computed from the point coordinates in the shape results. Figure 9 shows the selected features. The diagram in figure 10 shows the distances between pairs of features measured on the actual model and those computed by the factorization method. The measured distances between the steps along the right side of the roof (7.2 mm) were obtained by measuring five steps and dividing the total distance (36 mm) by five. The differences between computed and measured results are of the order of the resolution of our ruler measurements (one millimeter).

Part of the errors in the results is due to the use of orthography as the projection model. However, it tends to be fairly small for many realistic situations. In fact, it has been shown that errors due to the orthogrphic distortion are approximately about the same percentage as the ratio of the object size in depth to the distance of the object from the camera [Tom91].

4.2 Outdoor "House" Image Stream

The factorization method has been tested with an image stream of a real building, taken with a hand-held camera. Figure 11 shows some of the 180 frames of the building stream. The overall motion covers a relatively small rotation angle, approximately 15 degrees. Outdoor images are harder to process than those produced in a controlled environment of the laboratory, because lighting changes less predictably and the motion of the camera is more dif-

ficult to control. As a consequence, features are harder to track: the images are unpredictably blurred by motion, and corrupted by vibrations of the video recorder's head, both during recording and digitization. Furthermore, the camera's jumps and jerks produce a wide range of image disparities.

The features found by the selection algorithm in the first frame are shown in figure 12. There are many false features. The reflections in the window partially visible in the top left of the image move non-rigidly. More false features can be found in the lower left corner of the picture, where the vertical bars of the handrail intersect the horizontal edges of the bricks of the wall behind. We masked away these two parts of the image from the analysis.

In total, 376 features were found by the selection algorithm and tracked. Figure 13 plots the tracks of some (60) of the features for illustration. Notice the very jagged trajectories due to the vibrating motion of the hand-held camera.

Figures 14 and 15 show a front and a top view of the building as reconstructed by the factorization method. To render these figures for display, we triangulated the computed 3D points into a set of small surface patches and mapped the pixel values in the first frame onto the resulting surface. The structure of the visible part of the building's three walls has clearly been reconstructed. In these figures, the left wall appears to bend somewhat on the right where it intersects the middle wall. This occurred because the feature selector found features along the shadow of the roof just on the right of the intersection of the two walls, rather than at the intersection itself. Thus, the appearance of a bending wall is an artifact of the triangulation done for rendering.

This experiment with an image stream taken outdoors with the jerky motion produced by a hand-held camera demonstrates that the factorization method does not require a smooth motion assumption. The identification of false features, that is, of features that do not move rigidly with respect of the environment, remains an open problem that must be solved for a fully autonomous system. An initial effort has been seen in [BB91].

5 Occlusions

In reality, as the camera moves, features can appear and disappear from the image, because of occlusions. Also, a feature tracking method will not always succeed in tracking features throughout the image stream. These phenomena are frequent enough to make a shape and motion computation method unrealistic if it cannot deal with them.

Sequences with appearing and disappearing features result in a measurement matrix W which is only partially filled in. The factorization method introduced in section3 cannot be applied directly. However, there is usually sufficient information in the stream to determine all the camera positions and all the three-dimensional feature point coordinates. If that is the case, we can not only solve the shape and motion recovery problem from the incomplete measurement matrix W, but we can even hallucinate the unknown entries of W by projecting the computed three-dimensional feature coordinates onto the computed camera positions.

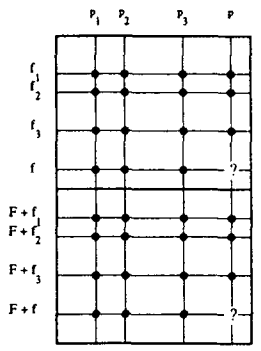


Figure 2: The Reconstruction Condition. If the dotted entries of the measurement matrix are known, the two unknown ones (question marks) can be reconstructed.

5.1 Solution for Noise-Free Images

Suppose that a feature point is not visible in a certain frame. If the same feature is seen often enough in other frames, its position in space should be recoverable. Moreover, if the frame in question includes enough other features, the corresponding camera position be recoverable as well. Then from point and camera positions thus recovered, we should also be able to reconstruct the missing image measurement. Formally, we have the following sufficient condition.

> **Condition for Reconstruction**: In the absence of noise, an unknown image measurement pair (u_{fp}, v_{fp}) in frame f can be reconstructed if point p is visible in at least three more frames f_1, f_2, f_3, and if there are at least three more points p_1, p_2, p_3 that are visible in all the four frames: the original f and the additional f_1, f_2, f_3.

Referring to Figure 2, this means that the dotted entries must be known to reconstruct the question marks. This is equivalent to Ullman's result [Ull79] that three views of four points determine structure and motion. In this subsection, we prove the reconstruction condition in our formalism and develop the reconstruction procedure. To this end, we notice that the rows and columns of the noise-free measurement matrix W can always be permuted so that $f_1 = p_1 = 1$, $f_2 = p_2 = 2$, $f_3 = p_3 = 3$, $f = p = 4$. We can therefore suppose that u_{44} and v_{44} are the only two unknown entries in the 8×4 matrix

$$W = \left[\frac{U}{V}\right] = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & ? \\ v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & ? \end{bmatrix}.$$

Then, the factorization method can be applied to the first three rows of U and V, that is, to the 6×4 submatrix

$$W_{6\times4} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \end{bmatrix} \tag{16}$$

to produce the partial translation and rotation submatrices

$$\mathbf{t}_{6\times1} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \text{and} \quad R_{6\times3} = \begin{bmatrix} \mathbf{i}_1^T \\ \mathbf{i}_2^T \\ \mathbf{i}_3^T \\ \mathbf{j}_1^T \\ \mathbf{j}_2^T \\ \mathbf{j}_3^T \end{bmatrix} \tag{17}$$

and the full shape matrix

$$S = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \mathbf{s}_3 & \mathbf{s}_4 \end{bmatrix} \tag{18}$$

such that

$$W_{6\times4} = R_{6\times3}S + \mathbf{t}_{6\times1}\mathbf{e}_4^T$$

where $\mathbf{e}_4^T = (1, 1, 1, 1)$.

To complete the rotation solution, we need to compute the vectors $\mathbf{i}_4$ and $\mathbf{j}_4$. However, a registration problem must be solved first. In fact, only three points are visible in the fourth frame, while equation (18) yields all four points in space. Since the factorization method computes the space coordinates with respect to the centroid of the points, we have $\mathbf{s}_1 + \mathbf{s}_2 + \mathbf{s}_3 + \mathbf{s}_4 = 0$, while the image coordinates in the fourth frame are measured with respect to the centroid of just three observed points (1, 2, 3). Thus, before we can compute $\mathbf{i}_4$ and $\mathbf{j}_4$ we must make the two origins coincide by referring all coordinates to the centroid

$$\mathbf{c} = \frac{1}{3}(\mathbf{s}_1 + \mathbf{s}_2 + \mathbf{s}_3)$$

of the three points that are visible in all four frames. In the fourth frame, the projection of c has coordinates

$$\begin{aligned} a_4' &= \frac{1}{3}(u_{41} + u_{42} + u_{43}) \\ b_4' &= \frac{1}{3}(v_{41} + v_{42} + v_{43}) , \end{aligned}$$

so we can define the new coordinates

$$\mathbf{s}_p' = \mathbf{s}_p - \mathbf{c} \quad \text{for} \quad p = 1, 2, 3$$

in space and

$$\begin{aligned} u_{4p}' &= u_{4p} - a_4' \\ v_{4p}' &= v_{4p} - b_4' \end{aligned} \quad \text{for} \quad p = 1, 2, 3$$

in the fourth frame. Then, $\mathbf{i}_4$ and $\mathbf{j}_4$ are the solutions of the two 3×3 systems

$$\begin{aligned} \begin{bmatrix} u_{41}' & u_{42}' & u_{43}' \end{bmatrix} &= \mathbf{i}_4^T \begin{bmatrix} \mathbf{s}_1' & \mathbf{s}_2' & \mathbf{s}_3' \end{bmatrix} \\ \begin{bmatrix} v_{41}' & v_{42}' & v_{43}' \end{bmatrix} &= \mathbf{j}_4^T \begin{bmatrix} \mathbf{s}_1' & \mathbf{s}_2' & \mathbf{s}_3' \end{bmatrix} \end{aligned} \tag{19}$$

derived from equation (4). The second equation in (17) and the solution to (19) yield the entire rotation matrix R, while shape is given by equation (18).

The components a_4 and b_4 of translation in the fourth frame with respect to the centroid of all four points can be computed by postmultiplying equation (7) by the vector $\eta_4 = (1, 1, 1, 0)^T$:

$$W\eta_4 = RS\eta_4 + \mathbf{t}\mathbf{e}_4^T\eta_4 .$$

Since $\mathbf{e}_4^T\eta_4 = 3$, we obtain

$$\mathbf{t} = \frac{1}{3}(W - RS)\eta_4 . \tag{20}$$

In particular, rows 4 and 8 of this equation yield a_4 and b_4. Notice that the unknown entries u_{44} and v_{44} are multiplied by zeros in equation (20).

Now that both motion and shape are known, the missing entries u_{44}, v_{44} of the measurement matrix W can be found by orthographic projection (equation (8)):

$$\begin{aligned} u_{44} &= \mathbf{i}_4^T\mathbf{s}_4 + a_4 \\ v_{44} &= \mathbf{j}_4^T\mathbf{s}_4 + b_4 . \end{aligned}$$

The procedure thus completed factors the full 6×4 submatrix of W and then reasons on the three points that are visible in all the frames to compute motion for the fourth frame.

Alternatively, one can start with the 8×3 submatrix

$$W_{8\times3} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \\ v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \\ v_{41} & v_{42} & v_{43} \end{bmatrix} . \tag{21}$$

In this case we first compute the full translation and rotation submatrices, and then from these we obtain the shape coordinates and the unknown entry of W for full reconstruction.

In summary, the full motion and shape solution can be found in either of the following ways:

1. *row-wise extension*: factor $W_{6\times4}$ to find a partial motion and full shape solution, and propagate it to include motion for the remaining frame (equations (19)). This will be used for reconstructing the complete W by row-wise extension.

2. *column-wise extension*: factor $W_{8\times3}$ to find a full motion and partial shape solution, and propagate it to include the remaining feature point. This will be used for reconstructing the complete W by column-wise extension.

5.2 Solution in the Presence of Noise

The solution propagation method introduced in the previous subsectioncan be extended to $2F \times P$ measurement matrices

with $F \geq 4$ and $P \geq 4$. In fact, the only difference is that the propagation equations (19) for row-wise extension and those for column-wise extension become overconstrained. If the measurement matrix W is noisy, this redundancy is beneficial, since equations (19) can be solved in the Least Square Error sense, and the effect of noise is reduced.

In the general case of a noisy $2F \times P$ matrix W the solution propagation method can be summarized as follows. A possibly large, full subblock of W is first decomposed by factorization. Then, this initial solution is grown one row or one column at a time by solving systems analogous to those in (19) in the Least Square Error sense.

However, because of noise, the order in which the rows and columns of W are incorporated into the solution can affect the exact values of the final motion and shape solution. Consequently, once the solution has been propagated to the entire measurement matrix W, it may be necessary to refine the results with a steepest-descent minimization of the residue

$$\|W - RS - \frac{1}{P}\mathbf{t}\mathbf{e}_P^T\|$$

(see equation (7)).

There remain the two problems of how to choose the initial full subblock to which factorization is applied and in what order to grow the solution. In fact, however, because of the final refinement step, neither choice is critical as long as the initial matrix is large enough to yield a good starting point. We illustrate this point in the next sectionof experiments.

6 More Experiments

We will first test the propagation method with image streams which include substantial occlusions. We first use an image stream taken in a laboratory. Then, we demonstrate the robustness of the factorization method with another stream taken with a hand-held amateur camera.

6.1 "Ping-Pong Ball" Image Stream

A ping-pong ball with black dots marked on its surface is rotated 450 degrees in front of the camera, so features appear and disappear. The rotation between adjacent frames is 2 degrees, so the stream is 226 frames long. Figure 16 shows the first frame of the stream, with the automatically selected features overlaid.

Every 30 frames (60 degrees) of rotation, the feature tracker looks for new features. In this way, features that disappear on one side around the ball are replaced by new ones that appear on the other side. Figure 17 shows the tracks of 60 features, randomly chosen among the total 829 found by the selector.

If all measurements are collected into the noisy measurement matrix W, the U and V parts of W have the same fill pattern: if the x coordinate of a measurement is known, so is the y coordinate. Figure 18 shows this *fill matrix* for our experiment. This matrix has the same size as either U or V, that is, $F \times P$. A column corresponds to a feature point, and a row to a frame. Shaded regions denote known entries. The fill matrix shown has $226 \times 829 = 187354$ entries, of which 30185 (about 16 percent) are known.

To start the motion and shape computation, the algorithm finds a large full submatrix by applying simple heuristics based on typical patterns of the fill matrix. The choice of the starting matrix is not critical, as long as it leads to a reliable initialization of the motion and shape matrices. The initial solution is then grown by repeatedly solving overconstrained versions of the linear system corresponding to (19) to add new rows, and of the system for the column-wise extension to add new columns. The rows and columns to add are selected so as to maximize the redundancy of the linear systems. Eventually, all of the motion and shape values are determined. As a result, the unknown 84 percent of the measurement matrix can be hallucinated from the known 16 percent.

Figure 19 shows two views of the final shape results, taken from the top and from the side. The missing features at the bottom of the ball in the side view correspond to the part of the ball that remained always invisible, because it rested on the rotating platform.

To display the motion results, we look at the $\mathbf{i}_f$ and $\mathbf{j}_f$ vectors directly. We recall that these unit vectors point along the rows and columns of the image frames f in $1, \ldots, F$. Because the ping-pong ball rotates around a fixed axis, both $\mathbf{i}_f$ and $\mathbf{j}_f$ should sweep a cone in space, as shown in Figure 20. The tips of $\mathbf{i}_f$ and $\mathbf{j}_f$ should describe two circles in space, centered along the axis of rotation. Figure 21 shows two views of these vector tips, from the top and from the side. Those trajectories indicate that the motion recovery was done correctly. Notice the double arc in the top part of figure 21 corresponding to more than 360 degrees rotation. If the motion reconstruction were perfect, the two arcs would be indistinguishable.

6.2 "Cup and Hand" Image Stream

In this subsectionwe describe an experiment with a natural scene including occlusion as a dominant phenomenon. A hand holds a cup and rotates it by about ninety degrees in front of the camera mounted on a fixed stand. Figure 22 shows four out of the 240 frames of the stream.

An additional need in this experiment is figure/ground segmentation. Since the camera was fixed, however, this problem is easily solved: features that do not move belong to the background. Also, the stream includes some nonrigid motion: as the hand turns, the configuration and relative position of the fingers changes slightly. This effect, however, is small and did not affect the results appreciably.

A total of 207 features was selected. Occlusions were marked by hand in this experiment. The fill matrix of figure 24 illustrates the occlusion pattern. Figure 23 shows the image trajectory of 60 randomly selected features.

Figures 25 and 26 show a front and a top view of the cup and the visible fingers as reconstructed by the propagation method. The shape of the cup was recovered, as well as the rough shape of the fingers. These renderings were obtained, as for the "House" image stream in subsection4.1,

by triangulating the tracked feature points and mapping pixel values onto the resulting surface.

7 Conclusion

The rank theorem, which is the basis of the factorization method, is both surprising and powerful. Surprising because it states that the correlation among measurements made in an image stream has a simple expression *no matter what the camera motion is and no matter what the shape of an object is*, thus making motion or surface assumptions (such as smooth, constant, linear, planar and quadratic) fundamentally superfluous. Powerful because the rank theorem leads to factorization of the measurement matrix into shape and motion in a well-behaved and stable manner.

The factorization method exploits the redundancy of the measurement matrix to counter the noise sensitivity of structure-from-motion and allows using very short inter-frame camera motion to simplify feature tracking. The structural insight into shape-from-motion afforded by the rank theorem led to a systematic procedure to solve the occlusion problem within the factorization method. The experiments in the lab demonstrate the high accuracy of the method, and the outdoor experiments show its robustness.

The rank theorem is strongly related to Ullman's twelve year old result that three pictures of four points determine structure and motion under orthography. Thus, in a sense, the theoretical foundation of our result has been around for a long time. The factorization method evolves the applicability of that foundation from mathematical images to actual noisy image streams.

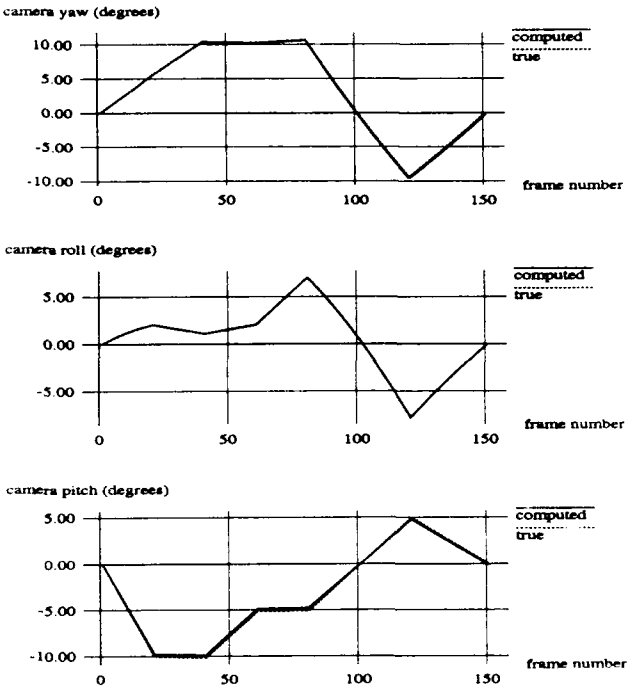


Figure 4: True and computed camera yaw, roll, pitch.

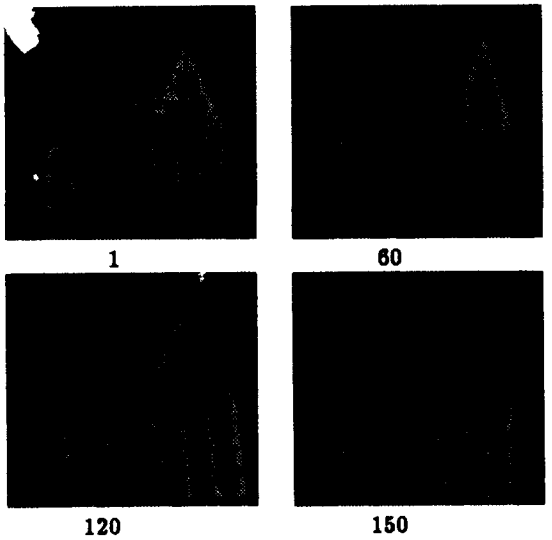


Figure 3: Some frames in the sequence. The whole sequence is 150 frames.

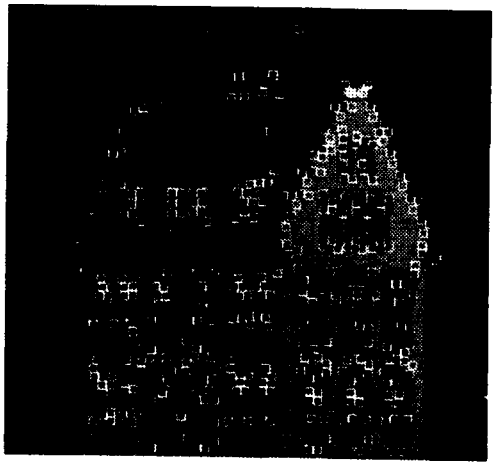


Figure 5: The 430 features selected by the automatic detection method.

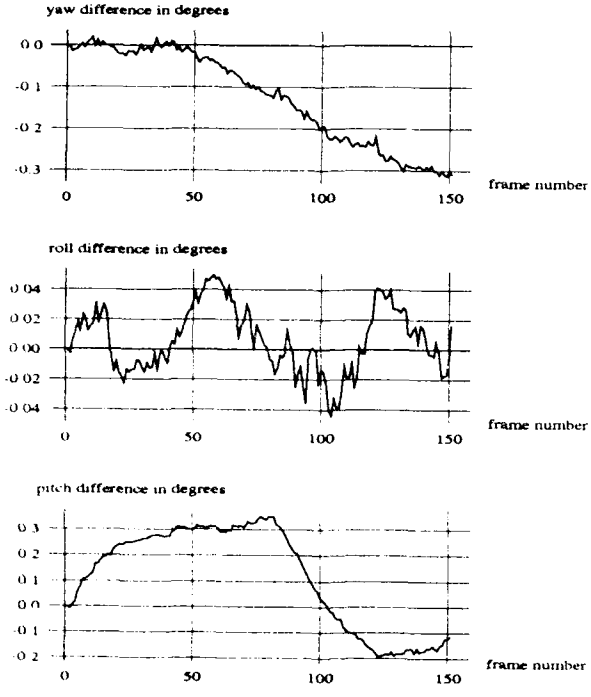


Figure 6: Blow-up of the errors in figure 4.

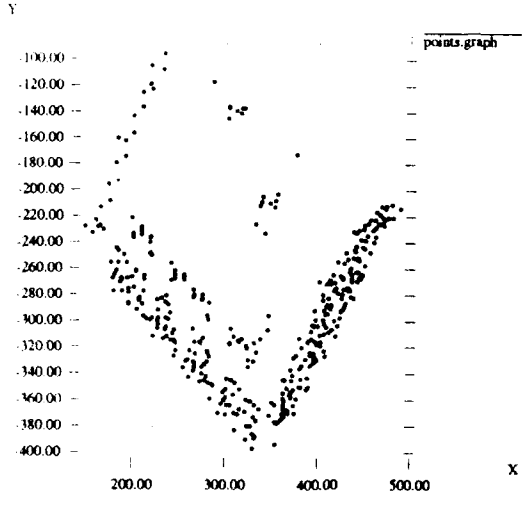


Figure 7: A view of the computed shape from approximately above the building (compare with figure 8).

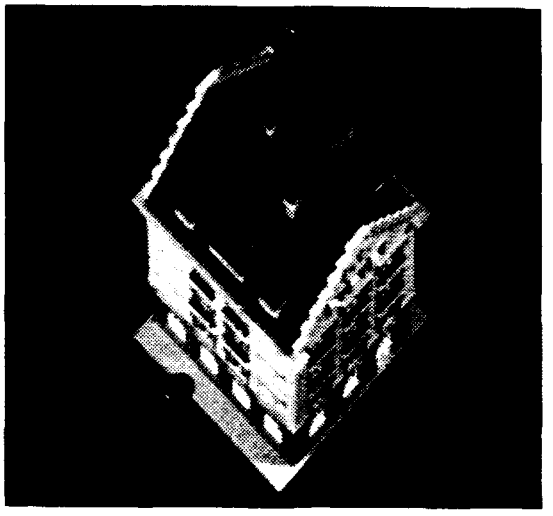


Figure 8: A real picture from above the building, similar to figure 7.

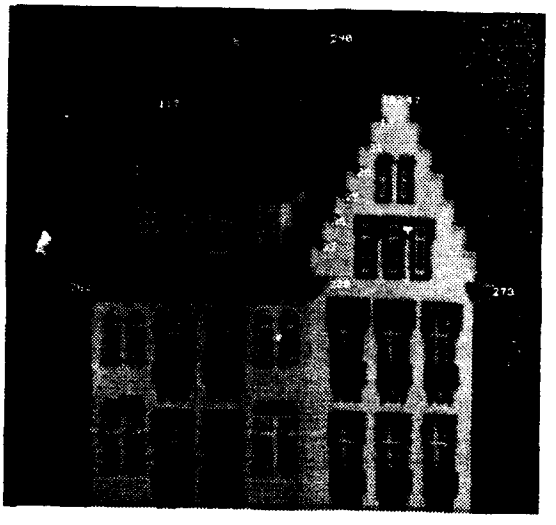


Figure 9: For a quantitative evaluation, distances between the features shown in the picture were measured on the actual model, and compared with the computed results. The comparison is shown in figure 10.

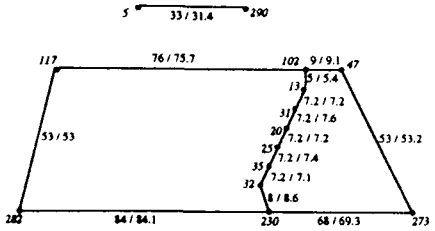


Figure 10: Comparison between measured and computed distances for the features in figure 9. The number before the slash is the measured distance, the one after is the computed distance. Lengths are in millimeters. Computed distances were scaled so that the computed distance between features 117 and 282 is the same as the measured distance.

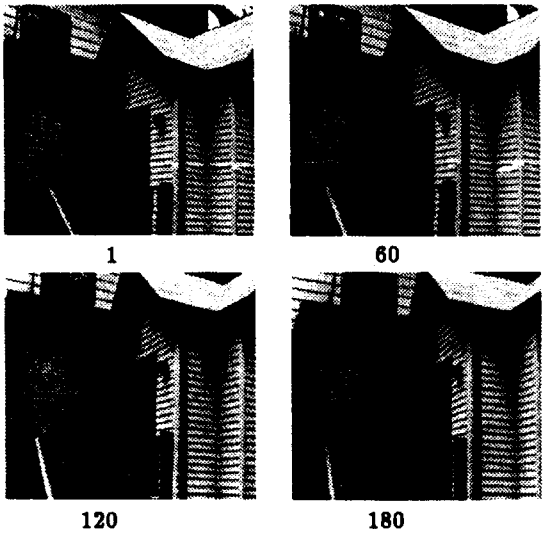


Figure 11: Four out of the 180 frames of the real house image stream.

Figure 12: The features selected in the first frame of the real house stream (figure 11)

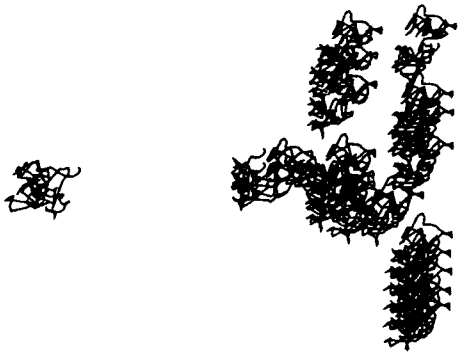


Figure 13: Tracks of 60 randomly selected features from the real house stream (figure 11.)

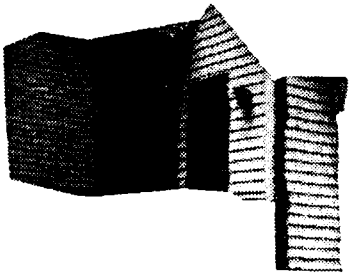


Figure 14: A front view of the three reconstructed walls, with the original image intensities mapped onto the resulting surface.

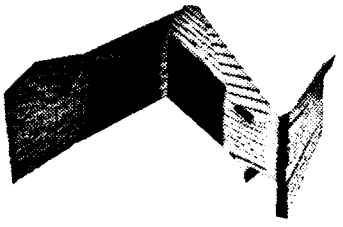


Figure 15: A view from above of the three reconstructed walls, with image intensities mapped onto the surface.

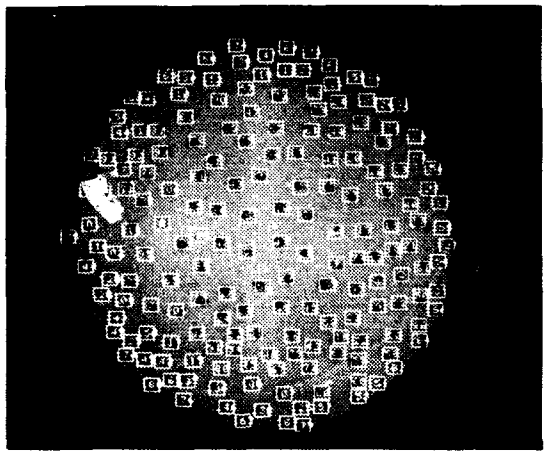


Figure 16: The first frame of the ping-pong stream, with overlaid features.

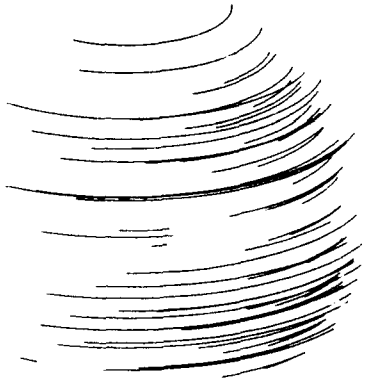


Figure 17: Tracks of 60 randomly selected features from the stream of figure 16.

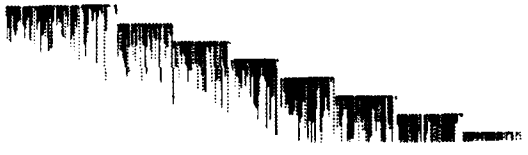


Figure 18: The fill matrix for the ping-pong ball experiment. Shaded entries are known.

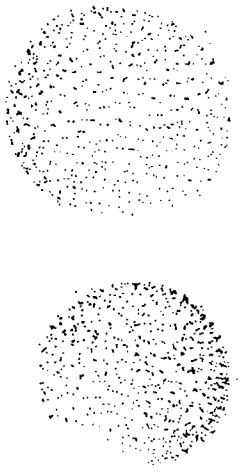


Figure 19: Top and side views of the reconstructed ping-pong ball.

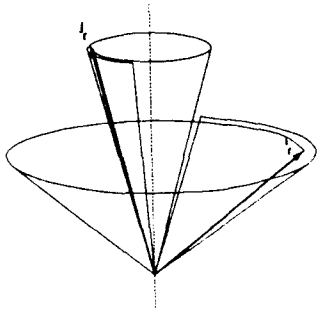


Figure 20: Rotational component of the camera motion for the ping-pong stream. Because rotation occurs around a fixed axis, the two mutually orthogonal unit vectors $\mathbf{i}_f$ and $\mathbf{j}_f$, pointing along rows and columns of the image sensor, sweep two 450-degree cones in space.

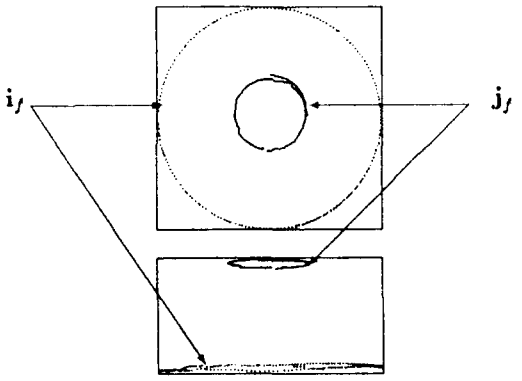


Figure 21: Top and side views of the $\mathbf{i}_f$ and $\mathbf{j}_f$ vectors identifying the camera rotation. See Figure 20.

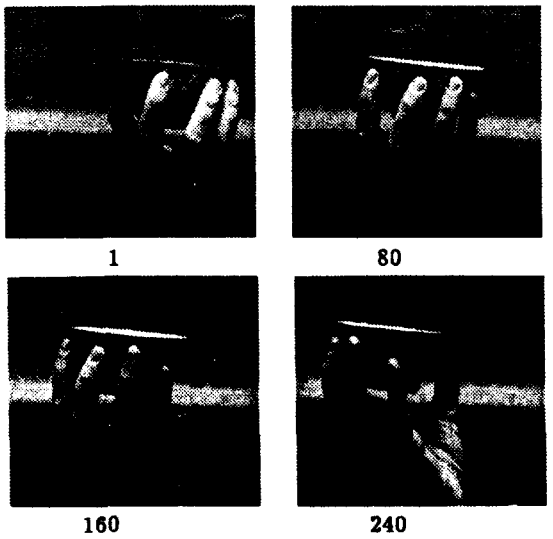


Figure 22: Four out of the 240 frames of the cup image stream.

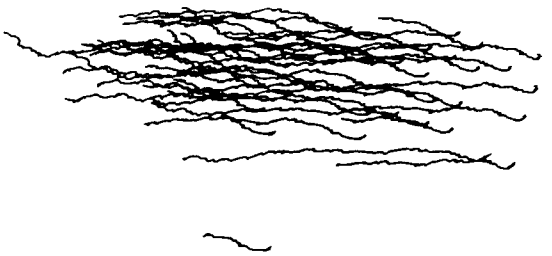


Figure 23: Tracks of 60 randomly selected features from the cup stream.

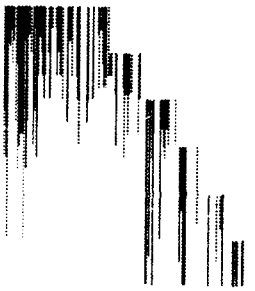


Figure 24: The 240 × 207 fill matrix for the cup stream (figure 22). Shaded entries are known.

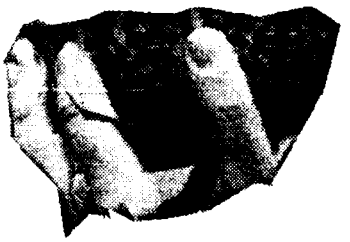


Figure 25: A front view of the cup and fingers, with the original image intensities mapped onto the resulting surface.

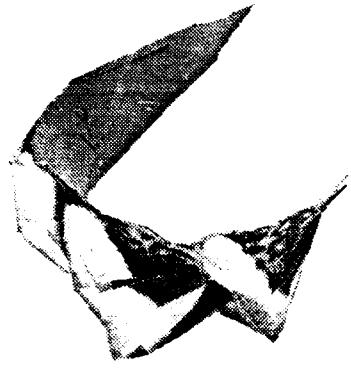


Figure 26: A view from above of the cup and fingers with image intensities mapped onto the surface.

References

[Adi85] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Pattern Analysis and Machine Intelligence*, 7:384–401, 1985.

[BB91] Terrance E. Boult and Lisa G. Brown. Factorization-based segmentation of motions. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 179–186, October 1991.

[BBM87] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[BCC90] T.J. Broida, S. Chandrashekhar, and R. Chellappa. Recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26(4):639–656, July 1990.

[BH83] A R. Bruss and B. K. P. Horn. Passive navigation. *Computer Vision, Graphics, and Image Processing*, 21:3–20, 1983.

[DA90] Christian H. Debrunner and Narendra Ahuja. A direct data approximation based motion estimation algorithm. In *Proceedings of the 10th International Conference on Pattern Recognition*, pages 384–389, Atlantic City, NJ, June 1990.

[DA91] Christian H. Debrunner and Narendra Ahuja. Motion and structure factorization and segmentation of long multiple motion image sequences. Technical Report UI-BI-CV-5-91, CSL, Univ. of Illinois, Urbana-Champaign, IL, 1991.

[GR71] G. H. Golub and C. Reinsch. *Singular Value Decomposition and Least Squares Solutions*, volume 2, chapter I/10, pages 134–151. Springer Verlag, New York, NY, 1971.

[Hee89] Joachim Heel. Dynamic motion vision. In *Proceedings of the DARPA Image Understanding Workshop*, pages 702–713, Palo Alto, Ca, May 23-26 1989.

[HHN88] Berthold K. P. Horn, Hugh M. Hilden, and Shahriar Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5(7):1127–1135, July 1988.

[HJ89] David J. Heeger and Allan Jepson. Visual perception of three-dimensional motion. Technical Report 124, MIT Media Laboratory, Cambridge, Ma, December 1989.

[LK81] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981.

[MKS89] Larry Matthies, Takeo Kanade, and Richard Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–236, September 1989.

[Pra80] K. Prazdny. Egomotion and relative depth from optical flow. *Biological Cybernetics*, 102:87–102, 1980.

[RA79] J. W. Roach and J. K. Aggarwal. Computer tracking of objects moving in space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):127–135, April 1979.

[SA89] Minas E. Spetsakis and John (Yiannis) Aloimonos. Optimal motion estimation. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 229–237, Irvine, California, March 1989.

[TH84] Roger Y. Tsai and Thomas S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces.

IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(1):13–27, January 1984.

[TK90a] Carlo Tomasi and Takeo Kanade. Shape and motion without depth. In *Proceedings of the Third International Conference in Computer Vision (ICCV)*, Osaka, Japan, December 1990.

[TK90b] Carlo Tomasi and Takeo Kanade. Shape and motion without depth. In *Proceedings of the DARPA Image Understanding Workshop*, pages 258–270, Pittsburgh, Pa, September 1990.

[TK91] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams: a factorization method - 2. point features in 3d motion. Technical Report CMU-CS-91-105, Carnegie Mellon University, Pittsburgh, PA, January 1991.

[TK92] Carlo Tomasi and Takeo Kanade. Selecting and tracking features for image sequence analysis. *submitted to Robotics and Automation*, 1992.

[Tom91] C. Tomasi. *Shape and Motion from Image Streams: a Factorization Method*. PhD thesis, CMU, September 1991.

[Ull79] Shimon Ullman. *The Interpretation of Visual Motion*. The MIT Press, Cambridge, Ma, 1979.

[WW85] A. M. Waxman and K. Wohn. Contour evolution, neighborhood deformation, and global image flow: planar surfaces in motion. *International Journal of Robotics Research*, 4:95–108, 1985.

A VLSI Smart Sensor for Fast Range Imaging [1]

We have built a range-image sensor that acquires a complete 28 × 32 range frame in as little as one millisecond. Using VLSI, sensing and processing are combined into a unique sensing element that measures range in a fully-parallel fashion. The accuracy and repeatability of the sensed data is 0.1% or better. In this paper, we review the cell-parallel method used, describe our VLSI implementation, outline procedures for calibrating the cell-parallel sensor and present some experimental results. We conclude by describing a second-generation range sensor integrated circuit which is now being tested.

[1] This research was done by Andrew Gruss, Shigeyuki Tada and Takeo Kanade, and was supported in part by an AT&T Foundation Grant, the National Science Foundation, under grant MIP-8915969, and the Defense Advanced Research Projects Agency, ARPA Order No. 7511, monitored by the NSF under grant MIP-9047590.

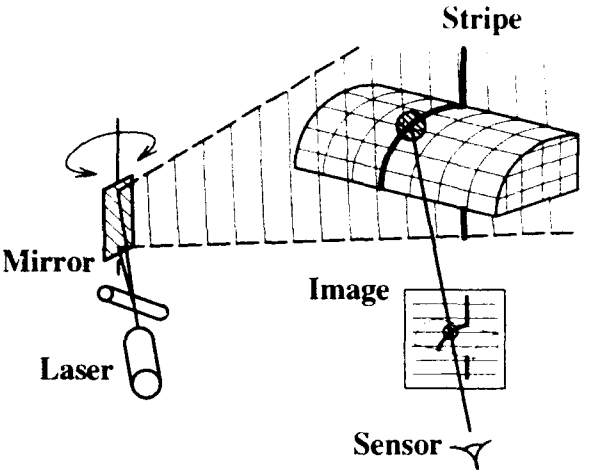


Figure 1: Traditional light-stripe range imaging.

1 Introduction

A cell-parallel implementation greatly improves the performance of a light-stripe range-imaging sensor[Gru91, KGC91, GKC91]. Though equivalent to conventional light-striping from optical and geometrical standpoints, cell-parallel light-stripe sensors incorporate a fundamental improvement in the range measurement process. As a result, the acquired range data is more robust and more accurate. Furthermore, range image acquisition time is made independent of the number of data points in each frame. By fully exploiting the capability of VLSI to both sense and process information, we have built a smart sensor that acquires a complete frame of 10-bit range image data in a millisecond.

2 A Cell-Parallel Approach to Light-Stripe Range Imaging

Range information is crucial to many robotic applications. A range image is a 2-D array of pixels, each of which represents the distance to a point in the imaged scene. Many techniques for the direct measurement of range images have been developed[Bes88]. Of these, the light-stripe methods have proven to be among the most robust and practical.

Fig. 1 illustrates the principle on which a light-stripe sensor is based. The scene to be imaged is lit by a stripe — a plane of light formed by fanning a collimated source in one dimension. The stripe is projected in a known direction using a precisely controlled mirror. When viewed by an imaging sensor, it appears as a contour which follows the profile of objects. The shape of this contour encodes range information. In particular, if projector and imaging sensor geometry are known, the distance to every point lit by the

shape to image motion directly, without using retinotopic input to our factorization method.

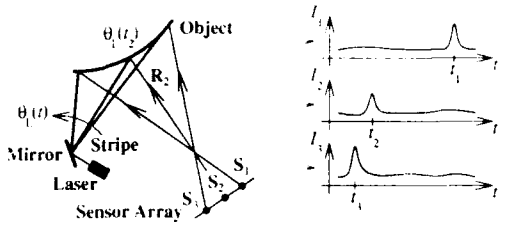


Figure 2: Cell-parallel light-stripe range imaging.

stripe can be determined via triangulation.

A conventional light-stripe range sensor builds a range image using a "step-and-repeat" procedure. A stripe is projected onto a scene, as described above, and one column of range image data is measured. The stripe is stepped to a new position and the process is repeated until the entire scene has been scanned.

Unfortunately, step-and-repeat implementations are slow. In order to build a complete range image using data from N stripe positions, N intensity images are required. The total time T_f^{Step} to acquire the range frame is

$$T_f^{\text{Step}} = N T_f^{\text{Video}} \quad (1)$$

Assuming $T_f^{\text{Video}} = 1/30$ second and $N = 100$, $T_f^{\text{Step}} = 3.3$ seconds is required.

The frame time of a step-and-repeat sensor has been improved by imposing additional structure on the light source. For example, the gray-coded sources used by Inokuchi[ISM84] reduce the factor of N in (1) to $\log_2 N$. However, achievable frame rates are still too slow and the fundamental problem remains — range frame time increases with spatial resolution.

2.1 The Cell-Parallel Method

The cell-parallel technique is an elegant modification of the basic light-stripe algorithm. The technique is a dynamic one, with time an important aspect of the range measurement process[ASP87].

Consider the geometry of a three-pixel, single-row cell-parallel range sensor, seen from above in Fig. 2. In the figure, the stripe plane is perpendicular to the page. The stripe is quickly swept across the scene from right to left, briefly illuminating object features. A sensing element, say S_2, monitors the light intensity I_2 returned to it along a fixed line-of-sight ray $\mathbf{R}_2$. When the position of the stripe is such that it intersects $\mathbf{R}_2$ at a point on the surface of an object, a "flash" will be observed by the sensing element.

Range to the object is measured by recording the *time* t_2 at which the flash is seen. The location of the stripe as a function of time is known because its projection angle $\theta_L(t)$ is controlled by the system. The "time-stamp" t_2 acquired by the sensing element measures the position of the stripe when its light is reflected back to the sensor. The three-dimensional coordinates of one object point are uniquely determined at the intersection of the line-of-sight ray $\mathbf{R}_2$ with the stripe plane at $\theta_L(t_2)$ on the surface of the object.

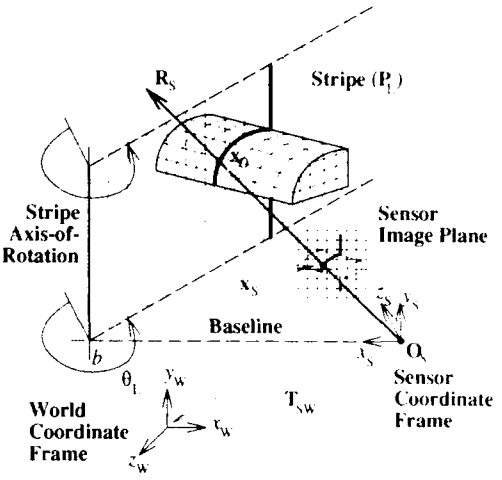


Figure 3: Cell-parallel system geometry

A sensor which collects a dense range image is formed by arranging identical sensing elements into a two-dimensional array. The cells of the array work in parallel, gathering a range image during a single pass of the light stripe. The time required to acquire the range frame is independent of its spatial resolution —

$$T_f^{\text{Cell}} = T_f^{\text{Stripe}} \quad (2)$$

The frame time T_f^{Stripe} of a cell parallel sensor is set by the bandwidth of the photo-receptor used in its sensing elements. Very high frame rates ($1/T_f^{\text{Stripe}}$) can be achieved. The photodiodes used in our cell design have bandwidth into the megahertz. They can detect a stripe moving at angular velocities in excess of 6,000 rpm.

2.2 Cell-Parallel System Geometry

Cell-parallel system geometry can be described using homogeneous coordinate transformations[BB82, NS79]. Referring to Fig. 3, the origin of the frame O_S is placed at the optical center of the imager. The stripe is a half-plane which radiates out from an axis-of-rotation aligned with the y-axis of the frame and passing through the point

$$\mathbf{x}_L = \begin{bmatrix} b & 0 & 0 & 1 \end{bmatrix}. \quad (3)$$

Stripe rotation θ_L is measured counter-clockwise about its axis when viewed from the positive y direction and defined to be zero when the stripe lies in the yz-plane. In a homogeneous representation, a plane is described in terms of a column vector $\mathbf{P}$ that satisfies the scalar product $\mathbf{xP} = 0$, where $\mathbf{x}$ is a homogeneous point that lies in $\mathbf{P}$. In the sensor coordinate frame defined above, the stripe plane is modeled

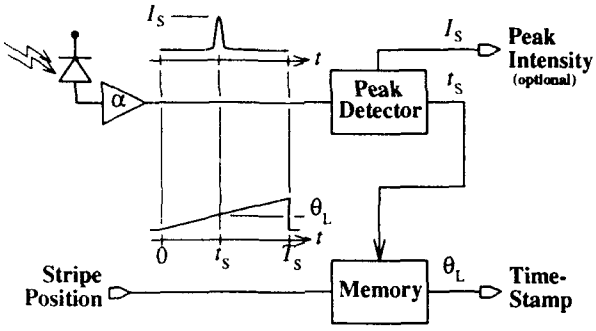


Figure 4: Basic sensing element block diagram.

in terms of b and θ_L as

$$\mathbf{P}_L = \begin{bmatrix} -\cos\theta_L \\ 0 \\ \sin\theta_L \\ b\cos\theta_L \end{bmatrix}. \quad (4)$$

The position $\mathbf{x}_S = (x_S, y_S, z_S)$ of a sensing element on the sensor image plane defines the line-of-sight ray $\mathbf{R}_S$. The parametric equation for a line in three dimensions is used to represent $\mathbf{R}_S$ as

$$\mathbf{x} = \frac{\tau}{\tau_S}(\mathbf{x}_S - \mathbf{O}_S) + \mathbf{O}_S \quad (5)$$

where $\tau_S = \|\mathbf{x}_S\| = \sqrt{x_S^2 + y_S^2 + z_S^2}$. The line parameter τ, when normalized by τ_S, is simply the distance along $\mathbf{R}_S$ measured from $\mathbf{O}_S$ heading toward the object.

The point of intersection $\mathbf{x}_O$, between the stripe and the line-of-sight, is found by solving $\mathbf{x}\mathbf{P}_L = 0$ for τ:

$$\tau = \frac{b\tau_S}{x_S - z_S\tan\theta_L}. \quad (6)$$

In the coordinate frame of the sensor, this point is

$$\mathbf{x}_O = \left[\ \frac{\tau}{\tau_S}x_S \quad \frac{\tau}{\tau_S}y_S \quad \frac{\tau}{\tau_S}z_S \quad 1\ \right]. \quad (7)$$

Thus, the 3-D position $\mathbf{x}_O$ of imaged object points can be recovered from the scalar distance measurement τ.

3 VLSI Range Sensor

A practical implementation of the cell-parallel range imaging algorithm requires a smart sensor — one in which optical sensing is local to the required processing. Silicon VLSI technology provided the means for building such a sensor.

Fig. 4 summarizes the operation of elements in the smart cell-parallel sensor array. Functionally, each must convert light energy into an analog voltage, determine the time at which the voltage peaks and remember the time at which the peak occurred.

3.1 A 28 × 32 Cell-Parallel Sensor Chip

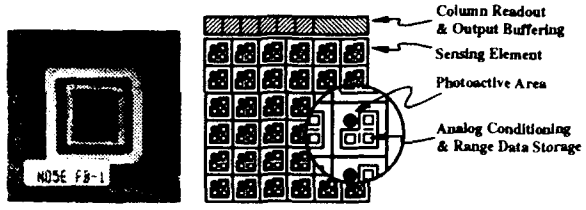


Figure 5: Range sensor integrated circuit.

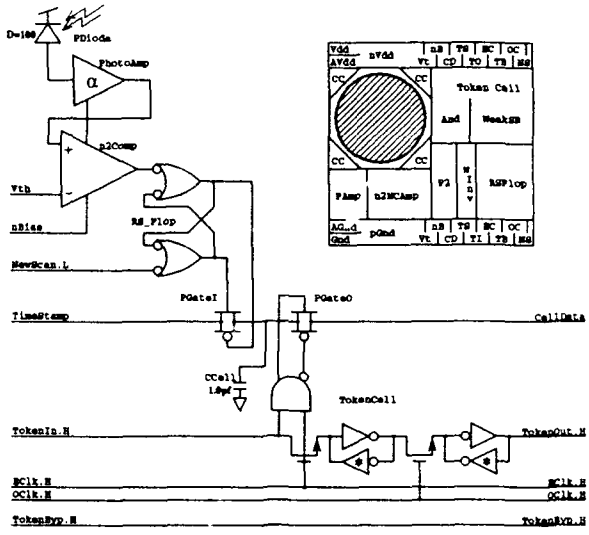


Figure 6: Sensing element circuitry.

The multi-pixel cell-parallel range sensor we have developed is shown in Fig. 5. This chip consists of 896 sensing elements arranged in a 28 × 32 array. It was fabricated using a 2 μm p-well CMOS, double-metal, double-poly process and measures 9.2 mm × 7.9 mm (width × height). Of the total 73 mm^2 chip area, the sensing element array takes up 59 mm^2, read-out column-select circuitry 0.37 mm^2 and the output integrator 0.06 mm^2. The remaining 14 mm^2 is used for power bussing, signal wiring, and die pad sites.

3.2 Sensing Element Design

The architecture chosen for the range sensing elements is shown in Fig. 6. Areas of interest in the diagram include the photo-receptor (`PDiode`), the photo-current transimpedance amplifier (`PhotoAmp`), threshold comparison stage (`n2Comp`), stripe event memory (`RS_Flop`), time-stamp track-and-hold circuitry (`PGateI/CCell`) and cell read-out logic (`PGateO/TokenCell`).

In operation, sensing elements cycle between two phases — *acquisition* and *read out*.

During the acquisition phase, each sensing element implements the cell-parallel procedure of Fig. 4. The photodiode within a cell monitors light energy reflected back from the scene. Photocurrent output is amplified and continuously compared to an external threshold voltage `Vth`. When photoreceptor output exceeds this threshold, the "stripe-detected" latch in the cell is tripped. The value of the

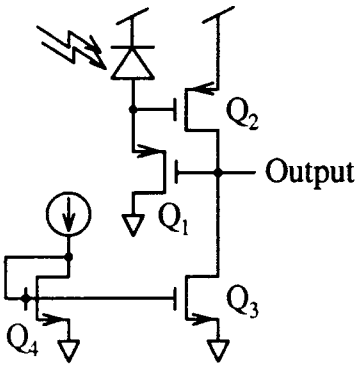


Figure 7: Non-linear transimpedance amplifier.

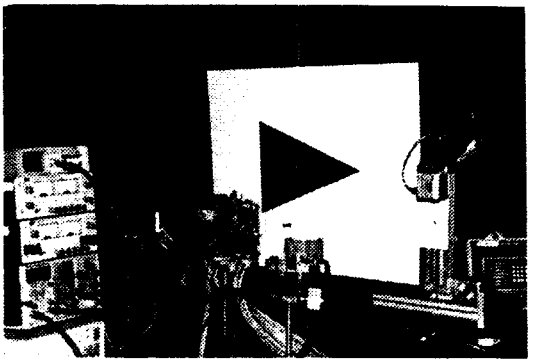


Figure 8: The cell-parallel range-finding system.

time-stamp voltage at that instant is held on the capacitor `CCell`, recording the time of the stripe detection.

The acquisition phase is synchronized with stripe motion and ends when the stripe completes its scan. At that time, the array sensing elements recorded a range image in the form of held time-stamp values. This raw range data must now be read from the chip.

A time-multiplexed read-out scheme off loads range image data in raster order through a single chip pin. One bit of token state is passed through the sensing element array, selecting cells for output. Dual n/p-transistor pass gate structures are used throughout the time-stamp data path. They permit the use of rail-to-rail time-stamp voltages, maximizing the dynamic range of the analog time-stamp data.

3.3 Stripe Detection

One of the more challenging aspects of the cell design involved the circuitry which detected the stripe.

A photodiode forms the light sensitive area within each cell. This diode is a vertical structure, built using the n-substrate as the cathode and the p-well of the CMOS process as the anode. An additional p^+ implant, driven into the well, reduces the surface resistivity of the anode and increases the device bandwidth.

The non-linear transimpedance amplifier of Fig. 7 was a key element of the sensor cell design. Reflected light from the swept stripe source generates nano-amp photo-current pulses and thus a very high-gain amplifier is required to convert this current into a usable voltage. In addition, very little die area could be devoted to photo-current amplification if cell area was to be kept small. The three transistor amplifier design of Fig. 7 satisfies both requirements. Its logarithmic transfer characteristic provides freedom from output saturation even when input light levels vary over several orders of magnitude. The output rise-time of photodiode/amplifier test structures in response to a stripe was measured to be a few microseconds.

3.4 Analog Signal Processing

Analog signal processing techniques played an important role in the design of this smart sensor. As shown in Fig. 6, sensing elements use analog circuitry to amplify the photo-current, to detect the stripe and to record the per-cell time-stamp information. Stripe timing is represented in analog form as a 0-5 V sawtooth broadcast to all cells of the array. This allowed the time-stamp value to be stored as charge on the 1 pf capacitor within each cell. The digital equivalent of latching a count into a multi-bit register would be significantly larger in area and would require that the digital time-stamp counters run during the acquisition phase. Thus, analog processing kept cell area small and minimized digital switching noise during photo-current measurements in the acquisition phase.

Table 1: CELL-PARALLEL SENSOR SYSTEM SUMMARY

Baseline	300 mm	
Laser Source	Laser Diode (Collimated)	
	Wavelength	780 nm
	Output Power	30 mW
	Stripe Width	1 mm
	Stripe Spread	40° (3 dB)
Sweep Assembly	Rotating Mirror	
	Sweep Angle	40°
Sensor Optics	1/2"-Format CCD Zoom Lens	
	Focal Length	12.5 to 75 mm
	f-number	$f/1.8$
A/D Precision	12 bits	

4 Prototype Range Image Sensor

The 28 × 32 element VLSI sensor prototype described in the previous section was incorporated into the light-stripe range system shown in Fig. 8. System components visible in the photograph include (from the left) the stripe generation assembly, the VLSI sensor chip and its interface electronics, a calibration target and the 3-DOF positioning system. Table 1 provides details of the configuration shown.

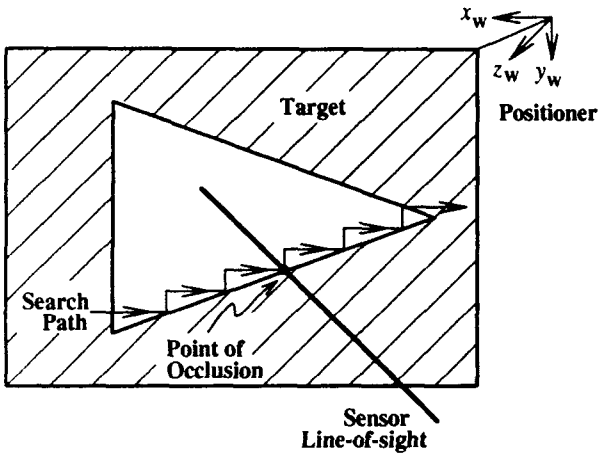


Figure 9: Line-of-sight measurement.

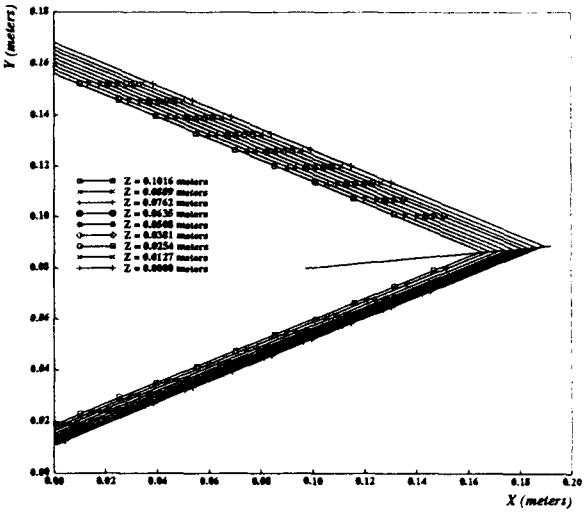


Figure 10: Cell (13,15) measured line of sight.

5 Cell-Parallel Sensor Calibration

Calibration provides the complete specification of system geometry necessary for converting cell time-stamp data into range images. Two sets of calibration parameters must be measured. First, 3-D sensor chip geometry and optical parameters must be measured — the *imager model*. Next, a mapping between time-stamp values θ_S and distance r for all sensing elements is developed — the *stripe model*.

5.1 Imager Model Calibration

This method measures component model geometry using reference objects, manipulated in the sensor's field of view with an accurate 3-DOF (degree of freedom) positioning device. The following two-step procedure is used (Fig. 3):

- the line-of-sight rays $\mathbf{R}_S$ for a few cells are measured, and
- a pinhole-camera model is fit to measured line-of-sight rays in order to approximate line-of-sights for all sensing elements.

A planer target out of which a triangular hole has been cut as shown in Fig. 9 is used to map out sensing element line-of-sight rays. The target is mounted on the positioner so that its surface is parallel to the world-xy plane.

A single 3-D point on the line-of-sight of a particular sensing element is found as follows. The target is moved to some z-position in world coordinates and held. The bottom edge of the triangular hole is located by moving the target around in x and y as indicated in Fig. 9. When a small motion in either x or y causes a large change in the time-stamp value reported by the cell, occlusion of the line-of-sight at an edge of the triangular cut is indicated.

Once many points along the bottom edge are located, a line, known to lie in the plane of the target, is fit. The location of the top edge is found in a similar fashion. The intersection of the top and bottom edge lines define one 3-D point that lies on the cell's line-of-sight. A number of these points are located by moving the target in z and repeating the process. The line-of-sight for a single cell can then be identified by fitting a 3-D line to these points. Experimental data from the calibration of one sensing element's line-of-sight is shown in Fig. 10.

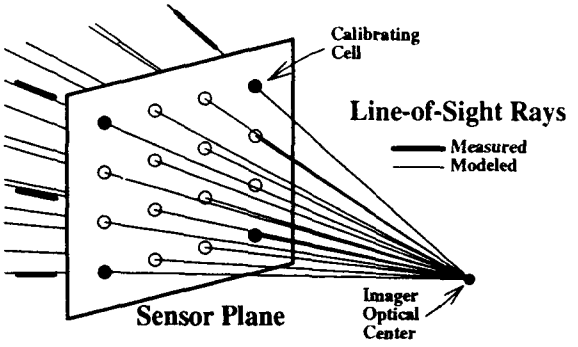


Figure 11: "Pinhole" line-of-sight approximation.

Mapping the line-of-sight rays for all 896 sensing elements in this manner is too time consuming. In practice, line-of-sight information is measured for 25 cells, evenly spaced in a 5 grid. The geometry of the remaining cells is approximated using a pinhole-camera model.

The pinhole-camera model[WCH90] constrains all sensing element line-of-sight rays to pass through a single point focus of expansion at the optical center of the camera. Fig. 11 graphically illustrates the process. Sensing element locations are assumed to lie in some *sensor plane*, at locations evenly spaced in a 2-D grid on the plane. Eleven model parameters must be determined that identify the transformation matrix $\mathbf{T}_{SW}$ and the geometry of the the sensor plane. A least-squares procedure is used to fit pinhole-model parameters to line-of-sight information measured in the first calibration step. Imager model geometry is now fully calibrated.

5.2 Advanced Imager Model Calibration

Unfortunately, calibration of the imager model via line-of-sight measurement is not suitable for use outside of the laboratory environment. "One-at-a-time" measurement of sensing element geometry, as outlined above, is slow and cumbersome.

We are developing a faster, more precise method for imager model calibration. In this new calibration method, the 3-DOF positioning system is replaced with a liquid crystal display (LCD) mask that need only be accurately positioned along one degree of freedom. The LCD mask is used to define precise black-and-white images that are "seen" by the range sensor. The method relies on intensity image information, measuring geometry through analysis of reference object images[ABA+87].

The LCD mask is placed between a diffuse planer target and sensor chip at a known position and is backlit by shining the system stripe source on the planer target. The pattern displayed on the LCD forms a black-and-white image on the sensor. Only illuminated sensing elements will latch the stripe-detected condition (Section 3-3.2). A single-bit intensity image is derived by identifying the time-stamp output of illuminated sensing elements.

Sensing element line-of-sight geometry is found by varying the LCD mask pattern in a controlled fashion. For example, a circular pattern, whose 3-D center is known, can be projected. A calibration point is found by measuring the 2-D location of this circle's center in the intensity image returned by sensor. Additional calibration data is measured by varying the position of the circle on the LCD mask and the position of the LCD along z_S. Also, by measuring the center different radii of the circle at a fixed position, we can compensate for the low spatial resolution of the current sensor. The new sensor chip design, discussed in Section 7, returns multi-bit intensity image data which further assists imager geometry calibration.

Use of the LCD mask significantly reduces the time required to perform imager-model calibration. In the previous method, two edges of a triangular hole had to be mapped out, via accurate back-and-forth movement, in order to yield a single calibration point. In the new method, one calibration point is measured from a single LCD-generated pattern without mechanical X-Y movement. Precise calibration of the low-spatial resolution range sensor is possible because high-precision patterns are generated by the LCD mask.

The use of an LCD mask to project precise 2-D patterns has application beyond the calibration of our light-stripe range sensor. For example, this technique could be used to assist more traditional camera calibration procedures or to present training data to image-based neural net systems. LCD displays have several advantages over CRT displays for applications like these — they are fast, they are static (not refreshed), and they form images which are stable and well defined.

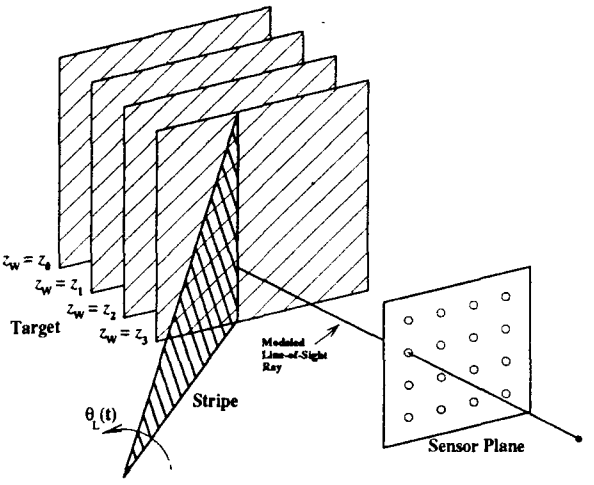


Figure 12: Time-stamp calibration.

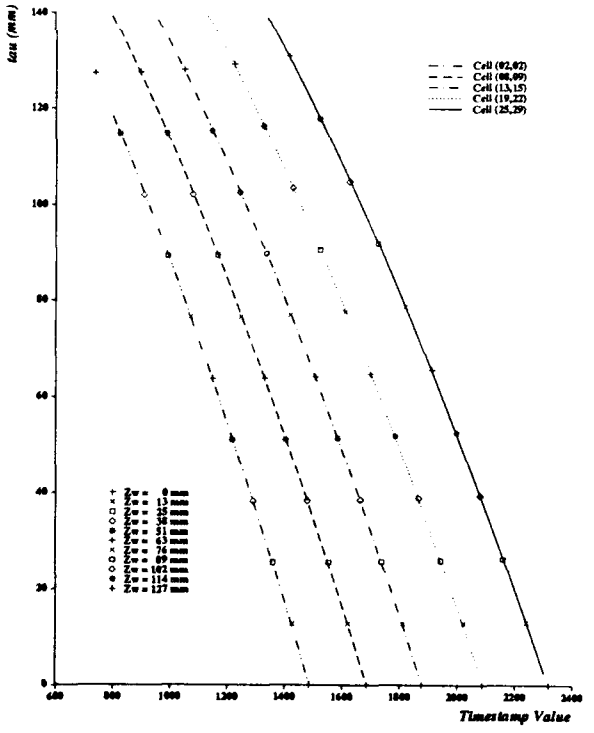


Figure 13: Time-stamp calibration result.

5.3 Stripe Model Calibration

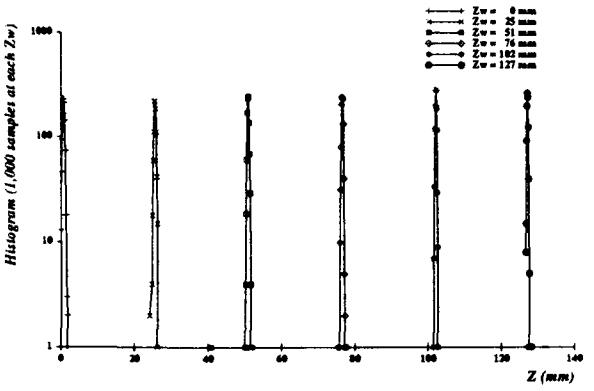


Figure 14: Cell (13,15) range-data histograms.

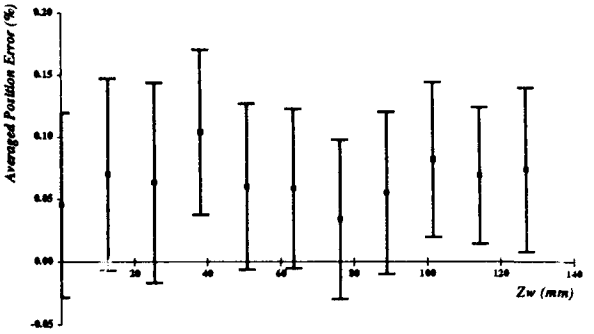


Figure 15: Range data accuracy and repeatability.

The second part of the calibration procedure determines the mapping between time-stamp data and range along all sensing element line-of-sight rays. As shown in Fig. 12, a planer target with no hole replaces the target used in step one. The new target is held at a known world-z position, parallel to the xy plane, and time-stamp readings θ_S from all sensors are recorded. This process is repeated for many z positions. Using this information, the function which maps cell time-stamp values θ_S into line-of-sight distance r for each sensing element is approximated by fitting a parabola to each. Experimental data, showing the fitted r verses θ_S functions for several sensing elements, is shown in Fig. 13. Calibration of the cell-parallel range sensor is now complete.

6 System Performance

6.1 Range Accuracy and Repeatability

The quality of the range data produced by the cell-parallel range sensor was measured by holding a planer target at a known world-z position with the 3-DOF positioning device. In the experimental setup, the world-z axis heads almost directly toward the sensor with the $z_W = 0$ point roughly 500 mm away. Analog time-stamp values from the sensor array were digitized, using a 12-bit analog-to-digital converter (A/D), and recorded for 1,000 trials. Light-stripe sweep (acquisition phase) time for each scan was 3 msec.

A histogram of the range data reported by one cell is plotted in Fig. 14. The horizontal axis represents the digitized time-stamp value, converted to world-z distance via the calibration model. Data for six world-z positions are combined in this plot. The vertical axis shows the number of times (plotted logarithmically), out of the 1,000 trials, that the sensing element reported that world-z distance. The sharpness of each peak is an indication of the stability (repeatability) of the range measurements.

Averaged statistical data for 25 evenly-spaced sensing elements is plotted in Fig. 15. In order to measure accuracy and repeatability, the position of the target, as reported by the cell-parallel sensor, is compared to the actual target z position. The "boxed" points in the plot represent the mean absolute error, expressed as a fraction of the world-z position and averaged for the 25 elements at z_W. One standard deviation of "spread", also normalized with z_W, is shown ($\pm$) above and below each box.

The experiments show the mean measured range value to be within 0.5 mm at the maximum 500 mm z — an accuracy of 0.1%. The aggregate distance discrepancy between world and measured range values remains less than 0.5 mm over the entire 360 mm to 500 mm z range. The cell-parallel sensor repeatability is found by computing the standard deviation of the distance measurements. The measured repeatability of histogram data is less than 0.5 mm — 0.1% at the maximum 500 mm positioner translation. The 0.5 mm repeatability decreases with the distance to the sensor — essentially with the slope of the time-stamp to distance mapping function (Fig. 13).

6.2 Range Image Acquisition

Fig. 16 shows a wire-frame representation of one 28 × 32 range image produced by the sensor. The imaged object is the cup shown in the figure, approximately 80 mm in diameter at its opening and 80 mm high. The range sensor is looking directly at the object from a distance of 500 mm. The viewpoint of the plot is at a point directly above the optical center of the sensor. The complete range image was acquired during a 3 msec stripe scan. The intersection points of the wire-frame plot are positioned on cell line-of-sight rays at the measured distance along the ray and the focus of expansion is located in front of the cup. Thus, the smaller "squares" represent object surface patches closer to the sensor. This is opposite the manner in which straight perspective would make an object with a grid painted on it appear, and at first glance gives the false impression that the "mold" used to make the cup has been imaged.

The curved smooth front surface of the object is clearly visible in the range data. The 20 mm handle of the cup is readily distinguished, as is the planer background behind the cup. The curved surface of the object halfway down the cup directly across from the bottom of its handle includes a slight shift of the wire-frame. The imaged cup is slightly narrower at its base by about 2 mm. The cell-parallel sensor is measuring this small 3-D feature at the 500 mm object distance.

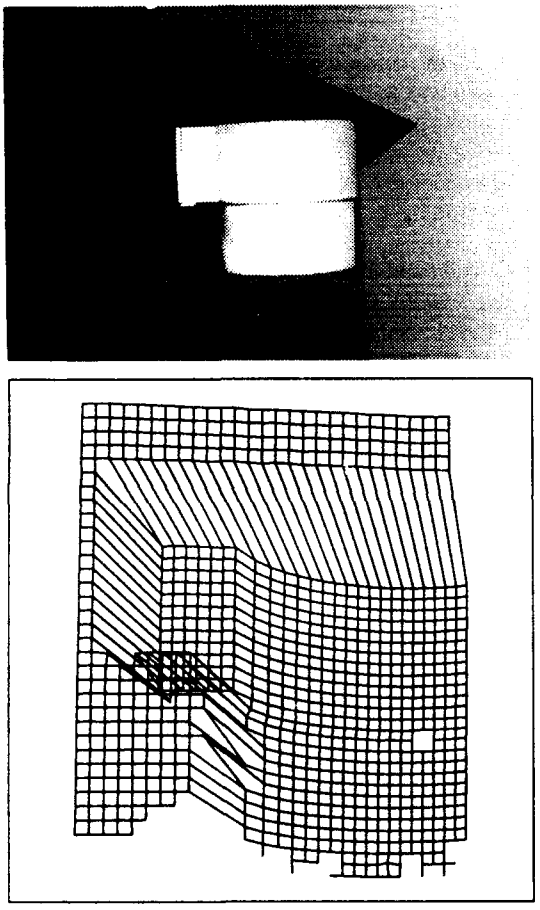


Figure 16: Range data wire frame.

Table 2: CELL-PARALLEL SENSOR PERFORMANCE SUMMARY

Spatial Resolution	28 × 32
Frame Time	Up to 1 msec
Operating Distance	350 to 500 mm
Accuracy	< 0.5 mm
Repeatability	< 0.5 mm

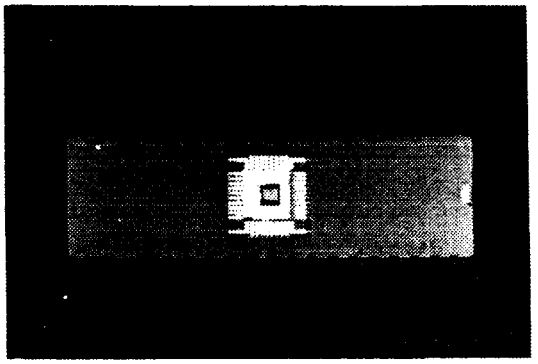


Figure 17: Second-generation range sensor integrated circuit.

6.3 Sensor Performance Summary

A summary of the cell-parallel sensor system performance is given in Table 2.

7 A Second Generation Sensing Element

A second-generation implementation of the light-stripe sensor array has been fabricated. This new chip, seen in Fig. 17, incorporates several advantages over the first design. The die area of the new cell, shown in Fig. 18, is 216 μm × 216 μm, 40% smaller than that of the cells of the first-generation sensor (photoreceptor area has been kept constant). Stripe detection is done in a more robust manner and range data read-out circuitry has been simplified. In addition, the new cell provides a means to record and read out the value of the peak intensity seen when it acquires a range data sample. The peak intensity information provides a direct measure of scene reflectance because stripe output power is known and distance to the object point is measured. In addition, the availability of intensity information allows for efficient sensor calibration (Section 5-5.2).

Peak detection is done using the circuit of Fig. 19. Operation of the circuit is straightforward. The source following transistor Q_p enables capacitor C_l to track the rising intensity input voltage transitions. No path is provided for C_p to discharge when photoreceptor output transitions downward. At the end of a scan, the largest intensity reading observed will be held. Stripe detection is easily accomplished by comparing the peak-intensity value V_f with the amplified photodiode output V_s. When V_s falls below the V_f, the output from the comparator is used to record a time-stamp

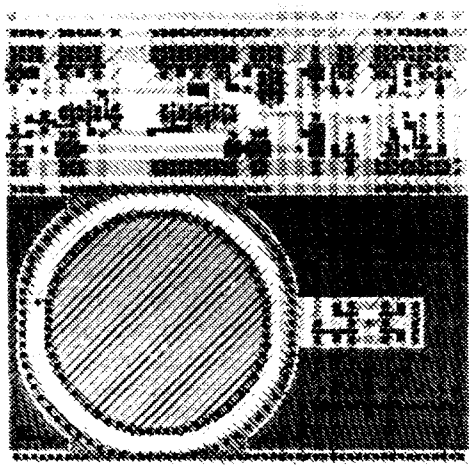


Figure 18: Second-generation sensing element layout.

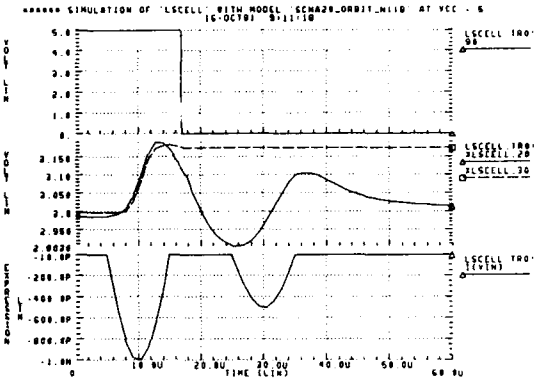


Figure 20: Second-generation sensing element simulation result.

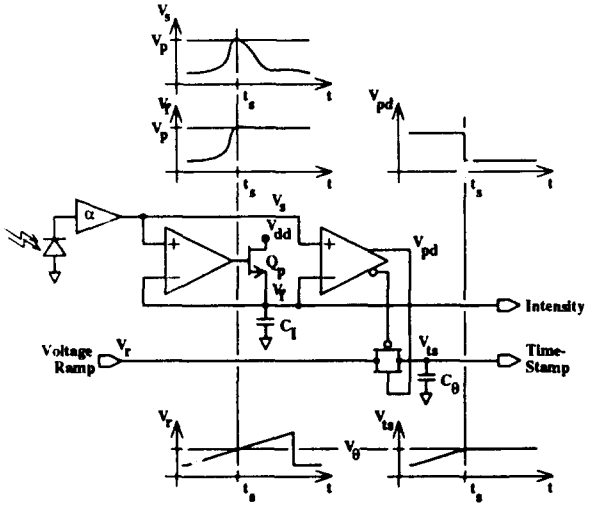


Figure 19: Second-generation sensing element circuitry.

value.

Using *Spice*[HSp90], operation of of the second-generation sensing element design was simulated. The simulation results are plotted in Fig. 20. The output from the peak-following circuit XLSCELL.30 acts as a dynamic threshold for each cell, replacing the externally applied global threshold of the first-generation design (Section 3-3.2). Comparator input offset mismatch made setting a global threshold level, valid for all cells in the array, difficult. Thus, stripe detection is made more robust by this modification. In addition, the "true" peak detection of the new design provides better quality range data because the new stripe detection scheme identifies the location of the peak in time more accurately than simple thresholding.

The peak-intensity value held within the second-generation cell is an important artifact of the ranging process and, in the new design, is provided as an additional sensing element output. The illumination source in the system, the stripe, is of known power. Intensity reduction from $1/r$-type losses can be accounted for because range to the object is measured. The intensity value therefore provides a direct measure of scene reflectance properties at the stripe wavelength. It is an image aligned perfectly with range readings from the cell array.

The area in each cell dedicated to time-stamp read out is much smaller in the new design. Direct addressing of the cell to be read, using row and column selects, eliminates the token state necessary in the first-generation design. The $N \times M$ array is read using N row select lines and M column select lines. A given cell is enabled for read out by asserting the row and column select lines that correspond to the location of the cell in the array. The two-level bus hierarchy has been maintained, however, to keep bus loading at a minimum. The area savings of the new read selection method has made cell area of the second-generation design smaller despite the additional peak detection circuitry.

8 Conclusion

We have presented the design and construction of a very high-performance range-imaging sensor. This sensor acquires a complete 28×32 range-data frame in a few milliseconds. Its range accuracy and repeatability were measured to be less than 0.5 mm on average at half-meter distances. The success of this implementation can be attributed to the use of a VLSI smart sensor methodology that allowed a practical implementation of the cell-parallel technique.

While the advantages of processing at the point sensing have been advocated by many, few practical smart-sensor implementations have been demonstrated. The cell-parallel range imager presented here bridges the gap between smart sensor theory and practice, demonstrating the impact that the smart sensor methodology can have on robotic perception systems, like automated inspection and assembly tasks.

Smart VLSI-based sensors, like the high-speed range image sensor presented here, will be key components in future industrial applications of sensor-based robotics.

References

[ABA+87] M. D. Altschuler, K. Bae, B. R. Altschuler, J. T. Dijak, L. A. Tamburino, and B. Woolford. Robot vision by encoded light beams. In Takeo Kanade, editor, *Three-Dimensional Machine Vision*. Kluwer Academic Publishers, 1987.

[ASP87] K. Araki, Y. Sato, and S. Parthasarathy. High speed rangefinder. In *Optics, Illumination, and Image Sensing for Machine Vision II*, volume 850, pages 184–188. SPIE, 1987.

[BB82] Dana H. Ballard and Christopher M. Brown. *Computer Vision*. Prentice-Hall, Inc., 1982.

[Bes88] Paul J. Besl. Range imaging sensors. Research Publication GMR-6090, General Motors Research Laboratories, March 1988.

[GKC91] A Gruss, T. Kanade, and L. R. Carley. Integrated sensor and range-finding analog signal processor. *IEEE Journal of Solid-State Circuits*, 26(3):184–191, March 1991.

[Gru91] Andrew Gruss. *A VLSI Smart Sensor for Fast Range Imaging*. PhD thesis, Carnegie Mellon University, November 1991.

[HSp90] Meta-Software, Inc., 1300 White Oaks Road, Campbell, CA 95008. *HSPICE User's Manual*, h9001 edition, 1990.

[ISM84] S. Inokuchi, K. Sato, and F. Matsuda. Range imaging system for 3-D object recognition. In *Proceedings of 7th International Conference on Pattern Recognition*, pages 806–808, Montreal, Canada, July 1984.

[KGC91] T. Kanade, A. Gruss, and L. R. Carley. A very fast VLSI rangefinder. In *Proceedings of the 1991 IEEE International Conference on Robotics and Automation*, pages 1322–29, Sacramento, CA, April 1991.

[NS79] William M. Newman and Robert F. Sproull. *Principles of Interactive Computer Graphics*. McGraw-Hill Book Company, 2nd. edition, 1979.

[WCH90] J. Weng, P. Cohen, and M. Herniou. Calibration of stereo cameras using a non-linear distortion model. In *Proceedings of the 10th International Conference on Pattern Recognition*, pages 246–253, Atlantic City, NJ, June 1990. IEEE Computer Society Press.

A Multiple-baseline Stereo Method [1]

Abstract

This paper presents a stereo matching method which uses multiple stereo pairs with various baselines to obtain precise distance estimates without suffering from ambiguity.

In stereo processing, a short baseline means that the estimated distance will be less precise due to narrow triangulation. For more precise distance estimation, a longer baseline is desired. With a longer baseline, however, a larger disparity range must be searched to find a match. As a result, matching is more difficult and there is a greater possibility of a false match. So there is a trade-off between precision and accuracy in matching.

The stereo matching method presented in this paper uses multiple stereo pairs with different baselines generated by a lateral displacement of a camera. Matching is performed simply by computing the sum of squared-difference (SSD) values. The SSD functions for individual stereo pairs are represented with respect to the inverse distance (rather than the disparity, as is usually done), and then are simply added to produce the sum of SSDs. This resulting function is

[1] This research was performed by Takeo Kanade and Masatoshi Okutomi and was supported by the Defense Advanced Research Projects Agency (DOD) and monitored by the Avionics Laboratory, Air Force Wright Aeronautical Laboratories, Aeronautical Systems Division (AFSC), Wright-Patterson AFB, Ohio 45433-6543 under Contract F33615-87-C-1499, ARPA Order No. 4976, Amendment 20. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of DARPA or the U.S. government.

called the SSSD-in-inverse-distance. We show that the SSSD-in-inverse-distance function exhibits a unique and clear minimum at the correct matching position even when the underlying intensity patterns of the scene include ambiguities or repetitive patterns. An advantage of this method is that we can eliminate false matches and increase precision without any search or sequential filtering.

This paper first defines a stereo algorithm based on the SSSD-in-inverse-distance and presents a mathematical analysis to show how the algorithm can remove ambiguity and increase precision. Then, a few experimental results with real stereo images are presented to demonstrate the effectiveness of the algorithm.

1 Introduction

Stereo is a useful technique for obtaining 3-D information from 2-D images in computer vision. In stereo matching, we measure the disparity d, which is the difference between the corresponding points of left and right images. The disparity d is related to the distance z by

$$d = BF\frac{1}{z}, \tag{1}$$

where B and F are baseline and focal length, respectively.

This equation indicates that for the same distance the disparity is proportional to the baseline, or that the baseline length B acts as a magnification factor in measuring d in order to obtain z. That is, the estimated distance is more precise if we set the two cameras farther apart from each other, which means a longer baseline. A longer baseline, however, poses its own problem. Because a longer disparity range must be searched, matching is more difficult and thus there is a greater possibility of a false match. So there is a trade-off between precision and accuracy (correctness) in matching.

One of the most common methods to deal with the problem is a coarse-to-fine control strategy [1] – [5]. Matching is done at a low resolution to reduce false matches and then the result is used to limit the search range of matching at a higher resolution, where more precise disparity measurements are calculated. Using a coarse resolution, however, does not always remove false matches. This is especially true when there is inherent ambiguity in matching, such as a repeated pattern over a large part of the scene (eg., a scene of a picket fence). Another approach to remove false matches and to increase precision is to use multiple images, especially a sequence of densely sampled images along a camera path [6] – [9]. A short baseline between a pair of consecutive images makes the matching or tracking of features easy, while the structure imposed by the camera motion allows integration of the possibly noisy individual measurements into a precise estimate. The integration has been performed either by exploiting constraints on the EPI [6][7] or by a sequential Kalman filtering technique [8][9].

The stereo matching method presented in this paper belongs to the second approach: use of multiple images with different baselines obtained by a lateral displacement of a camera. The matching technique, however, is based on the idea that global mismatches can be reduced by adding the sum of squared-difference (SSD) values from multiple stereo pairs. That is, the SSD values are computed first for each pair of stereo images. We represent the SSD values with respect to the inverse distance $1/z$ (rather than the disparity d, as is usually done). The resulting SSD functions from all stereo pairs are added together to produce the sum of SSDs, which we call SSSD-in-inverse-distance. We show that the SSSD-in-inverse-distance function exhibits a unique and clear minimum at the correct matching position even when the underlying intensity patterns of the scene include ambiguities or repetitive patterns.

There have been stereo techniques that use multiple image pairs taken by cameras which are arranged along a line [10][11][12], in the form of a triangle [13][14][15] (called trinocular stereo), or in the other formation [16]. However, all of these techniques, except [10] and [16], decide candidate points for correspondence in each image pair and then search for the correct combinations of correspondences among them using the geometrical consistencies that they must satisfy. Since the intermediate decisions on correspondences are inherently noisy, ambiguous and multiple, finding the correct combinations requires sophisticated consistency checks and search or filtering. In contrast, our method does not make any decisions about the correspondences in each stereo image pair; instead, it simply accumulates the measures of matching (SSDs) from all the stereo pairs into a single evaluation function, ie., SSSD-in-inverse-distance, and then obtains one corresponding point from it. In other words, our method integrates *evidence* for a final decision, rather than filtering intermediate *decisions*. In this sense, Tsai [16] employed strategy very similar to ours: he used multiple images to sharpen the peaks of his overall similarity measures, which he called JMM and WVM. However, the relationship between the improvement of the similarity measures and the camera baseline arrangement was not analyzed, nor was the method tested with real imagery. In this paper we show both mathematical analysis and experimental results with real indoor and outdoor images, which demonstrate how the SSSD-in-inverse-distance function based on multiple image pairs from different baselines can greatly reduce false matches, while improving precision.

In the next section we present the method mathematically and show how ambiguity can be removed and precision increased by the method. Section 3 provides a few experimental results with real stereo images to demonstrate the effectiveness of the algorithm. Section 4 presents conclusions.

2 Mathematical Analysis

The essence of stereo matching is, given a point in one image, to find in another image the corresponding point, such that the two points are the projections of the same physical point in space. This task usually requires some criterion to measure similarity between images. The sum

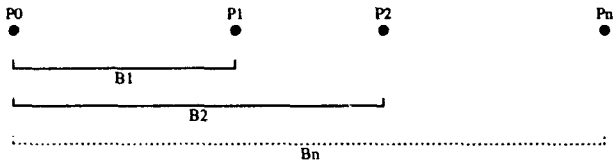


Figure 1: Camera positions for stereo

of squared differences (SSD) of the intensity values (or values of preprocessed images, such as bandpass filtered images) over a window is the simplest and most effective criterion. In this section, we define the sum of SSD with respect to the inverse distance (SSSD-in-inverse-distance) for multiple-baseline stereo, and mathematically show its advantage in removing ambiguity and increasing precision. For this analysis, we use 1-D stereo intensity signals, but the extension to two dimensional images is straightforward.

2.1 SSD Function

Suppose that we have cameras at positions $P_0, P_1, \ldots, P_n$ along a line with their optical axes perpendicular to the line and a resulting set of stereo pairs with baselines $B_1, B_2, \ldots, B_n$ as shown in figure 1. Let $f_0(x)$ and $f_i(x)$ be the image pair at the camera positions P_0 and P_i, respectively. Imagine a scene point Z whose distance is z. Its disparity $d_{r(i)}$ for the image pair taken from P_0 and P_i is

$$d_{r(i)} = \frac{B_i F}{z}. \tag{2}$$

We model the image intensity functions $f_0(x)$ and $f_i(x)$ near the matching positions for Z as

$$\begin{aligned} f_0(x) &= f(x) + n_0(x) \\ f_i(x) &= f(x - d_{r(i)}) + n_i(x), \end{aligned} \tag{3}$$

assuming constant distance near Z and independent Gaussian white noise such that

$$n_0(x), n_i(x) \sim N(0, \sigma_n^2). \tag{4}$$

The SSD value $e_{d(i)}$ over a window W at a pixel position x of image $f_0(x)$ for the candidate disparity $d_{(i)}$ is defined as

$$e_{d(i)}(x, d_{(i)}) \equiv \sum_{j \in W} (f_0(x+j) - f_i(x + d_{(i)} + j))^2, \tag{5}$$

where the $\sum_{j \in W}$ means summation over the window. The $d_{(i)}$ that gives a minimum of $e_{d(i)}(x, d_{(i)})$ is determined as the estimate of the disparity at x. Since the SSD measurement $e_{d(i)}(x, d_{(i)})$ is a random variable, we will compute its expected value in order to analyze its behavior:

$$\begin{aligned} & E[e_{d(i)}(x, d_{(i)})] \\ &= E\left[\sum_{j \in W} (f(x+j) - f(x + d_{(i)} - d_{r(i)} + j) \right. \\ &\quad \left. + n_0(x+j) - n_i(x + d_{(i)} + j))^2\right] \\ &= \sum_{j \in W} (f(x+j) - f(x + d_{(i)} - d_{r(i)} + j))^2 + 2N_w\sigma_n^2, \end{aligned} \tag{6}$$

where N_w is the number of the points within the window. For the rest of the paper, $E[\,]$ denotes the expected value of a random variable. In deriving the above equation, we have assumed that $d_{r(i)}$ is constant over the window. Equation (6) says that naturally the SSD function $e_{d(i)}(x, d_{(i)})$ is *expected* to take a minimum when $d_{(i)} = d_{r(i)}$, i.e., at the right disparity.

Let us examine how the SSD function $e_{d(i)}(x, d_{(i)})$ behaves when there is ambiguity in the underlying intensity function. Suppose that the intensity signal $f(x)$ has the same pattern around pixel positions x and $x + a$,

$$f(x+j) = f(x+a+j), \qquad j \in W \tag{7}$$

where $a \neq 0$ is a constant. Then, from equation (6)

$$E[e_{d(i)}(x, d_{r(i)})] = E[e_{d(i)}(x, d_{r(i)} + a)] = 2N_w\sigma_n^2. \tag{8}$$

This means that ambiguity is expected in matching in terms of positions of minimum SSD values. Moreover, the false match at $d_{r(i)} + a$ appears in exactly the same way for all i; it is separated from the correct match by a for all the stereo pairs. Using multiple baselines does not help to disambiguate.

2.2 SSD with respect to Inverse Distance

Now, let us introduce the *inverse distance* ζ such that

$$\zeta = \frac{1}{z}. \tag{9}$$

>From equation and (2),

$$d_{r(i)} = B_i F \zeta_r \tag{10}$$

$$d_{(i)} = B_i F \zeta, \tag{11}$$

where ζ_r and ζ are the real and the candidate inverse distance, respectively. Substituting equation (11) into (5), we have the SSD with respect to the inverse distance,

$$e_{\zeta(i)}(x, \zeta) \equiv \sum_{j \in W} (f_0(x+j) - f_i(x + B_i F \zeta + j))^2, \tag{12}$$

at position x for a candidate inverse distance ζ. Its expected value is

$$E[e_{\zeta(i)}(x, \zeta)] = \sum_{j \in W} (f(x+j) - f(x + B_i F(\zeta - \zeta_r) + j))^2 + 2N_w\sigma_{n_i}^2. \tag{13}$$

Finally, we define a new evaluation function $e_{\zeta(12\cdots n)}(x, \zeta)$, the sum of SSD functions with respect to the inverse distance (SSSD-in-inverse-distance) for multiple stereo pairs. It is obtained by adding the SSD functions $e_{\zeta(i)}(x, \zeta)$ for individual stereo pairs:

$$e_{\zeta(12\cdots n)}(x, \zeta) = \sum_{i=1}^{n} e_{\zeta(i)}(x, \zeta). \tag{14}$$

Its expected value is

$$E[e_{\zeta(12\cdots n)}(x, \zeta)] = \sum_{i=1}^{n} E[e_{\zeta(i)}(x, \zeta)]$$

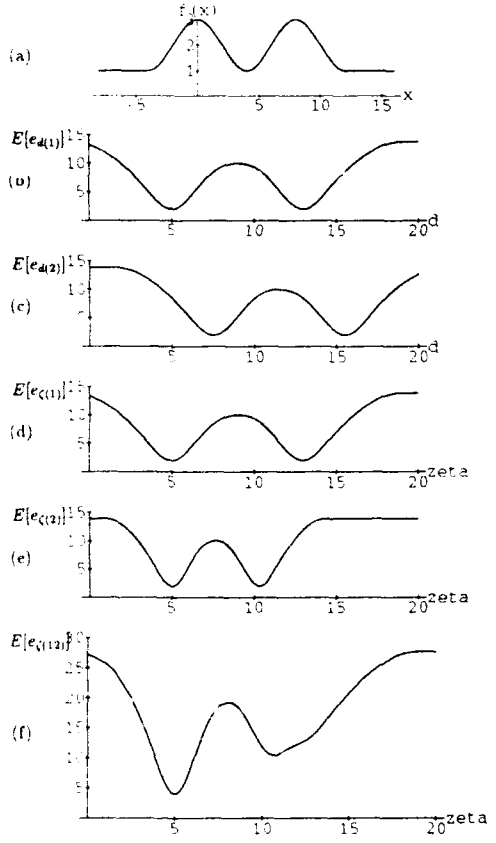


Figure 2: Expected values of evaluation functions: (a) Underlying function; (b) $E[e_{d(1)}]$; (c) $E[e_{d(2)}]$; (d) $E[e_{\zeta(1)}]$; (e) $E[e_{\zeta(2)}]$; (f) $E[e_{\zeta(12)}]$

$$= \sum_{i=1}^{n} \sum_{j \in W} (f(x+j) - f(x + B_i F(\zeta - \zeta_r) + j))^2 + 2n N_w \sigma_n^2. \quad (15)$$

In the next three subsections, we will analyze the characteristics of these evaluation functions to see how ambiguity is removed and precision is improved.

2.3 Elimination of Ambiguity (1)

As before, suppose the underlying intensity pattern $f(x)$ has the same pattern around x and $x + a$ (equation (7)). Then, according to equation (13), we have

$$E[e_{\zeta(i)}(x, \zeta_r)] = E[e_{\zeta(i)}(x, \zeta_r + \frac{a}{B_i F})] = 2N_w \sigma_n^2. \quad (16)$$

We still have an ambiguity; a minimum is expected at a false inverse distance $\zeta_f = \zeta_r + \frac{a}{B_i F}$. However, an important point to be observed here is that this minimum for the false inverse distance ζ_f changes its position as the baseline B_i changes, while the minimum for the correct inverse distance ζ_r does not. This is the property that the new evaluation function, the SSSD-in-inverse-distance (14), exploits to eliminate the ambiguity. For example, suppose we use two baselines B_1 and B_2 ($B_1 \neq B_2$). >From equation (15)

$$\begin{aligned} E[e_{\zeta(12)}(x, \zeta)] &= \sum_{j \in W} (f(x+j) - f(x + B_1 F(\zeta - \zeta_r) + j))^2 \\ &+ \sum_{j \in W} (f(x+j) - f(x + B_2 F(\zeta - \zeta_r) + j))^2 \\ &+ 4N_w \sigma_n^2. \end{aligned} \quad (17)$$

We can prove that

$$E[e_{\zeta(12)}(x, \zeta)] > 4N_w \sigma_n^2 = E[e_{\zeta(12)}(x, \zeta_r)] \quad \text{for } \zeta \neq \zeta_r. \quad (18)$$

(refer to appendix A) In words, $e_{\zeta(12)}(x, \zeta)$ is *expected* to have the smallest value at the correct ζ_r. That is, the ambiguity is likely to be eliminated by use of the new evaluation function with two different baselines.

We can illustrate this using synthesized data. Suppose the point whose distance we want to determine is at $x = 0$ and the underlying function $f(x)$ is given by

$$f(x) = \begin{cases} cos(\frac{\pi}{4}x) + 2 & \text{if } -4 < x < 12 \\ 1 & \text{if } x \leq -4 \text{ or } 12 \leq x. \end{cases} \quad (19)$$

Figure 2 (a) shows a plot of $f(x)$. Assuming that $d_{r(1)} = 5$, $\sigma_n^2 = 0.2$, and the window size is 5, the expected values of the SSD function $e_{d(1)}(x, d_{(1)})$ are as shown in figure 2 (b). We see that there is an ambiguity: the minima occur at the correct match $d_{(1)} = 5$ and at the false match $d_{(1)} = 13$. Which match will be selected will depend on the noise, search range, and search strategy. Now suppose we have a longer baseline B_2 such that $\frac{B_2}{B_1} = 1.5$. >From equations (6) and (10), we obtain $E[e_{d(2)}]$ as shown in figure 2 (c). Again we encounter an ambiguity, and the separation of the two minima is the same.

Now let us evaluate the SSD values with respect to the inverse distance ζ rather than the disparity d by using equations (12) through (15). The expected values of the SSD measurements $E[e_{\zeta(1)}]$ and $E[e_{\zeta(2)}]$ with baselines B_1 and B_2 are shown in figures 2 (d) and (e), respectively (the plot is normalized such that $B_1 F = 1$). Note that the minima at the correct inverse distance ($\zeta = 5$) does not move, while the minima for the false match changes its position as the baseline changes. When the two functions are added to produce the SSSD-in-inverse-distance, its expected values $E[e_{\zeta(12)}]$ are as shown in figure 2 (f). We can see that the ambiguity has been reduced because the SSSD-in-inverse-distance has a smaller value at the correct match position than at the false match.

2.4 Elimination of Ambiguity (2)

An extreme case of ambiguity occurs when the underlying function $f(x)$ is a periodic function, like a scene of a picket fence. We can show that this ambiguity can also be eliminated.

Let $f(x)$ be a periodic function with period T. Then, each $e_{\zeta(i)}(x, \zeta)$ is expected to be a periodic function of ζ with the period $\frac{T}{B_i F}$. This means that there will be multiple

(a) (b)

Figure 3: "Town" data set: (a) Image0; (b) Image9

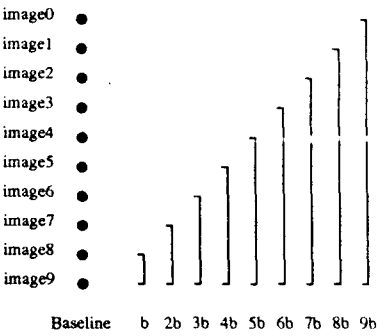


Figure 4: "Town" data set image sequence

minima of $e_{\zeta(i)}(x,\zeta)$ (i.e., ambiguity in matching) at intervals of $\frac{T}{B_iF}$ in ζ. When we use two baselines and add their SSD values, the resulting $e_{\zeta(12)}(x,\zeta)$ will be still a periodic function of ζ, but its period T_{12} is increased to

$$T_{12} = LCM\left(\frac{T}{B_1F}, \frac{T}{B_2F}\right), \qquad (20)$$

where $LCM()$ denotes Least Common Multiple. That is, the period of the expected value of the new evaluation function can be made longer than that of the individual stereo pairs. Furthermore, it can be controlled by choosing the baselines B_1 and B_2 appropriately so that the expected value of the evaluation function has only one minimum within the search range. This means that using multiple-baseline stereo pairs simultaneously can eliminate ambiguity, although each individual baseline stereo may suffer from ambiguity.

We illustrate this by using real stereo images. Figure 3(a) shows an image of a sample scene. At the top of the scene there is a grid board whose intensity function is nearly periodic. We took ten images of this scene by shifting the camera vertically as in figure 4. The actual distance between consecutive camera positions is 0.05 inches. Let this distance be b. Figure 3 shows the first and the last images of the sequence. We selected a point x within the repetitive grid board area in image9. The SSD values $e_{\zeta(i)}(x,\zeta)$ over 5-by-5-pixel windows are plotted for various baseline stereo pairs in figure 5. The horizontal axis of all the plots is the inverse distance, normalized such that $8bF = 1$. Figure 5 illustrates the trade-off between precision and ambiguity in terms of baselines. That is, for a shorter baseline, there are fewer minima (i.e. less ambiguity), but the SSD curve is flatter (i.e. less precise localization). On the other hand, for a longer baseline, there are more minima (i.e. more ambiguity), but the curve near the minimum is sharper; that is, the estimated distance is more precise if we can find the correct one.

Now, let us take two stereo image pairs: one with $B = 5b$ and the other with $B = 8b$. In figure 6, the dashed curve and the dotted curve show the SSD for $B = 5b$ and $B = 8b$, respectively. Let us suppose the search range goes from 0 to 20 in the horizontal axis, which in this case corresponds

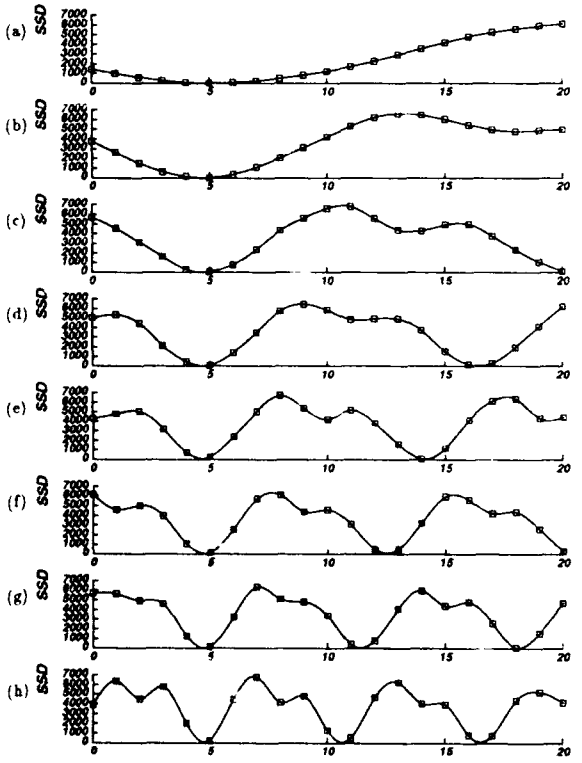


Figure 5: SSD values vs. inverse depth: (a) $B = b$; (b) $B = 2b$; (c) $B = 3b$; (d) $B = 4b$; (e) $B = 5b$; (f) $B = 6b$; (g) $B = 7b$; (h) $B = 8b$. The horizontal axis is normalized such that $8bF = 1$.

to 12 to ∞ inches in distance. Though the SSD values take a minimum at the correct answer near $\zeta = 5$, there are also other minima for both cases. The solid curve shows the evaluation function for the multiple-baseline stereo, which is the sum of the dashed curve and the dotted curve. The solid curve shows only one clear minimum; that is, the ambiguity is resolved.

So far, we have considered using only two stereo pairs. We can easily extend the idea to multiple-baseline stereo which uses more than two stereo pairs. Corresponding to

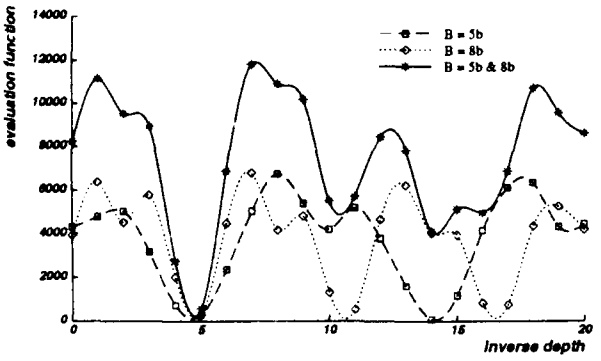


Figure 6: Combining two stereo pairs with different baselines

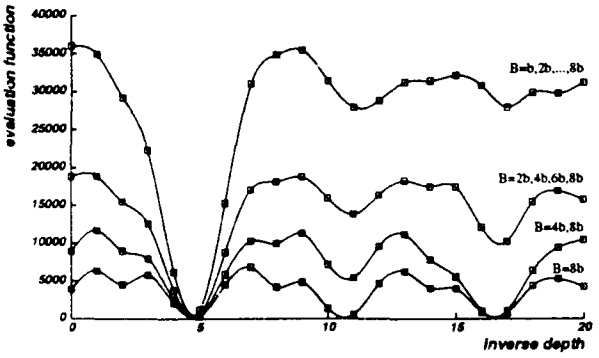


Figure 7: Combining multiple baseline stereo pairs

equation (20), the period of $E[e_{\zeta(12\cdots n)}(x,\zeta)]$ becomes

$$T_{12\cdots n} = LCM\left(\frac{T}{B_1F}, \frac{T}{B_2F}, \ldots, \frac{T}{B_nF}\right) \quad (21)$$

where $B_1, B_2, \ldots, B_n$ are baselines for each stereo pair.

We will demonstrate how the ambiguity can be further reduced by increasing the number of stereo pairs. >From the data of figure 4, we first choose image1 and image9 as a long baseline stereo pair, ie. (1) $B = 8b$. Then, we increase the number of stereo pairs by dividing the baseline between image1 and image9, i.e. (2) $B = 4b$ and $8b$, (3) $B = 2b$, $4b$, $6b$ and $8b$, (4) $B = b$, $2b$, $3b$, $4b$, $5b$, $6b$, $7b$ and $8b$. Figure 7 demonstrates that the SSSDs-in-inverse-distance shows the minimum at the correct position more clearly as more stereo pairs are used.

2.5 Precision

We have shown that ambiguities can be resolved by using the SSSD-in-inverse-distance computed from multiple baseline stereo pairs. The technique also increases precision in estimating the true inverse distance. We can show this by analyzing the statistical characteristics of the evaluation functions near the correct match.

>From equations (3), (10), and (12), we have

$$e_{\zeta(i)}(x,\zeta) = \sum_{j\in W}(f(x+j) - f(x+B_iF(\zeta-\zeta_r)+j) + n_0(x+j) - n_i(x+B_iF\zeta+j))^2. \quad (22)$$

By taking the Taylor expansion about $\zeta = \zeta_r$ up to the linear terms, we obtain

$$f(x+B_iF(\zeta-\zeta_r)+j) \approx f(x+j)+B_iF(\zeta-\zeta_r)f'(x+j). \quad (23)$$

Substituting this into equation (22), we can approximate $e_{\zeta(i)}(x,\zeta)$ near ζ_r by a quadratic form of ζ:

$$\begin{aligned} & e_{\zeta(i)}(x,\zeta) \\ \approx & \sum_{j\in W}(-B_iF(\zeta-\zeta_r)f'(x+j) \\ & +n_0(x+j) - n_i(x+B_iF\zeta+j))^2 \\ = & B_i^2F^2a(x)(\zeta-\zeta_r)^2 + 2B_iFb_i(x)(\zeta-\zeta_r) + c_i(x), \end{aligned} \quad (24)$$

where

$$a(x) = \sum_{j\in W}(f'(x+j))^2 \quad (25)$$

$$b_i(x) = \sum_{j\in W} f'(x+j)(n_i(x+B_iF\zeta+j) - n_0(x+j)) \quad (26)$$

$$c_i(x) = \sum_{j\in W}(n_i(x+B_iF\zeta+j) - n_0(x+j))^2. \quad (27)$$

The estimated inverse distance $\xi_{r(i)}$ is the value ζ that makes equation (24) minimum;

$$\xi_{r(i)} = \zeta_r - \frac{b_i(x)}{B_iFa(x)}. \quad (28)$$

Since $E[b_i(x)] = 0$, the expected value of the estimate $\xi_{r(i)}$ is the correct value ζ_r, but it varies due to the noise. The variance of this estimate is:

$$\begin{aligned} Var(\xi_{r(i)}) &= \frac{Var(b_i(x))}{B_i^2F^2(a(x))^2} \\ &= \frac{2\sigma_n^2}{B_i^2F^2a(x)}. \end{aligned} \quad (29)$$

Basically, this equation states that for the same amount of image noise σ_n^2, the variance is smaller (the estimate is more precise) as the baseline B_i is longer, or as the variation of intensity signal, $a(x)$, is larger.

We can follow the same analysis for $e_{\zeta(12\cdots n)}(x,\zeta)$ of (14), the new evaluation function with multiple baselines. Near ζ_r, it is

$$\begin{aligned} e_{\zeta(12\cdots n)}(x,\zeta) \approx & \left(\sum_{i=1}^{n}B_i^2\right)F^2a(x)(\zeta-\zeta_r)^2 \\ & +2F\left(\sum_{i=1}^{n}B_ib_i(x)\right)(\zeta-\zeta_r) + \sum_{i=1}^{n}c_i(x). \end{aligned} \quad (30)$$

The variance of the estimated inverse distance $\xi_{r(12\cdots n)}$ that minimizes this function is

$$Var(\xi_{r(12\cdots n)}) = \frac{2\sigma_n^2}{\left(\sum_{i=1}^{n}B_i^2\right)F^2a(x)}. \quad (31)$$

when its light is reflected back to the sensor. The three-dimensional coordinates of one object point are uniquely

where $\mathbf{x}$ is a homogeneous point that lies in $\mathbf{P}$. In the sensor coordinate frame defined above, the stripe plane is modeled

>From equations (29) and (31), we see that

$$\frac{1}{Var(\zeta_{r(12\cdots n)})} = \sum_{i=1}^{n} \frac{1}{Var(\zeta_{r(1i)})}. \qquad (32)$$

The inverse of the variance represents the precision of the estimate. Therefore, equation (32) means that by using the SSSD-in-inverse-distance with multiple baseline stereo pairs, the estimate becomes more precise. We can confirm this characteristic in figures 6 and 7 by observing that the curve around the correct inverse distance becomes sharper as more baselines are used.

3 Experimental Results

This section presents experimental results of the multiple-baseline stereo based on SSSD-in-inverse-distance with real 2D images. A complete description of the algorithm is included in Appendix B.

The first result is for the "Town" data set that we showed in figure 3. Figures 8 (a) and (b) are the distance map and its isometric plot with a short baseline, $B = 3b$. The result with a single long baseline, $B = 9b$, is shown in figure 9. Comparing these two results, we observe that the distance map computed by using the long baseline is smoother on flat surfaces, i.e., more precise, but has gross errors in matching at the top of the scene because of the repeated pattern. These results illustrate the trade-off between ambiguity and precision. Figure 10, on the other hand, shows the distance map and its isometric plot obtained by the new algorithm using three different baselines, $3b$, $6b$, and $9b$. For comparison, the corresponding oblique view of the scene is shown in figure 11. We can note that the computed distance map is less ambiguous *and* more precise than those of the single-baseline stereo.

Figure 12 shows another data set used for our experiment. Figures 13 and 14 compare the distance maps computed from the short baseline stereo and the long baseline stereo: the longer baseline is five times longer than the short one. For comparison, the actual oblique view roughly corresponding to the isometric plot is shown in figure 15. Though no repetitive patterns are apparent in the images, we can still observe gross errors in the distance map obtained with the long baseline due to false matching. In contrast, the result from the multiple-baseline stereo shown in figure 16 demonstrates both the advantage of unambiguous matching with a short baseline and that of precise matching with a long baseline.

4 Conclusions

In this paper, we have presented a new stereo matching method which uses multiple baseline stereo pairs. This method can overcome the trade-off between precision and accuracy (avoidance of false matches) in stereo. The method is rather straightforward: we represent the SSD values for individual stereo pairs as a function of the inverse distance, and add those functions. The resulting function, the SSSD-in-inverse-distance, exhibits an unambiguous and sharper minimum at the correct matching position. As a result there is no need for search or sequential estimation procedures.

The key idea of the method is to relate SSD values to the inverse distance rather than the disparity. As an afterthought, this idea is natural. Whereas disparity is a function of the baseline, there is only one true (inverse) distance for each pixel position for all of the stereo pairs. Therefore there must be a single minimum for the SSD values when they are summed and plotted with respect to the inverse distance. We have shown the advantage of the proposed method in removing ambiguity and improving precision by analytical and experimental results.

Acknowledgment

The authors would like to thank John Krumm for his useful comments on this paper. Keith Gremban, Jim Rehg and Carol Novak have read the manuscript and improved its readability substantially.

A SSSD-in-inverse-distance for Ambiguous Pattern

Proposition: Suppose that there are two and only two repetitions of the same pattern around positions x and $x + a$ where $a \neq 0$ is a constant. That is, for $j \in W$

$$f(x+j) = f(\xi+j), \qquad \text{if and only if } \xi = x \text{ or } \xi = x + a. \qquad (33)$$

Then, if $B_1 \neq B_2$, for $\forall \zeta, \zeta \neq \zeta_r$,

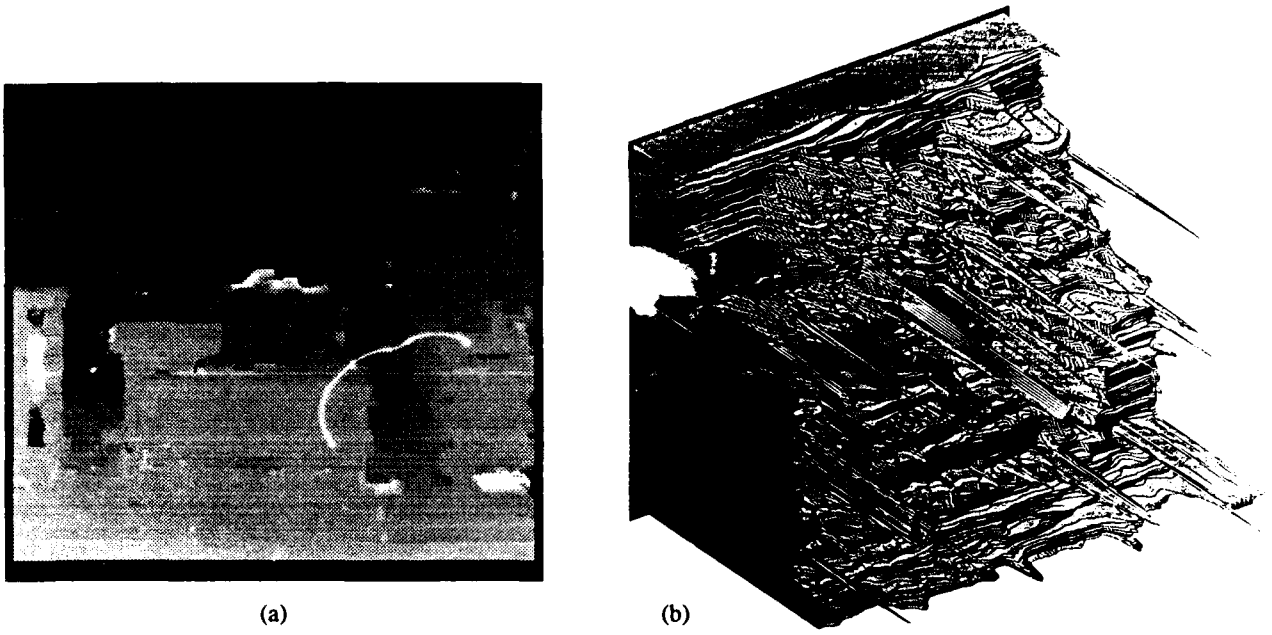
$$\begin{aligned} & E[e_{\zeta(12)}(x,\zeta)] \\ = & \sum_{j\in W} (f(x+j) - f(x + B_1F(\zeta - \zeta_r) + j))^2 \\ + & \sum_{j\in W} (f(x+j) - f(x + B_2F(\zeta - \zeta_r) + j))^2 + 4N_w\sigma_n^2 \\ > & 4N_w\sigma_n^2 = E[e_{\zeta(12)}(x,\zeta_r)]. \end{aligned}$$

Proof: Tentatively suppose that for $\exists \zeta_f, \zeta_f \neq \zeta_r$,

$$\begin{aligned} & \sum_{j\in W} (f(x+j) - f(x + B_1F(\zeta_f - \zeta_r) + j))^2 \\ + & \sum_{j\in W} (f(x+j) - f(x + B_2F(\zeta_f - \zeta_r) + j))^2 \\ = & 0. \end{aligned} \qquad (35)$$

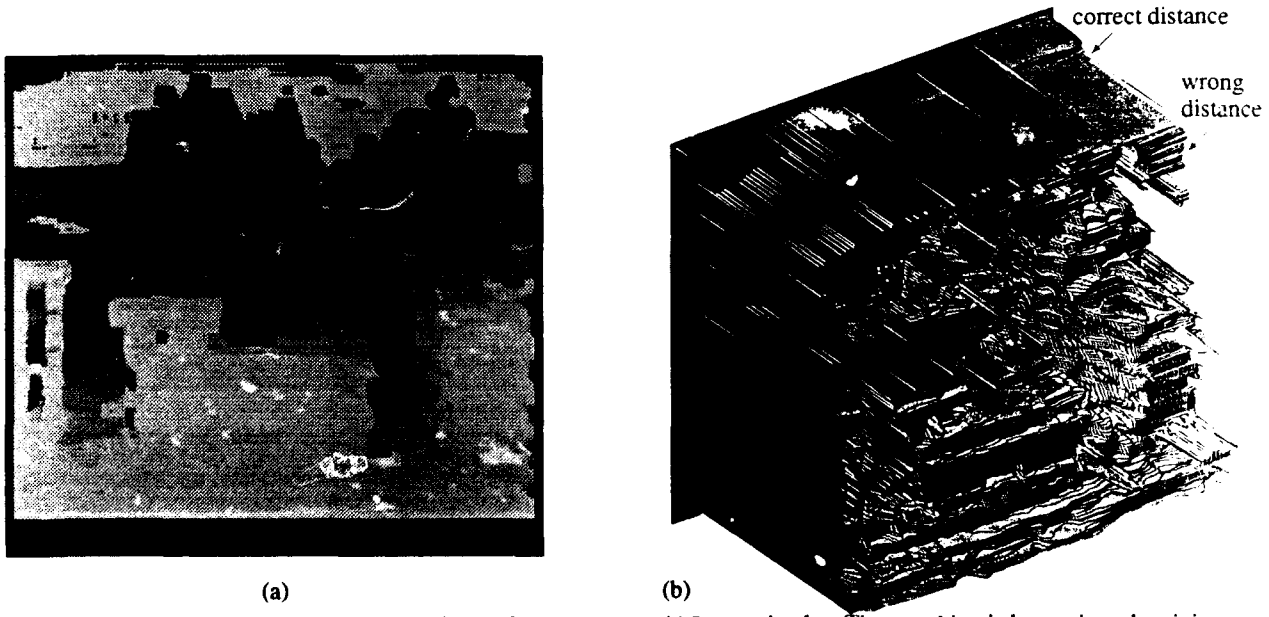
Then, it must be the case that

$$\begin{aligned} f(x+j) &= f(x+a_1+j) \\ \text{and } f(x+j) &= f(x+a_2+j). \end{aligned} \qquad (36)$$



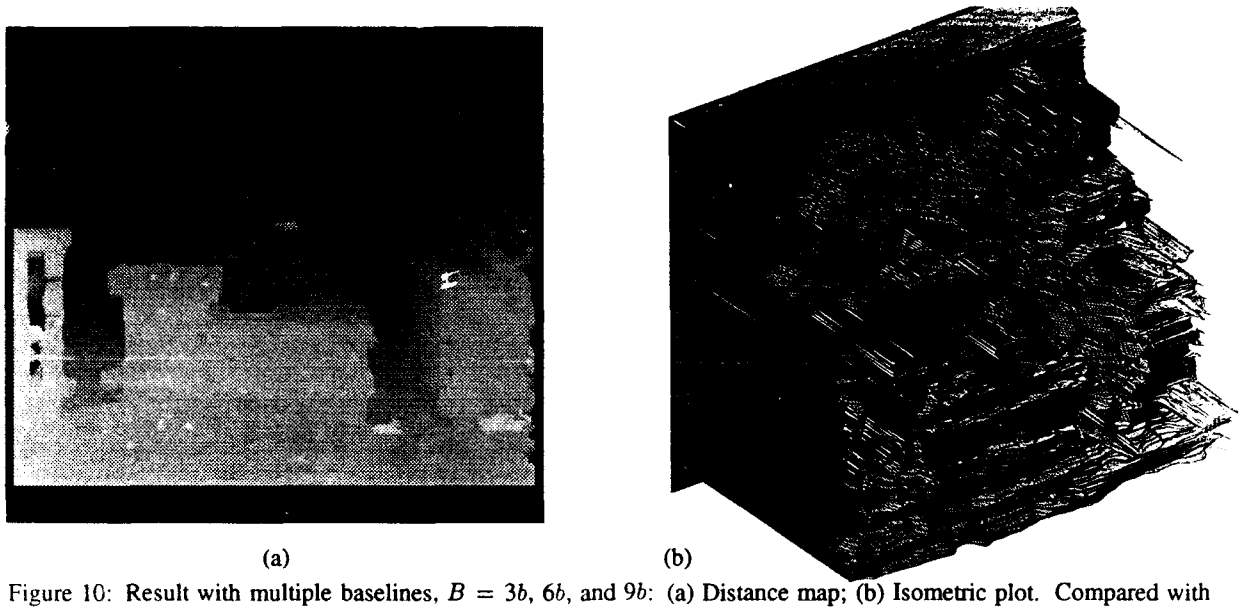
(a) (b)

Figure 8: Result with a short baseline, $B = 3b$: (a) Distance map; (b) Isometric plot of the distance map from the upper left corner. The matching is mostly correct, but very noisy.



(a) (b)

Figure 9: Result with a long baseline, $B = 9b$: (a) Distance map; (b) Isometric plot. The matching is less noisy when it is correct. However, there are many gross mistakes, especially in the top of the image where, due to a repetitive pattern, the matching is completely wrong.



(a) (b)

Figure 10: Result with multiple baselines, $B = 3b$, $6b$, and $9b$: (a) Distance map; (b) Isometric plot. Compared with figures 8(b) and 9(b), we see that the distance map is less noisy and that gross errors have been removed.

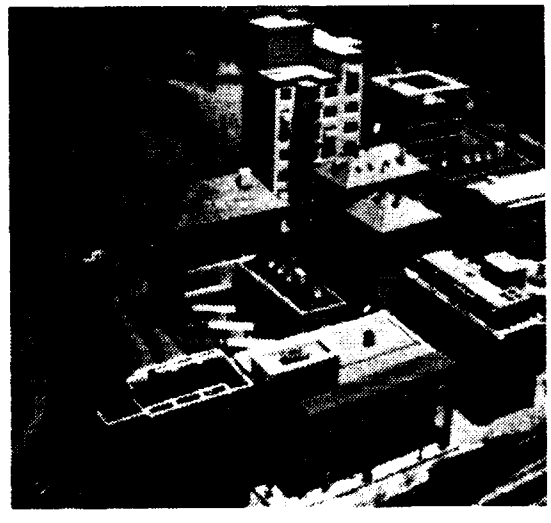
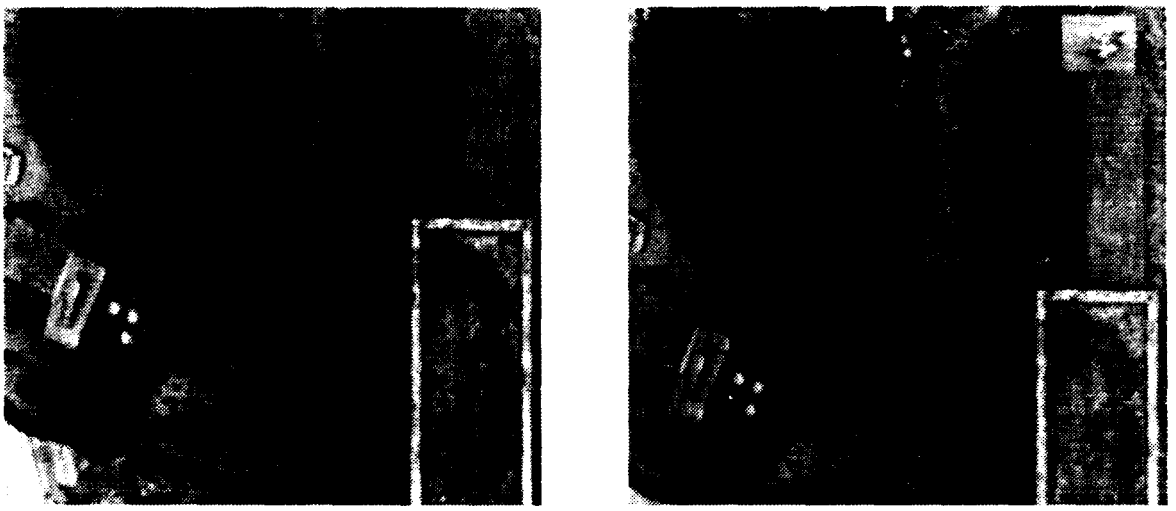
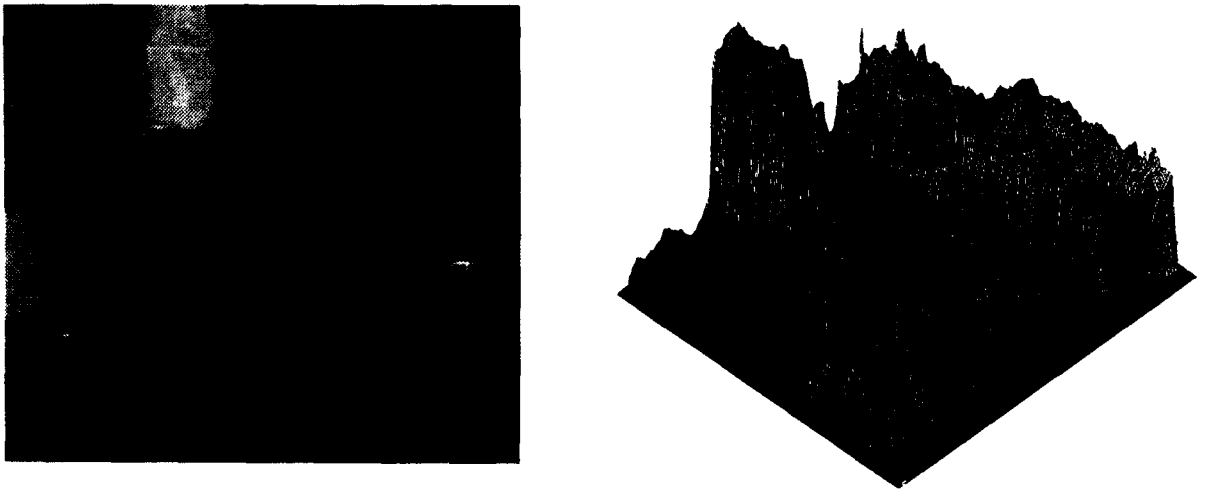


Figure 11: Oblique view



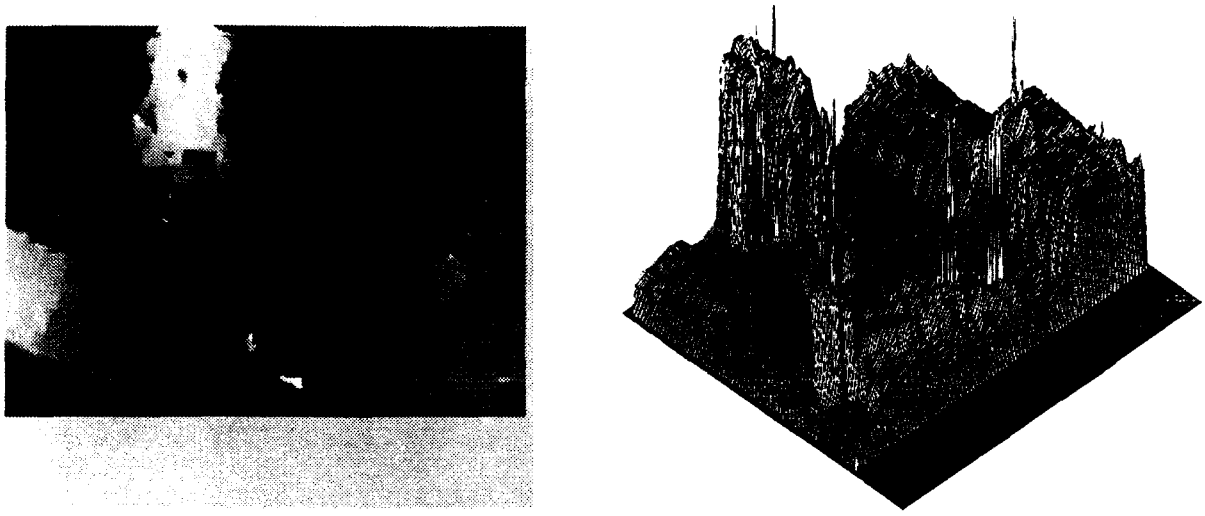
(a) (b)

Figure 12: "Coal mine" data set, long-baseline pair



(a) (b)

Figure 13: Result with a short baseline: (a) Distance map; (b) Isometric plot of the distance map viewed from the lower left corner



(a) (b)

Figure 14: Result with a long baseline: (a) Distance map; (b) Isometric plot

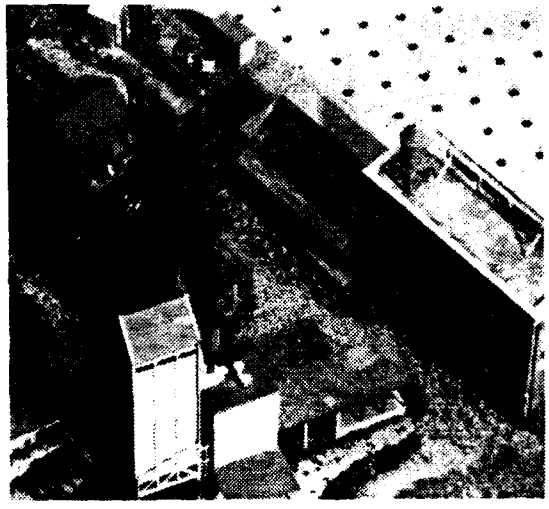
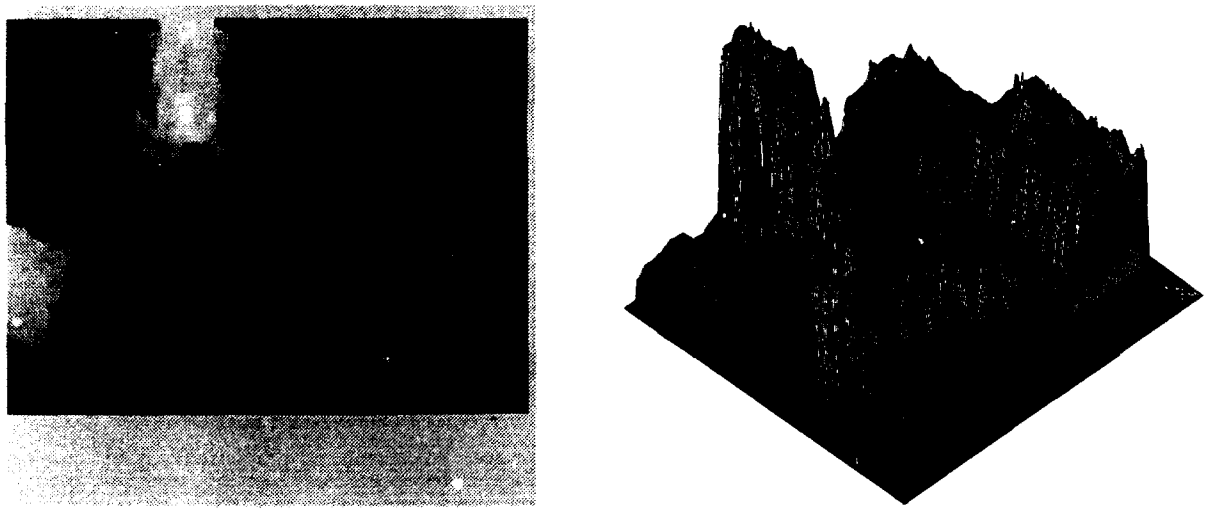


Figure 15: Oblique view



(a) (b)

Figure 16: Multiple baselines: (a) Distance map; (b) Isometric plot

for $j \in W$, where

$$\begin{aligned} a_1 &= B_1 F(\zeta_f - \zeta_r) \\ a_2 &= B_2 F(\zeta_f - \zeta_r). \end{aligned}$$

Since $B_1 \neq B_2$ and $\zeta_r \neq \zeta_f$,

$$a_1 \neq a_2. \tag{37}$$

So, we have

$$f(x+j) = f(\xi + j), \qquad \text{for } \xi = x, x + a_1, \text{ or } x + a_2. \tag{38}$$

Since this contradicts assumption (33), equation (35) does not hold. Its left hand side must be positive. Hence (34) holds.

B Multiple-Baseline Stereo Algorithm

We present a complete description of the stereo algorithm using multiple-baseline stereo pairs. The task is, given n stereo pairs, find the ζ that minimizes the SSSD-in-inverse-distance function,

$$SSSD(x, \zeta) = \sum_{i=1}^{n} \sum_{j \in W} (f_0(x+j) - f_i(x + B_i F \zeta + j))^2. \tag{39}$$

We will perform this task in two steps: one at pixel resolution by minimum detection and the other at sub-pixel resolution by iterative estimation.

Minimum of SSSD at Pixel Resolution

For convenience, instead of using the inverse distance, we normalize the disparity values of individual stereo pairs with different baselines to the corresponding values for the largest baseline. Suppose $B_1 < B_2 < \cdots < B_n$. We define the baseline ratio R_i such that

$$R_i = \frac{B_i}{B_n}. \tag{40}$$

Then,

$$B_i F \zeta = R_i B_n F \zeta = R_i d_{(n)}, \tag{41}$$

where $d_{(n)}$ is the disparity for the stereo pair with baseline B_n. Substituting this into equation (39),

$$SSSD(x, d_{(n)}) = \sum_{i=1}^{n} \sum_{j \in W} (f_0(x+j) - f_i(x + R_i d_{(n)} + j))^2. \tag{42}$$

We compute the SSSD function for a range of disparity values at the pixel resolution, and identify the disparity that gives the minimum. Note that pixel resolution for the image pair with the longest baseline (B_n) requires calculation of SSD values at sub-pixel resolution for other shorter baseline stereo pairs.

Iterative Estimation at Sub-pixel Resolution

Once we obtain disparity at pixel resolution for the longest baseline stereo, we improve the disparity estimate to sub-pixel resolution by an iterative algorithm presented in [12][17]. For this iterative estimation, we use only the image pair $f_0(x)$ and $f_n(x)$ with the longest baseline. This is due to a few reasons. First, since the pixel-level estimate was obtained by using the SSSD-in-inverse-distance, the ambiguity has been eliminated and only improvement of precision is intended at this stage. Second, using only the longest-baseline image pair reduces the computational requirement for SSD calculation by a factor of n, and yet does not degrade precision too significantly.

In the experiments shown in section 3, we used the following algorithm for sub-pixel estimation: Let $d_{0(n)}$ be the initial disparity estimate obtained at pixel resolution. Then, a more precise estimate is computed by calculating the following two quantities:

$$\Delta d_{(n)} = \frac{\sum_{j \in W} (f_0(x+j) - f_n(x + d_{0(n)} + j)) f_n'(x + d_{0(n)} + j)}{\sum_{j \in W} (f_n'(x + d_{0(n)} + j))^2} \tag{43}$$

$$\sigma^2_{\Delta d_{(n)}} = \frac{2\sigma_n^2}{\sum_{j \in W} (f_n'(x + d_{0(n)} + j))^2}, \tag{44}$$

The value $\Delta d_{(n)}$ is the estimate of the correction of the disparity to further minimize the SSD, and $\sigma^2_{\Delta d_{(n)}}$ is its variance. We iterate this procedure by replacing $d_{0(n)}$ by

$$d_{0(n)} \leftarrow d_{0(n)} + \Delta d_{(n)} \tag{45}$$

until the estimate converges or up to a certain maximum number of iterations.

References

[1] D. Marr and T. Poggio. A theory of human stereo vision. In *Proc. Roy. Soc. London*, volume vol. B 204, pages 301–328, 1979.

[2] W. E. L. Grimson. Computational experiments with a feature based stereo algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(1):17–34, January 1985.

[3] Stephen T. Barnard. Stochastic stereo matching over scale. *International Journal of Computer Vision*, pages 17–32, 1989.

[4] M.J. Hannah. A system for digital stereo image matching. *Photogrammetric Engineering and Remote Sensing*, 55(12):1765–1770, Dec 1989.

[5] Jer-Sen Chen and Gerard Medioni. Parallel multiscale stereo matching using adaptive smoothing. In *ECCV90*, pages 99–103, 1990.

[6] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1), 1987.

[7] Masanobu Yamamoto. The image sequence analysis of three-dimensional dynamic scenes. Technical Report 893, Electrotechnical Laboratory - Agency of Industrial Science and Technology, Tsukuba, Ibaraki, Japan, May 1988.

[8] Larry Matthies, Richard Szeliski, and Takeo Kanade. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[9] Joachim Heel. Dynamic motion vision. In *Proceedings of the DARPA Image Understanding Workshop*, pages 702–713, Palo Alto, Ca, May 23-26 1989.

[10] B. Wilcox. Telerobotics and Mars Rover research at JPL. Lecture at CMU, Oct. 1987.

[11] Hans P. Moravec. Visual mapping by a robot rover. In *Proc. IJCAI'79*, pages 598–600, 1979.

[12] Larry Matthies and Masatoshi Okutomi. A bayesian foundation for active stereo vision. In *SPIE, Sensor Fusion II: Human and Machine Strategies*, pages 62–74, November 1989.

[13] M. Yachida, Y. Kitamura, and M. Kimachi. Trinocular vision: New approach for correspondence problem. In *Proc. ICPR*, pages 1041–1044, 1986.

[14] Victor J. Milenkovic and Takeo Kanade. Trinocular vision using photometric and edge orientation constraints. In *Proceedings of the Image Understanding Workshop*, pages 163–175, Miami Beach, Florida, December 1985.

[15] N. Ayache and F. Lustman. Fast and reliable passive trinocular stereo vision. In *Proc. ICCV'87*, pages 422–426, 1987.

[16] Roger Y. Tsai. Multiframe image point matching and 3-d surface reconstraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(No.2), March 1983.

[17] Masatoshi Okutomi and Takeo Kanade. A locally adaptive window for signal matching. In *Proc. of Int'l Conf. on Computer Vision*, December 1990.

Building and Using Scene Representations in Image Understanding

H. Harlyn Baker*
Artificial Intelligence Center
SRI International
Menlo Park, CA 94025, USA

1. SUMMARY

The task of having computers able to understand their environments through direct imaging has proved to be formidable. With its beginnings about 30 years ago (1), the field of computer vision has grown as a major part of the pursuit for artificial intelligence. Most elements of this pursuit - language understanding, reasoning and planning, speech - are very difficult challenges, but vision, with its high dimensionality of space, time, scale, color, dynamics, and so forth, may be the most challenging. Early attempts to develop computer vision focused on restricted situations in which it was feasible to provide the computer with fairly complete descriptions of what it would encounter. In such cases, single images provided the sensory information for analysis. As the domains of application grew, the requirements for more competent descriptions of the world increased. Dealing with three-dimensional (3D) dynamic structures (the real world) from 3D dynamic platforms (we humans) calls for greater capabilities on both the analysis and synthesis sides of the issue. The analysis side is the processing of sensory data for such tasks as recognition and navigation, and a number of techniques are discussed here for dealing with these two-, three-, and higher-dimensional data. The synthesis side is the construction of 'internal' descriptions of what is seen in the environment - constructed now so that they may be used subsequently for the above tasks. This latter issue is the underlying theme we pose in this paper - developing representations from vision that will later enable effective automated operation in our 3D dynamic environments.

2. INTRODUCTION

Vision, which appears so easy for all of us, has proved to be an extremely complex task when addressed with computers. Despite early expectations in the field for realization of machine vision capabilities, it has grown to occupy a large proportion of the continuing artificial intelligence research effort. Understanding the coarse structure, let alone the nuances, of our environment continues to be a large and, in many parts, elusive challenge.

*The SRI research discussed here has been sponsored by DARPA under contracts DACA-76-85-C-0004, DACA-76-90-C-0021, and DACA-76-92-C-0003, and by Fujitsu System Integration Laboratory.

2.1 Knowledge for Analysis

A major component of the vision efforts seen today still parallels approaches taken throughout the years - the building in to the system of specific knowledge of the domain it will encounter. Vision does not take place without memory. As sighted individuals, we have a great deal of expertise, accumulated over years of observing and interacting with our 3D dynamic environments. Undoubtedly, certain capabilities appear with us at birth. Experience, however, and the memory that it accumulates, is equally critical to our performance. It enables us to rapidly and robustly interpret situations and events, recognize the familiar, and react opportunely to what we see. Since experience appears so necessary to our performance, it seems essential that a computer charged with seeing also have access to some equivalent sort of background knowledge. Although seldom enunciated, how this knowledge is given to the system, how it is represented, and how it is used in analysis of the visual imagery turn out to be principal issues in computer vision.

These knowledge issues occur at all levels of the analysis, from deciding what useful information from small parts of individual images to extract for subsequent processing (e.g., brightness values, gradients, contour elements), to considering what is relevant for identifying a striding distant silhouette as one's Uncle Bob. At some levels of the analysis there are generally accepted definitions of the knowledge that is appropriate (for example, the use of spatial-frequency-tuned filters), but, mostly, very little is understood and very little is agreed upon about these matters.

2.2 Representational Limitations

My discussion here relates to this knowledge-source issue. I phrase it as building and using computational representations in the task of understanding what is presented in an image of a scene. I present a number of pieces of work, indicating the capability they were designed to provide, the role of this capability in a vision system, and the level of initial-state knowledge provided to the system along with its ability to augment this through time. The main point I draw out is that all computer vision systems begin with an alphabet of operational primitives used to represent the image data. They have a vocabulary of combinations of these that they can deal with for scene interpretation. The capability of the system is set by its expressive power in this vocabulary, while its utility in a broader context is determined by the breadth of these definitions and its ability to grow beyond their limiting bounds. The

latter issue pushes up against generic 'learning,' an area of artificial intelligence probably unparalleled in both its potential and the ratio of its promise to its realization.[1] However, the issue of a system's repertoire of expression – its ability to build representations from imaged data and use them in understanding the visual situation – provides a key measure of its contributions: its contribution in solving the particular problem it addresses as well as its contribution to the computer vision task in general.

Two major determinants of the capabilities of a vision system are (1) the modes of imaging used, and (2) the elements on which it bases its analysis. In the next section I will provide a reference framework for these by discussing the principal modes of image data acquisition (single images, binocular stereo, and dynamic sequences) and the two choices for processing styles – homogeneous versus structured. The comparisons of image understanding systems I make in the following sections will be framed by these categories.

3. IMAGING MODALITIES

Imagery for scene analysis comes in three principal forms: monocular views; binocular views, and multi-image sequences of views – looking at a photograph, looking with your two eyes without being able to move your head, and the general situation of two eyes on a mobile head. Each form of data contributes differently to the scene representation and image understanding tasks.

3.1 Dynamic Scenes

Image sequences may provide information about scene dynamics (other moving objects), or give differing perspectives on a scene viewed as the sensor moves around. This is a mode of operation that people are clearly very capable of using, as we observe our dynamic world and move around in it, exploring. The relatively new area of 'active' vision (as in a sensor that adjusts its perspective to satisfy its requirements) studies acquiring and exploiting these sorts of data. Since, from the viewpoint of survival, anything that is in motion in our vicinity is of special interest to us, the analysis of dynamic imagery may be expected to play a critical part in a computer vision system.[2] Taking the more active role in data acquisition – moving around and collecting information from a variety of perspectives – leads to considerably more robust and more precise scene measurements. The cost is considerably more processing.

3.2 Binocular Viewing

What a single moving sensor does not provide is precise 3D measurement of moving objects. To determine the three-space position of an object requires seeing it from several (at least two) known perspectives simultaneously. A moving object viewed by a moving sensor is viewed from only one perspective at any instant.

Binocular views, image pairs captured simultaneously from different locations (as the eyes provide), can give sufficient information to enable 3D interpretation of both static and dynamic elements of a scene. That is, simple triangulation (back projection) can be applied to corresponding points in two images from known viewing positions to determine the location of the observed point in three-space. The biggest problem in stereo – one that has been with us from the beginning – is developing reliable techniques for determining which point in one image corresponds to a point in the other. This is the 'correspondence' problem – matching elements[3] between views. Although static binocular viewing is unusual – in human vision most binocular perception is dynamic – it is certainly effective, as viewing Figure 5 (subsection 6.3.3) will show. Depth is a powerful aid to scene understanding.

3.3 Single Images

With a stationary sensor viewing a nonchanging scene, a single snapshot view may be all that is available, and alone must be the basis for scene interpretation. That humans can operate with such a deficiency of information, for example in viewing photographs, lacking dynamics and explicit three-dimensionality, reveals the power of our processing and the value of memory and experience.

Most early theses in computer vision dealt with analysis of single images, and their failings immediately taught us the lesson of extensibility. Lacking access to the rich information of depth and motion, systems for single-image analysis were initialized with specific knowedge of the simple objects with which they could deal, and had no way to grow beyond this aside from reprogramming.

If all that is presented is a single image, and never in the context of a dynamic sequence, any interpretation will have to forego explicit temporal or 3D analysis. Since we presumably do not begin life with explicit knowledge of 3D structures, such as houses and cars, yet develop understanding of them over time (with both stereo and temporal data available), it is inconceivable that memory could operate without temporal analysis.

3.4 Processing Elements

A distinction with the different modes of operation that will be contrasted throughout this article is the choice of analytic element used in the analysis – image pixels or 'higher-level' features such as contrast edges or extended contours. These are often termed pixel-based and feature-based processing. At the pixel level, image intensity values are treated in an undifferentiated way, and the resulting representation is often termed "retinotopic" for its resemblance to a retinal layout. Feature-based processing and description works with a distinguished subset of the image information, and leads to scene descriptions that are more sparse but, through better localization, are also more precise. Although in truth this dichotomy is more of a continuum, I will exclusively consider the latter as *structured* abstractions from the imagery – the features will be edge elements or parts of contours.

[1] The question of learning is probably at the root of the question of intelligence.

[2] An immediate question with such analysis lies in what is being tracked through the dynamic sequence, and we will return to a discussion of this.

[3] A variety of choice of 'element' have been developed.

4. SINGLE IMAGE ANALYSIS

A common task in computer vision is to identify or classify items in a single image taken of some scene. For example, the task may be to identify and assemble components of a small machine, or to identify targets in an aerial view of a military installation. Clearly, single snapshot images of such a scene will lack 3D and dynamic information. The processing must rely on some comparison of what the computer expects to see with descriptions it extracts from the single image.

At the pixel level, the comparison may aim to group parts of the scene based on textural and other classifications. For example, a region that exhibits high spatial intensity variation (texture) may be classified as vegetation if the scene is expected to contain vegetation. Homogeneous regions may be sky if, again, the domain is known to be a natural scene out of doors. Anticipated relations between classified regions may provide use of mutual consistency to make the interpretation more robust. For example, if sky must be above vegetation, which is generally above the ground, then these spatial relations should be required of the classified regions. The major determinants of the capability of the system are the quality of the classifiers and the suitability of the relations. One may appreciate that determining effective classifications and relationships, valid across a wide range of realistic situations, might be difficult.

At the feature level, 2D shape descriptors are typically extracted from such imagery, for example straight lines, curves, and smooth contours, grouped into contiguous pieces. Some previous automated or interactive process has led to the development of a 'model vocabulary' – a set of feature groupings that can be composed together to represent the range of objects anticipated in the scene. Recognition involves comparing the extracted features (e.g., lines, arcs) and their interrelationships with those represented by the models.

What is probably most important to observe in this single-image analysis is that the processing must be preceded by defining what is expected to be seen in the images. Since 3D shape and motion are not available to the analysis, recognition must be based solely on the 2D information that can be obtained.

4.1 Interpretation through Pixel Classification

Strat (2) has demonstrated an impressive capability at interpreting natural scenes with a pixel-based classification system along the lines outlined above. He points out that most recognition schemes are based on geometric representations and matching of discrete features, yet natural scenes are neither well described by geometry nor characterized by specific localizable features. Taking a more eclectic approach, he develops a battery of filters that attempt to classify image regions, and builds a relational network among these descriptors. What brings the classifiers together is 'context' – the expected relationships between labeled components. These contexts are established manually in advance of any processing, and are individually constructed for specific domains.

By making the recognition context sets very specific, for example identifying 'foliage against sky' rather than simply 'foliage,' they can be made more reliable. At the same time, generic contexts can be defined that may be satisfied when more specific ones cannot. Context sets may include components that are both positive (for example, tree trunks tend to be vertical), and negative (ground cannot extend above the skyline). A variety of grouping and segmentation techniques are used over a variety of scales to produce candidate scene region labelings – estimates of pixel groupings (similar intensity or color), similar texture, horizontal or vertical orientation, line-like structure, and so forth. Robust operation is attained through use of overlapping or redundant filters. For example, sky may be either an untextured homogeneous region of high intensity or an area of smoothly varying general brightness above most other areas in the image. Cliques – mutually consistent sets of classifications – are sought over the image. The clique providing the greatest reliability and coverage is chosen as the best interpretation of the scene.

Using an auxiliary knowledge representation system (the Core Knowledge System, CKS (3)), a sequence of images may be processed, accumulating and sharing constraints from their individual interpretations. This, together with a coarse use of stereo (4), enables Strat's system to build up a rough symbolic 3D map of the area being viewed.

The examples Strat presents are in outdoors scenes of trees, rolling hills, and pathways. Figure 1 shows a 3D reconstruction of an outdoor scene analyzed with this system.

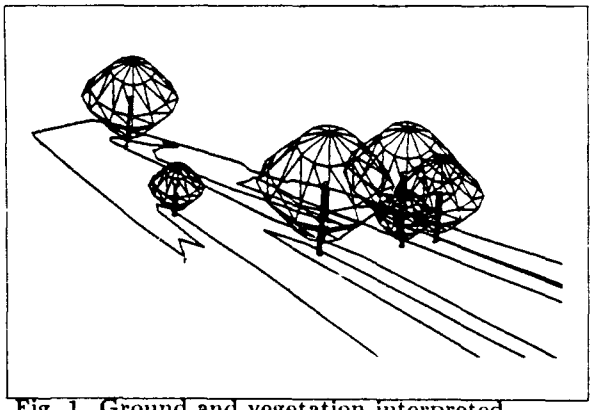


Fig. 1. Ground and vegetation interpreted from a single image.

While demonstrating a good capability at classifying image components in domains where the relationships have been prespecified, this approach is unlikely to provide the depth of interpretation needed for general scene understanding. One factor in this is that the system would require a significantly larger vocabulary of objects with increasingly tight constraints on their interpretation to distinguish, for example, among different types of trees or, more critically, to recognize specific trees, such as the one with a broken branch on the top of a certain hill. This requires geometric understanding rather than an understanding of certain relationships. In addition, no mechanism is presented for abstracting the required rules from the data. If one wants the system to show a utility beyond simple domains, this generative aspect is essential, and geometry probably cannot be avoided. Nevertheless, relational measures are generally missing from geometric-based recognition systems, and the use of this relational

approach in a partnership with the more metric approach of shape- and structure-based techniques should lead to more reliable operation for both.

4.2 Shape from a Single Image

A difficulty in trying to obtain information about shape or 3D structure from a single image is that a particular single image could arise from an infinity of scene configurations. The simplest example of this is an image of the image itself, where there is clearly no three-dimensionality to be observed, only interpreted. Interpretation requires knowledge, including knowledge of the physics of the imaging process and the local implications of intensity variation with respect to the shape of the imaged surface. Nevertheless, we all have the ability to interpret single images as 3D scenes, and there has been considerable effort in the field to develop similar capabilities in the computer. Using iterative optimization techniques and models of illumination, reflectance, and variations including albedo, Leclerc and Bobick (5), and others, have demonstrated the ability to recover surface height from simple measures on the imagery.

That such analysis cannot be guaranteed correct is apparent from its fundamental assumptions. The interplay of reflectances and shadowing could cause havoc with the modeling, which presumes fairly simple relationships between light source and reflecting surface. Any variation is interpreted as either surface shape or simple albedo change. Such shading analysis probably will have its greatest use where other depth measurement techniques, such as binocular stereo, have insufficient information to operate, yet can provide 3D constraint to limit ambiguity.

4.3 Models in Interpreting Single Images

Undoubtedly, much of the world is quite well described geometrically or by discriminable aspects of coloring, texture, or structure. Since the world is three-dimensional, a critical element of scene analysis must be the ability to represent and recognize 3D objects. In these cases, recognition may be attained by locating specific scene features and comparing their parameters with those chosen in advance to represent specific objects. Recognition, here, may be viewed as searching through a set of 3D object descriptions and finding the mapping of position, orientation, and scale that provides the most satisfactory correspondence. Aside from the selection of feature descriptors and the inevitable question of how to acquire the object descriptions in the first place, the major challenge in this work is effective search through the potentially enormous set of match possibilities.

Two pieces of research can highlight the approaches taken to this shape-based or structural recognition. While addressing 3D recognition, each uses information from single images for its recognition. The first represents objects as integrated networks of 3D points. The second provides coverage of the 3D situation by storing a range of representations, each pertaining to a small set of viewing perspectives.

4.3.1 3D Models with Image Matching in 2D

Huttenlocher and Ullman (6) introduced the term 'alignment' - a method to match stored models with features obtained from a view of a scene. In their work, the features – both in the scene and in the model – are two-dimensional contours (each classified by its shape) and their endpoints, if a straight contour, or midpoints otherwise. A model is a set of 3D points forming triangles (planar facets), and the contours of which they are part. Alignment is the process of selecting pairs of corresponding triangles (from the model base and from the imagery) and using the transformation implied by their match to map the rest of the contour description. The transformations are simple translations, rotations, and scalings. Estimating the goodness of fit of the resulting transforms enables selection of a 'best' interpretation.

4.3.2 2D Models and Image Matching

Chen and Mulgaonkar (7) address the problem of model-matching using 2D image data in a more methodical and practical manner. While using a related approach to the matching - hypothesizing 'alignment' transforms and mapping the related constraints for validation with the data, the detail of their strategy offers considerable advantage.

Two characteristics of their work stand out. First, they build their models in a semiautomated way by showing the system parts from various perspectives and under different lighting conditions. Model acquisition is a crucial and potentially[4] very time-consuming component of setting up a recognition task, and a which technique that automates this using the results of its own analysis immediately has more utility. Each model is structured as a set of classified contour elements – straight and curved segments – ordered by their relevance to the matching task. Features that are detectable most often in the training set and are found most likely to be correctly identified in the data are ranked higher in importance. These should be the first to be sought in the matching. This 'learning' strategy enables each model to be organized in a manner that is most effective for establishing its presence or absence in the scene. In effect, a model is a sequence of instructions for validating an object's presence in the image – it is a program.

Their representational system is 2D, and a single object will be composed of several perspective models, with each covering a small range of viewing angles – plus or minus perhaps 15 degrees in each direction. This is not as satisfying a solution as building a unified 3D model of each object; however, it has practical advantages in that it simplifies both the modeling task and recognition.

The system was developed and demonstrated on an industrial assembly operation, involving about two dozen parts, and has since been used for identifying objects in a dynamic context (see subsection 6.3.3).

4.4 Prospect Beyond Single Images

The techniques described above have relied primarily, if not totally, on 2D information, both in their models and in their image understanding. The use of 3D information for model representation and recognition has had less and generally more recent investigation. The principal difference in these works arises from the necessity of obtaining 3D information from the scene. This cannot be done from

[4] "potentially" because very few object recognition system have any sizeable model repertoire

single images, and requires either active ranging (for example, structured lighting, sonar, radar) or at least two simultaneous perspectives from passive sensors such as cameras.

This step to three dimensions lays the foundation for the distinction I wish to make in approaches to image understanding. If the system has no recourse to 3D temporal or spatial information, then its knowledge is limited to what the developer programs in: if the system has an ability to integrate information across space or time, then it can begin to meaningfully augment its knowledge base. Acquisition of this 3D information is the focus of the next two sections.

5. SCENE MODELING FROM STEREO

Image pairs, providing two perspectives of a scene, provide the data for inferring the range to points in a scene. This is termed binocular 'stereo' processing, after its resulting solid three-space description of the scene. The goal of stereo analysis is to obtain the best estimate possible of the range to points in the scene. 'Best' may depend on a number of requirements, including speed. The point to observe about these systems, however, is that they have some knowledge about the state of the world they are looking at – knowledge that serves to constrain the solution they present – and they have the common goal of developing a 3D description of the scene. It is common in stereo research to produce a range map, but very uncommon to do anything further with it, for example, navigating or controlling a robot arm.

Once the camera position and correspondences are known, estimating the range to some feature in the scene is a simple matter of triangulation. An effective mechanism for limiting the cost of determining these correspondences lies in using the 'epipolar constraint.' Knowing the two camera relative positions and attitudes enables definition of the expected pattern of disparity on the images. For cameras directed in parallel, the disparities will only be lateral, while for converging cameras the patterns will be radial. This camera information is used to shape the search window for possible corresponding elements, so it both reduces ambiguity and decreases computational cost.

5.1 Pixels versus Features

Within stereo processing, two major approaches are taken in selecting correspondences, one based at the pixel level and the other at the feature level. The objective within the two is the same, however – recovering the 3D structure of the scene as represented by the 3D location of its components. The main distinction lies in what constitutes these 'components.'

5.2 Scene Geometry from Image Pair Pixels

In pixel-based stereo processing, the objective is to label each point in an image (where possible) with a range value. If the relative positions of the cameras are known and corresponding pixels can be found in the two views, then relative range can be estimated directly by triangulation. Absolute range comes from knowing absolute camera displacements. The techniques used for solving the correspondence problem generally involve correlation – estimating the similarity between image regions in the two views. This similarity is usually measured as a local difference in intensity value between corresponding parts of the two images, with secondary constraints being introduced to enforce global consistency. The former, local measure, uses a small support function – typically a square or circular region centered on a pixel – with the similarity being either a simple sum-of-squared differences (SSD), or a correlation coefficient measure. The correlation coefficient measure may be normalized to eliminate the effect of linear variations that might arise, for example, from viewing at different times of the day, under differing light conditions, or with separate automatic gain adjustments on the two cameras.

In SSD matching, the expression to be minimized at any pixel (x, y) is:

$$SSD_{x,y} = \sum_{r_x, r_y} [I_L(x+r_x, y+r_y) - I_R(x+d_x+r_x, y+d_y+r_y)]^2$$

where (d_x, d_y) is a displacement from the source image pixel $I_L(x, y)$, and (r_x, r_y) defines a region of integration in the destination image, $I_R(x + d_x, y + d_y)$. This sum may be weighted to diminish the effect of brightness variance with radius. The vector (d_x, d_y) with minimal sum $SSD_{x,y}$ is selected as the image of the pixel at (x, y) in the second frame.

In normalized correlation, optimization is based on the measure:

$$E = \frac{\sum_{r_x, r_y} [I_L(x, y) - \hat{I}_L][I_R(x, y) - \hat{I}_R]}{\sqrt{\sum_{r_x, r_y} [I_L(x, y) - \hat{I}_L]^2 \sum_{r_x, r_y} [I_R(x, y) - \hat{I}_R]^2}}$$

where $\hat{I}$ is the mean brightness over the image region (r_x, r_y) centered at (x, y).

5.2.1 Normalized Cross Correlation

A typical approach to pixel-based stereo analysis is that of Hannah(4). Here, normalized correlation provides the matching metric, and processing in a resolution hierarchy provides a global consistency constraint. This use of a resolution hierarchy is fairly common in computer vision. It involves building a pyramid-like structuring of the image data, with the bottom level being the full-dimensioned image, and successively higher levels being the half-resolution versions of the one below them. The top level is a small, very highly reduced, and subsampled version of the original image – it has only very low spatial frequency components, with the higher frequencies being removed by the successive averagings.

A strategy often used in computer stereo vision is to match coarse features first (low spatial frequencies), and then use the results at this scale to constrain finer scale matching (higher spatial frequencies).[5] Beyond this constraint, Hannah also requires that her correspondences are the same in left-to-right matches as they are in right-to-left matches. Analysis of the correlation coefficient and

[5] It is always possible to show images in which such an arbitrary direction of progression will give the wrong answer.

an autocorrelation measure enables this process to ignore matches that have insufficient evidence for reliable estimation. This has the benefit that hallucinations, such as giving range to the sky, do not occur often. This technique, however, is costly in computation.

5.2.2 Stochastic Stereo

An alternate that is particularly suitable for implementation on a SIMD parallel processor is a stochastic method, developed by Barnard, using a simulation of the physical process of annealing to enforce global consistency (8). This method uses a composite similarity measure – image intensity difference and a gradient constraint that biases the solution in favor of a flat disparity map. The stochastic element enters the analysis in the way the individual difference measures are combined in looking for a global solution for the image pair. As in annealing, the system is injected with energy (heat), allowed to cool, heated up again – although less – then cooled again, repeating until there is very little change between these heat/cool cycles. The measured change is this similarity measure – a weighted sum of intensity difference and implied disparity gradient for the selected pixel matches. The different 'heat' settings allow a varying range of disparity adjustments in the pixel matching.

The measure minimized for optimization in stochastic stereo is:

$$E_{ij} = \sum_{ij} (|\Delta I_{ij}| + \lambda |\nabla D_{ij}|),$$

with $\Delta I_{ij} = I_R(i, j + D_{ij})$, where I_L and I_R are the left and right brightness values, and ∇D_{ij} is the gradient of the associated disparity estimate; λ balances the brightness and smoothness constraints.

Even when a parallel processor is used, the cost of iteration makes this a fairly time-consuming technique. Images of size 512 by 512 pixels require about 10 minutes of processing time on an 8000-processor Connection Machine (CM).

5.2.3 Real Time SSD Matching

A third technique worth examining for its simplicity and effectiveness is an SSD method implemented on both a 16000-processor CM and on a coarse-grained (5-processor) i860 parallel processing system (9). Much effort was invested in making this process run as rapidly as possible to support real-time control, and it can perform stereo matching on images 256 pixels square at about 40 Hz on the CM and 10 Hz on the i860 configuration. The SSD phase gives velocity estimates for each pixel, mode analysis of this velocity distribution selects the major discrete motions, and an adjustment phase tracks regions over time. It has been used to control a robotic arm in tasks such as maintaining centered view on pedestrians and on another robot arm.

5.2.4 Considerations

Both of these parallel approaches share a common drawback. They process only in integer units of disparity, so deliver just a small number of bits of range resolution. In the case of the stochastic stereo, this was about 5 bits (32 levels), while with the SSD method it was about 3 bits (8 levels). Any change in this precision incurs added computational cost. Hannah's method delivered subpixel correlation measures, and was precise down to small fractions of a pixel unit.

5.3 Structured Stereo Processing

Another approach to stereo analysis for obtaining 3D information about a scene involves the processing of not pixel values but abstracted features – contour elements as produced by zero-crossing operators. Marr and Poggio, Baker, and Mayhew and Frisby were the early developers of this feature-based approach to stereo matching.

Marr and Poggio (10), later joined by Grimson (11), worked with zero crossings of the Laplacian of a Gaussian (LOG), and progressed from large Gaussians to small Gaussians in a hierarchic-pyramid manner. Matches obtained at the coarse level constrained the possible matches at finer levels. A consistency measure was implemented by insisting that disparities over a small region were identical. An unfortunate artifact of this is that their results tend to represent the scene as planar chunks at different ranges. Mayhew and Frisby (12), later joined by Pollard (13), used a figural continuity constraint to enforce connectivity of depth estimates for LOG features that were connected in projection. They also used peaks and troughs of this signal, presenting evidence from psychophysics supporting human use of these in vision, and introduced a variation of the scale analysis of Marr and Poggio – looking for consensus in neighboring bands rather than in successive coarse-to-fine levels. Baker (14) used a form of figural continuity as well, and followed his feature matching (extrema of intensity gradient related to zeros of the LOG) with constrained intensity matching to provide a dense range map. Grimson used a surface-fitting technique to interpolate between matched features to estimate this map.

The fact that feature-based stereo results in sparse range measures has been raised as a criticism. Dense results are preferred. Feature-based approaches have greater precision, however, as they focus on the more localizable parts of the imagery. Scale processing is felt to be a key to providing dense results. Pixel-based techniques have been more easy to implement on SIMD parallel processors, so they may have an inherent advantage for real-time development.

Much other research has addressed pixel-based and feature-based stereo, including using a third camera to provide an ambiguity-resolving perspective and introducing other constraints (a recent survey paper covers much of this area well (15)). Among some dozen and a half systems evaluated competitively a few years ago (16), Hannah's system was ranked first across a majority of the categories (17).

5.4 Differential Techniques: Motion and Range

A different approach to disparity estimation has been developed for motion processing – optic-flow analysis – where the objective is to estimate movements in a scene (18). Under certain conditions these techniques may also be used for stereo range estimation. Two principal points distinguish this work from pixel- and feature-based

matching approaches. First, the presumption is that there is very little difference from one image to the next - motion processing allows this, whereas typical stereo has a sufficiently large baseline that images may differ significantly. Second, differential techniques are used that do not depend on feature localization in the image.

5.4.1 Optic-Flow Analysis

Horn and Schunk (19) developed the brightness-constancy constraint, which relates variation of intensity between successive images with the underlying variation in the scene. The principle behind this differential technique is that derivatives of the spatiotemporal intensity data indicate rate of image change. If the image change is due only to camera displacement, then simple derivative convolutions on the spatiotemporal intensity data can be used to estimate scene distances. If the change is due to scene motion, then the technique estimates velocities. Since the expression for the variation at a single point is underconstrained, the solution involves a least-squares approximation that integrates over some local neighborhood, and this makes the result sensitive to the density of discrete motions in the vicinity. The estimates are best where there is strong local texture (surface detail) with a single velocity. Where the texture is weak (there is little distinctive detail) or the local vicinity contains more than one motion (such as occurs at object boundaries), the estimate can be rather meaningless. Despite this, the results tend to be generally credible.

With the differential approach, image disparity (or velocity) (d_x, d_y) at frame t can be determined by minimizing the following expression:

$$\sum_{r_x, r_y} [d_x I'_x(x, y, t) + d_y I'_y(x, y, t) + I'_t(x, y, t)]^2,$$

where I'_x, I'_y, and I'_t are spatial and temporal derivatives of image intensity $I(x, y, t)$.

The summation is again taken over a local region of the image (r_x, r_y). One finds the least-squares solution, in closed form, by taking derivatives of this expression with respect to d_x and d_y. The least-squares estimate is given by:

$$\mathbf{d} = -\mathbf{M}^{-1}\mathbf{b},$$

where

$$\mathbf{M} = \begin{pmatrix} \sum I'^2_x & \sum I'_x I'_y \\ \sum I'_x I'_y & \sum I'^2_y \end{pmatrix},$$

and

$$\mathbf{b} = \begin{pmatrix} \sum I'_x I'_t \\ \sum I'_y I'_t \end{pmatrix}.$$

This expression has minimum error when

$$d_x I'_x + d_y I'_y + I'_t = 0,$$

that is, when the observed image gradient vector (I'_x, I'_y, I'_t) is orthogonal to the observed disparity (or velocity) vector $(d_x, d_y, 1)$. Figure 2 shows the optic flow computed for the motions of a sedan and van against a stationary background, the imagery of which is shown at the top of Figure 5.

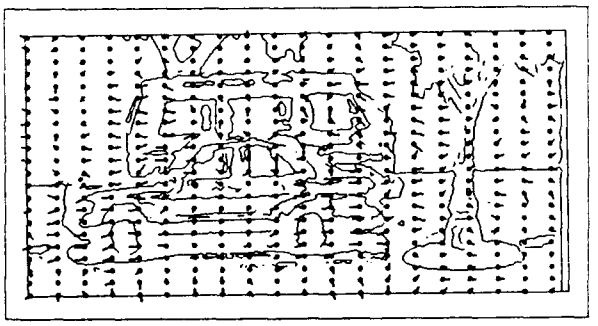


Fig. 2. Optic Flow for Moving Sedan and Van.

5.4.2 Hierarchic Optic-Flow Computation

Hanna has presented a method for extending the applicability of the gradient-based technique to images with significant variation between frames (20). This operates through a hierarchic-pyramid analysis, beginning with low-resolution coarsely sampled imagery, and progressing through to the full resolution data. A unit of pixel measure in the coarse imagery corresponds to a 2^n by 2^n pixel region at highest resolution n levels finer, so a gradient computed at this single unit can identify the predominant motion over that much larger window. Recursive processing of this motion estimation followed by image remapping - to bring the corresponding image locales into alignment for the next gradient analysis - may be viewed as delivering the n-bit motion vector a bit at a time, starting from the highest-order bit. What is important to note is that with this hierarchic approach, gradient-based optic flow can also be used for stereo range estimation - large disparities are handled by the coarser scales. The major difficulty remains, however, that there can be no guarantee this coarse-to-fine progression will give correct results. A small feature that is moving to the left while the predominant region motion at a coarse level moves to the right will be 'mapped' in the wrong direction for being detected at any of the succeeding levels.

An iterative remapping method very similar to Hanna's was used much earlier by Quam in his *hierarchical warp stereo* process (21). The matching metric in this work was correlation, rather than gradient-based optic flow.

5.5 Issues in Stereo Processing

A number of questions must follow any depth recovery process, such as: Are there measures of confidence associated with individual estimates? Is the result conclusive? Are there errors of omission (gaps) or commission (range estimates where there can be none)? Does the process deliver a description of objects or just an array of numbers that represent a range 'map?' How relevant is the resulting description to the intended use? Since the purpose of range recovery is tied to some other task, such as understanding the scene or moving about in it, these questions can determine the utility of the whole exercise.

One of the principal dissatisfactions in stereo analysis has been in its reliability. Perhaps 90% of a scene can be adequately modeled with the above techniques, but the remaining 10% failure can make the results almost unusable. Higher reliability is needed before one can trust an autonomous device for guidance. There is very little opportunity to obtain better accuracy when presented

with only two perspectives of a scene. Ambiguities are difficult to detect, and cannot be resolved without the introduction of more information. This information has often taken the form of *a priori* knowledge about scene and object types (for example, that the scene contains static opaque rectilinear structures).

Better additional information that is not domain specific, is provided by "trinocular stereo," which involves acquiring a third view of the scene. This was first introduced by Burr (22) and later followed by Faugeras's group in France (23). This third view, if noncollinear with the other two, provides a second epipolar constraint that can disambiguate potential match uncertainties.

Almost without exception, stereo techniques have difficulty in correct handling of occlusion (where a feature does not have a match in the corresponding view), image reversals (where feature left-to-right ordering is inverted between views), transparency (where multiple ranges are associated with individual view points), and canopy phenomena (where there are a few predominant and quite different depth ranges over a small region of the view). These are significant issues for depth estimation and natural scene interpretation.

A more general comment on two- or three-view stereo is that the resulting descriptions are not of the same quality as those we perceive when we as humans observe a scene. Stereo results look like cut-outs, with a series of ranges computed for certain directions of the camera. The same can be observed in looking at a stereo pair of photographs - the perception is likely to have a flat, disjoint, and chunky appearance. The perception we have under natural conditions is more continuous and connected, and this results from our ability to observe in the continuum through time. We change our viewing position to suit our demands for fill-in and clarification, and integrate information through active control of the viewing process, such as obtaining a description of some novel 3D object by grasping it and manipulating it before the eyes.

6. SCENE MODELING FROM SEQUENCES

Recent approaches to 3D vision have addressed this processing of image sequences, where a sequence comprises many views from different positions. This more closely resembles the operation of the human system, where we observe with eyes that are free to move, collecting information from various perspectives. This multiple-view approach could provide considerably more complete descriptions of a scene, revealing, for example, what the back side of an object looks like, and could do so with much less ambiguity. Aside from restricted cases, however, it has proved difficult to exploit this extra data in the coherent manner required. One of the problems lies in organizing and maintaining coherent descriptions of the rather massive amount of data involved - sequences could be hundreds of frames long, or more.

6.1 Correspondence Through Time

Sequence processing shares many of the computational issues of stereo. The principal problem in stereo processing has been identified as putting into correspondence, accurately and reliably, features that appear in two views of a scene. Determining the correspondence is an ill-posed problem: ambiguity, occlusion, image noise, and other influences resulting from the differing appearance of objects in the two views make feature matching difficult. In sequence analysis, where rapid image sampling produces images that change little from one to the next, matching is less problematic. In some approaches this is taken to an extreme, with sampling sufficiently rapid that images vary smoothly between views. The following sections describe how this temporal continuity has been developed and exploited for robust tracking and estimation of scene features.

6.2 Pixel-Based Sequence Analysis

As was the case with stereo analysis (cross-correlation and gradient analysis), there are two principal approaches to pixel-based motion analysis. In correlation, the objective is to determine for each pixel in one frame, its image in the next frame. Techniques as described in section 5.2 are used for this. SSD is more typical than normalized correlation in sequence analysis. With temporal sampling sufficiently fine that brightness changes are of a smaller magnitude than changes due to motion, there is little requirement for accommodating to varying illumination. With the optic-flow approach, on the other hand, explicit matching is avoided, and motion is derived directly through differential analysis, as described in section 5.4.

Another problem both correlation and optic-flow analyses encounter is that they are designed for pair-wise computation rather than for sequential tracking. Since they are referenced on the center of a pixel in one image, their displacements are not easily chained with precision through a sequence. Range estimates will be imprecise over a short baseline, so the reliability and precision obtainable for matches over a long baseline become crucial questions.

Pixel-based and point-based reconstruction techniques, where they have been developed to the stage of integrating measures over a sequence (for example, (24, 25)), do not exploit the continuity of observations. Rather, they treat observations from different perspectives as disjoint, and pool them in (more or less estimation-theoretic) volume sets.

A recent innovation - the use of a singular value decomposition procedure - uses intermediate feature trackings to synthesize a long baseline through many small changes. It recovers both the shape and motion observed in transformation of a rigid body (26). The tracking employed uses an autocorrelation measure to select distinctive image features (in a spirit similar to that of Hannah). By tying observations together through the sequence, it obtains the benefits of a large baseline with the reduced error of small-increment image variation.

A difficulty with local-support integration techniques (pixel-based approaches in general) is that when the local region of integration overlaps different range distributions, the estimate may be quite meaningless. Since these bounding areas are of particular interest in most 3D tasks - such as grasping and navigating - this deficiency can be quite severe. The issue is particularly salient in motion analysis, where an intermediate velocity estimate is much

more misleading than an intermediate range estima... Intelligent window shaping may improve the situation, although at significant cost (27).

6.3 Structured Processing – EPI Analysis

There is much more in an image sequence than is being processed by techniques such as those described above. Selecting only highly localizable features leads to sparse scene descriptions, while use of the full image contents, as in optic-flow and correlation approaches, leads to much uncertainty, weak localization, and fragmented tracking. An alternative exists in utilizing the three-space correlate of 2D image contours. The motivation of this 'structured' approach to sequence analysis is that dynamic imagery has both spatial and temporal structure, while pixel-based techniques represent neither and must determine them both during its operation. Pixel-based techniques compute the temporal structure by 'tracking' features using correlation or optic-flow analysis, and determine the spatial structure by grouping results after temporal tracking. And yet the structure is there in the data.

Epipolar Plane Image (EPI) Analysis is such a technique that holds particular promise for scene reconstruction (28). It integrates throughout the data acquisition and has several major advantages over other approaches, such as not requiring correlation or any similar matching strategy, and dealing explicitly with spatial and temporal continuity. The features utilized are at object and texture discontinuities, so do not involve integration across different range distributions. This technique was the first to exploit small increments over a large integrated continuous baseline for the ideal mix of reliability and precision in motion analysis. The geometry and intuition of imaging in this situation are a little unusual, so I will review the implications of the generally used epipolar constraint in the context of sequence processing.

6.3.1 Epipolar Geometry

In Figure 3 (left), a camera is shown at two different positions along a linear path. At each of the sites the camera is looking at right angles to the path, and a feature such as P will appear displaced to the right in the second view with respect to the first. This displacement is along the projection of the plane formed by P and the two camera centers. This plane is termed an "epipolar plane." For a continuing sequence of such images, the point P will stay on the same image scan line from frame to frame. Because of this epipolar structuring, we can confine our depth analyses in right-angled linear motions to single sets of scan lines. Figure 4 shows a volume formed by stacking up the data collected in an image sequence and slicing horizontally to reveal such a set of scan lines. The pattern of streaks in this slice makes the lateral displacement character quite apparent and their interpretation quite direct: Near features have streaks with low slopes, more distant features have higher slope. Stereo processing of s... a scene would correspond to comparing features between, say, the first and the last frame, or the first and last line of this image. The continuity evidenced here takes the uncertainty out of the matching process. Analysis of these slice images, termed epipolar-plane images (EPI images) after their composition from samples of a single epipolar plane, led to an effective technique for estimating the range to features in a scene.

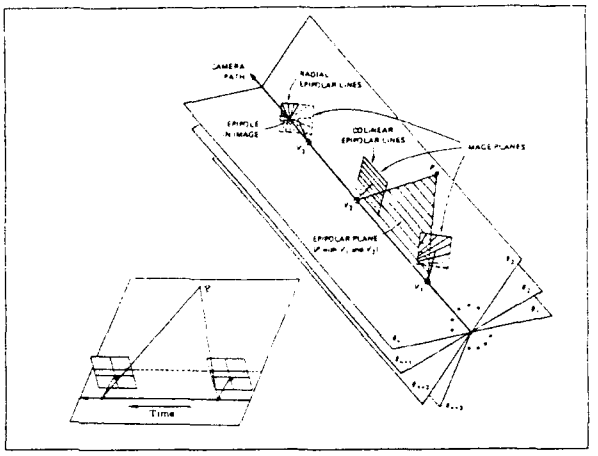


Fig. 3. Epipolar Configuration for Moving Camera.

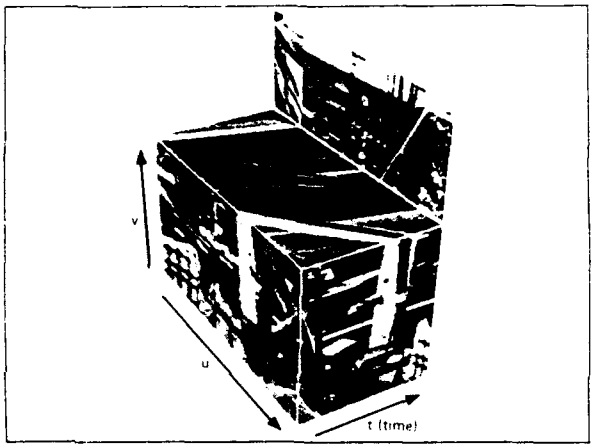


Fig. 4. Spatiotemporal Image Volume.

6.3.2 Spatiotemporal Manifolds

To expand the technique to more complex viewing situations such as nonlinear and varying-velocity camera paths with varying camera orientations, as would be found when a human moves through a scene (Figure 3 (right) shows patterns of epipolar lines that arise for linear motion and varying view direction), it was necessary to generalize the geometric representations used. In the earlier work, EPI-based linear features – representing the evolution of individual features over time – were detected and processed. In generalizing the approach, *spatiotemporal manifolds* – representing the time evolution of whole spatial contours – were constructed and used in inferring scene structure (29).

This reformulation brought another advantage: Representing the time-evolution of contours rather than individual features would produce connected 3D space curves rather than isolated points. Grouping of scene measures into meaningful and related structures remains one the largest problems in vision. Since even the most reliable and precise depth map is only another input to the scene-understanding process, any technique that can deliver direct segmentation and grouping information with its measures will have a great impact on the use and reliability of its data.

6.3.3 Tracking and Identification

Figure 5 shows a composite development in tracking and identification using the spatiotemporal manifolds for feature localization in space and time, and the 2D modeling facility of Chen (7) for object recognition. The figure shows in successive steps the strongest zero-crossing contours in three adjacent frames (the first and last of which are shown at the top), with the final view showing the results of identifying a van and sedan in these data. The bottom of the figure shows the models used in the recognition. These were constructed in a earlier training phase. An added benefit in this figure is that it demonstrates the value of stereo in perception: The paired figures are presented for crossed-eye viewing and, when fused into a single percept, will reveal a considerably more coherent interpretation, one that may be impossible to obtain monocularly.

6.4 Stereo and Motion

Undoubtedly, simultaneous stereo and motion analysis must be obtained for us to hope to achieve the capabilities of the human mobile-binocular system. Stereo is essential, as motion can only compute range to stationary objects and for known camera motion. At the same time, motion and sequence analysis are essential, as the active element in exploring an environment, both for modeling it and for navigating through it, cannot be met from a single perspective or even a set of predetermined perspectives. While the number of research efforts addressing stereo and motion analysis is small (9, 24, 25, 30), a coherent approach to integrating these two related modalities will be essential to capturing the true three-dimensionality of our environment. Figure 6 shows an integration of this sort of stereo range estimation and sequence processing operating on a field of rocks. The initial description (middle) is refined from subsequent views resulting in better definition on object 3D shape (bottom). The computational requirements for this data-intensive challenge are now being met by multi- and parallel-processors, with a number of research groups investigating stereo sequence analysis in high-performance computing environments.

6.5 Recognition of 3D Shape

The techniques described above have addressed the issue of obtaining estimates of scene 3D structure from two or more views. The major purpose of this is to provide the third dimension for tasks involving recognition and navigation. Unfortunately, very little has been done in using the 3D estimates produced. An early effort that took on this problem was my modeling research in Edinburgh (31). Models of 3D shape were constructed through analysis of objects observed rotating about a known axis. Using a 3D alignment technique, models built from current imagery were compared with models stored in the training phase, and the closest 3D fit was selected as the match.

Although more refined techniques have been developed in the interim, for example the work of Szeliski (32) in building 3D representations using rotation, the majority of research in 3D model matching has used either very simple representations, such as rectilinear blocks (33), or direct ranging techniques, such as provided by structured light or laser devices (34). Where 3D objects have been recognized, they have rarely been modeled by the same process used for their recognition. An exception to this lack of acquisition and use of 3D information in computer vision is in autonomous navigation systems (35, 36), although most systems use active ranging. Some of these systems are capable of extracting 3D scene features and then using these in obstacle-avoiding traversal of the area. Again, however, the representations tend to be simple (boxes, points) and not adequate for representing anything of the sophistication and detail of our environments. A good review of 3D object description techniques may be found in a paper by Besl (37). Some of the works he cites address the issue of model building within a recognition context.

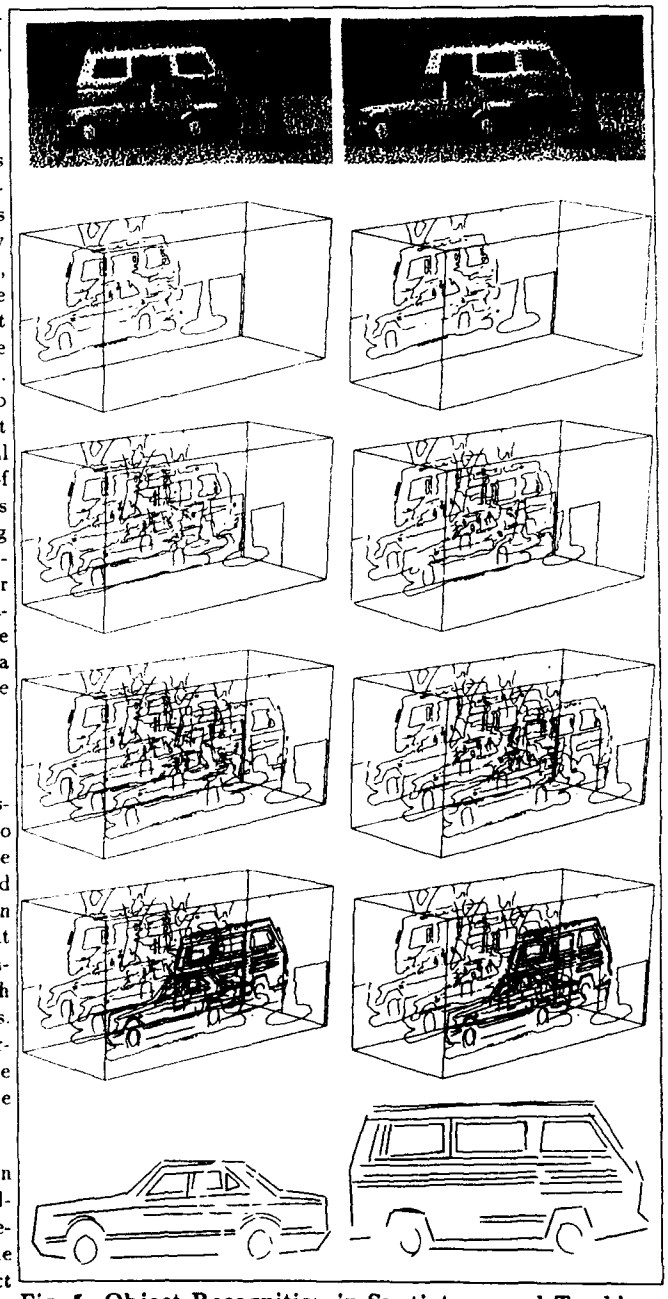


Fig. 5. Object Recognition in Spatiotemporal Tracking.

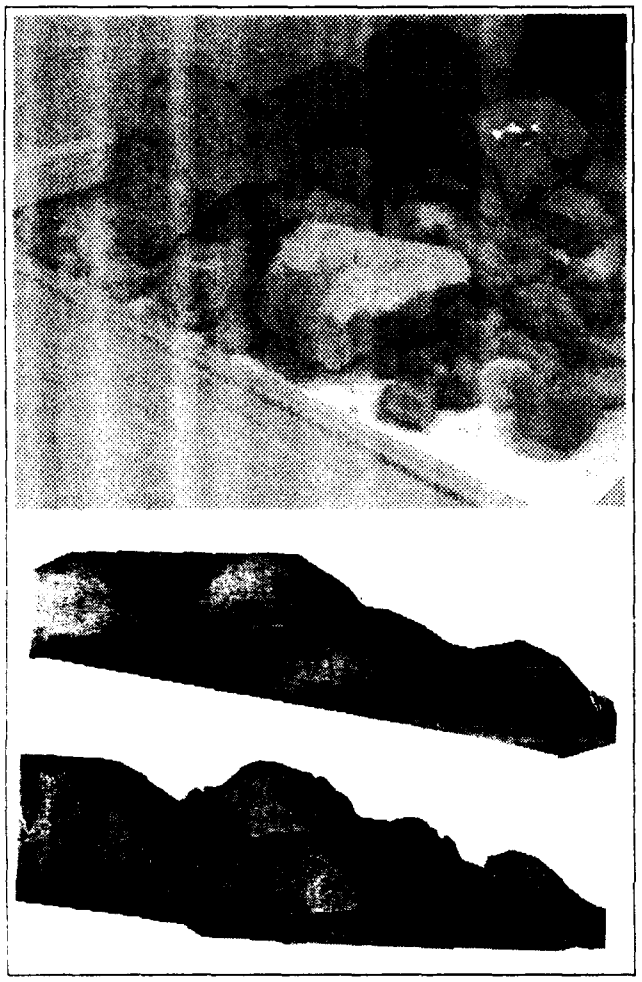


Fig. 6. Refined Scene Model from Stereo Sequence.

7. CONCLUDING REMARKS

A system that is to operate in the real world – that is, to find its way around and interact with other processes in the environment – must be able both to use information about the scene and to derive information during its operations through use of its sensors. This building and using of information in scene analysis, both geometric and otherwise, is an essential element for autonomous operation. Given sufficiently expressive modeling, single images will be adequate for interpretation, but to capture these models requires developing temporal and stereo integration techniques, and ones that encompass both geometric and relational information about objects and their surroundings. The alternative – programming in advance whatever is to be seen – cannot deliver the flexible capabilities needed for operation in the relatively unstructured and unconstrained domains in which we hope to operate our vision systems.

When looking at the challenge of precision operation in a world with the complexity of ours, we can see we have come a long way, yet still have considerably more to accomplish. Techniques for analysis over scale, 2D and 3D object modeling, optic-flow and spatiotemporal analyses, combining with object recognition using 2D and 3D geometric and relational descriptors, are leading us in the direction of attaining these capabilities.

References

[1] Roberts, L. G. (1965). "Machine Perception of Three-Dimensional Solids," *Optical and Electro-Optical Information Processing*, MIT Press.

[2] Strat, T. M., and M. A. Fischler (1991). "Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, No.10, 1050–1065.

[3] Smith, G., and T. M. Strat (1987). "Information Management in a Sensor-Based Autonomous System." *Proc. DARPA Image Understanding Workshop*, 170-177.

[4] Hannah, M. J. (1980). "Bootstrap Stereo," *Proc. DARPA Image Understanding Workshop*, 201–208.

[5] Leclerc, T. G., and A. F. Bobick (1991). "The Direct Computation of Height from Shading," *Proc. Computer Vision and Pattern Recognition*, Maui, Hawaii, 552–558.

[6] Huttenlocher, D. P., and S. Ullman (1988). "Object Recognition Using Alignment," *Proc. DARPA Image Understanding Workshop*, 370–379.

[7] Chen, C-H., and P. G. Mulgaonkar (1992). "Automatic Vision Programming," *Computer Vision, Graphics and Image Processing: Image Understanding*, Vol. 55, No.2, 170–183.

[8] Barnard, S. (1989). "Stochastic Stereo Matching Over Scale," *Intl. Jour. Computer Vision*, Vol. 1:1, 17–32.

[9] Woodfill, J. L., and R. D. Zabih (1991). "An Algorithm for Real-time Tracking of Non-Rigid Objects," *Proc. American Assoc. Artificial Intelligence*, Anaheim, CA., 718–723.

[10] Marr, D., and T. Poggio (1979). "A Computational Theory of Human Stereo Vision," *Proc. Royal Society of London*, Vol. B204, 301–328.

[11] Grimson, W. E. L. (1981). "A Computer Implementation of a Theory of Human Stereo Vision". *Proc. Royal Society of London*, Vol. B292, 217–253.

[12] Mayhew, J. E. W., and J. P. Frisby (1981). "Psychological and Computational Studies Towards a Theory of Human Stereopsis," *Artificial Intelligence*, Vol. 17, 349-385.

[13] Pollard, S. B., J. E. W. Mayhew, and J. P. Frisby, (1981). "PMF: A Stereo Correspondence Algorithm Using a Disparity Gradient Limit," *Perception*, Vol. 14, 449–470.

[14] Baker, H. H., and T. O. Binford (1981). "Depth from Edge and Intensity Based Stereo," *Proc. Seventh Intl. Joint Conf. Artificial Intelligence*, Vancouver, B.C., 631-636.

[15] Dhond, U. R., and J. K. Aggarwal (1989). "Structure from Stereo – A Review," *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 19, No.6., 1489–1510.

[16] Gülch, E. (1988). "Results of Test on Image Matching of ISPRS Working Group III/4," *Intl. Archives of Photogrammetry and Remote Sensing*, Vol. 27, III, 254–271.

[17] Hannah, M. J. (1988). "Digital Stereo Matching Techniques," *Intl. Archives of Photogrammetry and Remote Sensing*, Vol. 27, III, 280–293.

[18] Heeger, D. J. (1988). "Optical Flow Using Spatiotemporal Filters," *Intl. Jour. Computer Vision*, Vol. 1:4, 279–302.

[19] Horn, B. K. P., and B. G. Schunk (1981). "Determining Optical Flow," *Artificial Intelligence*, Vol. 17, 185–203.

[20] Hanna, K. J. (1991). "Direct Multi-Resolution Estimation of Ego-Motion and Structure from Motion," *IEEE Workshop on Visual Motion*, New Jersey, 156–162.

[21] Quam, L. H. (1983). "Hierarchical Warp Stereo," *Proc. DARPA Image Understanding Workshop*, 149–155.

[22] Burr, D. J., and R. T. Chien (1977). "A System for Stereo Computer Vision with Geometric Models," *Proc. Fifth Intl. Joint Conf. Artificial Intelligence*, Cambridge, Mass., 583.

[23] Ayache, N., and F. Lustman (1987). "Fast and Reliable Passive Trinocular Stereovision," *Proc. Intl. Conf. Computer Vision*, London, 422–427.

[24] Grosso, E., G. Sandini, and M. Tistarelli (1989). "3-D Object Reconstruction Using Stereo and Motion," *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 19, No.6., 1465–1477.

[25] Fua, P., and P. Sander (1992). "Reconstructing Surfaces from Unstructured 3D Points," *Proc. DARPA Image Understanding Workshop*, San Diego, 615–625.

[26] Tomasi, C., and T. Kanade (1991). "Factoring Image Sequences into Shape and Motion," *IEEE Workshop on Visual Motion*, New Jersey, 21–28.

[27] Okutomi, M., and T. Kanade (1992). "A Locally Adaptive Window for Signal Matching," *Intl. Jour. Computer Vision*, Vol. 7:2, 143–162.

[28] Bolles, R. C., H. H. Baker, and D. H. Marimont (1987). "Epipolar-Plane Image Analysis: An Approach to Determining Structure from Motion," *Intl. Jour. Computer Vision*, Vol. 1:1, 7–55.

[29] Baker, H. H., and R. C. Bolles (1989). "Generalizing Epipolar-Plane Image Analysis on the Spatiotemporal Surface," *Intl. Jour. Computer Vision*, Vol. 3:1, 33–50.

[30] Zhang, Z., O. D. Faugeras (1992). "Three-Dimensional Motion Computation and Object Segmentation in a Long Sequence of Stereo Frames," *Intl. Jour. Computer Vision*, Vol. 7:3, 211–241.

[31] Baker, H. H. (1976). "Three-Dimensional Modelling," *Proc. Fifth Intl. Joint Conf. Artificial Intelligence*, Cambridge, Mass., 649–655.

[32] Szeliski, R. (1990). "Shape from Rotation," *Proc. Computer Vision and Pattern Recognition*, Maui, Hawaii, 625–630.

[33] Lowe, D. L. (1990). "Integrated Treatment of Matching and Measurement Errors for Robust Model-Based Motion Tracking," *Proc. Intl. Conf. Computer Vision*, Osaka, 436–440.

[34] Chen, C-H., and A. C. Kak (1989). "A Robot Vision System for Recognizing 3-D Objects in Low-order Polynomial Time," *IEEE Trans. System, Man, and Cybernetics* Vol. 19, No.6, 1535–1563.

[35] Ayache, N., and O. D. Faugeras (1989). "Maintaining Representations of the Environment of a Mobile Robot," *IEEE Trans. Robotics and Automation*, Vol. 5, No.6, 804–819.

[36] Iyengar, S. S., and A. Elfes, editors (1991). Autonomous Mobile Robots: Perception, Mapping, and Navigation, IEEE Press, Washington.

[37] Besl, P. J., and R. C. Jain (1985). "Three-Dimensional Object Recognition." *Computing Surveys*, Vol. 17, No.1, 75–145.

SILICON VISION: ELEMENTARY FUNCTIONS TO BE IMPLEMENTED ON ELECTRONIC RETINAS

B. ZAVIDOVIQUE & T. BERNARD

Une "rétine intelligente" est un dispositif associant de manière intime une couche optoélectronique à des moyens de calcul. Le rapprochement "acquisition/transformation des données" favorise l'émergence d'un nouveau type d'interaction entre traitements massifs analogiques et digitaux. Nous listons donc, pour discussion, plusieurs tentatives de calcul analogique, voire neuronal, dans le cadre du processus de vision. Mais l'analogique ne suffit pas à rendre les rétines réellement "intelligentes". Si bien que nous décrivons une couche supplémentaire de traitement itératif cellulaire booléen, plausible dans de telles machines de vision "à dimension humaine", évaluée à travers quelques exemples.

Vision - capteurs intelligents intégrés - traitement cellulaire et neuronal - opérateurs visuels de base - implantation analogique vs digitale

A smart retina is a device which intimately associates an optoelectronic layer with processing facilities. The rapprochement between acquisition and processing is particularly suited for the emergence of novel kinds of interaction, between analog and digital massive computations. Therefore, several attempts of analog, possibly neural, computations linked to the vision process are listed and discussed. But analog is not enough for really smartening retinas . Then, an additional plausible coat of cellular boolean iterative processing in these "human size" vision machines is described, and commented on through examples.

Vision - integrated smart sensors - cellular and neural processing - basic vision operators - anlog vs digital implementations

B.Z. Université Paris XI - IEF - Bât. 220 91405 ORSAY cedex France et ETCA/CREA, 16 bis, Ave Prieur de la Côte d'Or, 94114 Arcueil cedex France - email zavido @ etca.etca.fr
T.B. ETCA/CREA 16 bis, Ave Prieur de la Côte d'Or, 94114 Arcueil cedex France - email VG @ etca.etca.fr

I - A GLANCE AT VISION

Visual perception performed by computers is usually decomposed as a chain of processes, as shown on Fig.1.

Figure 1 : Classical Visual Perception.

Low-level image processing is meant to extract pertinent informations like edges and regions, depths, movements... However, in most realistic enough robot vision applications only candidate-feature subsets are extracted at this level. Then these parts remain to be cleaned, gathered and organized into features which are 2-D projections of some at least 3-D phenomenon. So, at low level, the processed objects (images) are characterized by their 2-D topology, the local nature of inter-pixel correlation, and the a priori even distribution of information among pixels. Processes are thus shift-invariant with supports limited to small neighborhoods. They can hence take great advantage of specific computer architectures featuring massive spatial parallelism and simple processor interconnections.

Once the information from the original image has been filtered and concentrated into structural or semantic knowledge, the 2-D topology disappears. This is where high-level processing starts. The objects become arbitrary graphs, whose processing poses serious connectivity and/or programmability problems on multiprocessor architectures.

Let us underline the clear semantic gap between the so-called low and high level processings : as soon as it is somewhat fancy, any feature extraction has to be controlled by a more intelligent procedure which takes advantage of explicit description of an object model, or structure, or situation... While not compensating for this gap in a permanent and fundamental manner*, the "smart retina" concept brings a solution; it is at least a technological solution, but some of its instances show cheering features of optimality, when they are embedded in the context of the whole pattern recognition process.

Now, current robotics is not only moving towards involving complicated senses such as vision or aerial acoustics but it aims at associating several of them within sensor fusion schemes. Theoretical results like the so called "multiarmed bandit" theorem tend to prove that it is worth implementing some local computing power closer to sensors, when the communication bandwidth necessary for control is already causing problems.

This makes another reason to focus on smart retinas, vision being likely to play, as in the human case, a major part in robot perception.

A smart retina is a device which intimately associates an optoelectronic layer with some processing facility. The closeness definitely suggests a VLSI implementation approach, possibly monolithic. But, so far, only elementary feature extraction, up to limited object identification, has been proved technologically feasible.

In that case, why should "smart retina" imply "integrated retina"? Here is a non exhaustive list of possible answers:

- vision usually means immense amounts of input data
- the current state of wiring technology causes the signal/noise ratio to fall drastically at circuit output
- in any case, changing the computing topology is often very power consumming
- the tradeoff to be made between precision and quantity of information is likely to benefit from massive loose computational style rather than the common precise computational style
- analog to digital conversion is a waste in many respects:
  .. there is a loss of information due to conversion,
  .. there is a loss in speed and functionality (artificially added operations to calculus),
  .. exploiting the natural correlation in images will require rebuilding the initial topology,
  .. it puts processing apart from data flow
- real external conditions for vision require fast feedback loops (from adapting to light, up to feature extraction)

To propose a more definitive answer, we first give a slightly more precise definition together with

first properties (§II), we then explain some very primitive examples (§IIIa) to illustrate:

- first, the concept of smart retinas
- second, the input-output problem

In these examples, the outside world is simplified (either exhaustively described or in translation). Then, a bit of analog processing followed by a uniform result gathering performs the intended task, and only one or two global outputs are produced.

However, the preceding experiences suggest potential benefits from "analog thinking" when an algorithmic concept comes to cohabit with analog implementations of early vision processes. Descriptions of analog phenomena inside the system provide a language which helps to drastically compact any design, and enforces some interesting improvements at the algorithmic level. This fact is illustrated in (§IIIb) by comparisons between implementations of the convolution or other basic operations like differentiation. Indeed, in less toy-like cases than § IIIa's, current robot vision does not allow routine actions in such a direct manner and

* there is no clear evidence, however, that this gap be anything but artificially added by techniques.

anyway, such actions would be triggered on a larger set of parameters.

This shows integrating is not enough, even associated with analog thinking, hence introducing the concept of "rough vision", based on separating the structure of the image from the semantics it refers to. It applies first to object recognition thanks to neighborhood combinatorial logic which is easy enough to implement on retinas. Logical implies binary, but in this process the adapted binarization will be made a true processing operation, possibly a feature extraction and not only an A/D conversion. This is described and commented on in § IV before conclusion.

II- THE "RETINA CONCEPT" : A PANACEA ?

Let us define more precisely "smart retinas" as tentative "human-size" vision machines, intimately associating optoelectronic devices with analog-to-digital converters and (minimal) digital processors to be integrated on monolithic (CMOS) circuits.

Such circuits can be viewed then as stacks of "3" intermixed functional layers :

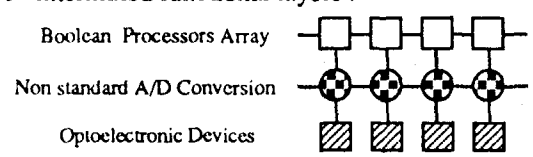


Figure 2 : The "Retina" circuit (cross section).

From a VLSI point of view, a Retina structure is up-to-date. It exploits today's abilities of submicronic technologies to allow a rapprochement between acquisition and processing (up to few 100's x 100's elementary processors, with few dozens transistors each, can be gathered on a monolithic circuit using a 1µm CMOS technology). The intimate association of different functional layers however is subject to strong topological constraints. These are suggested to be naturally satisfied on fig.1.

While certainly related to existing biological visual systems (but still very far and caricatural), the "retina concept" features numerous and fruitful advantages considering § I:

- The classical serial bottleneck separating acquisition from processing is replaced by a parallel conversion layer. Instead of artificially breaking and then reconstructing the 2-D topology (because of limited I/O bandwidth), the analog-to-digital conversion is harmoniously "sandwiched" between analog acquisition and digital processing.

- A/D conversion is non-standard but well managed. Image sequences are known to be locally correlated both in the space and time domain. This can be advantageously exploited to encode the analog image flow into compact digital representations. For the sake of topology, this naturally leads to a binary image representation, most often based on a one-to-one mapping between analog and binary pixels. If the sole spatial correlation is taken advantage of, the analog-to-binary encoding procedure is called "halftoning". We will show in section IVb this can be neatly implemented in silicon. But a tradeoff occurs: more pixels for less grey levels, or the opposite.

- Though halftoning can be considered as an unavoidable quantization operation implying a loss of information, which has to be minimized with respect to some peculiar signal processing criterion (as we do in section IV), it actually acts as an information filter, which can enhance specific early vision features, such as edges, regions, movements, optical flow, depth... (cf .[Mea 88] &[Hut88]). Processing inside the retina thus appears as a close cooperation between an analog layer and a boolean one.

- The analog information representation, right after acquisition, is so heavy that arbitrary interactions between pixels cannot be implemented easily to be programmable. Only information processing structures provided with a highly physical meaning that map straight into silicon, leave some hope to avoid the burden of storing, duplicating and moving analog pixels.

- By massive parallelization of both information flows and processings, operations inside the retina are brought closer in space and time. This emphasizes the interest of bidirectional (instead of only bottom-up) information flows, because the top-down feedback can be fast enough to ensure some convergence properties. For example, a complex problem like matching successive images of a moving scene, is reduced to its simpler expression when the sampling frequency is high enough.Another example is neural interactions between analog and boolean layers.

Thanks to these advantages, it becomes possible to output meaningful results in accordance with the claim of smartness, but due to technology, there still remains an additional price to pay: either to deal with very specific applications or to particularize vision in some other manner like restricting it to a rough type (see § IVa). On top of that, the above list shows anyway a need for a fair share of analog contribution to meet the constraints of rapidity and compacity as imposed by real time robot vision. This makes the layers in fig.2 become the 3 mousqueteers of robot vision as they are actually four, being joined by an analog processing layer of prime importance. We now analyze significant research results within that perspective, prior to detailing more of our own work.

III - ANALOG ELECTRONICS AND RETINAL FUNCTIONS

IIIa - Specific attempts

As far as we know, the first significant attempt to introduce some intelligence within the sensor chip goes back to [Lyo81] with the desire for a high-reliability mouse (used to track the movement of a workstation user's hand) with no moving parts. As the "optical mouse" is downward looking at the special pattern of a pad on which it is moved around, motion is detected and measured. The "optical mouse" is a mostly digital sensor used in a very cooperative environnement : an hexagonal grid. However, important features like the local automatic gain control (AGC) are already present through the use of self-timed circuit techniques and mutually inhibating light sensors. The tracking algorithm, which compares 2 successive 4x4 images is based on a case by case approach, dealing with the 900 possibilities of image couples.

The theme of motion detection on uniformly moving scene has generated a fair amount of work since then. In [Bis84], the stress is put on high resolution 1-D motion detection, in order to determine 3-D motion from several sensors. In [Tan84], a "paperless" version of the optical mouse is integrated, to deal with less cooperative environnements. An image of an arbitrary scene is sensed by the array of photodiodes, stored and correlated with the next image taken on the next cycle. The position of maximum correlation indicates the relative motion of the image during the time between samples. A global AGC is used, and correlation computations are both analog and digital. Finally, a fully analog and time-continuous version has been integrated, as described in [Tan88] and [Mea88], that makes full use of global collective neural computations to output the velocity vector of the image.

For these applications, the output problem is implicitely solved because only one or two global informations about the scene are actually extracted from the sensor chip. This is also the case for sensors that deal with simple target tracking applications, like following the brightest spot on an image [DeW88] or following a spot among other bright spots [Umm89], and for which only a couple of coordinates have to be output.

However, early vision, which takes full advantage of collective computation based on only local connections within VLSI circuits, generally does not change the topology of the processed objects : an image is transformed into another image. In this context, CCD technologies can support a large family of linear operations, particularly needed for spatial and temporal convolutions as in [Bea89]. These operators can be completed by simple saturation based non-linearities as thresholding or magnitude comparison as done in [Eid88]. Early vision has also been integrated in standard CMOS technologies, from compact spatio-temporal differentiation in the "silicon retina" described in [Siv87] and [Mea88], up to expensive optical flow computation in [Hut88].

At last, various approaches try to deal more or less successfully with the problem of outputting the information present on the image. In [Gin88], only the areas of interest are output from the sensor. The image may be also binarized or halftoned as in [Mar89]. Three-dimensionnal integration as presented in [Kio88] and [Kat86], is also a possible way allowing the superposition of different processing levels on the input image, and thus allowing the output of only high-level compact information.

IIIb - A more structured approach towards vision

Transducing light into current.

Standard CMOS technologies are well-adapted to visible light detection : when an optical signal impinges on a p-n junction operated under reverse bias, the depletion region[1] serves to separate photogenerated electron-hole pairs, and an electric current flows in the external circuit. *This light-matter interaction has to be considered as the very start of the vision process. Several configurations using different devices are available, of which the choice is not neutral and can be more or less adapted to the subsequent hardware and/or software vision layers.*

The simplest light detector is the photon flux integration mode photodiode used in CCD cameras. It is simply constructed by diffusing a highly n-doped area at the surface of a p-type substrate (an NMOS technology is sufficient). After being initially reverse biased, the junction capacitance is discharged by the photogenerated current. At the end of the exposure, the voltage decrease is about exponentially related to the illumination level and integration time : $\log[V(t)/V(0)] \propto - \Phi.t$.

When response speed is not critical, but power is needed, a natural byproduct of the CMOS process [Mea88] can be used : the vertical bipolar transistor. The base is an isolated section of well, the emitter is a diffused area in the well, and the collector is the substrate. Electron-hole pairs are generated at the well-substrate interface where the p-n junction is reverse-biased. For every photogenerated majority carrier arriving into the thin base (from the collector), about a thousand minority carriers pass through it (from emitter to collector) before the necessary recombination finally occurs : this is the phototransistor action. This natural current gain can be used before subjecting the signal to any noise from subsequent amplification stages. *It can also be*

[1] When a p-n junction is formed between two oppositely doped semiconductor, a charge depleted region appears at the interface in which very high electric fields are encountered. Instead of getting recombined, electron-hole pairs generated in this zone are violently separated.

controlled making the vision sensitivity possibly dynamically shifted.

Incident light on a region of the surface of a semiconductor is also known to cause a local change in that region's conductivity. As noticed in [Her89], *this effect can be exploited to construct a global representation of incident images, which possibly allows faster pattern recognition processes by implicitely solving the image output problem.*

Logarithmic representation of illumination intensity.

In order to properly operate in outdoor scenes (say from moonlit to sunlit scenes), electronic photoreceptors must give meaningful outputs over several orders of magnitude of illumination intensity. The linear light to intensity conversion occurring within depleted devices like photodiodes and phototransistors thus must be followed by some further non-linear conversion. Moreover, as pointed out in [Mea88], *it is very desirable to make the voltage difference between two points depend only on the contrast ratio between the two corresponding points in the image. Indeed, in a simply modeled scene, this contrast ratio is a ratio between reflectances, which are independent of the relative illumination level.* This mathematically implies the use of an exponential law. Fortunately, exponential phenomena exist in a semiconductor like silicon : the appearance of the source-to-drain channel in MOS transistors is ruled by the Fermi-Dirac distribution (stastistical physics & Boltzmann law) which ensures that charge carrier concentrations within the channel depend exponentially on the gate voltage along about a half volt wide interval, which is called the weak inversion (or subthreshold) region. This has been used by [Mea88] where the current from a phototransistor is fed into two diode-connected MOS transistors in series operating in the weak inversion region, and providing a 0.2 volt output voltage decrease per decade increase in current (see Fig.3).

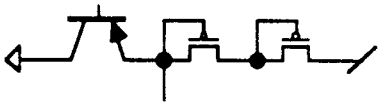


Figure 3 : Logarithmic Photoreceptor.

Using the MOS transistor in the weak inversion region to exploit its exponential behavior is a first example of the search (among the wide variety of analog VLSI phenomena) for adequate non linear operators, *which are finally the ones to extract the important information from the input image signal.* Among others, non linearities that easily map into silicon are the square law, the sigmoïd function, saturation and hysteresis phenomena, and comparison operators. For example, hysteresis inverters are fundamental devices in the "analog toolbox" as shown in [Ber88] and [Smi89].*These non-linearities play key roles from simple yet time consuming operations like thresholding up to advanced neural optimization algorithms like neural halftoning [Ber90] or optical flow computation involving line processes [Hut88].*

Linear functions.

Besides the use of tricky non linear devices, analog implementations of vision processes rely on the existence of a "library" of (hopefully) compact cells that embed more regular transformations, such as storage, duplication, addition, substraction, multiplication but also piecewise linear functions like the absolute value, and more generally conditional functions like the maximum or minimum functions. However, implementations depend on wether the input signal is a voltage, a current or a charge. One of the skill of the designer is to find the right information supports to embed a particular vision algorithm efficiently. *This is nothing but the equivalency for type conversion of variables in programmed image processing !*

The charge domain, taking full advantage of CCD processes [Boy70] in which a charge can be stored or spatially shifted at negligible loss, is unsurprisingly suited to linear image processing [Tie74]. Charge mixing or sharing are the basic operations for additive functions, we will see in the next paragraph how they can naturally implement very useful spatial convolutions. But substraction can also be implemented thanks to 3-D coupling as used in [Fos84] : besides the usual lateral coupling used in charge transfer devices, the vertical coupling between the charge on the electrode and the charge in the channel embeds a natural differencing phenomenon. Charge splitting, which is equivalent to multiplying by a positive coefficient less than one, can also be implemented as explained in [Ben84]. If CCD's are used in conjunction with active CMOS transistors, they can implement up to charge magnitude comparison and non destructive sensing and amplification (cf [Col87] & [Fos87]). Time delaying is also easily embedded as it is controlled by external clocking sequences : this is a definite advantage for motion detection applications. However, clocking requirements and difficulties to implement non-linear operators in the charge domain, other than saturation nonlinearities, suggest that currents and voltages are indispensable alternative system variables for the analog implementation of vision processes.

Linearity in the current/voltage domain looks less natural since operators generally involve the use of MOS transistors, possibly associated with bipolar transistors (BiCMOS technology), all of which are all but linear. Ranges of linearity are consequently narrower than in the charge domain, with widths possibly as small as 0.2V in the case of [Mea88]. A common operation is the duplication of a signal, illustrated on fig.4, either by a current mirror or by a voltage follower. As can be noticed, the price to pay for the same

operator can seriously differ, depending on the type of the input signal.

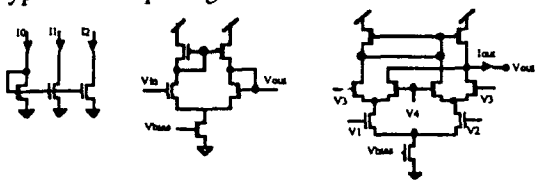


Figure 4 : Current Mirror (I2 = I1 = I0), Voltage Follower (Iout = 0 ⇒ Vout = Vin) and Gilbert Multiplier (Iout ∝ Ibias • [V1-V2] • [V3-V4])

Another important operation is the four-quadrant multiplication that can be implemented thanks to the "two-stage" differential pair shown on fig.4, and known as the CMOS version of the Gilbert multiplier [Gil68] : A triple product is actually performed, between two algebraic quantities (V_1-V_2) and (V_3-V_4) and a positive quantity I_{bias}, which is the current flowing in the lower transistor and set by V_{bias}. However, image processing often involves the interaction of larger sets of input signals. The fundamental autocorrelation properties of images are responsibie for the central importance of smoothing and differentiating operators in both the spatial and temporal domain. As far as motion detection is concerned, electronic time constants must fit the time scale of motion events in the observed scene : unfortunately, the largest RC constants that can be controlled in silicon are smaller than 0.1ms = 10MΩ x 10pF, which is too fast for our real world. This problem can be avoided by discretizing time, or using peculiar controllable resistive circuits such as the one presented in [Siv87]. After this general and brief presentation of a starting repertoire of general analog operators that can be used in "analog vision", we now present a few examples where physical laws inherent in electronic have met the operating or computionnal need of certain aspects of vision.

Gaussian Spatial Convolution.

Gaussian kernels have been shown to be of primary importance in edge detection algorithms (cf [Can86]). Thanks to the Central Limit Theorem, the repeated binomial convolution of a signal or an image is a good approximation to gaussian filtering. Sharing and halving charge packets is easily performed in the charge domain, particularly with the help of charge coupled devices. So binomial convolution can be performed in a CCD imaging array clocked by an unconventional method as described in [Sage85] and generalized in [MIT88]. Fig.5 shows a novel 2-D CCD convolution cell to be used in an hexagonal tiling. The boundary of the cell is indicated by a shaded area. The structure of the cell is simplified : after a certain clock sequence, charges are transferred from bucket to bucket according to the arrows.

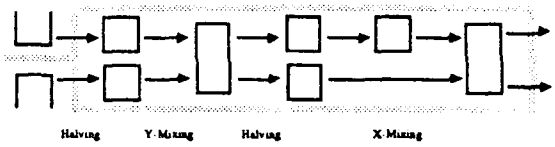


Figure 5 : 2-D Parallel and Pipelined Binomial Convolution

The left part of the cell performs a binomial convolution along the vertical axis, while the right part convolves along the horizontal axis in a manner which is similar to implementations of FIR filters in classical signal processing pipelined architectures. The image is input column after column on the left side. The final network's heigth matches the number of rows in the image, while its width depends on the gaussian kernel's variance to be implemented. *Finally an input image is massively convolved in a parallel pipelined fashion, and the I/O problem is degenerated from 2-D to 1-D. Moreover, the variance σ of the gaussian kernel can be controlled by using a partial width of the network, hence adapting the resolution which the image is processed at.*

Whereas the choice of the binomial filter is just one efficient way to iteratively approach the gaussian shape, there are other diffusion or relaxation processes that are more typical of fundamental electric equilibria found in VLSI, and that we present now.

Diffusion-Based Spatial Convolution.

Static image processing is fundamentally based on spatial interactions between pixels or sub-structures that are more or less far apart in the processed image. This corresponds to the structural approach of vision, which can actually take place at every level of vision. When performed at the lowest level anyway, these spatial interactions are extremely computationally intensive and would definitely benefit from "natural" physical interaction phenomena.

When statistically considered, images have to be processed in a shift-invariant manner, without privileging any particular direction. Moreover, it makes sense to weaken their interaction as pixels get further apart from each other. We are thus looking for a shift-invariant phenomenon allowing the isotropic but radially decreasing diffusion of a physical quantity towards its neighborhood. This can be implemented thanks to current diffusion in resistive materials, which is a linear process : if a current is injected at some point of a resistive sheet of conductive material featuring a uniform surfacic leakage resistance towards some source of potential (e.g. ground), the induced voltage profile or impulse response is indeed a rotation-invariant kernel (cf [Ber88]) whose radial shape is given by the first modified Bessel function : $V(r) \propto K_0(r)$, where r is an absolute normalized radius. Before discussing *the relevance of the "diffusion kernel" shape for vision purposes*, let us characterize it more precisely. To get some physical intuition about

$K_0(r)$, we can consider the current diffusion in the adjacent dimensions : 1-D and 3-D. For a resistive line $V(r) \propto \exp(-r)$, and for a resistive volume $V(r) \propto \exp(-r)/r$. As expected $K_0(r)$ shows an intermediate behavior that we can precise thanks to equivalent forms for small and large arguments : $K_0(r) \tilde{}_0 -\log(r)$ and $K_0(r) \propto \tilde{}_\infty \exp(-r)/\sqrt{r}$.

In a VLSI circuit however, we are bound to spatially discretize this current diffusion process onto a resistive ladder network of the type shown n fig.6. This network is shift-invariant. Horizontal resistors are called diffusion resistors with value R_d. Vertical resistors are connected to ground, and called leakage resistors with value R_l. Input injected currents X_i diffuse all over the network contributing to the output node voltages V_j. This process is linear such that we get $V=K*X$, where K is a characteristic convolution kernel depending on the sole ratio R_l/R_d. *If this ratio is variable, this is truly a multiresolution facility which is available to the analog vision algorithm designer ! Recent devolopments about the use of wavelets (cf [Mal89] & [Mal90]) in image processing still enhance the importance of such a feature.*

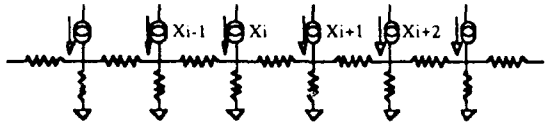


Figure 6 : A Resistive Diffusion Network (1-D version).

In the 1-D case, the kernel voltage profile is simply exponential (as in the continuous model), that is $K(r) \propto \exp(-r)$ or $K(x) \propto \exp(-|x|)$ because Kirschoff laws can be written in a recurrent manner. In the 2-D case however, there is no closed form giving K(x,y). There are actually at least 2 network topologies that can be used : either rectangular or hexagonal. The continuous model proves useful to understand the asymptotic behavior (towards ∞). Unlikewise, close to 0, that is for the central pixel on which the unity current is injected and for its neighbors, infinite voltages forecasted by the continuous model vanish ; the node voltages are finite and have to be estimated thanks to iterative algorithms.

It is fairly easy however to derive analytically K^{*-1}, the inverse of K for convolution (regardless of the dimension or the network topology) which in turn yields FT(K), the Fourier transform of K (with K considered as a distribution). *This is a door to understanding the effect of the discrete current diffusion in terms of frequential analysis.*

By expressing Kirschoff laws for each node of a rectangular 2-D extension of the network shown on fig.6, we get $(\forall i \in Z)\ (\forall j \in Z)$

$$X_{i,j} = (1/R_l+4/R_d).V_{i,j} - 1/R_d.(V_{i-1,j}+V_{i+1,j}+V_{i,j-1}+V_{i,j+1})$$

Figure 7: Laplacian Kernels in the rectangular case Δ_r and the hexagonal case Δ_h.

By using the dirac distribution δ and the rectangular laplacian $\Delta_r = 4.\delta_{0,0} - \delta_{-1,0} - \delta_{1,0} - \delta_{0,-1} - \delta_{0,1}$ (shown on fig.7), Kirschoff laws yield :

$X = (\delta/R_l+\Delta/R_d)*V$, where $*$ stands for convolution. But $V = K*X$, so :

$$\mathbf{K^{*-1} = (\delta/R_l+\Delta/R_d) \quad (1)} .$$

We can now switch to the frequency domain to get the periodic Fourier transform of K^{*-1} and finally K, with frequency coordinates ω_x and ω_y :

$$FT(K^{*-1}) = 1/R_l+4/R_d.[\sin^2(\omega_x/2)+\sin^2(\omega_y/2)]$$

$$\Rightarrow \mathbf{FT(K) = (1/R_l+4/R_d.[\sin^2(\omega_x/2) +\sin^2(\omega_y/2)])^{-1} \quad (2')}$$

We have just been characterizing 2-D "diffusion kernels" in many aspects. We have now gathered enough information about them to show their relevance for vision purposes.

Within recent years, much work has been devoted to the optimization of smoothing diffusion kernels allowing the removal of noise before edge detection. Beside the "gaussian hegemony" mentionned in the previous section, exponential filters have also been proved in [She86] and [She87], to be optimal for a multiedge model. Now, when a straight edge is convolved by a 2-D diffusion kernel K, K is actually projected according to the direction perpendicular to the edge into ... an exponential filter ! The edge detection capabilities of the "silicon retina" described in [Mea88] are the straightforward application of this property. We have also proposed (but not implemented) a more sophisticated edge detection algorithm implementation based on diffusion kernels in [Bel88].

We will also show in § IVb that diffusion kernels are particularly suited to the halftoning problem, that is the analog-to-binary conversion of images, as mentioned in [Ber90].

Though the fully 2-D parallel implementation of diffusion kernels seems much more "natural" than that of gaussian kernels, there remains a few difficulties to solve before it can be really mapped into silicon. As previously mentioned , it is very desirable to implement controllable resistors (at least the leakage resistors which are the less numerous) in order to benefit from an analog multiresolution facility. This apparently requires the use of active resistors. The natural compacity of the diffusion network allows a large number of pixels to be integrated on the same circuit, however it also raises severe power consumption problems. Using transistors in the weak inversion region is a potential solution to lower current

values, as explained and applied in [Mea88]. In that case, resistors are controlled thanks to the tunable transconductance of a CMOS differential amplifier used as a unity-gain follower. Yet, the linearity range is not larger than 200mV. *When the device gets saturated, it turns out to perform a simple but automatic segmentation of the input image by preventing two neighbor pixels from exchanging more than a fixed current upper bound.*

However, considering the uncertainty on each transistor characteristics in the weak inversion region (up to an equivalent gate voltage uncertainty of a few tens of millivolts), the linear range narrowness seems more undergone than desired : it requires dynamic selfcorrecting circuitry or static a posteriori analog compensation by EPROM-like techniques[2], all of which may be area-consumming. Further more, such analog voltage precision seems to prevent the cohabitation with digital layers which requires external clocks, inducing significant amounts of noise.

We have been studying an alternative solution to the implementation of diffusion filters based on an unconventionnal use of switched capacitors (cf [Ber88] and [Ber90]). This approach leads to reasonnable power consumptions : To give an order of magnitude, if a (fairly large) 1pF capacitor was to be charged and discharged from 0V to 5V at a 1MHz frequency at every pixel site, a 100x100 pixels retina would demand a power of about 0.1W. However, either it requires an analog CMOS process providing a double polysilicium layer, or "slightly" non-linear p-n junction capacitances have to be used (cf [Ber89]). *In the latter case, it is amazing to notice how many roles the same simple device can play : a strip of n-diffusion over the p-substrate will be used a) to connect two pixels, b) to act as a switched capacitor and c) to convert light into current.*

Finally, a globally better precision can be achieved with comparable silicon area, partially because capacitors are really easy-to-use bidirectional media to perform *"type conversion"* between charges and voltages.

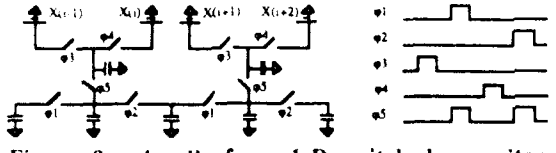


Figure 8 : 4 cells from 1-D switched capacitor diffusion network and associated clocking cycle.

Fig.8 shows how a 1-D image X, input through voltages sources, can be convolved by a diffusion kernel on a switched capacitor network. Horizontal and vertical capacitors are called respectively diffusion and leakage capacitors. A few peculiarities have to be emphasized. The convolution is only asymptotically obtained after a sufficient number of elementary switching cycles. About 10 are necessary to reach a 0.1% precision when $C_d=C_l$. The ouput voltages are somewhat immaterial since only half of them are available each time clock $\varphi 5$ is high in the clock cycle. "Neurons" (pixels) are indeed separated according to their parity. This iterative aspect allows to share a single leakage capacitor between a pair of odd and even neurons. This neatly generalizes to 2-D, where neurons are separated in a checkerboard fashion. Now only the elementary cycle is presented on fig.8. *Though capacitances have static values, a discrete multiresolution facility is recovered thanks to the use of more complex cycles in order to obtain narrower diffusion kernels or even different types (e.g. gaussian-like) of kernels at no further implementation cost !*

We have just been comparing different implementations of regular diffusion networks. However, when resistors can be separately and dynamically controlled, resistive networks can have much broader early vision applications (cf [Hor86],[Koc86],[Hut88] and [Koc89]). The price to pay is area, but also algorithm complexity : for example, negative resistors, which are area-consumming, can also pose convergence problems.

From edge to motion detection

The above examples have made tangible the intuition that vision can be fruitfully thought about in an analog manner. But even more exciting are the unifying "short cuts" that simple analog devices, within a continuous range of operating conditions, can provide between usually well separated vision concepts.

The silicon retina described in [Mea88] is an examplary case embedding into a regular resistive and capacitive network both edge and motion detection, in a tunable manner. A schematic and linear version is shown on fig.9.

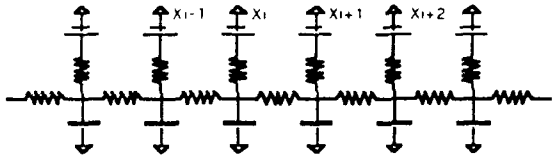


Figure 9 : A 1-D linearized version of the "silicon retina" (cf [Mea88])

The resistive part is just an equivalent version of the current diffusion network shown on fig.6, but inputs are now voltages, which dynamically represent the light intensity falling on each pixel (this was actually an intermediate step of the metamorphosis of the resistive network shown on fig.6 into the switched capacitor network shown on fig.8). The equivalence is a direct consequence of the Norton-Thevenin theorem. Besides, one capacitor has been added to each network node, in order to perform temporal differentiation. The outputs of the network are the voltages across the

[2]Such techniques provide long term analog storage of charges.

leakage resistors. The spatial and temporal network behavior is described by its space and time constants. The space constant depends on the sole ratio R_d/R_l (if diffusion resistance are cut, R_d gets infinite and the space constant becomes 0), whereas the time constant varies linearly with R_l and R_d. *So the same simple network used with different resistance values can continuously switch from edge to motion detection. Beyond this linearized view of the "silicon retina", the devices saturability also plays a significant role in the overall computation.*

From mean to median filtering

The saturation of a unity gain follower, when used as a resistor between the output node and the input node (which appears as a voltage source), can be clearly interpreted from a vision point of view when used in a "follower aggregator circuit" (cf [Mea88]) as shown on fig.10 (The G_i are the respective conductances of the voltage followers in their linear region).

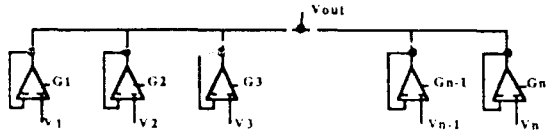


Figure 10: Follower aggregation circuit.

As explained in [DeW88], if all the V_i voltages are within the same 200 mV wide interval, all the voltage followers are operated in their linear region. As the sum of the currents at the output node must be zero, a weighted mean of the input voltages is computed: $V_{out} = \Sigma G_i.V_i \; / \; \Sigma G_i$.

On the other hand, if the V_i voltages are too further apart from each other, a large majority of voltage followers will be saturated, that is they will act as current sources. The saturation current is known to be proportional to the transconductance G_i. If all voltage followers were saturated, the
final output voltage would be such that :

$$\sum_{V_i<V_{out}} G_i = \sum_{V_i>V_{out}} G_i$$

This computation defines a weighted median.

Finally the quantities on which the computation is performed appear to be the conductances G_i. They are set by the bias voltage of the differential amplifiers, and can represent the incident light as is the case in [DeW88]. On the other hand, the input voltages are used to control the type of computation. If a spatially increasing profile of voltages is input to the network (such that voltages differences $V_{i+1}-V_i$ are constant), V_{out} will naturally indicate the area on which the incident light is maximal. *Depending on the slope of the voltage profile, the precise value of the "pointer" V_{out} will result of a weighted mean (small slope) or weighted median (large slope) or a tunable combination of both, in order, for example, to perform an adequate noise removal on the input image.*

IIIc - Does "analog" mean "smart enough" ?

We have just been browsing from the most specific analog attempts to integrate vision up to more structured approaches, putting in evidence may be unexpectedly strong relationships between analog techniques and "high level" vision concepts. We have illustrated the versatile power of analog hardware within VLSI circuits, but also its limitations due to technological and more generally physical constraints, which, for example, can make the cohabitation with digital hardware uneasy.

However, very few people have proposed even partial solutions to solve the output problem for general enough applications. Many research groups in the field do claim that this problem of input output in vision is smartly solved thanks to windowing i.e. reducing the field of processing, then the number of processed pixels, by approximately two orders of magnitude. Thus processing inside the shrunk data may be more sophisticated. They dangerously underestimate the control problem of positioning the window, now well-known as the problem of "narrow in wide angle", or of attention focusing. In the research about multisensor fusion, most proposed solutions to it ask for advanced stochastic control (Bar84, Mer88) or extended linear filtering(Bar89). Other smart attempts closer to smart sensors deal with fovealisation (multiresolution in silicon) and or active vision i.e. short loop between camera actuators and data processors to come up with natural regularisation.

IV - YET ANOTHER MESH ARRAY SMART SENSOR?

IVa - Rough vision

In order to get to some programmable or adaptative recognition, on top of analog thinking we still had to adapt the retina concept jointly from the technical point of view of the implementation, and the more fundamental one of vision.

On the technical ground:

- As far as the digital layer is concerned (the top one on fig.2), the choice of a binary image representation is the crux of the matter. First, the maximization of computational power at fixed implementation cost is likely to strongly benefit from the boolean nature of the quantized images. The complexity of a processor as a function of the number of bits it processes is at least quadratic (e.g. for a multiplication operation). By its deep homogeneity, the binary representation

(1bit/pixel) allows the use of really "bare" monobit processors (about only 25 transistors). Their interconnection with their four closest neighbors turns the top layer into a cellular mesh array that can implement any shift-invariant boolean function (cf. Gar88). The larger the function support, the longer its computation. The function support is indeed scanned thanks to iterative image shifting. So supports are practically limited to local neighborhoods. This is why we have called those boolean operators : NCP's, standing for Neighborhood Combinatorial Processings.

- NCP's are well-adapted to low-level image processing. More generally, NCP's allow the implementation of a "rough but complete" type of vision, for which NCP algorithms results can be output from the retina in a concentrated fashion (such as the image integral, higher order moments, or sparse pixel coordinates) thus avoiding a potential communication bottleneck with the external world.

- Last but not least, the binary representation provides a fruitful duality between operators and objects. Any NCP can be simply interpreted as the alternative recognition of a set of boolean patterns. Now, on the one hand, any image portion inside the retina is a potential NCP pattern. On the other hand, any pattern can be processed as an image inside the retina. This confers autoprogrammation abilities on the retina, which are of particular interest for tracking purposes (Gar88).

On the vision ground:

The magic in the previous section becomes the halftoning process which makes the whole NCP concept available and sensible. Now there is again certainly something to pay for it. Let us explain right away the trade-off hiding behind a "rough but complete" vision, by giving first more formal definitions and properties.

1) NCP's (Neighborhood Combinatorial Processings) are exactly the shift-invariant operators on binary images. We have concisely defined them using set theory, where binary images can be represented as finite subsets of Z^2. $FP(Z^2)$ standing for the set of finite subsets of Z^2 (binary images), NCP $t_{U,V}$ is defined thanks to two parameters, $V \in FP(Z^2)$ and $U \subset P(V)$ (set of the subsets of V), by

$$FP(Z^2) \rightarrow FP(Z^2)$$

$$t_{U,V} I \rightarrow t_{U,V}(I) = \{ z \in Z^2 / (-z+I) \cap V \in U \}$$

2) NCP's are stable through the composition operation o :

$\forall V_1 \in FP(Z^2)$, $\forall U_1 \subset P(V_1)$, $\forall V_2 \in FP(Z^2)$,

$\forall U_2 \subset P(V_2)$,

$t_{U1,V1} \circ t_{U2,V2}$ is an NCP $t_{U,V}$ whose parameters are

$V = V_1 \oplus V_2$ and $U = t_{U1,V1}^{-1}(U_2)$.

Therefore, NCP's can be decomposed along a noise and distorsion tolerant structure revealing process, according to the scheme shown on fig.11. We note $\perp$ this decomposition operation based on a context specific pattern base, as detailed in (§IVc)

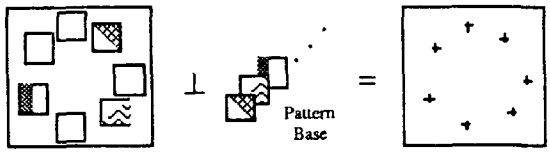


Figure 11: N.C.P. Functional Decomposition.

So, if all semantics or context handling is "subcontracted" to a controller which could be nothing more than a boolean pattern base manager, then in many well delimited cases (up to target tracking and more!) recognition is merely a tolerant dot pattern matching at some point generalizing both the notion of interest (say area of) and multiresolution. Figure 12 displays some suggestive graphic examples:

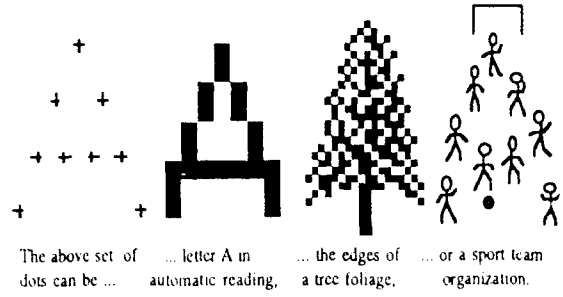


Figure 12 : Structure, Semantics and Multiresolution ...

It is easy to understand that such considerations hold only for very restricted cases, making up the "rough" vision. The direct counterpart of the rough character of the retina vision is its completeness, i.e. the ability to carry out vision processes from acquisition to decision (cf fig.1).

This highly pragmatic tradeoff remains most valuable compared to other potentially monolithic and complete vision systems, such as pattern recognition neural networks. As far as Hopfield networks are concerned, it is currently admitted that at least 10 neurons are required per basin of attraction. Similar properties hold for the Hebb's rule. Now VLSI technologies currently limit the number of highly interconnected neurons on the same circuit from a few tens up to a few hundreds (when interconnection tricks are exploited). So the number of patterns that can be recognized by today's integrated neural networks is bound to a few tens, and it is not likely to increase significantly but if a radical mutation occurs to solve the "interconnection" problem. On the contrary, the Retina concept makes a better use of today's integrating facilities. Due to the "vision

roughness", there is no need for more than about a hundred of patterns, that are to be provided by a robust enough controller. Pattern recognition is certainly slower than when performed by analog neural networks, since computations are iterated inside the Retina. However it is so easy for the retina to pass from one context to another by changing the pattern base, whereas neural networks have to enter a long learning phase.

If integrated neural pattern recognition is still several orders of magnitude ahead, a neural approach however is of immediate interest for simpler and more regular operations like non-standard A/D conversions within the Retina context. The section IVb explains why, displaying an exemplary application. As already mentioned in § II, the filtering associated to halftoning does influence NCP to be used and determines the "retina vision". So in § IVc, we finally come to grey level picture processings inside the retina.

IVb - Analog-to binary conversion and halftoning

Again, the whole structure and in particular the conversion layer can take full advantage of the computational abilities of highly interconnected analog networks. In particular, the homogeneity of the binary representation is determinative. The even distribution of information over all bits (each one will support an information of physically equivalent importance) has a direct influence on the "energetic landscapes" used in early vision optimization problems. This especially prevents local minima from being too shallow and hence improves the performances of neural computations. A well-known counter-example is the 4-bit A/D converter studied in [Tan86] and [Smi86] where the presence of such undesirable local minima is put in evidence.

Halftoning techniques deal with the bilevel rendition of continuous tone pictures. The retina structure requires a fast and parallel halftoning technique with good fidelity at low implementation cost! Unfortunately, among usual halftoning techniques, none meets all these constraints. A state of the art can be found in [Bil83] and [Uli88]. Error diffusion methods, considered to be the best, are inherently sequential, hence unappropriate. Ordered dither (cf [Bay73]) is the only "cheap" parallel technique, but with quite a poor fidelity.

We have dealt with halftoning as a first general-purpose milestone for the conversion layer of our retina , towards a more advanced vision system. As reported in previous work [Ber88] analog neural networks provide a very attractive alternative to the halftoning problem.

The energy approach

The retina structure provides a one-to-one mapping between analog (bottom layer on fig.2) and binary pixels (top layer). So, for any site in the retina array, whose index is k ($k \in Z^2$, where Z is the integer set), an analog signal $X(k) \in [0,1]$ is received from a photosensitive device and a binary information $B(k) \in \{0,1\}$ is produced by the halftoning conversion.

We want to keep B close to X according to a tonal/spatial fidelity criterion. We choose to minimize a frequency-weighted squared error between X and B. Through Parseval equality, it is mathematically equivalent to perform the minimization of the following quadratic energy E (. stands for image dot product and * for convolution product) :

$$E = 1/2 \, . \, [\, L*(B-X)\,] \, . \, [\, L*(B-X)\,]$$

L must be considered as an intermediate convolution kernel whose coefficients are related to the above frequency weights through Fourier transform. We mainly use kernel K = L*L, which is of immediate meaning for the actual implementation of the procedure.

As shown in [Ber90], local minima of E prove to be fixed points of a compact evolution equation :

$$B \rightarrow \mathbf{Hinv}_{K(0)} \circ [K*(B-X)] \qquad (2)$$

Hinv, which stands for Hysteresis Inversion, appears as a fundamental non-linearity in the "analog toolbox". It is illustrated on fig.13. The hysteresis cycle width is responsible for the convergence properties of the whole network.

Figure 13 : Hysteresis inversion : a fundamental non-linearity.

Along with compactness, the choice of a diffusion based neural interconnection satisfies two natural physical constraints in the world of images : shift-invariance and isotropy. No halftoning technique has ever gathered both properties. Based on threshold matrices, ordered dither methods (cf. [Bay73]) ignore both of them which contributes to their poor spatial and tonal fidelity. Currently considered as the best, random 2-D error diffusion methods (cf [Uli88]) are shift-invariant but naturally anisotropic due to the raster order of processing, triggering the appearance of undesirable correlated artifacts. So, unlike the other techniques, our method features sine qua non properties to reach a really high fidelity. Only its isotropy is imperfect due to rectangular grids not being radially symmetric.

Moreover, the corresponding minimized quadratic energy can be advantageously interpreted in the frequency domain, where it has an exact and simple mathematical expression, regardless of the dimension (1-D or 2-D for us). Fig. 14 displays some interesting samples. Due to

the decreasing shape of their Fourier transform, diffusion kernels are able to keep faithfully low frequencies by hiding the quantization noise within higher frequencies.

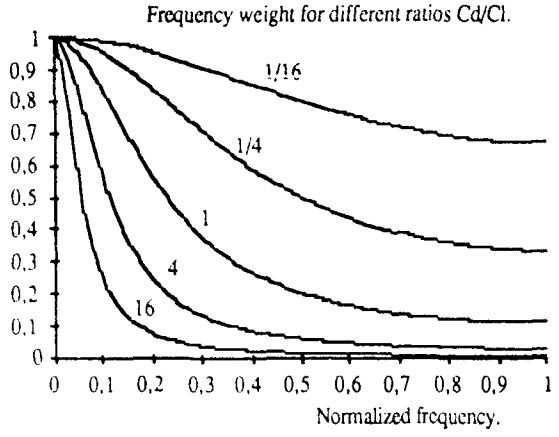


Figure 14 : Frequency Weights for various Kernels K.

But are these curves optimal for halftoning purposes ? The answer is in the affirmative.

Tonal resolution is the only potential weakness of shift-invariant halftoning techniques (like ours). Its separate (but constrained) optimization with respect to kernel K is not likely to spoil an already excellent existing spatial resolution. Now if we restrict ourselves to 1-D constant images, Σ-Δ modulation can be shown both to be optimal for halftoning purposes and to perform the optimization of the MSE between the integrals $\int X(k)$ and $\int B(k)$, with k varying in Z. This again justifies previous attempts (cf [Uli88]) to extend Σ-Δ modulation to 2-D. A major contribution of our work is that we have done so without introducing an arbitrary order on Z^2 (unlike existing 2-D error diffusion methods).

Let us note $\delta(k)$ the dirac distribution in site k, $D=\delta(1)-\delta(0)$ the derivation filter, and $\Delta=D*D=2.\delta(0)-\delta(-1)-\delta(1)$ the laplacian filter. Besides, a -1 exponent means the inverse for convolution. Σ-Δ modulation on constant images thus appears as the minimization of the following frequency weighted MSE criterion :

$$\| D^{-1}*(B-X) \|_2 = 1/2\ [\Delta^{-1}*(B-X)].(B-X) \qquad (3)$$

Though physically unrealizable, (3) has a sense from a formal calculus point of view and turns all the closer to (2) as we show K^{-1} to be a slightly modified laplacian filter ! :

Picture Processing Examples.

The shape of the diffusion kernel K is derived from Kirschhoff laws. Using ratio C_d/C_l (switched capacitor) instead of R_l/R_d, we get : $K^{-1} = C_d/C_l.\Delta + \delta$ (see § diffusion based convolution in IIIb)

If we spread kernel K by making C_d/C_l larger and larger, (2) becomes asymptotically equal to (3) and global minima of (2) become optimally halftoned images. The relationship $K^{-1} = C_d/C_l.\Delta + \delta$ actually characterizes resistive diffusion networks regardless of the dimension. However, when kernel K gets wider, the local minima of (2) become more numerous and subsequently of a lesser quality. The problem is that the neural optimization can get stuck in any of them : this is the very limitation of our method. We need to make a trade-off between the quality of criterion (2) and the quality of its local minima. After having extensively experienced the procedure, it empirically appears that suitable ratios C_d/C_l go from 2 to 8.

Resistive & switched capacitor implementations.

Equation (3) is so neat that the choice of K is definitely the crux of the matter. We have insisted in the previous section on the key role played by simple resistive networks (as presented on fig.6) for a highly compact implementation of appropriate shift-invariant synaptic weights. So, much of the work is done, and the transcription of the transformation equation (2) into the resistive electronic circuit shown on fig.15 is straightforward. The resistive implementation proves extremely simple and regular. The switched capacitor implementation is detailed in [Be-90].

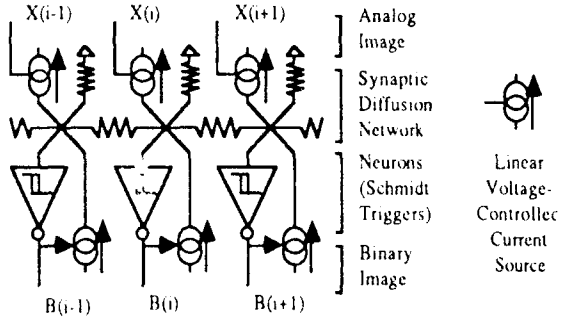
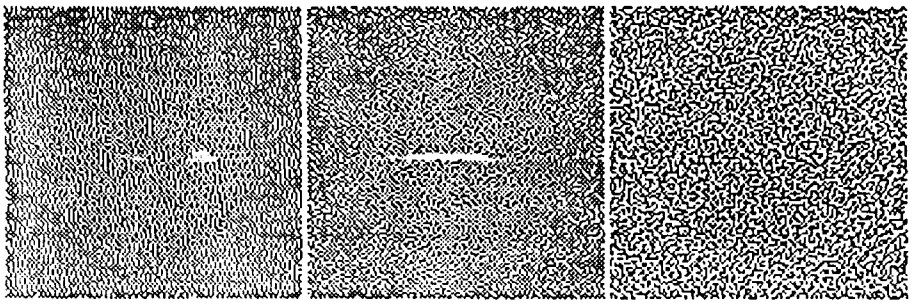


Figure 15 : 1-D resistive neural halftoning network.

Synthetic Sunset
(1) (2) (3)
black=0 white=1
sun=0.46 sky=0.54

(1) Traditional error diffusion with Floyd and Steinberg filter.
(2) Our method with "inverted laplacian" spatial filters.
(3) Failure of gaussian filters.



IVc - More about NCP and retinian visions

To begin with, let us explain the way combinatorial boolean operators can be used on thresholded images (see [Pre79] and [Ros82]). A template (element of U) is determined thanks to two parameters O and Z, which are two disjoint subsets of V: the template [V, O, Z] is the set of all subsets of V which include O and are disjoint from Z. It can be conveniently represented by a picture displaying 1's at the sites of O, 0's at the sites of Z and "don't care" at the sites of V which belong neither to O nor to Z. In this case, the application of the NCP with parameters $\{V, ([V, O_i, Z^i])i\}$ to a binary picture I (considered as a subset of ZxZ) is the following subset of ZxZ:

$$t(I) = \{ z \in ZxZ, (-z+I) \cap V \in \cup_i [V,O_i,Z_i].$$

Now, let us explain how to use a NCP sequence for boolean template matching. First, consider a small binary picture P included in a rectangular window R. The picture t(I) which results from the application of the NCP whose parameters are (R, [R,P, R\P])to a binary picture I is given by the following equation:

$$t(I) = \{z \in ZxZ, (-z+I) \cap R \in [R,P, R \setminus P] \} = \{z \in ZxZ, (-z+I) \cap R = P\}$$

This means that the pixels of t(I) are located at the sites z whose neighborhood (z+R) matches pixel by pixel the template [R,P, R\P]. Of course, if one performs this matching process to match copies of a window of an acquired picture P in an acquired picture I, then the resulting picture will be black i.e. no match will occur. Thus, one needs a way to handle some similarity relation between templates. A conventional template matching approach is to define some similarity measure between pictures [Bar 72]. Now, the point is that as NCP operate through logical operations exclusively, to compute some numerical distance with them is not very welcome, and thus one has to rely on some geometric similarity.

A first approach consists in substituting to the template [R,P,R\P] the template. [R,Pn,(R\P)n], where Pn and (R\P)n are the erosion of respectively P and (R\P) by a square of size n. The pictures this template matches, are geometrically similar to [R,P,R\P].

A more sophisticated approach relies in the continuous plane RxR, where R is the set of real numbers, on Hausdorff's distance. Between two compact subsets of RxR it is given by the following equation:

$$I(A,B)=\inf\{\varepsilon \in R, (B \oplus D_\varepsilon) \supset A \text{ and } (A \oplus D_\varepsilon) \supset B\}$$

where $\oplus$ is the Minkowski's sum and D_ε a disk of radius ε.

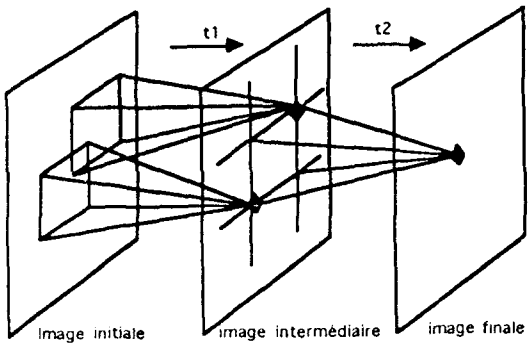
Thus, the Hausdorff distance of A and B is less than ε as soon as $(B \oplus D_\varepsilon) \supset A$ and $(A \oplus D_\varepsilon) \supset B$. By analogy, consider an elementary square S_n of size n. Then we will mark the points Z where

$$(z+P) \oplus S_n \supset (z+V) \cap I, (I \oplus S_n) \supset (z+P)$$

This does not exactly check whether the Hausdorff distance between (z+V) ∩ I and (z+P) is less than n, but this approximation gives good results and remains easy to compute on the fly.

To go further, we want to introduce some structural similarity between templates while still relying on NCP operations. For that purpose, let us choose two square windows R_1, R_2 such that $R = R_1 \oplus R_2$. Let G be a regular square grid included in R_2. Now, consider the windows extracted at the sites of G in P, i.e. for each site z of G, let W_z be $R_1 \cap (-z+P)$. Let T_z be the template $[R_1, W_z, R_1 \setminus W_z]$ and t_1 the NCP defined by the template $(T_z)_{z \in G}$. Besides, let t_2 be the NCP defined by the template $[R_2,G,\varnothing]$.

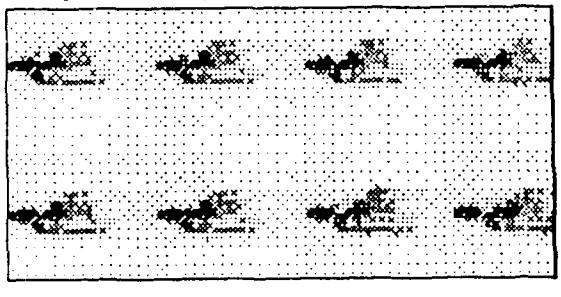
Now, let us choose the grid step and the size of R_1, such that in the one hand the windows W_{z1}, W_{z2} in G overlap and such that $P \supset \cup_{z \in G}(z+W_z)$.



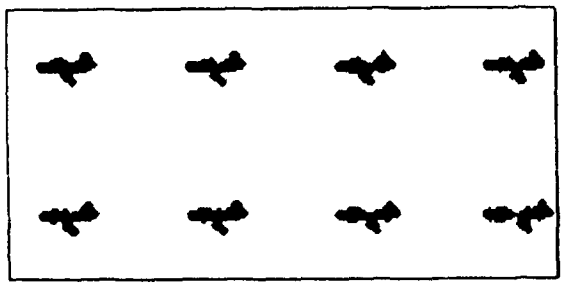
Structural NCP decomposition

Through the successive application of t_1 and t_2, one matches pictures which are generated by swapping the windows (W_Z) between the sites of G (fig.11 and 16): for real pictures, most often the only permutation which meets overlaping and P-covering results in P. Now let us introduce some geometrical similarity between t_1 templates as previously. Then introduce some structural similarity by matching points which are located in the neighborhood of G sites and by allowing some of G sites to have no match. For this purpose, the picture t_1 (I) resulting from the application of t_1 to a picture I is dilated by Sn before the application of t_2. Moreover, t_2 is modified to allow that no match occur at a small number of G sites (introduction of "don't care").

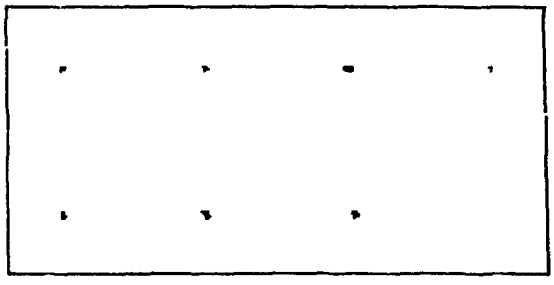
Thus, a unique process of NCI. decomposition into a map product, holds an elastic match between patterns. Moreover, this process may be iterated according to stuctural picture complexity. Examples on tank pictures are given below.



Initial sequence of half-toned pictures



Result of t_1: first NCP iteration



Result of t_2: second NCP iteration

Numbers of suitable operators can be implemented, not only straight recognition. Of course, as it is, the retina can perform combinatorial cellular logic operations [Pre79]. These include erosion, dilation, and their iterations as opening, closing, ... All operations relying on template matching are easily implemented too: they include primarily binary edge detection, shrinking and thinning.Other useful primitives like binary propagation [Duf86], turn out to require supplementary memory points.

The addition of extra memory points (one or two per PE) allows implementing number of other algorithms which are better (fully and systematically) investigated considering a precise designed device. Now, the power of a full preprocessing stage for binary pictures towards statistical pattern recognition could be reached thanks to a global counter. It allows the computation of the area of patterns and thus combining geometric operators with counting yields the full range of numerical features as area, intercept number, connectivity number, and also various histograms and granulometries. After illustrating that point, through a non trivial example, let us show how to perform a counter in the smart sensor itself.

Ex.1: an NCP pseudo-euclidian skeletonization

A local operation as the pseudo euclidian skeletonization may be done inside a smart sensor. In the algorithm described in [Lev 75], height templates Ti are given (A1, B1,..., A4, B4), and must be applied successively.

00.	.00	.1.	.1.
000	1.0	.11	0..
011	110	110	011
.1.	110	.1.	011
11.	. .0	000	0.1
.1.	.1.	.00	00.
A1	A2	A3	A4
B1	B2	B3	B4

For one iteration, all the points of an image I corresponding to the template Ti must be removed to perform the image J ($\neg$, l, & stand respectively for negation, logical or and logical and):

$$J = I \,\&\, (\, T_i(I)\,) = I\ (\, T_i(I)\,) = t1\ (t2\ (I))$$

So, this operation is the composition of two NCP $t1_1$ and $t2_i$, defined as:

$t1 = \neg$
$t2i = \neg + T_i$

The application of the eight templates Ti is implemented by NCP composition. It makes the main loop of this pseudo-euclidian skeletonization to be performed by our smart retina.

Ex. 2 : An NCP counter

In the resulting image of counter algorithm, all the black pixels will be concentrated upon a border of the sensor. To count the number of black pixels, we only use the output of the number of black points along its edges.

The projection of the binary picture I upon the bound B of the sensor, translates all the black pixels with a given direction GD, up to the resulting image J, where all the black pixels are concentrated on B. This algorithm is presented in [Tof87]. Here is a NCP equivalency. For a projection from east to west, NCP p is as:

$p = 10\,x + x11 = \mu 1 + \mu 2$

Templates μ 1 and μ2 represent respectively a progression of one unit to the right, and the meeting with an obstacle. The projection consists of iterating p, up to a constant image.

All the Freeman vector projection may be given by rotation of p. These projections will use a reduced support (3x3 pixels). The projection p' from north-west to south-east is

$$p' = \begin{matrix} 1xx \\ x0x \\ xxx \end{matrix} + \begin{matrix} xxx \\ x1x \\ xx1 \end{matrix}$$

/	→	\	↓	/	←	\
	↑					
p1	p0	p7	p6	p5	p4	p3
	p2					

Elementary projections pi.

If no border constraint exists, black propagated pixels will progressively disappear (translation effect of μ1). Contrarily, if one border B is black, B will be an obstacle (effect of μ2)

If n is the number of pixels of I, and $L = \sqrt{n}$ the width of the retina, the number of iterations is $\sqrt{n}$.

If we consider a composition of projection pi, then the stability region of semi-planes will be the intersection of stability regions R_{pi} of each pi. Now, map multiplying elementary projections pi to concentrate all the black pixels of the binary picture I upon a bound of the retina, consists in operating one or more cycles of n projections $c_k = (p_{1,k} o .. o p_{n,k})$, and iterating c_k until convergence. The choice of the projections is critical, because there are invariant pictures under two projections. For instance, the stability region of projections p5 and p7 is not empty [Rei85].

The convergence will be obtained with projection without stability region, and a good convergence is experimentally got with c0 and c1 cycles. In that case, the counter algorithm needs four cycles, c0, c1, c0, c1.

$c0 = p0 \circ p7 \circ p6$
$c1 = p4 \circ p5 \circ p6$

Real pictures are not well captured by thresholding; and introducing a threshold does not fit exactly the flavor of autonomy. To perform recognition from grey level images, two avenues make sense a priori:

• to rend automatically a picture under a fonn of black and white compact regions. Such a blackening process is again a cellular automaton implementable as NCP ([Rei88]) which can be added directional properties to. It allows to execute ali previously defined operators although tolerance in decomposing is harder to justify. But, learning vanishes here in a way or is drastically changed up to contradict our approach of direct learning by the image itself.

• to generalize the NCP decomposition algorithm to multilevel images so as to analyze directly halftoned images. A new step is required: to extract key structures related to grey levels, grey level sets or density gradients... Then, recognition comes as before from the control of key juxtaposition, which at that point fits perfectly a search for optimal equilibrium between B-coding (halftoning) and NCP. The approach relies on detecting regions as they gather some repartition of grey levels, knowing that a given halftoning process greatly constrains the possible repartitions. Particular NCP's made of sub-templates which get the same density in templates are true spatial counters, and give a hint on grey level repartition inside a region. Technically a marge is introduced again under the form of don't care pixels in the sub-templates. This fuzzyfication is shown to result into a potential spatial shift of key-templates. So, in practice, if templates T as given through windows, are subdivided into wj's which number of occurences are rendered by a given dot configuration Mj, the tolerance on grey level configurations is made of both don't care pixels in Wj and little shifts in Mj. We illustrate the results by tracking the same tanks as before.

First picture 's grid

Results after M and W

V - CONCLUSION.

While technologically realistic, the rapprochement between acquisition and processing within smart sensors opens doors towards peculiar types of interaction between analog and digital computations. The technological constraints however are strong enough to impose a pragmatic approach for setting the analog/digital balance along with the overall performances of the sensor. From this point of view, our Retina tries to be exemplary. Its vision is particularized to allow the use of really bare boolean processors, and consequently the monolothic integration of a significant number of them (100x100 in 1µm CMOS technology). Besides, the "roughness" in the image representation (1 bit/pixel) is compensated for by analog processing on the acquired image, which exploits natural correlation properties of the images. Neural techniques are of great interest for such purposes as shown in the halftoning case. They can also be used to enhance particular early vision features thus leading to more specific retinas.

More generally, there is an unsurprising need at every level of vision for arranging non linearities, function of knowledge and recognition to be performed. Allowing analog layers to cooperate intimately with programmable binary layers (binary on a first phase?) certainly is a good solution, at least in vision which can make do with quite spacially limited connections. Analog suggests rather isotropic communications, where, at most, natural nonlinearities are taken advantage of, while digital suggests more complex interconnections by iterating or programming, hence possibly premeditated anysotropy and nonlinearities.

But, may be the most important aspect of research in the field of analog vision is that concepts or paper work MUST one day be confronted with actual implementation. Though it is an expensive approach, technological constraints impose some sound realism, in front of algorithmic claims. In this confrontation, "silicon" proves to be a most valuable source of inspiration, as it might be translating some fundamental laws where physics encompasses information processing.

VI - REFERENCES

[Bar72] "A class of algorithms for fast digital image registration" D.I. Barnea, H.F. Silverman -IEEE Trans. on Computer, 21 - Féb.72

[Bar84] "Dual controlguidance for simultaneous identification and interception" Birminwal K;, Bar-Shalom Y." Automatica, Vol 20, N°6, 1984

[Bar89] "Estimation and multitarget-multisensor tracking: principles and techniques - Bar-Shalom Y. Conf. DCAN Toulon - Juillet1990

[Bay73] "An Optimum Method for Two-Level Rendition of Continuous Tone Pictures" B. E. Bayer. IEEE Int. Conf. Commun. 1973.

[Bea89] "Time and space multiplexing focal plane convolvers", P.R.Beaudet, SPIE Vol.1071, Optical Sensors and Electronic Photography, 1989.

[Bel88] "A diffusion-based edge detector", E. Belhaire, P. Garda, T. Bernard, B. Zavidovique, 22nd Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, November 1988.

[Ben84] "Charge packet splitting in charge domain devices", S.S.Bencuya and A.J.Steckl, IEEE Transactions on Electron Devices, ED-31(10), pp1494-1501, 1984.

[Ber88] "A family of analog neural halftoning techniques", T. Bernard, P. Garda, A. Reichart, B. Zavidovique and F. Devos, EUSIPCO - 4th European Signal Proc. Conf., Grenoble, 1988.

[Ber88] "Design of a halftoning integrated circuit based on analog quadratic minimization by non-linear multistage switched capacitor network", T. Bernard, P. Garda, A. Reichart, B.Zavidovique and F. Devos, Proc. of the 1988 Int. Symp. on Circuits and Systems, 1988.

[Ber89] "Convolueur imageur *electronique*", T. Bernard, French patent N° 89.16552, dec. 1989.

[Ber90] "A neural halftoning algorithm suiting VLSI implementation", T. Bernard, P. Garda and B.Zavidovique, Proc. of the 1990 IEEE ICASSP, 1990.

[Ber90] "About the adjective neural, when applied to smart sensors", T. Bernard, B.Zavidovique, Proc. of the 1990 IEEE ICPR, 1990.

[Bil83] "On the Error Diffusion Technique for Electronic Halftoning.C. Billotet-Hoffmann and O. Bryngdahl. Proc. SID Vol 24/3 pp 253-258, 1983.

[Bis84] "The self tracker : a smart optical sensor on silicon", G.Bishop and H.Fuchs, in P.Penfield (ed), Proc. of Advanced Research in VLSI, p.57, MIT, 1984.

[Boy70] "Charge coupled semiconductor devices", W.S.Boyle and G.E.Smith, Bell System Journal, Vol.49, p587, 1970.

[Can86] "A computational approach to edge detection", J.F.Canny, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.8, No.6, pp 679-698, 1986.

[Col87] "High performance charge packet comparison for analog CCD applications", R.E.Colbeth et. al., Columbia University, CTR Tech. Rep., 1987.

[DeW88] "A two-dimensional visual tracking array", S.P.DeWeerth and C.A.Mead, Proc. 1988 MIT Conf. on VLSI, pp259-275, MIT Press, Cambridge, Mass.

[Duf86] "Cellular logic image processing" M.J.B. Duff, J.F. Fountain, Academic Press, London 86

[Eid88] "CCD focal plane array analog image processor", E.S.Eid and E.R.Fossum, Proc. SPIE Vol. 977 Real Time Signal Processing XI, San Diego, CA, 1988.

[Fos84] "A linear and compact charge-coupled charge packet differencer replicator", E.R.Fossum and R.C.Barker, IEEE Trans. on Electron Devices, ED-31(12), pp1784-1789, 1984.

[Fos87] "Charge-coupled computing for focal plane image preprocessing", E.R.Fossum, Optical Engineering, September 1987, Vol.26 No.9, pp916-922.

[Fos89] "Architectures for focal plane image processing", E.R.Fossum, Optical Engineering, August 1989, Vol.28 No.8, pp865-871.

[Gar88] "Analog VLSI and Neural Systems, Silicon Retina (Chapter 15)" P.Garda, A.Reichart, H.Rodriguez, F.Devos, B.Zavidovique, 9th ICPR, Rome, October 1988.

[Gar85] "Yet another mesh array smart sensor ?" P. Garda, B. Zavidovique, F. Devos - CAPAIDM 85 - Miami - November 18-20 1985

[Gil68] "A precise four-quadrant multiplier with subnanosecond response", B.Gilbert, IEEE Journal of Solid State Circuits, SC-3:365, 1968.

[Gin88] "Adapted sensitivity / intelligent scan imaging sensor chips", R.Ginosar and Y.Y.Zeevi, SPIE Vol.1001, Visual Communications and Image Processing'88, 1988.

[Her89] "New focal plane architecture and transform for [illegible] target recognition", J.E.Hershey, R.Liberati and [illegible] Hammer, Applied Optics, Vol. 28, No. 18, 15 september 1989.

[Hop82] "Neural Networks and Physical Systems with Emergent Collective Computational Abilities" J.J. Hopfield. Proc. Nat. Acad. Sc.; USA, vol 79, pp 2554-2558, 1982.

[Hor86] "Robot vision", MIT Press, Cambridge Mass., 1986.

[Hut88] "Computing motion using analog and binary resistive networks", J. Hutchinson, C. Koch, J. Luo and C.A. Mead, Computer, March 1988, vol.21, n°3.

[Kat86] "Three-dimensionnal integrated sensors", S.Kataoka, IEDM86 Tech. Dig., pp361-364.

[Kio88] "Monolithic character recognition system implemented as prototype intelligent image sensor by 3D integration technology", K.Kioi, S.Toyoyama and M.Koba, IEDM88 Tech. Dig., pp66-69.

[Koc86] "Analog "Neuronal" Networks in Early Vision" C.Koch, J.Marroquin, A.Yuille.

[Koc86] "Analog "Neuronal" Networks in Early Vision", C.Koch, J.Marroquin and A.Yuille, Proc. Natl. Acad. Sci. USA83, 4263-4267, 1986.

[Koc89] "Resistive networks for computer vision : a tutorial", C.Koch, in : "An introduction to neural and electronic networks", S.F.Zornetzer, J.L.Davis and C.Lau, eds., Academic Press, 1989.

[Lev75] "Parallel thinning of binary pictures", Arcelli, Cordella, Levialdi - Electronics letters, Avril 75, Vol 11 N°7

[Lyo81] "The optical mouse, and an architectural methodology for smart digital sensors", Proc. of the Carnegie-Mellon University Conference on VLSI Systems and Computations, p.1, 1981

[Mal89] "Review of multifrequency channel decomposition of images and wavelets models", S.Mallat, IEEE Trans. on Acoustic Speech and Signal Processing, Dec.1989 .

[Mal90] "Signal characterization from multiscale edges", S.Mallat and S.Zhong, Proc. of the 1990 IEEE ICPR, 1990 .

[Mar89] "An intelligent image sensor based on two-dimensional cellular automata", P.Tsalides, A.P.Marriot, P.J.Hicks and A.Thanailakis, IEEE COMPEURO'89 Conf. Proc., pp3.99-3.101

[Mea85] "A sensitive electronic photoreceptor", C.A.Mead, 1985 Chapell Hill Conf. on VLSI, pp463-471, Computer Science Press, Rockville, Maryland.

[Mea88] "Analog VLSI and Neural Systems", C.A. Mead., Addison Wesley 1988.

[Mea88]"Computing motion using analog and binary resistive networks", C. Mead. Addison Wesley 1988.

[Mer88] "Techniques probabilistes d'intégration et de contrôle de la perception en vue de son exploitation par le système de décision d'un robot - Thèse 1988

[Mit88] Private communications about ongoing research at MIT, 1989.

[Pre79] "Basics of cellular logic with some applications in" K. Preston and al, Proc IEEE - Vol 27, N° 5, May 79

[Rel88] "Aspects algorithmiques d'une vision fruste d'un robot embarquable" Thèse 1988

[Ros82] "Digital Picture processing" A. Rosenfeld , C. Kak - Academic Press 82

[Sag85] "A high speed analog two-dimensional gaussian image convolver", J.P.Sage, OSA Topical Meeting on Machine Vision, FD5/1-4, Incline Village, Nevada, March 1985.

[Ser82] "Image analysis and matheatical morphology" J. Serra, Academic London 82

[She86] "An optimal linear operator for edge detection", J.Shen and S.Castan, Proc. of IEEE CVPR'86, Miami, 1986.

[She86] "Edge detection based on multiedge models", J.Shen and S.Castan, Proc. of the SPIE'87 Cannes, 1987

[Siv87] "Real-time visual computations using analog CMOS processing arrays", M.A.Sivilotti, M.A.Mahovald and C.A.Mead, Proc. 1987 Standford Conf. on Advanced Research in VLSI, pp295-312, MIT Press, Cambridge, Mass.

[Smi86] "Practical Design and Analysis of a Simple Neural Optimization Circuit" M.J.S.Smith and C.L.Portmann. IEEE Trans. on CAS, Vol33, No5, May 1986.

[Smi89] "Practical Design and Analysis of a Simple Neural Optimization Circuit", M.J.S.Smith and C.L.Portmann, IEEE Trans. on CAS, Vol36, No1, Junuary 1989.

[Tan84] "A correlating optical motion detector", J.E.Tanner and C.Mead, in P.Penfield (ed), Proc. of Advanced Research in VLSI, p.57, MIT, 1984.

[Tan86] "Decision Circuit, and a Linear Programming Circuit" D.W.Tank and J.J.Hopfield.Proc. Natl. Acad. Sci. USA83, 4263-4267, 1986.

[Tan88] "Integrated optical motion detection", J.E.Tanner, PhD Thesis, Dept. of Comp. Sc., California Institute of Technology, S223:TR:86, 1986.

[Tie74] "Intracell charge transfer structures for signal processing", J.J.Tiemann et. al., IEEE Trans. on Electronic Devices, Vol. ED-21, p.300, 1974.

[Tof87] "Cellular automata machines, A new environment for the modeling" T. Toffoli, N. Margolus, MIT Press 1987

[Uli88] "Dithering with Blue Noise" R.A. Ulichney. Proc. of the IEEE, vol 76, pp 56-80, 1988.

[Umm89] "Implementing gradient following in analog VLSI", C.B.Umminger and S.P.DeWeerth, Proc. of the Decennial Caltech Conf. on VLSI, pp195-208, MIT Press, 1989.

3D Computer Vision for Navigation/Control of Mobile Robots

G. B. Garibotto
S. Masciangelo
Elsag Bailey spa, R&D Department
Via Puccini 2,
I-16154 Genova
Italy

1 SUMMARY

This note is aimed to investigate how much visual sensors may be effective in supporting autonomous navigation of mobile robots. Although in practical realizations, with robustness and reliability constraints, it is always necessary to integrate multi sensor modalities, the discussion here is just limited to analyze computer vision advantages and disadvantages, with particular attention to:

- a binocular stereo vision module for obstacle detection, with no precise calibration (reactive process to operate at fast rate, from 5 to 10 Hz.).
- trinocular stereovision based on segment primitives for the reconstruction of free space for navigation, in which case an accurate calibration procedure is requested.
- landmark detection for self-positioning and orientation of the mobile vehicle, using perspective invariants, for indoor navigation.

Some comments are also provided on computer vision architectures to support real time implementations. A real-time front end vision subsystem is described, being able to compute 3D segment based stereovision at 5Hz and segment token tracking at 10 Hz. Finally, some demo arrangements are briefly referred, where an intense experimentation of such results is in progress, as a test bed for different industrial applications.

2 INTRODUCTION

The interest in free-ranging mobile robots is no more limited to the classical industrial AGV market, but is increasing in a wide range of potential applications requiring great operational flexibility in less structured environments. Hence, it turns out that typical external sensors, guidance methodologies and control architecture are no more satisfactory for the new set of challenging requirements.

Passive computer vision has been traditionally considered non-competitive against other sensors due to the high cost and lack of robustness of the algorithms, but the recent progress in theoretical issues, availability of special hardware architectures and the increase in complexity of applicative tasks and scenarios make computer vision a key technology also from an industrial exploitation point of view.

This paper is intended to give an overview of the research activities of Elsag Bailey in the field of visual navigation. Particular emphasis is given to the experimental evaluation of the different approaches and a critical analysis of engineering trade-offs which make it possible to implement computer vision techniques in real applications.

A further goal of this work is to discuss how to insert different perception, planning and control modules in a coherent logical architecture and how to implement this architecture on real time hardware.

Visual navigation modules can be classified in many ways: a classical approach consists in considering the operative range, that is the distance of the workspace from the vehicle, which leads to split the general navigation task in three levels: long-range, intermediate-range and short-range. A different but related taxonomy concerns the temporal updating rate of each module, according to real time requirements in real applications.

An alternative approach [1] suggests to consider visual competencies instead of modules, that is to decompose the navigation system in *behaviour layers* instead of *functional modules*. This idea, as discussed in [2], embodies some advantages such as a more direct integration of perception and actuation.

The paper is organized as follows: the next section presents the applicative scenario and introduces the experimental evaluation criteria, sections 4 to 6 describe visual modules and techniques, from the short range to global navigation. Each part refers to experiments and industrial evaluation with respect to alternative solutions, including some literature references.

3 APPLICATIVE SCENARIO AND TECHNOLOGY EVALUATION

Industrial AGVs (Autonomous Guided Vehicles) are already an in-use technology, with known limits and problems. Vision is likely to provide the basis for the second generation AGVs, the so-called "free ranging" AGVs. Currently AGVs navigate using the inductive guidance principle, that implies expensive and unflexible buried wires, or following reflective tape sealed on the floor, that does not resist to the harsh conditions of the industrial environment.

Safety is achieved by ultrasound belts, which limit the vehicle maximum speed and create problems of encumbrance in cramped environments. Moreover, certain types of obstacles like holes, steps, smooth surfaces, thin metallic objects such as chair legs, are not detected at all, underlining the limits of current technology.

GEC Electrical Projects marketed on a Caterpillar vehicle [3] one of the few commercially available free ranging AGV, that will be considered as a reference for the experimental evaluation of our passive vision based system. GEC vehicle makes use of triangulation laser systems with retro-reflective bar-coded targets spread all over the workspace. Security is achieved through IR proximity sensors and mechanical bumpers. The main reported drawbacks includes the loss of maneuvering capability in constrained environments due to the encumbrance of the bumpers, the necessary limit to the maximum velocity due to the short operative range for a reliable IR obstacle detection, the difficulty to operate in scarcely structured or cluttered environments, such as warehouse or in lorry loading, where targets could be occluded or difficult to be placed. Moreover the process of docking workstations or loading/unloading in unconstrained conditions are tasks still too hard for standard technologies.

A novel, promising market sector potentially interested in advanced mobile robots is represented by Service Robotics [4]. Service robotics refers to a novel concept and usage of industrial robots in tasks that are not highly repetitive and not too much constrained. Service robots therefore require much more intelligence, flexibility and sensory capabilities than their industrial ancestors and the application opportunities and potential markets of this emerging technology lie outside the domain of traditional industrial robots.

Mobile robots with relatively simple locomotion can be used in indoor environments to automate routine transport activities. The main examples include hospitals where samples, specimens, medicines and meals have to be carried around, and large offices, banks or postal offices where mail, documents and other items have to be transported through corridors, hallways and other pre-assigned routes. Specifications for these mobile robots include free ranging capabilities, flexibility in reconfiguring pre-planned routes, safety even in peopled areas, and a simple man-machine interface.

Helpmate© from TRC [4] is one of the first service indoor robot in use. It exploits multiple sensors to achieve the required autonomy: ultra-sounds are used for safety and guidance (wall following), flashing IR lamps and a CCD camera are arranged to form a structured light obstacle detector. Monocular passive vision is also used to maintain the heading direction by following the ceiling lamps in long and homogeneous corridors. Algorithms and system architectures presented below will be evaluated against *generic* tasks, but representative of the mentioned application classes.

4 SAFETY LEVEL: GROUND PLANE OBSTACLE DETECTION

The safety level refers to the capability of detecting unexpected, possibly moving, objects which can obstruct the navigation path. An obstacle can be defined as everything with a positive or negative height with respect to the ground level, whose amount exceeds the robot capability to overcome it. Negative heights refers to holes, stairs and any abrupt interruption of the ground, which is as dangerous for navigation as any other obstacle.

The general problem definition is usually completed by a few simplifying hypotheses:

- the vehicle moves on a flat floor;
- the tilt angle between the cameras and the floor is known and constant.

In the domain of indoor navigation those constraints are usually verified, therefore algorithms are still valid in operative conditions as well.

A generalization of the *obstacle detection* problem including also navigation planning and control aspects is called *obstacle avoidance*, that is the robot capability to plan and execute locally a trajectory to overcome the obstacle and recover the originally planned path. In the following we focus on the sensory technologies and algorithms to address these two problems.

Obstacle detection modules, regardless the adopted sensory technology, have to be evaluated with reference to some established design specifications and performance parameters:

- **Fast computation**: the module response rate affects, together with the field-of-view (FOV) of the sensor, the vehicle cruise velocity, which is a major system parameter.
- **Interface with planning**: some modules just detect obstacles, others return an estimation of their positions and dimensions to be fed to a planner in order to compute an avoidance trajectory.
- **Robustness and reliability**: a safety module must be highly reliable. False alarms just delay

navigation but failures in detecting objects affects the vehicle integrity and the safety of people around. Crucial parameters to evaluate are the dependency on the obstacle appearance (shape, colour, texture) and the algorithm sensitivity to drifts of the a priori hypotheses (flat floor, set-up angles, etc.).

Obstacle detection and avoidance are deemed to be critical in autonomous navigation, therefore there exist many different approaches, using passive vision, laser, ultrasonics, IR proximity sensors or some combination of them, to solve the problem but none is considered fully satisfactory. Here we try to demonstrate that passive vision is a feasible and powerful sensor compared to alternative current technologies and can be the core of a safety subsystem.

Proposed approaches range from binocular stereo to monocular dynamic systems. Binocular stereo systems [6, 5] reconstructs the world in order to detect 3D structures in an alarm zone ahead the robot within the FOV. The knowledge of the position of the ground plane with respect to the cameras is commonly used to speed up processing and to focus on 3D data not lying on the ground.

4.1 A stereo Ground Plane Obstacle Detector

The algorithm, originally developed at the University of Genoa [7], is based on a fast comparison between the current stereo disparity and a reference disparity map of the ground floor.

An automated off-line procedure is necessary to produce a reference map of the ground floor, which is supposed flat. However there is no need of an explicit calibration of the stereo rig parameters as requitred by stereo matching algorithms.

The calibration process consists of a correlative stereo algorithm, based on a coarse-to-fine correlation procedure. The disparity map is computed iteratively and averaged by including new stereo views of some random patterns placed on the ground floor, until the variance of the disparity points is low enough. During on-line operations, to check the presence of an obstacle inside the selected windows a correlation approach is used.

The left image of the stereo pair is subdivided in square patches of size 16 × 16; each one has an expected disparity value given by the pre-computed disparity map of the ground floor. Making the correlation between a patch of the left image and the correspondent patch on the right image shifted of the expected *ground plane* disparity it is possible to verify whether an upstanding object violates the expected match of the two image patches. In practice, the usual stereo matching process is reversed: instead of correlating many patches to detect the right disparity for each patch, it is used the a priori knowledge of the disparity in the "no obstacle" case to check whether the correlation is good, otherwise a collision alarm is generated (see Figure 1).

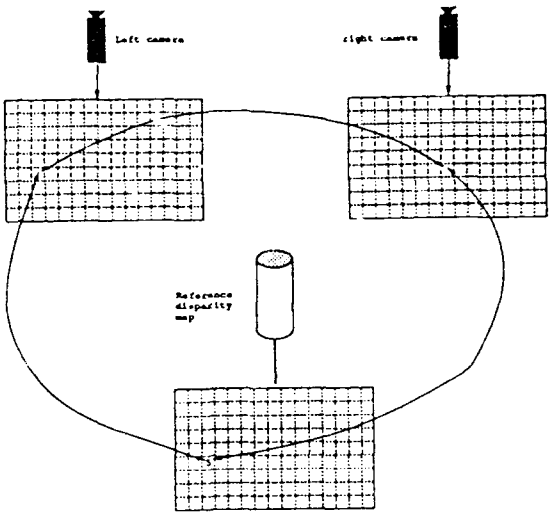


Figure 1: On-line obstacle detection mechanism: the disparity map of the ground floor is used to select the patches in the stereo pair to correlate.

This approach solves the problem of obstacle detection very efficiently and rapidly even if the 3D structure of the obstacle is not explicitly reconstructed and, therefore, a local map of the free-space cannot be available for path planning.

Actually a qualitative obstacle avoidance strategy has been implemented: it is possible to roughly evaluate the position of the obstacle by looking at the image parts where the expected disparity has been violated, and to decide whether the occlusion is on the left, on the right or straight ahead of the vehicle.

4.2 Real-time parallel implementation

The pressing computational performance requirements, estimated in about 10 Hz to cope with the standard speeds of mobile robots, leads to the need for a dedicated hardware implementation of the GPOD algorithm. Currently two real time implementations are available: at the University of Genoa on a VDS 7001 Eidobrain workstation, equipped with a special image processing board where the kernel of the algorithm has been microcoded, and at Elsag Bailey on the multiprocessor EMMA2© where the algorithm has been parallelized.

The Eidobrain image processing board supports the contemporary acquisition of a stereo pair and high communication throughput among frame buffers and the Arithmetic Unit. Therefore, although sequentially implemented, the algorithm runs at 10 Hz.

A parallelisation study, preliminary to the development of a more appropriate hardware front-end, has been conducted on the MIMD EMMA2 computer. A three-processor module is involved in the computational part of the algorithm. Each of the 3 Intel iAPX286 performs the same task, by means of a data partitioning approach. The computation of the correlation value is speeded up by a custom mathematic coprocessor, made by Elsag Bailey, associated to each processing element.

There is also another level of temporal parallelism: a pipeline scheme allows the master processor to control acquisition of a new stereo pair while the previous one is still in the processing phase.

This implementation runs at about 4 Hz, due to delays on the transmission of images on the system bus, which is not a video bus, but guarantees the parallel processing of the whole images and, therefore, an increased reliability as compared to the sequential version which stops the raster scan as soon as a single patch detects an alarm.

4.3 Technical evaluation of the GPOD

The requirements of a safety module for navigation are very strict in terms of robustness if it has to be integrated on a real vehicle, particularly in application involving the presence of people.

Basically we can recall the following advantages:

- the method allows fast implementations, up to 10 Hz, even on a limited amount of hardware, and good computational performances leading to safe navigation at a relatively high speed of the vehicle;
- the algorithm does not require complex, time-consuming or frequent re-calibration procedures and so it may be continuously run, without human intervention;

  vision based correlative stereo permits to navigate in constrained environments and detect thin metallic obstacles (such as stool legs) and smooth edges which typically are critical for ultrasonic sensors;

and the following drawbacks:

- the success rate depends on the amount of texture on the obstacle. Complete absence of texture or pictorial evidences causes a failure as, for instance, in front of a white wall. However, this criticism is valid for any passive vision system and can be easily removed by using some active sensor, such as IR or ultrasounds, in combination with vision.
- polished floors with particular illumination conditions, prevent a correct behaviour since highlights on the floor hold a disparity, as opposed to markings on the ground plane, and violates the prerecorded disparity map constraints, generating false alarms. The use of polarizing filters on the cameras improves the performance by cutting down some highlights. Anyway, the problem is not completely solved because polarizing filters are optimised on a particular incidence angle and cannot entirely remove these effects.
- the implemented process is without memory and does not support common path planning algorithms. Such purely reflexive navigation strategy can cause problems while maneuvering in narrow environments.

5 Exploratory level: free space map building and local path planning

The task is to build local representations of the robot environment to map free space which can be used to plan and update suitable trajectories to reach a selected target position. The final goal of this task is to improve incrementally this 2D map by including new data acquired by visual sensors and keeping memory of the past viewpoints. Of course a prerequisite is to perform such a process quickly enough to support real-time navigation. The present implementation described in the paper is performed at discrete steps, by stopping the vehicle and exploring the scene to do map integration and decide the next robot action.

The obtained 2D representation is local both in space and time with no semantic information. It is just a boundary of the free space around the robot, to provide the current state of the environment, including unforeseen events or unpredictable objects and obstacles. This local representation is passed to the higher level, slower process, which is supposed to plan a safe medium range trajectory. Otherwise, this information can be sent directly to a remote station and displayed to the human operator, for teleguidance control supervision. This is a very simple and reliable way to close the loop at a higher level, on the basis of a very narrow bandwidth channel. An example of this approach is briefly referred in the following sections.

Different approaches are referred in the literature to compute this local map. In [8] a volumetric reconstruction of the scene is obtained through dense stereo correlation. Voxels are integrated in the vertical direction and the results are then projected onto the floor, with selected resolution, to achieve an occupancy map of the environment. Major limitations of this approach are the computation cost of the volumetric reconstruction and the large amount of data produced, which require additional compression of information to find out free space in front of the vehicle. In fact it is always necessary to reach a compromise between the required resolution and a manageable size of the volume of data.

The approach proposed here consists in computing

sparse 3D segments which are representative of visible features in the scene, using a suitable stereo arrangement and then projecting to the floor the most relevant part of them. In fact these data are cut between a lower value (a few centimeters above the floor) and a higher value (slightly above the height of the robot). In this case we assume the ground plane to be almost flat. Segment primitives are considered appropriate to describe an indoor environment with man-made objects and furniture. Of course appropriate lighting conditions are required to provide the necessary image contrast for feature detection. In the following the adopted stereovision process is briefly recalled as well as the real-time processing architecture which has been realized to implement it at rates faster than 1 Hz.

5.1 Trinocular stereovision

A trinocular stereovision approach [9], based on the matching of line segment tokens has been implemented for depth computation. The preprocessing is arranged in a *pipeline* fashion, that is, a sequence of cascaded algorithms each one elaborating the output of the previous stage.

The major processing steps are:

- non-maxima suppression edge detection as an extension of the original Canny approach [10];
- edge linking using a two-step procedure for list making in a raster scanning and fusion and merging of the generated edge lists (G.Giraudon).
- polygonal approximation of edge chains using a modification of a Sklansky approach.

The stereo algorithm is based on three cameras placed at the vertices of a almost equilateral triangle, and roughly converging to a common fixation area. The processing chain of the trinocular stereovision process is recalled in figure 2.

The matching algorithm follows a prediction/-verification scheme; at first, a match hypothesis between two segments from two different views is created on the basis of geometrical criteria; then, the position of the corresponding segment on the third image is predicted. A global validation procedure is finally used, by including additional constraints of regularity and smoothness in the reconstructed 3D scene, and discarding ambiguous matches.

A precise calibration of this arrangement is a key point for the success of stereo matching. The third camera is primarily used for consistency check of match hypotheses and the main advantages of this approach, with respect to binocular solutions, are:

- the implementation of stereo matching is simpler and faster,
- the system is more robust against ambiguous situation.

Besides, also 3D reconstruction is improved by reducing data uncertainty from three different viewpoints.

5.2 Real-time processing architecture

The hardware architecture, depicted in figure 3, reflects the algorithmic structure. This front-end unit has been developed within the framework of the ESPRIT Project P940. This computer vision machine is called DMA from the acronym of the project itself *Depth and Motion Analysis* and is used in other Telerobotic experiments as described in [11].

The video-bus for image transfer at video-rate is the Datacube MAXBUS, which connects all modules dealing with raster image data. The system bus for data transfer, system control, and host interface is the VME bus; all the boards are connected to it and follow the interfacing and arbitration VME standard.

Edge detection is implemented at TV rate according to Canny's approach. Two boards have been produced: the former is composed by 4 FIR building blocks (LSI logic L64240); the latter implements, on dedicated hardware, the "Non-maxima Suppression" algorithm.

The edge linker board is based on 2 fixed point digital signal processors (Analog Devices ADSP-2100) with 2 piggy-back coprocessors to provide fast implementation of a set of primitives (detection and analysis of 8-connected edge pixels and memory occupancy checks).

Polygonal approximation and trinocular stereo matching make use of symbolic information instead of image data. Moreover the stereo matching algorithm structure requires different data partitioning, among the DSPs working in parallel, at the various steps of the process. For these reasons the two algorithms reside on a flexible multi-DSP architecture based on Motorola DSP56000. Data flow control among the different DSPs and the execution of sequential processing steps are performed by a standard 68020 CPU, which in this case plays also the role of *master* board. A very powerful floating-point multi-DSP board, containing 4 DSP96002 from Motorola has been realized on a double-Europe VME card. This unit is particularly effective in 3D reconstruction and high level floating point computation. A Token Tracker module is also available on a single DSP (ADSP2100) board and is able to perform segment feature tracking in a temporal sequence at a maximum rate of 10 Hz.

The software architecture of the machine can be described by the following levels:

- the core of the system can be represented as a *state machine* where each state represents a single DMA function (acquisition, FIR, edge detection, etc). The state machine works as a task allocator: it selects the different drivers of the DMA boards according to the DMA process sequence

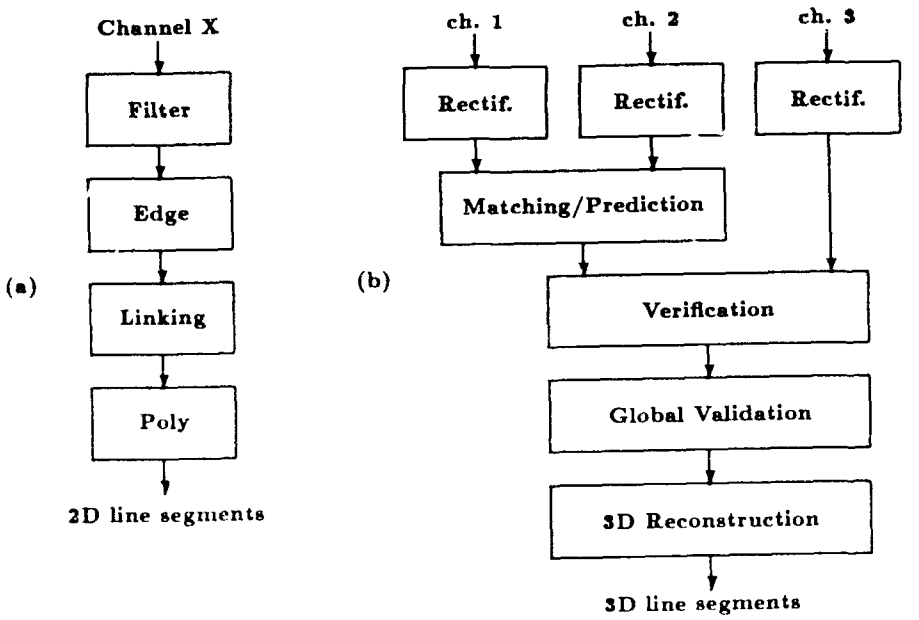


Figure 2 Trinocular stereo vision; a)preprocessing chain for each vision channel; b) stereo matching algorithm.

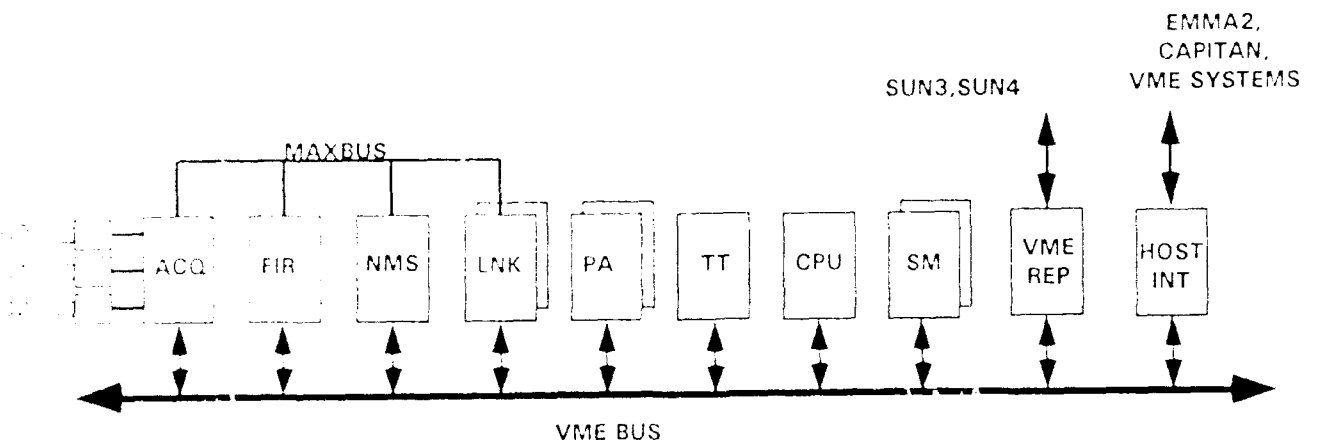


Figure 3: The hardware front-end block diagram

Module	SW impl. (Sun3)	HW impl.
FIR filtering + Edge Detection	39.4 s (filter 11X11)	40 ms
Edge Linking	4.1 s (6500 edges)	120 ms (4 DSPs, 10000 edges)
Polig. Approx.	11.4 s (6500 edges)	200 ms (4 DSPs, 500 segm.)
Stereo matching	30 s (140 segm.)	200 ms (4 DSPs, 200 segm.)
Token tracking	10 s. (250 segm)	100 ms (1 DSP, 250 segm.)

Table 1: Computational performance of the different processing modules compared to a software implementation on a Sun3 workstation

required by the application program, loads the correct parameters, coordinates the pipeline activation of the modules.

- A portion of the control system is dedicated to the MD56 multi-DSPs boards, that can be considered as a MIMD machine since each DSP can host different applicative programs, exploiting the available synchronization and communication primitives. Moreover the 68020 CPU acts as the master processor of the MD56 multiprocessing system, hosting the main of the applicative software (polygonal approximation and stereo matching so far).
- Finally there is the interface towards the host environment, composed by a communication protocol between the DMA machine and the user interface running on the host workstation and a command interpreter, which decodes the instructions received from the host

Table 1 refers the computation time required by the individual processing modules, as compared to a software implementation on a SUN3 workstation. Such results refer to the processing of typical scenes in our laboratory environment (mechanical pieces and indoor scenes).

5.3 Free space computation as the upper envelope of the computed 3D segments

As already mentioned, the basic idea consists in projecting the reconstructed 3D segments onto the floor (known by calibration) and then process them to obtain the free-space navigation map. There are different ways to do that. One approach is referred in [12] where a 2D Delaunay triangulation on the ground floor is used, to better organize the available data

A first step of processing consists in simplifying the bunch of the projected segments to avoid local clusters and intersections, which badly affect the triangulation process. This Delaunay triangulation is also performed as a support for further higher level processing. In fact in [12] the empty triangles, corresponding to free space, are easily identified, through visibility constraints. The corresponding graph, formed by such triangles, is used to generate collision free trajectories for the robot. Moreover, this representation is particularly suitable for an updating process. In fact, when new sensory data are acquired from stereovision, the ground floor map is updated by including new segments into the Delaunay triangulation and the process is iterated. An example of the reconstructed map and planned path is shown in fig. 4 corresponding to a recent on-line demonstration of the system at INRIA, in Nice.

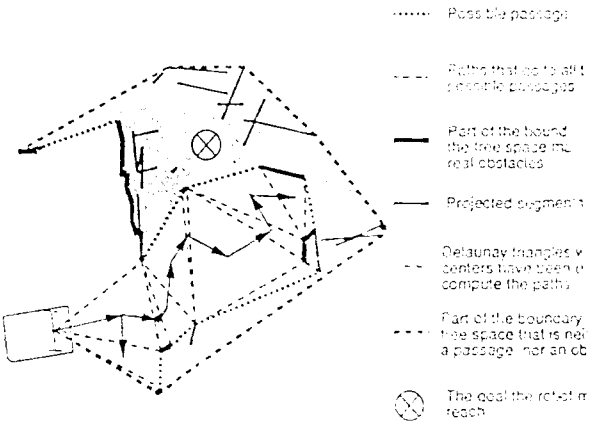


Figure 4: Example of a path computed from the graph formed by the free Delaunay triangles.

Another approach, which has been investigated in [13] consists in performing 3D interpolation of the reconstructed 3D segments in the scene, through a Constrained Delaunay triangulation (CDT). The purpose here is to recover a planar surface approximation of the objects close to the robot, using visibility constraints, as a series of triangular patches whose sides include the extracted 3D stereo segments. The navigation map is obtained by projecting onto the floor all possible paths across those triangular patches and merging them in a lower radial boundary (LRB), computed from the current position of the robot, which is the origin of the polar map. This is definitely the most complete and robust approach for the free space computation, since it makes use of the full perceived stereo information, although at the price of a high computational complexity. Actually an efficient algorithm for 3D interpolation has been implemented as a 2D

Delaunay triangulation on the image plane [14] and real time performance may be easily foreseen on suitable processing architectures (it takes about 10 seconds on a standard SUN3 workstation). To simplify

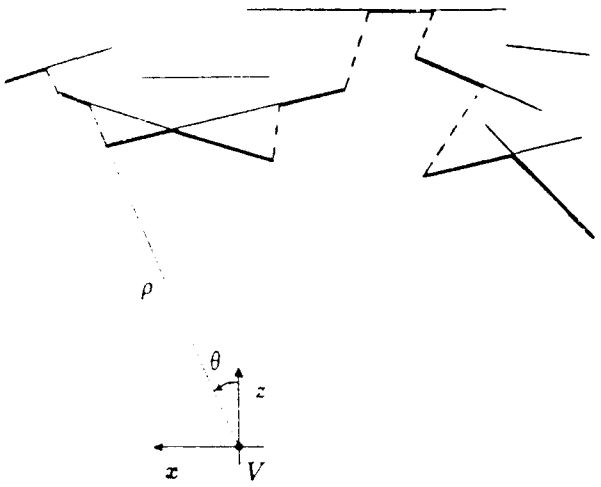


Figure 5: Computation of the Lower Radial Boundary (LRB), by polar scanning around the viewpoint V.

this situation, a suboptimal scheme has been adopted in our experiments, by computing directly the LRB of the free space, without any surface interpolation of the scene. This is obtained by a polar scanning, around the reference viewpoint on the mobile robot, of all projected segments as shown in fig. 5. The process is incremental and is based on a module which performs the fusion of two LRB's from the same viewpoint. Actually a single segment may be considered as a special case of a LRB with a small radial extension. The implemented algorithm for the fusion is based on the *sweepline* technique applied to the intervals determined by the endpoints of all segments and their intersections. The theoretical computational complexity of the algorithm is estimated to be quadratic with the number of segments, although from experimental results a linear dependence has been found.

Fig. 6 shows the reconstructed map for a scene of our lab with a chair, a desk and an industrial robot. The line segments in the map have different meanings. Solid lines correspond to real edge segments detected by stereovision. Dashed lines are virtual boundaries due to visibility constraints, since nothing is visible beyond them. As such no decision can be taken on the free space available in such areas and a next stereo reconstruction from another viewpoint is necessary to improve both the density of the scene and the confidence in the reconstructed map. Actually some irregularities are detectable in the map expecially for those features which are far away from the robot position, where the stereovision process is less accurate. Anyway, the obtained map is quite sufficient to plan a safe trajectory and reach another position from which to explore again the environment.

The availability of the previously described hardware for 3D stereovision at high speed permits an intense experimentation of this tool in a teleguidance mode of operation, as referred in section 7.

6 Global navigation: Landmark detection and self-positioning

A common approach to global navigation, that is the capability to perform complex and long missions autonomously, consists in programming the robot to follow a predetermined path by *dead reckoning*, using landmarks or beacons to correct errors in the position estimate. Dead reckoning is the estimate of the robot position and orientation from measurements of wheel motion (*odometry*). Odometry alone does not guarantee to accomplish the navigation task since it suffers from several sources of inaccuracy such as wheel slippage, therefore, an external sensor, able to reset every now and then odometric errors is necessary.

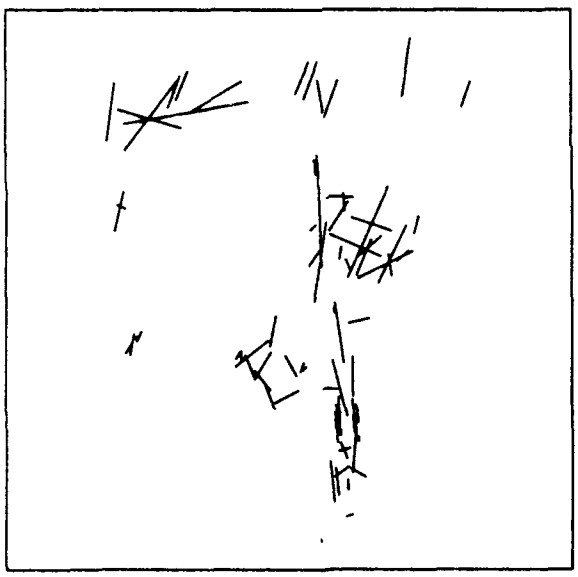
Industrial AGVs use generally active beacons in shopfloor applications, such as IR laser scanner and bar-coded retroreflective targets [3]. On the contrary, we claim that in non-industrial indoor environments (offices, hospitals) a valid alternative approach is represented by passive vision which does not need potentially dangerous laser emissions and high cost for the installation of the devices.

The passive vision approach relies upon *landmarks*, that is known scene entities which allow to recover the robot position and orientation from their appearance onto the image (or images). Landmarks can be natural entities or objects already present in the environment whose position and image appearance can be recorded by the robot through a *learning by showing* procedure. This approach, followed by [5] and [15], is the most general and challenging since does not require any intervention onto the environment. A more conservative but reliable alternative consists in the installation of pre-designed landmarks in order to simplify their recognition and pose computation.

Another way to classify passive vision-based self location techniques is on the basis of the technique for the estimation of the landmark position:

- stereo-based 3D feature extraction and model matching (2 or 3 cameras);
- triangulation of features detected and matched in multiple images through robot motion [15] (1 camera);
- monocular model-based perspective backprojection of the landmark (1 camera).

Our approach relies on the 3D pose recovery of a preselected landmark from the perspective inversion of



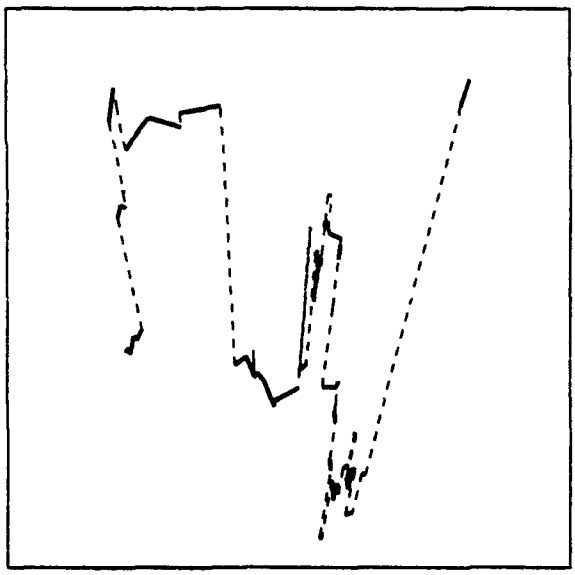


Figure 6: a. The original scene; b. The scene map after projection of the 3D line segments onto the ground floor; c. The Lower Radial Boundary of the freespace.

its projection on a single image. The main advantages over the other self-location methods are:

- there is no need to match features among different images;
- no complex generic object recognition is required, since the landmark recognition is performed by a dedicated procedure;
- the a priori map is very synthetic since there is no need for a complete description of the environment in geometric terms; in fact a list of landmark positions suffices;
- processing of a single image for each self-positioning operation;
- no triangulation is required and, therefore, a less dense landmark distribution in the environment is necessary, since there is just one landmark for each recalibration point.

6.1 Landmark design and the relative self-positioning algorithm

Even if the fundamental property of a landmark is the possibility to successfully apply a perspective inversion procedure to its image, other desirable characteristics should be the following:

- *detectability* in the image by a fast and robust algorithm;
- *robustness* with respect to partial occlusions;
- easy and reliable *discrimination* among different instantiations of the same landmark type;
- the achievable *accuracy* must be good enough to allow the reset of odometry errors;

As such a simple and promising landmark to investigate is a circle, producing in the sensor image an elliptical edge.

From a mathematical point of view, the problem of the perspective inversion of an ellipse generated by a circle in the space, is reduced to find out those planes whose intersections with the cone over the ellipse and with vertex in the origin are circles (see fig. 7). We can only determine the normal to the right planes, and not the distance from the origin, because parallel sections of a cone are all similar geometric entities. The a priori knowledge of the landmark radius value allows us to choose among the parallel planes which one corresponds to the actual case and, therefore, to estimate the landmark-to-robot absolute distance.

Avoiding special cases, there are two possible normals for every ellipse, i.e. two possible sets of parallel planes: this intrinsic perspective ambiguity is solved by making the assumption that landmarks lie on walls, that is surfaces perpendicular to the navigation floor, whose pose with respect to the camera can be calibrated.

A key point is the existence of a a robust and reliable method to extract elliptic arcs from image contours.

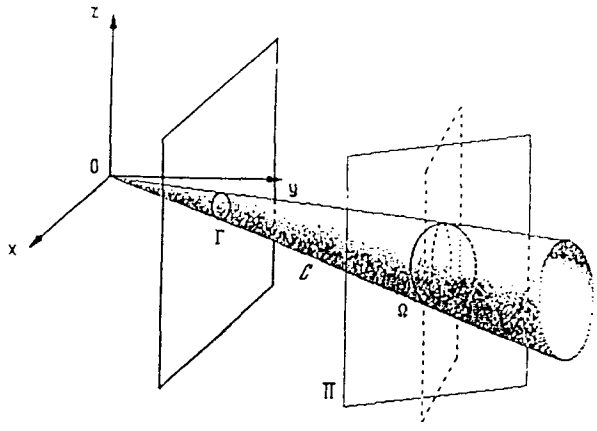


Figure 7: 3D circle and corresponding projected image ellipse.

The approach, outlined in fig. 8, is characterised by a preliminary stage of geometric reasoning on the segments coming from the polygonal approximation of the edge chains of the image. As such it is possible to deal successfully with *outliers* and noise of real scenes [16]. Then, an ellipticity test is carried out on candidate chains of segments in order to select the contours which can be fitted by an ellipse equation.

In this way the 3D position of the robot is computed with respect to a frame of reference centered on the current landmark. Hence, it is necessary to fully identify such landmark in order to provide a global positioning of the vehicle in the navigation map. Unfortunately a landmark consisting of a single circle cannot guarantee a unique identification, therefore a more complex configuration is proposed: the circular annulus (see fig. 6.1).

An invariant physical feature of the landmark is a good candidate to be used in identification, the problem being how to measure it from images. By means of the ellipse perspective inversion algorithm it is possible to compute the linear relation between the radius of a circle and the distance of its centre from the camera pinhole; therefore, if we observe two different concentric circles we are always able to compute the ratio of their radii. If such concentric circle pairs with different radius ratios are used as landmarks, the ratio between the inner and the outer circle can then be extracted independently of the robot pose and used for identification. The two concentric circles forming the landmark have different purposes: the outer is used to determine the pose of the camera with respect to it; the inner is used to identify the landmark by the radius ratio.

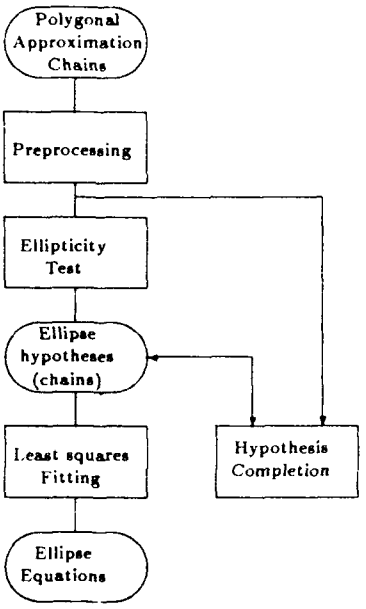


Figure 8: Flowchart of the ellipse detection algorithm

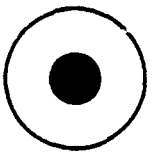


Figure 9: The concentric circles which forms the landmark

6.2 The landmark based navigation strategy

Mission plans describing possible robot paths are sequences of points of interest that the robot has to reach. Each one has a local reference system attached to it and at least a recognisable landmark with known position in this local frame. With respect to these landmarks the robot can estimate its values of position and orientation in the environment.

During navigation, self-positioning is performed whenever, according to odometry data, the robot should have reached the supposed destination position. In this case the robot stops and, using its knowledge about the environment, turns on itself trying to acquire the landmark in the field of view of the camera.

Through landmark identification and its perspective inversion, the mutual rough position estimate is computed and the resulting state vector of the robot is passed to the pilot module in charge of planning the route towards the next point of interest listed into the mission file. If the odometric errors lead the robot outside the landmark visibility region, the landmark detection module communicates its failure and the robot rotates on its own axis in order to search for it. Moreover, the system robustness is improved by the ability to recognise each single landmark so that even if the robot get lost, he can recover his mission by searching for the nearest landmark visible in the camera field of view.

7 A comprehensive demonstration of visual navigation

Within the framework of the European research project ESPRIT P2502 (VOILA) an experimental platform for robotic navigation has been set up. The general architecture is based on the following elements:

1. the TRC Labmate© mobile platform, controllable via an RS-232 serial port. The vehicle is equipped with odometric sensors.
2. Three CCD cameras mounted on an appropriate rig;
3. EMMA2, an ELSAG-made multiprocessor [?], that provides parallel processing capabilities;
4. a PC 486 equipped with a frame grabber for monocular scene analysis, directly connected to EMMA2 which acts as the application supervisor;
5. the already described DMA vision front-end, again connected to EMMA2 through a dedicated parallel interface.
6. A host minicomputer (Q-bus and VMS operating system) to be used as host for EMMA2.

7.1 Description of the demonstration

This demonstration is primarily intended to exploit a Teleguidance mode of operation supported by remote visual perception. It is worthwhile to stress the practical relevance of many short term applications where the presence of the human operator in the loop cannot be removed.

Three visual navigation functionalities are demonstrated showing different levels of integration between the human operator and the robot.

According to the kind of operator interface and the competencies of the vehicle three subdemonstrations are experimented:

(i) Direct Teleguidance;

(ii) Landmark-based Teleguidance;

(iii) Exploration and map building.

7.2 Direct Teleguidance

This demonstration shows the possibility to inspect or control an indoor environment with a mobile platform. It is not necessary to have a model of the environment or to build a global map of it.

The two principal actors of the demonstration are the autonomous mobile robot and a human operator. The architecture of the demonstration must clearly distinguish between the local site, where is the human operator and the remote site, where the mobile robot works.

One CCD camera provides the operator with a display of the remote site. Pure teleoperation is limited to the interactive choice of the navigation goal through a joystick, to select a target point on the displayed scene, as shown in fig. ??.

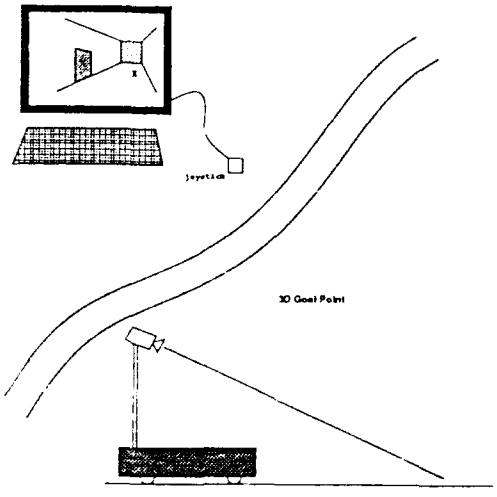


Figure 10: The direct teleguidance concept: the operator clicks onto the computer screen the position that the robot must reach autonomously.

This point is backprojected onto the floor, using some a priori knowledge about the set-up. Then, it becomes the goal of the mobile vehicle, which has to navigate to it without any additional intervention of the human operator, unless some special events occur.

During the local navigation to the subgoal the vehicle will be completely autonomous and will detect the presence of unexpected obstacles. The task of obstacle detection will be performed by the ground plane obstacle detector (GPOD) algorithm. When an obstacle is detected the robot avoids it and tries to recover the original path using odometry. Finally, at the end of the robot action, the human operator resumes the system control and decides a new subgoal.

7.3 Landmark-based Teleguidance

Landmarks are very useful also in a Teleguidance scheme. The operator's job is simplified if the workspace is synthetically described in terms of predefined landmarks. The robot mission can be controlled at the *Task Level* by issuing commands like go from landmark x to landmark y.

Moreover, the presence of the operator at a supervision level can be exploited for recovering from unforeseen situations without aborting the mission. In particular, the operator can correct the vehicle orientation whenever the odometric drifts prevent the camera from framing the expected landmark or solve high level ambiguities in the recognition phase.

7.4 Exploration and map building

In this demonstration the robot utilises the capability to recover the free space in order to plan safe trajectories towards a given goal avoiding unknown obstacles.

Here the three cameras are set up in stereo configuration and connected to the DMA machine real time stereovision system which provides a wireframe 3D reconstruction of the scene.

The demonstration shows a mobile robot which reaches a goal specified by the operator, finding out autonomously a collision free trajectory without any a priori knowledge about the environment. At the end of the run, a freespace map is available proving the ability not only to navigate but also to explore the scene.

As the field of view of the stereo rig is relatively small, it is necessary to get a panoramic view of the environment by panning the stereo rig through a robot rotation.

8 Conclusion

The paper refers on the use of artificial vision tools to support autonomous navigation of mobile robots for indoor applications. Even if we look at the challenging scenario of service robotics, the considered examples here are referred to a teleguidance mode of operation, which is typical of hostile environment applications and surveillance tasks. In this case, the human operator acts as a mission supervisor at an appropriate level, depending also from the degree of autonomy and safety of the robot action.

In practical situations the mobile robot will be necessarily equipped with multiple sensors (lasers, IR, ultrasounds, tactile bumpers, etc.) beside vision, to obtain the more appropriate solution for the specific problem at hand.

This paper is not intended to promote any particular industrial or commercial product, nor to address a precise application task. Besides, its aim is to investi-

gate potential advantages, and limitations, of passive vision using ordinary TV cameras in different configurations, to provide different levels of perception competencies.

The first level is that of safety, to detect and avoid static and moving obstacles and allow the vehicle to move also in peopled areas. The second one is the exploratory level, to compute the free space available around the robot, and apply a short term strategy of navigation planning. A further level of competence is that of self orientation with respect to the environment, using landmark recognition and 3D positioning. The most promising control scheme to fully exploit this hierarchy of competencies is the *subsumption* architecture which is implemented here on a multiprocessor machine.

Finally the problem of real-time processing is considered, with the description of a modular hardware front-end unit, able to perform 3D stereovision at a very fast rate (over 1 Hz).

The achievement of these results has been possible only through a fruitful cooperation with many advanced research teams from Universities and from Industries in Europe, within the framework of the ESPRIT programme. Most of these modules are already integrated in our development experimental system, which represents a very powerful and flexible environment for industrial exploitation of such advanced research results.

Acknowledgments

This research has been partially developed within ESPRIT projects P940 on real-time vision and P2502 on high level vision. In particular we would like to mention the fruitful cooperation with the research teams of prof. O.D.Faugeras (INRIA, Nice), prof. G.Sandini (DIST, Genova), prof. V.Torre (DIF, Genova).

References

[1] Brooks R. , " A Robust Layered Control System for a Mobile Robot", IEEE Journal of Robotics and Automation, vol.RA-2, n.1, March 1986.

[2] Arrighetti S., Ferrari F., Masciangelo S., Sandini G., "Implementation of a Subsumption Architecture for Mobile Robotics on the EMMA2 Multiprocessor", Proc. of BARNAIMAGE 91, Barcelona (Spain).

[3] Robins M.P., "Free-ranging automatic guided vehicle system", GEC Review, 2(2), 129-132.

[4] Engelberger J., "Robotics in Service", London, Kogan Page Ltd., 1989.

[5] Onoguchi K. , M. Watanabe, Y. Okamoto, " A Visual Navigation System Using a Multi-information Local Map", Proc of IEEE Conf. on Robotics and Automation, Cincinnati (Oh), 1990.

[6] Zheng Y., Jones D.G. , Billings S.A., Mayhew J.E.W., Frisby J.P., "SWITCHER: A stereo algorithm for Ground Plane Obstacle Detection", in Proc. of the fifth Alvey Vision Conference, Reading, September 1989.

[7] Ferrari F., E.Grosso, G.Sandini, M. Magrassi, "A Stereo Vision System for Real Time Obstacle Avoidance in Unknown Environment", IROS-90, Tokyo, July 1990.

[8] Grosso E., Sandini G., Tistarelli M. (1989), 3-D Object Reconstruction Using Stereo and Motion. IEEE Trans. on Systems, Man and Cybernetics, vol.18, no.6.

[9] Ayache N. and Lustman F., "Trinocular Stereo Vision for Robotics", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, N. 1, January 1991.

[10] Deriche R., "Using Canny's criteria to derive a recursively implemented optimal edge detector", International Journal of Computer Vision, 1(2): 167-187.

[11] Garibotto G., "A Real-Time Multiprocessor Vision System", Proc. of Vision'90, Detroit, Michigan, November 1990.

[12] Buffa M., Faugeras O.D., Zhang Z., "Obstacle avoidance and trajectory planning for an indoors mobile robot using stereo vision and Delaunay triangulation", Proc. of the Round Discussion on Vision-based Vehicle Guidance, IROS-90, Tokyo, July 1990.

[13] Bruzzone E., Cazzanti M., De Floriani L., Mangili F., "Applying Two-dimensional Delaunay Triangulation for Stereo Data Interpolation", Proc. of ECCV'92 Santa Margherita Ligure, May 1992, Lecture Notes in Computer Science 588, Springer-Verlag.

[14] Bruzzone E., Garibotto G., Mangili F., "Three-dimensional Surface Reconstruction Using Delaunay Triangulation in the Image Plane", Proc. of Int. Workshop on Visual Form, Capri (Italy).

[15] Ferrari F. *et al.*, "A Practical Implementation of a Multilevel Architecture for Vision-based Navigation", Proc. of the Fifth ICAR'91, Pisa, June 1991.

[16] Masciangelo S. , " 3D Cues from a Single View: Detection of Elliptical Arcs and Model-based Perspective Backprojection", Proc. of BMVC 90, Oxford, September 1990.

[17] E. Appiani, B. Conterno, V. Luperini, L. Roncarolo, "EMMA2, a High-Performance Hierarchical Multiprocessor", IEEE MICRO pp.42-56, February 1989.

Machine Perception Exploiting High-Level Spatio-Temporal Models

E.D. Dickmanns

Aerospace Technology Department
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
D-8014 Neubiberg, Germany

ABSTRACT

A paradigm for machine perception is presented which takes time and 3D space in an integrated manner as the underlying framework for internal representation of the sensorially observed outside world. This world is considered to consist of material and mental processes evolving over time. The concept of state and control variables developed in the natural sciences and engineering over the last three centuries is exploited to find a new, more natural access to dynamic real-time vision and intelligence. A. Schopenhauer's conjecture of 'The world as evolving process and internal representation' (1819) is combined with modern recursive estimation techniques [Kalman 60] and some components from geometry and AI in order to arrive at a very efficient scheme for autonomous robotic agents dealing with evolving processes in the real world in real time. Application to autonomous mobile robots is discussed.

CONTENT

INTRODUCTION

Webster's Seventh New Collegiate Dictionary gives the following definitions of terms in connection with the word 'perception':

Perceive: 1. to attain awareness or understanding of, 2. to become aware of through senses. Percept: an impression of an object obtained by use of senses.
Perception: 1: consciousness; 2a: a result of perceiving; observation; 2b: a mental image: concept; 3a: awareness of the elements of environment through physical sensation; 3b: physical sensation in the light of experience; 4a: direct or intuitive cognition: insight; 4b: a capacity for comprehension.
Perceptual. relating to, or involving sensory stimulus as opposed to abstract concept.

These definitions clearly indicate a wide range of meanings, however, a close linkage to physical sensing in general and to vision in special (2b, 3b, 4a); 'objects' as 'elements of environment' are referred to, as well as to the fact that perception is a mentally based activity (3a to 4b). However, the bottom-up data processing aspects are emphasized more than abstract concepts. Definition 3b may be the most appropriate one in the context of machine perception; with regard to applications, 4b covers the task context (see also 'perceive' and 'percept').

'Understanding' or 'comprehending' includes knowledge about semantical relationship in the context of action sequences or goal functions to be optimized. So, perception gains its value in connection with control activities, or at least with preparations for future ones. Without the capability of control actuation, perception would be meaningless (and frustrating?).

Intelligent systems are capable of handling complex sets of goal functions over time and of taking advantage of processes happening in their environment for achieving their goals.

Because of its remote sensing capability, the sense of vision ist the major source of information in our natural environment. The state of development of microelectronics today allows to tackle machine vision as a very promising next step in the evolution of technology on Earth. This section is devoted to dynamic vision as one major component in machine perception for locomotion control.

THE DEVELOPMENT OF TECHNICAL VISION SYSTEMS

Computer vision has evolved from digital image processing over the last three decades. Therefore, it is usually embedded in a quasistatic framework of snapshot interpretation. On the contrary, biological vision systems seem to have developed for motion detection and control in an ever changing physical environment. Are the best suited methods for both tasks the same or are there fundamental differences?

In the Artificial Intelligence (AI) community the vision problem has initially been tackled as a quasistatic problem. Much effort has been devoted to the inversion of the perspective mapping process taking several (consecutive) frames into account; for a survey see [Nagel 83]. This does not take advantage of the temporal continuity conditions in the physical world to which all material processes are usually subjected.

In physics, especially in mechanics, powerful methods have been developed over the last three centuries in order to describe the observed behavior of material processes. In engineering, over the last three decades these methods have been supplemented by features well adapted for recursive digital data processing. Recursive in this context means that least squares data interpretation is achieved step by step as new data arrive. The discipline of systems dynamics evolved out of these activities encompassing aspects of several fields: from sensor technology, signal processing, control theory and design, actuator technology through dynamic behavior of systems.

In this article, the systems dynamics approach is applied to the field of visual dynamic scene understanding, motion control and intelligence. Off the beaten track of main stream research into computer vision, this approach has been developed over the last decade. Combining well proven engineering methods with knowledge from geometry (perspective mapping) and some new aspects of AI, a surprisingly powerful and efficient scheme for the general task of dynamic machine vision using distributed processing resulted. The basic connecting link is a very old idea which the German philosopher Arthur Schopenhauer conjectured more than 170 years ago ['Die Welt als Wille und Vorstellung', 1819, freely translated: The world as evolving process and internal representation].

Building on I. Kant's basic result from two centuries ago, which also formed the foundation for Schopenhauer's conjecture, namely that space and time are not attributes of objects but are carried into the world through our perception and analysis system, it was decided to represent space and time directly in the interpretation scheme. In addition, the constraint was deliberately imposed on the approach that it should work in real time, i.e. that the computational progress over time is directly linked to the progress of the physical process observed and controlled, and not limited by the present state of computer hardware performance. Of course, this confined the problems to be treated considerably in the early 80-ies. It had the members of the team look at problems in a different way, however, and both image processing and scene interpretation algorithms developed differently as compared to the results of other groups who worked under the paradigm that the increasing processing power of future miroprocessor generations will solve all the performance problems with respect to real time.

After a decade of steadily increasing complexity of the problems solved and with experience in five different problem areas, it seems timely to present the approach and the basic ideas behind it in a comprehensive way; the seven dissertations in which most of the material has been originally published are in German language and, therefore, not readily accessible to the general public. The survey article [Dickmanns and Graefe 88] triggered much interest which was one of the driving factors for writing this document.

The present article is intended as a general introduction to the '4D approach' for all those interested in machine vision applications in real world dynamical scenes. Emphasis is put on exploiting knowledge about the physical world and temporal processes; image sequences are nothing but discrete and systematically impoverished intermediate carriers of information about the spatio-temporal world. It is the main goal of the article to shift the paradigm for dynamic machine vision from more academic computer science to practical applications in physics and engineering and to the corresponding methods. Practitioners should find it particularly attractive to experience the direct connections from this modern, very promising field of development to well proven methods in conventional applied sciences.

Resorting to these tools, hopefully, will not have AI-researchers turn away immediately. It is the blend of methods which will lead to efficient machine intelligence systems.

LESSONS LEARNED FROM THE NATURAL SCIENCES, MATHEMATICS AND ENGINEERING

The intention of this approach is not primarily to generate some artificial counterpart of what is called intelligence, but to enable machines with complex sensory systems and the capability of self-controlled locomotion to get around in the real world in a meaningful way; by doing this, some kind of intelligence will emerge more as a side effect in a natural way.

In physics and the engineering sciences mankind has learned over the last centuries how to analyse and represent natural and artificial objects and processes in the environment efficiently. The condensed results of this longterm endeavor of interest to the field of dynamic vision are reviewed briefly in the following sections.

Three-dimensional (3D) space and time
Early geometricians, already millennia ago, discovered that the space we happen to live in can be exhaustively analysed using three independent coordinates. After the more modern French scientist Descartes the orthogonal ('Cartesian') coordinate systems in wide use today are named.

The relationship between space and time has been more obscure for a long time. It was Newton who in the 17-th century invented the differential calculus and applied it to motion analysis. This step in the natural sciences together with the introduction of the inverse square field of gravity brought about a revolution in motion understanding. After this step the geometrically known orbits of planets (Kepler's ellipses) could be linked to a few dynamical motion parameters. The time derivative of the moment of momentum (the second time derivative of position variables in cases of constant mass) was postulated to be proportional to forces, which in a gravity field were in turn linked to position.

The general description of this famous motion law, which despite modern theory of relativity is well justified in conventional mechanics still today, may be written in vector notation as ($^\circ = d(\,)/dt$)

$$\underline{x}^\circ = \underline{f}(\underline{x}, \underline{u}, \underline{p}, t) \ , \qquad (1)$$

where $\underline{x}$ is the state vector with n components, $\underline{u}$ the control vector of dimension r to be freely selected at each point in time, and $\underline{p}$ the parameter vector of dimension q characterizing the special problem. In each degree of freedom, since acceleration as the second time derivative is proportional to forces or moments, two state components (position and velocity) have to be taken into account. Therefore a particle moving freely in 3D space has to be described by 12 state variables, 6 for translation and 6 for rotation, 3 each for position and velocity. For motion in a plane, 6 state variables are sufficient.

It is the integral relationship from acceleration to velocity and from velocity to position which constitutes essential (implicit) knowledge about the temporal behavior of massive objects in the real world. We humans do not have to learn this knowledge consciously, since it is absorbed subconsciously during the first years of our lives while we learn to crawl and walk and to react to other moving objects or subjects properly. Some individuals develop a special skill in this respect; they are good sportsmen even though they may not be able to explicitly formulate how they behave. A wealth of knowledge about the real world is acquired and coded in our neural nets this way even though it is not yet known how.

3D shape and perspective mapping
A similar situation may prevail with respect to our 3D shape understanding through vision. Geometric mapping has been applied for many millennia in all cultures around the globe. Sensible theories about the vision process are less than one millenium old; a nice survey on early vision theories is given in [Lindberg 76]. The difficult problem in vision is that even though the input into data processing is a 2D matrix (spherically arranged in the eye or planar in a camera) the conscious interpretation should be spatial according to the relative physical positions of objects in the real world. For one single photographic snapshot this problem cannot be solved; much effort in computer vision has been devoted to the problem of how many different images are sufficient for uniquely reconstructing the spatial scene.

The law of perspective projection, according to which each visible particle emanates or reflects straight-line light rays from its spatial position to the receiver, is considered to be a sufficiently good model, discarding all side effects of real lenses and mapping devices.

The shape of real bodies has to be inferred from intensity distributions over its visible surfaces and their behavior over time during relative motion. Oftentimes, physical edges and region boundaries on the surface lead to intensity edges in the image plane which, when observed under steadily changing aspect conditions, may allow the proper spatial interpretation (shape from X).

For the representation of 3D shapes the engineering sciences have perfected a 2D representation scheme showing parallel projection views from three (or all six) mutually orthogonal directions. If the object has a plane of symmetry, two (four) of these viewing directions should preferably lie within this plane. One or two reference axes are usually chosen in such a way that the object is oriented in a functionally proper way under normal Earth gravity conditions (e.g. a car with all four wheels touching the ground plane). Nonunique interpretation possibilities (e.g. in concavities) may be disambiguated by special 2D cuts through these regions. A skilled and trained person can imagine the proper perspective view of this object from any aspect condition. For practical purposes, only approximately correct 3D views (to within a few percent accuracy) are often sufficient for object recognition; this can be achieved using relatively simple heuristics for fast and efficient computation of the perspective image given the 2D normal views. 2D shapes with smoothly curved contours and corners can be efficiently represented in a translation, rotation- and scale- invariant form by Normalized Curvature Functions (NCF) [Dickmanns 85] which in turn are easily measurable by tangency operations in the image plane.

Dynamical models of physical processes
The term 'dynamical model' in mechanics, systems dynamics and control theory means a generic differential equation description (like in eq. (1)) for some motion process. We confine the discussion here to motion of massive bodies, be it rigid or elastic. In the case of rigid bodies, classical mechanics has shown that the overall motion can be decoupled into translation of the center of gravity (cg) and rotation around the cg. In the case of elastic bodies some deformation may be superimposed which in the case of free motion usually is an oscillation around a reference shape.

For massive rigid bodies, the forces and moments acting on a specific body are usually very limited in magnitude leading to a characteristic motion behavior over time like a ball flying through the air in the gravity field; gravity and its secondary effects like friction in sliding or rolling motion as well as fluid dynamic drag predominate many motion processes in the real world. Once these basic influences are properly understood (internally represented by a model), a prediction of physical motion in 3D space becomes easy. Combining this with the perspective mapping knowledge of the previous section allows to predict motion appearing in the image plane. Note that for the motion in the image plane no similarly simple direct models can be given due to the nonlinear perspective mapping involved.

The use of dynamical models enforces the internal representation to be in space and time simultaneously (4D). Since the image sequence is discretized over time (50 or 60 Hz corresponding to a video cycle time T' of 20 or 16 2/3 ms), this basic cycle time T' or an integer multiple T thereof is used to transform the differential equation (1) into a difference equation leading to a state transition matrix **A** and a control input matrix **B**

$$\underline{x}[(k+1)T] = \mathbf{A}(\underline{x}, \underline{p}, kT) \cdot \underline{x}(kT) + \mathbf{B}(\underline{x}, \underline{p}, \underline{u}, kT) \cdot \underline{u}(kT), \quad (2)$$

which yield a very compact knowledge representation for the temporal evolution of physical processes in the real world. Note that in the second additive term on the right hand side the effect of control action is contained; this makes this type of representation especially attractive since it allows to include the intelligent motion control part into the prediction scheme. For more long term prediction, probably for investigating the effect of some future control time history of the own vehicle (maybe even several alternatives thereof) this eq. has to be evaluated as many times as requested into the future, thereby allowing a simple means for temporal reasoning. Entire action sequences may be investigated (simulated) this way before decision taking.

State and control variables, process parameters
In an efficient description of real world processes there are three types of variables involved:

1. Those which can be changed at any time at will: e.g. steering wheel turn rate of a car, voltage applied to an electromotor, force applied to an aircraft control stick, throttle position of an engine. These variables are called **control variables $\underline{u}(t)$**.
   Note that this definition is somewhat arbitrary: If the force applied to an aircraft control stick is such that the desired control stick position is reached before the aircraft starts moving in its eigenmodes, the control stick position could have been chosen as the control variable (as has been done with the engine throttle). The essential point is that the control motion has to have a dynamic behavior at least one order of magnitude faster than the controlled process.

2. Those variables which can not be changed directly but which only evolve over time: these are the socalled **state variables $\underline{x}(t)$**. Their evolution over time is as characteristic for an object in the temporal domain as shape is in the spatial domain. Exploiting this knowledge about moving objects in addition to shape constancy results in much more efficient recognition and tracking schemes for moving objects. Note that the spatial velocity components of objects are state variables in this sense; again, this is a strong argument for favoring an internal representation in 3D space and time via dynamical models.

3. Variables which are fixed over periods of time and which may be selected at some discrete point in time, including the system design phase: socalled **system parameters $\underline{p}$**. Typical examples are shift gear position in a car, landing flap position in an aircraft, switch positions etc. and the constants in the system matrices A and B. This set of system variables can be considered constant over time for short term motion behavior even though there may occur a slow change due to wear and tear or environmental effects like temperature or humidity.

Knowledge about a dynamical system is firstly coded in the set of parameters $\underline{p}$ and the structure of the matrices A and B as well as their numerical entries. Equally important in the temporal domain is, however secondly, knowledge of how the system is going to behave with respect to its state variables in response to some control input over time. Especially, the question of how a desired set of state components can be achieved efficiently by appropriate control input time histories is practically relevant; the entire field of 'optimal control theory and application' is devoted to this problem. Mathematicians have developed the calculus of variation for this purpose [Euler 1744] and the 'Maximum principle' [Pontryagin et al. 62], which especially in aerospace engineering but also in many other fields has important and widespread applications since the time that digital computers allow to solve the corresponding difficult numerical problems [Bryson, Ho 75].

To intelligent agents the control variables are of special importance since they constitute the only means through which any influence can be exerted on an evolving process in the real world. Discretely selectable parameters like a switch or flap position may be viewed as 'control parameters' and handled correspondingly. Controls in this sense are the extremely important parts of a system where 'a free will' working on information collected by sensors can exert an influence on the process under control. The provocative term 'free will' will be discussed later.

Feedforward and feedback control loops (cybernetics)
When an experienced person drives a car and wants to switch lane on a highway she or he implements an approximately sinusoidal steering wheel maneuver over time without thinking about it. The amplitude and the time rate are adjusted in such a way that the car finishes this maneuver approximately in the center of the new lane. This can be done in one smooth overall maneuver. A beginner, on the contrary, since unfamiliar with the behavior of the car, will tend to use small incremental control inputs and observe the reaction of the car which in turn will lead him to select the next control input step until the car will finally also end up in the new lane,

however, much later and without a smooth control time history. The experienced person since knowing the temporal response of the car to a 'feedforward control' time history made use of this knowledge leading to better performance; the beginner observing the actual discrepancy between desired and actual state used the difference in some way to feed the control input according to some rule (e.g. a constant factor times the negative difference).

By applying a 'feedback control law' the behavior over time of the controlled vehicle is fixed, but modified relative to the 'open loop'-behavior without any control input. The actuator need not be a person but may be some suitable technical subsystem like an electro-motor or an hydraulic actuator leading to an automated system.

Control engineering and mathematics have developed theoretical and numerical methods which allow designing closed-loop systems with complex eigenbehavior. Literature abounds in this field; just one among many others is [Kailath 80].

Dynamic sytems design
With the powerful digital microprocessors available today, combinations of event-triggered parameterized feedforward control time histories and robust feedback control laws for different subtasks allow the development of very flexible and high performance automatic systems.

Even though the theories developed are mostly based on the assumption of a linear system description, a very large percentage of the generally nonlinear 'plants' (the technical systems to which automation is applied) can be handled this way since linearisations around the actual reference point usually are sufficiently good approximations to the system, especially since feedback controllers keep the system actively in this domain by their functioning. By adding a system identification component, the temporal change of system parameters can be detected and the control scheme may be adjusted accordingly without human intervention.

Modern trends go towards coupling automatic control systems with expert systems in order to improve flexibility and robustness of the overall system under a wide variety of operating conditions. The system discussed in the sequel for real time machine vision may be subsumed under this category.

Kalman's recursive state estimation technique
For interpreting measurements, modern control systems theory has deviced an elegant scheme, how optimal estimates of the actual state of internally represented objects from the real outside world may be arrived at in an efficient way exploiting dynamical models about spatio-temporal relationships of the processes involved. It allows recovering the full state vector even in cases where only partial measurements of some output variables can be taken. These output variables have to be linked to the state variables by some smooth functional relationship. This scheme is extremely well suited to vision processes where the depth component is systematically lost during imaging and where partial occlusions are more the rule than an exception.

Measurements usually are noise corrupted. Therefore, good state estimation can only be achieved when processing many more data than are minimally required. A brief sketch of the historical development of this technique is given in the following subsections.

Gauss's model based least squares scheme for measurement interpretation: When the structure of the motion trajectory is known in advance like for ellipses in planetary motion around the central star, this knowledge can be used efficiently in order to smooth noisy measurement data. The mathematician K.F.Gauss has introduced the technique of fitting curves of known structure to noisy data by minimizing the sum of the squares of the residues. This has lead to much improved accuracies in orbit determination and general curve fitting.

Note, that this improvement is achieved by using solution curves of motion processes, and that a set of measurement data has to be batch processed at a time.

From generic solution curves to differential equation models: If the goal is to have good actual motion state estimates while motion is in progress one would like to have a scheme which gives an incremental update at each point in time when new data become available. If the process observed can be influenced by control input, no a priori structure for the solution curve can be given. In these cases, instead of exploiting solution curves the underlying generic differential equations are more appropriate. For the linear case with known noise statistics [Kalman 1960] has given a recursive least squares scheme which allows optimal state estimation from a reduced set of output measurements. Space does not allow to go into details here; the interested reader is referred to [Maybeck 79]. The known system structure of eq (2) allows to recover state components which are not directly measured by substituting structural knowledge for missing measurements, observability given. The error covariance matrix plays an important role in this process and may be exploited for the removal of outliers, thereby stabilizing the interpretation process.

The big advantage of this recursive state estimation scheme is that always only the last measurements are used for updating the best estimates without the need for storing previous data, which is especially rewarding in image sequence processing where each image comprises enormous amounts of data (10^5 to 10^6 Bytes). The result of all previous data is the present best estimate for the state vector of objects and the covariance matrix corresponding to a storage requirement in the order of magnitude 10^2 per object tracked.

Extended and sequential (numerically favorable) recursion schemes: In the case of nonlinear components in the system description, the socalled extended Kalman filter has been developed based on linearisations around the actual reference point.

In order to keep the covariance matrix symmetric, the upper triangle factorization UDU^T has been introduced [Bierman 75; Maybeck 79]. It is numerically more efficient and stable and is being widely used.

If the state update is computed every time one single measurement component is acquired, the use of two-dimensional arrays in the program may be reduced, leading to faster execution. In addition, this scheme allows an easy adjustment for image sequence processing in the case where - due to occlusion or some other cause - the number of measurement components varies from frame to frame. In our software, this feature has been adopted as a general standard [Wuensche 88, Christians 89, Mysliwetz 90].

Real-time vision, in our approach, is considered to be a measurement process with remote access to the systematically transformed object state (by perspective projection); identification of the object has to be achieved simultaneously with the determination of the motion state.

For image sequence processing, the recursive estimation scheme had to be further extended for the nonlinear perspective mapping of point and line features. In addition, the relationship between the dynamical model for eg-motion and the position and orientation of features on the surface of the body had to be incorporated. The resulting overall scheme will be described next.

STIMULI FROM PHILOSOPHICAL THOUGHTS

Humans with their capability of locomotion and complex information processing may be considered as very complex dynamical systems with a mental component by far not yet understood. Philosophers for millennia have tried to understand human performance in different fields. The natural sciences joined in this endeavor since more than three centuries in a more systematic fashion, but still one is way from having satisfactory answers, though considerable progress has been made recently with the help of information processing technology.

On the basis of Newton's laws of motion and the new understanding of time, Kant in the 18-th century clarified the situation in philosophy by his main works 'Critiques ..... [Kant 1780-ies] to a considerable extent. He separated space and time from attributes of objects granting the former ones a special basic quality. He also introduced a clear distinction between a material object (the 'thing by itself' = "das Ding an sich" (in German)) and a human's notion about this object. The succeeding 'Idealist'-philosophers at the turn from the 18-th to the 19-th century may have turned world interpretation 'upside-down' by giving ideas priority over matter and over the outside world; at least, this was Schopenhauer's impression. In an attempt to put the world from this position 'back onto the feet again', he speculated about the interdependence between the material processes in the world and mind. The basic idea behind the second part of his book title 'The world as will and **internal representation**' [Schopenhauer 1819] may be considered to be a major breakthrough in concepts about cognition.

This basic idea has been adopted as the focal point in our approach to machine vision irrespective of all previous philosophical and psychological controversy. It is not intended to get involved into this discussion as far as humans are concerned; however, this idea has been - probably for the first time - put to work in the context of cognitive machines.

Let us assume there is a material world to which an autonomous agent, say based on a conventional wheeled road vehicle, itself being part of this world, has limited access (with regard to physical state measurements). This may be achieved through a multi-sensor system encompassing properly calibrated odo- and velocimeters, sensors for control inputs, inertial sensors for translation (accelerometers) and rotation (angular rate and position sensors), a microphone for audio-input and imaging sensors in some spectral bands. All these signals are fed into a computer system with properly suited data processing programs.

The autonomous system is assumed to be endowed with all the relevant knowledge components discussed in the previous section. Provision has been taken that the engine is running, the sensory and motor control systems are operative and that there is enough computing power available for properly processing the sensory data; the computer system has access to the control actuation subsystems (even including voice output, say).

The yet open question is: Is it possible to generate an overall system capable of demonstrating a behavior which is qualitatively similar to that of intelligent humans?

THE INTEGRATED 4D APPROACH TO DYNAMIC VISION

The main goal of this approach from its beginning in the early 80-ies has been to take advantage of the full spatio-temporal framework for internal representation and to do as few reasoning as possible in the image plane and in between frames. Instead, temporal continuity in physical space according to some model for the motion of objects is being exploited in conjunction with spatial shape rigidity in this 'analysis-by-synthesis' approach.

Basic scheme

Dynamical models link time to spatial motion, in general. The shape models exhibit the spatial distribution of visual features on the surface which allow objects to be recognized and tracked. In order to exploit both types of models at the same time, the prediction error feedback scheme for recursive state estimation developed by Kalman and successors has been extended to image sequence processing by our group [Kalman 60; Wuensche 88]. There are so many publications on this approach that only a short summary will be given here (see e.g. the survey article [Dickmanns and Graefe 88]).

Figure 1 shows the resulting coarse overall block diagram of the vision system based on these principles. To the left, the real world is shown by a block; control inputs to the own vehicle may lead to changes in the visual appearance of the world either by changing the viewing direction or through egomotion. The continuous changes of objects and their relative position in the world over time are sensed by CCD-sensor arrays (shown as converging lines

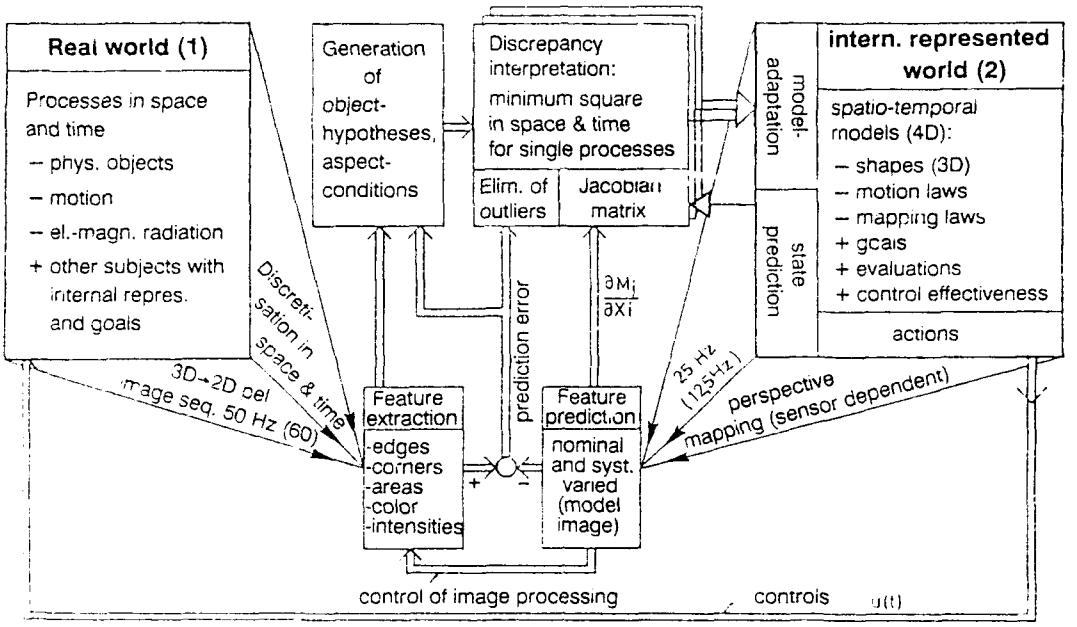


Figure 1. Basic scheme for 4D-image sequence understanding by prediction error minimization

to the lower right, symbolizing the 3D to 2D data reduction). They record the incoming light intensity from a certain field of view at a fixed sampling rate. By this imaging process the information flow is discretized in two ways: There is a limited spatial resolution in the image plane and a temporal discretization of 16 2/3 or 20 ms (due to the different video standards), usually including some averaging over time.

Instead of trying to invert this image sequence for 3D-scene understanding, a different approach by analysis through synthesis has been selected, taking advantage of the available recursive estimation scheme after Kalman. From previous human experience, generic models of objects in the 3D-world are known in the interpretation process. This comprises both 3D shape, recognizable by certain feature aggregations given the aspect conditions, and motion behavior over time. In an initialisation phase, starting from a collection of features extracted by low level picture element (pel) processing (lower center left in fig. 1), object hypotheses including the aspect conditions and the motion behavior (transition matrices) in space have to be generated (upper center left in fig.1). They are installed in an internal 'mental' world representation intended to duplicate the outside real world. After the philosopher K.Popper this is sometimes called 'world_2', as opposed to the real 'world_1'.

The initialisation is the most difficult part and has been solved for well defined simple problems only. A more general capability is being developed presently. It consists of both data driven bottom up and model driven top down components cooperating over time as discussed in the next section.

Once an aggregation of objects has been instantiated in the world_2, exploiting the dynamical models for those objects allows the prediction of object states for that point in time when the next measurements are going to be taken. By applying the forward perspective projection to those features which will be well visible, using the same mapping conditions as in the TV-sensor, a model image can be generated which should duplicate the measured image if the situation has been understood properly. The situation is thus 'imagined' (right and lower center right in fig. 1). The big advantage of this approach is that due to the internal 4D-model not only the actual situation at the present time but also the sensitivity matrix of the feature positions and orientations with respect to all state component changes can be determined, the socalled Jacobian matrix (upper block in center right, lower right corner). This need not necessarily be done by analytical means but may be achieved with little programming effort by numerical differentiation exploiting the mapping subroutines already implemented for the nominal case.

This rich information is used for bypassing the perspective inversion via recursive least squares filtering through feedback of the prediction errors of the features. Unfortunately, space does not allow to go into more details here (see [Dickmanns and Graefe 88]).

This approach has several very important practical advantages:

- no previous images need be stored and retrieved for computing optical flow or velocity components in the image plane as an intermediate step in the interpretation process,
- the transition from signals (pel data in the image) to symbols (spatio-temporal motion state of objects) is done in a very direct way, well based on higher level knowledge, the 4D world model integrating spatial and temporal aspects;
- intelligent nonuniform image analysis becomes possible, allowing to concentrate limited computing

resources to areas of interest known to carry meaningful information;

- the position and orientation of well visible features can be predicted and the feature extraction algorithms can be provided with information for more efficiently finding the desired ones; outliers can easily be removed thereby stabilising the interpretation process.

- viewing direction control can be done directly in an object-oriented manner.

Processing a variable number of features measured from frame to frame is alleviated by using the sequential filtering version. For improving numerical performance, the UD-factorized version of the square-root-filter is used [Bierman 75]. Details may be found in [Wuensche 88; Mysliwetz 90; Bierman 77; Maybeck 79]. By exploiting the sparseness of the transition matrix in the dynamical model a speedup may be achieved.

Two interpretation phases have to be distinguished: First the initialization phase when no previous knowledge about the scene is available, and second the continuous tracking phase, when objects have been recognized and their future behavior is being observed.

From features to physical objects in space and time
In the first phase, usually not time critical, like initialisation while at rest, regions in the image are systematically searched for feature groupings indicative of some known object (lower center of fig. 2). From the collection of features found, object hypotheses have to be generated as to which objects are being viewed under which aspect conditions.

Depending on the task context the higher levels to which the results of feature extraction are reported have to come up with hypotheses for generic objects fitting these data by proper parameter adjustment. Several such hypotheses will usually be generated. They allow to make specific predictions as to where which other features should be found if the hypothesis is correct. Checking these predictions over time, the best hypothesis will hopefully be arrived at by eliminating the less likely ones.

With this information, suitable dynamical models together with body-shapes and aspect conditions have to be instantiated in the recursive estimation loop (shaded blocks in center of figure 2, started by the right column of the inverted U-shaped outer frame). The dynamical models are then used to predict the cg-motion and body rotations around the cg. This information is combined with geometrical shape in order to determine the spatial position and orientation of well visible features. Their positions in the image plane are predicted and the feature extractors in the image processing system are directed to these regions and orientations ('geometric reasoning'-block in lower center right of fig. 2).

The differences between measured and predicted feature data are used in conjunction with the filter gain matrix in order to update the predicted state variables after removal of disturbances recognized (upper right center in fig. 2). The temporal sequence of errors is also used for checking the validity of the hypotheses underlying the actual recursive computation. If consistently poor predictions are obtained, the corresponding hypothesis has to be adjusted; this may concern shape components, parameters in the dynamical model or the complete model. This part up to now has been implemented in a rather rudimentary form. For more complex dynamical scenes than the ones treated up to now, an object oriented data base (in the computer science sense) for a variety of physical objects (in the common sense) has to be implemented; this work has just been started (upper right corner in fig. 2).

A dynamical model has to be instantiated for each physical object capable of being moved. In road vehicle guidance this is not only the ego-vehicle and other vehicles but also the road, the appearance of which varies while driving upon it, at least in the general case with horizontal and/or vertical curvature. This is indicated in fig. 2 by the perspectively shown multiple boxes in the recursive center part.

The state of several objects in conjunction with environmental parameters and the active goal function of the ego-vehicle constitute a situation, to be discussed below. After recognizing the situation (center of upper bar in fig. 2) control modes or actual control time histories may be selected and implemented in an efficient way.

Reflex-like egomotion behavior
Since in the internal representation scheme chosen both the spatio-temporal state variables and the controls at the disposal of the system are explicitly represented, it is straightforward to apply the concept of state variable feedback in order to obtain optimal behavior for well defined tasks. Modern control theory provides the proven background for this approach. For each class of tasks, like lane following, convoy driving etc. in visual road vehicle guidance, a special feedback control law tuned to the actual dynamic parameters of the vehicle yields a characteristic behavioral mode.

Since the computation required is but a matrix-vector-multiplication, this simple operation can be done additionally at the lower level where the recursive state estimation is performed, thereby alleviating the higher levels from any involvement in high frequency control computation; in addition, this eliminates the incremental time lag which would have been introduced by the communication between the hierarchical levels required. With this workload sharing the higher levels may run at considerably lower cycle times (limited only by the requested lumped reaction time delay to some event requiring control mode switching). For systems with dynamical capabilities in the range of humans, several hundred milliseconds reaction time delay may be acceptable, while the recursive state estimation with reflex-like feedback control may run at 40 to 120 ms cycle time (two to six video cycles) typically.

In case a new event in the outside world requires special action, like the detection of an obstacle in the lane at a certain look-ahead distance, the upper decision level may trigger some predefined feedforward control time history

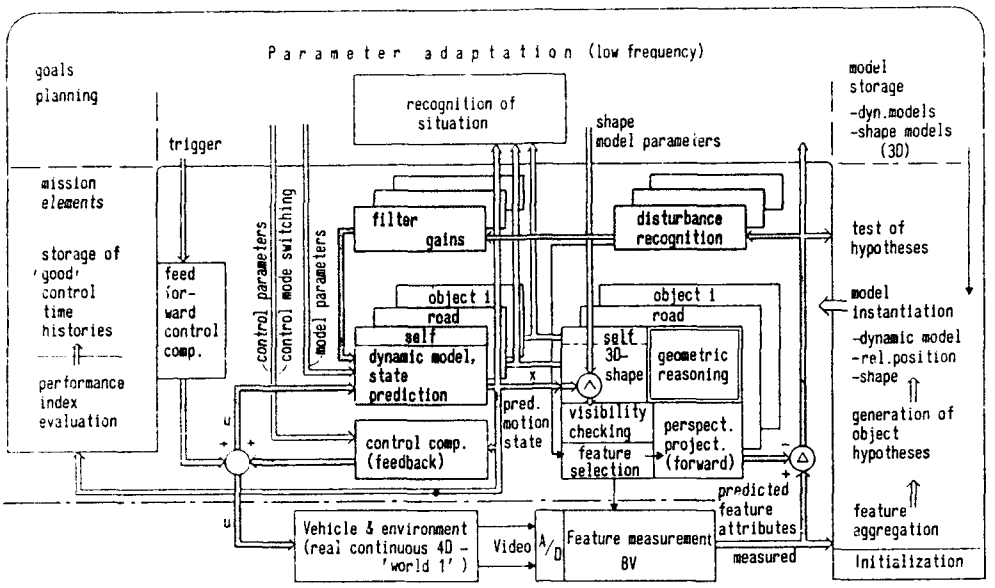


Figure 2. Gross flow chart of the 4D approach to real-time vision

(left in fig.2) with a set of parameters known to be able to deal with this new situation (for example either braking or lane changing).

The concepts up to this point have been implemented and proven to be very efficient computationally and robust enough for real world applications. The following sections deal with extensions under way and planned for the near future. The integrated 4D internal representation including time derivatives of state variables and the effect of control actuation over time yields a rich background for action planning and prediction of possible future evolution of the situation. Thus, based on fast forward simulation, temporal reasoning becomes relatively simple and complex situations may be handled in a straight forward manner.

Objects, subjects and situations

Before dealing in more detail with the notion of situations a brief review of the concept of subjects as introduced in [Dickmanns 89] will be given: Mobile entities in the observed outside world may be classified according to the fact whether or not they have the capability of activating some locomotion or perception system control at their disposal. There exists a large variety of systems with many shades of sophistication. Those which perform internal sensor data processing in such a way that control actuation is not directly coupled to measured data will be called 'subjects'. They are separated from the rest called objects (proper) because they require additional (internal or 'mental') state variables in order to completely describe their state. (Deliberately, no attempt is made to remove the grey zone implicit in this definition.)

For most real autonomous systems it will be impossible to determine their internal state completely. For most practical applications it will be sufficient to grossly know that part of the internal state of an autonomous partner which is relevant for the task at hand. This may be its actual 'view' of the situation, its actual goal function (or system of goal functions together with a likely control strategy) and its way of arriving at decisions in the situation as perceived.

Since usually all control decisions are based on more or less inexact estimates and since too many parameters of other systems are incompletely known, it seems wise to refrain from computing too detailed expectations of other subjects' behavior but only prepare reactions to the most likely ones; careful observation of the development of motion trajectories of the physical body of other subjects will give indications of its likely intentions. The most likely behaviors to be expected may be derived from decision and control strategies which oneself would adopt in the other subject's situation.

This way of defining a situation is in agreement with the one proposed in [Nagel 88]. Here however, the state of the objects and subjects is assumed to be known as good as possible through the recursive estimation scheme, and one is looking for a suitable control decision, the effect of which on the future evolution of the situation can be predicted by utilizing the dynamical models for all objects and subjects involved (assuming likely control inputs).

Mental states and intelligence

For an independent outside observer the internal representation of objects and their states in another subject constitute an increase in state variables of the entire system since the other subject may base control decisions on its actual 'view of the world'; these 'mental' states will then have their effect on the physical world when the resulting control action starts changing the real physical state of objects in the world. Therefore, these mental states are decisive factors in understanding situations; in the German language the word 'Wirklichkeit', usually translated as a synonym for 'reality', allows a different interpretation including these action-consequence effects: Ideas too may be part of 'reality' in the sense of 'Wirklichkeit' since they may effect changes in the evolu-

tion of processes in the real world. (The word 'wirken', from which Wirklichkeit is derived, means 'to effect changes or reactions'.)

Fixing the way how internal representations are arrived at, when sets of input data are given, therefore, is a decisive factor in the design and shaping of cognitive systems. [Maybe the hard core of human cultures, essentially, is an equivalent to this process on a very sophisticated level.] The richer an internal representation can be made by linking incoming data to predefined interpretation structures or to previously stored experience with different types of objects and subjects, the better will the system be able to deal with a variety of situations in the sense of achieving its goals despite perturbing factors. If rich interpretational schemes are available, a cognitive system may recognize situations or courses of actions from short subsequences, and it may be able to react early in an efficient, goal oriented way.

This capability seems to be at the core of the ancient definition of intelligence: The word 'intelligence'was claimed to have originated from the Latin verb 'interlegere' meaning to be able to read in between of lines: those facts or hints which are not explicitly written down but which can be concluded from the context. Translated to the more modern usage of the word this would mean that a system could be called intelligent if it is able to recognize an action or a process sequence, especially a future one, from partial observations only; given an early correct interpretation, such a system would be able to also act early and adequately and to have advantages over lower performance competitive systems. This interpretation seems to be in agreement with the general usage of the word intelligence in everyday life. Note that this interpretation is a quite natural outgrowth of the basic approach taking spatio-temporal representations and the definition of controls in this context into account.

Especially with the sense of vision it is possible to apprehend situations 'at a glance' if typical arrangements of objects and subjects and short but typical action fragments can be observed. This, however, is only possible if the temporal domain is adequately represented by proper models.

SYSTEM ARCHITECTURE BASED ON THE INTEGRATED 4D APPROACH

In our vision system the main sensors are two passive monocular imaging arrays (CCD-cameras, black and white) mounted on a two-axis-platform fixed to each other with a given relative orientation. Their viewing direction can be controlled by the interpretation system according to its needs in the actual context; the controller is integrated into the image processing system.

Based on the concepts discussed above the system developed also has a temporal structuring besides the usual structuring with respect to subtask hierarchies; both aspects will be discussed in the following subsections.

Temporal structuring

Video signal processing of course is linked to the 50 Hz video frame rate; this yields the basic cycle time of 20 ms for image feature extraction of which all slower cycles are integer multiples. The only faster cycle up to now is the viewing direction control for active vision and stabilization; it may use inertial angular rate signals at a small fraction of the video cycle time (typically 5 ms).

Recursive state estimation is done at the rate necessary for control computation: If the vision based automatic system is expected to have about the same dynamic range as the human operator, its corner frequency should be around 2 Hz. Taking sampled control theory into account, this results in a reasonable sampling frequency of 10 to 25 Hz yielding basic control cycle times from 2 to 5 video cycles (40 to 100 ms). The largest value means at a speed of 30 m/s (108 km/h) a new image every 3 meters, the smallest every 1.2 m. This is considered to be sufficient irrespective of the computing power available.

At this rate the complete physical state of all interesting objects is being recursively estimated. Using state feedback control laws, behavioral competences of the autonomous vehicle can be realized for different tasks and situations by simple matrix vector multiplication. This provides the vehicle with fast reflexlike behavioral modes without having to resort to the higher knowledge levels.

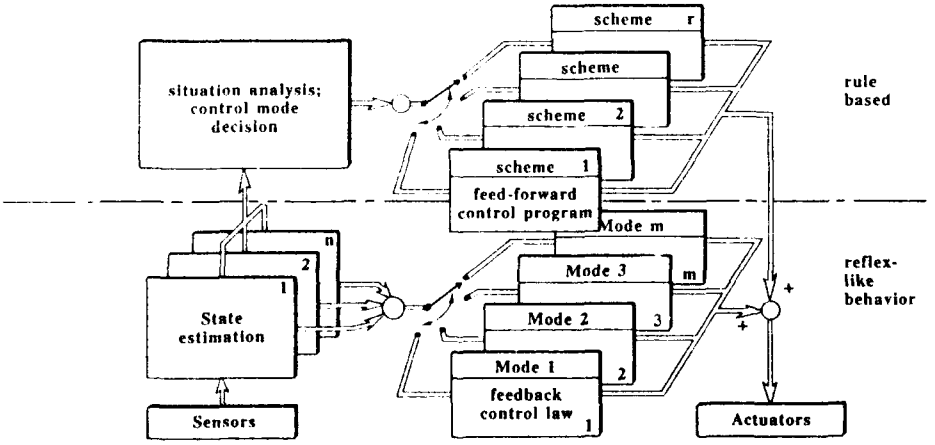


Figure 3. Selectable fast, reflex like feedback control determination with triggered feed forward components; situation dependent control mode decision

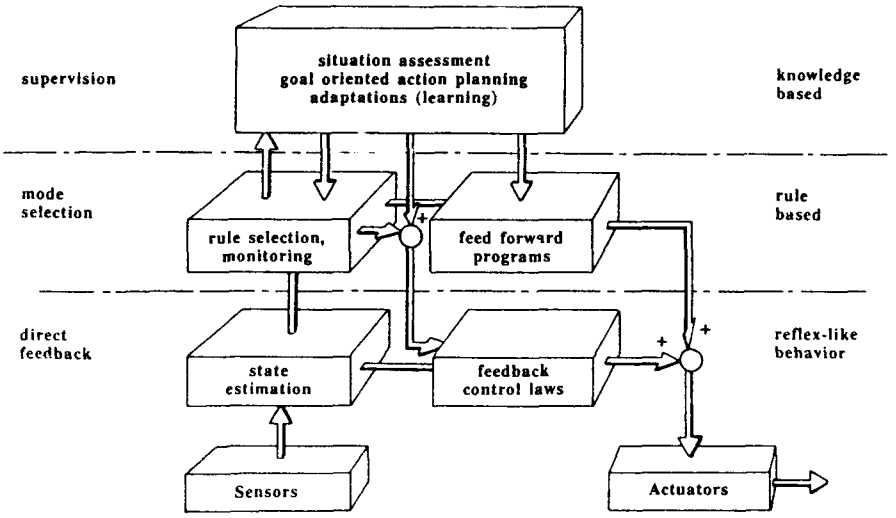


Figure 4. Hierarchical scheme for adapable fast control determination

Adding the capability of triggering proper control mode sequences as shown in figure 3 depending on simple situation indicators (some feature dependent rules), this may lead to yet relatively complex overall behaviors like lane driving with transitions to convoy driving or stopping and other combinations.

When such a pool of basic behavioral modes is available by the fast reacting lower levels the knowledge based higher levels may be allowed slower reaction times, perhaps down to the seconds-range. This figure would still be in agreement with average human performance.

In order to gain additional degrees of freedom for the complex visual perception task it may also be advisible to design overlapping specialised subtasks into the system which work at different time scales but at the same perception problem. One such task which is being studied in our system is the recognition of another object while in motion: There is one subtask which estimates the relative position and spatial speed components rather quickly (40 ms) taking only a very rough (2D) shape representation into account; a second subtask with a different group of processors tries to recognize the full 3D structure of the moving object at a much slower rate. Both may support each other by data or hypothesis exchanges.

On the upper knowledge based levels there is now more time for inferencing using background knowledge in the problem domain. At the same time, relevant environmental parameters may be evaluated and taken into account. In the normal behavioral modes the higher levels just have to monitor the performance of the overall system and to be alert to respond to new situations which may come up. Reaction times of several hundred milliseconds seem acceptable in comparison to human performance. Figure 4 shows the resulting hierarchical scheme.

Besides the different cycle times there is need for another temporal structuring in a (temporal) range sense. All measurements are taken and all controls are output in an exchange with the real world at the point 'here and now' in space and time, moving monotonically on the time axis. Contrary to the real world, the internal representation - also the temporal one! - can be halted and considered quasistatically. This is usually being done in logical considerations, leading to special problems when dealing with dynamical situations.

In figure 5 the internal representation density is shown in a qualitative way over the time axis. The sliding point 'here and now' is marked by the vertical line. In a temporal region around this line the internal representation of objects and the environment is kept and updated by recursive estimation exploiting stored knowledge about the processes observed in a fully dynamic spatio-temporal framework. Time histories of interesting state and control variables may be stored over a sliding short term interval in order to be able to recognize low frequency process charateristics which may be of advantage for longer term predictions into the future. Prediction density varies with the time range: For one prediction step, all state variables will be predicted in the framework of the recursive estimation scheme for each single dynamic object supporting prediction error minimization. Longer term predictions may be of interest only for some objects, maybe even for only a restricted set of variables (e.g. estimation of collision probability). In order to make reasonable predictions for other subjects it is necessary to recognize their intentions, i.e. their likely control time history application in the framework of some goal they seem to be striving for; because there are so many uncertainties when subjects are involved, predictions usually terminate in the near future.

A somewhat different situation prevails with respect to the past. Here, process time histories when properly measured and stored will allow retrospective analysis correlating control input data with observed state histories; this may be used to derive knowledge about the specific system under scrutiny or for accumulating statistical data about objects and processes.

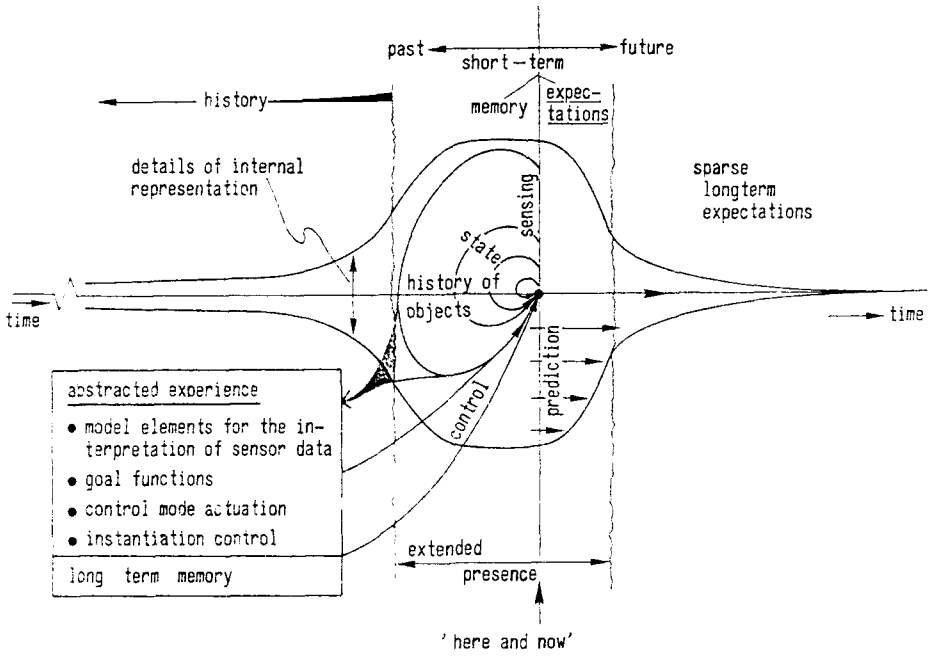


Figure 5. Temporal structuring for details of internal representation (qualitatively)

This temporal integration of perception is considered to be an essential component of learning temporal motion behavior like step responses and eigenfrequencies of objects and subjects in the real world.

From the representational point of view, it corresponds to establishing the link between the differential representation valid for the point 'here and now' and the integral representation of resulting maneuver elements based on some stereotypical control input time history. The result of parameterized stereotypical control actions can thus be represented by a few symbolic parameters linking by a maneuver element two discrete states temporally well apart; an agent capable of understanding these symbols in connection with dynamical models and the temporal integration procedure may manipulate a set of these elements in a quasistatic manner into a proper sequence in order to achieve some overall mission. This is the approach usually taken in AI motion planning, however, very often without caring about the underlying dynamical control aspects.

For fast, efficient and smooth control of processes in the real world this underlying (in biological systems mostly implicit) knowledge has to be exploited; the 4D-approach provides exactly this link (which our human neural net builds up during early phases of (nonintelligent) life in childhood).

Up to now the designer has built these capabilities into our technical systems. However, no principial difficulty can be seen in providing a more advanced system with the proper tools available in the engineering community for developing this on their own.

These activities may run in parallel on additional processors using software packages developed in the field of control engineering, system analysis and systems identification; the resulting parameters may be used in the decision and control processes thereby allowing adaptations to changing situations and environmental parameters (for example roads on a winter afternoon turning from wet to icy).

In the long run, even more deeply structured temporal activities may be considered: Given the availability of proper software, the system may work on stored data time histories during periods where computing power is not needed for actual locomotion control (in parking condition). Several alternative control time histories and the resulting values of the goal function may be evaluated by simulation with the dynamical model available, for the situation considered. This 're-thinking' of situations with a reference outcome meanwhile known, may lead to changes in decision parameters for future action, constituting one component of learning. Another form may be the retrospective comparison of maneuvers performed in similar situations with different control options showing the relative performance achieved; this would be the learning of appropriate behavioral decisions.

Typically during this process, the amount of data to be stored is reduced considerably leading to condensed descriptions of system characteristics (class properties, learning about facts and appropriate behavioral parameters). These characteristics, usually, are no more state variable time histories but system and control parameters or condensed average state descriptions (e.g. mean values, variances).

In this way, the 'present awareness subsystem' based on differential representations in the 4D-approach working around the point 'here and now' (central blob in figure 5) can be exploited in several directions by the knowledge

based subsystem shown in the rectangular box to the lower left; the latter one represents integral effects derived from experience over time for specific situations and tasks.

Expectation based data fusion
When a complex perception system fed by different sensors with different delay times in the data processing pipeline has to deal with the real world, control decisions should be taken based on situation assessment for one single point in time. A control output to the real world can only be effected at the temporal point 'now'.

Knowing what the time delay in the control actuation sequence from decision taking to real world implementation is, and having temporal (dynamical) models for the process to be controlled available, it seems to be wise to exploit these models for making predictions of object and subject states exactly for the point of control implementation. If all measurement takings are geared to the same point kT, an especially efficient system design results.

The different time delays in the data paths may now be compensated by corresponding numbers of prediction steps applying the object specific dynamical models. With redundant data sets the Kalman filter approach allows recursive least-squares-error data interpretation exploiting knowledge both about the real world process and about the various measurement subprocesses. Removal of outliers exploiting the covariance matrix helps stabilizing the interpretation.

Hierarchical structuring
With respect to behavior control, in fig. 4 the resulting hierarchical scheme has been given. Table 1 shows the hierarchical structuring with respect to measurement and scene recognition aspects. No special low level image preprocessing is performed; instead, the algorithms for feature extraction on the basis of controlled correlation [Kuhnert 88; Mysliwetz 90] are designed in such a way as to exhibit good noise reduction properties. Mainly, edge element and corner features have been used up to know. There is no final decision made with respect to 'optimal' features based on bottom up data only; accepted features for object interpretation are selected on the basis of an overall 'Gestalt'-idea derived from perspective mapping of an internal 3D shape representation (second line from bottom in table 1). At the single object level, time is introduced via the dynamical models for 4D representation; up to now, no interframe differencing as in optical flow has been applied. The future has to show whether this type of image sequence processing will be necessary at all. (It is well known that nature in its biological systems does make use of it; this has triggered quite a bit of activities in this area also for technical vision systems. Whether and under which circumstances this is advantageous has yet to be determined). In our approach a 'virtual optical flow' for features is computed on the basis of the internal spatio-temporal representation and perspective forward projection.

The levels discussed up to now have been implemented in the image sequence processing system BVV_2 [Graefe 85; Mysliwetz 90] and more recently in a transputer network [Thomanek, Dickmanns 92; Behringer et al. 92]. The scene understanding (upper) part in table 1 has been implemented on a PC-AT in the past and has been ported onto a transputer system also. From several objects and environmental data the situation is recognized and checked against the requirements for task achievement. If no special action is needed the system continues in its present mode; if some change of the operational mode becomes necessary a replanning is performed and the resulting mode change is triggered.

The control output is fed back to the internal representation via the prediction step, updating all the lower levels, thereby adjusting the measurement and interpretation process to the actual state.

This frequent and fast traversion both bottom up and top down in the interpretation scheme assures efficient exploitation of both high level knowledge and most recent measurement data.

The gross flow chart corresponding to table 1 has been discussed already as figure 2 above. It has been arranged in such a way that the procedural recursive state estimation techniques using control engineering methods form the core of the figure while the more knowledge based higher level activities are grouped around this center showing the interaction paths.

A different viewpoint for subdivision showing other facets of the same system has been given at the end of [Dickmanns and Graefe 88]; the completely autonomous simulation capability inherent in this approach, and referred to already above, may even work without any sensory input normally being the driving factor.

	activity level	processors	operation	result
scene understanding	control level	MPS	compute expectations control viewing direction apply vehicle control	→ action
	↑		↑	
	task level	MPS	relative goal state evaluation	→ planning, decisions
	↑		↑	↑
	object level	MPS	→ situation assessment parameter adaption	→ situation
state estimation	↑		↑	↑
	feature level	4D-OP	→ feature aggregation	→ objects in space/time
	↑		↑	↑
	pel level	PP	→ feature extraction	→ features in image plane

Table 1. Modular processing structure for complex tasks

Stored data may possibly be taken as starting points or as reference trajactories to study variations around; interesting questions with respect to 'mind' and 'dreams' may come up.

EXPERIMENTAL RESULTS

The general scheme of dynamic machine vision and expectation based perception discussed above has been developed during parallel application to four different areas, after the idea had come up around 1980 in connection with the problem of visually balancing an inverted pendulum on an electrocart [Meissner, Dickmanns 83]. The first application oriented problem was planar docking of a reaction propelled air cushion vehicle with three fully independently controllable degrees of freedom [Wuensche 86, 88] simulating autonomous spacecraft docking. The second area was road vehicle guidance to be discussed in somewhat more detail below. The third one was birdlike autonomous landing approaches for conventional aircraft under visual flight conditions; this may be of interest for unmanned vehicles or as basis for an electronic copilot and will also be briefly discussed below.

Autonomously guided vehicles for transportation tasks on the factory floor are the fourth application area; in this context, the capability of landmark navigation has been developed and demonstrated [Hock 91]. Autonomous visual guidance of helicopters has been tackled in 1992.

Road vehicle guidance

The application area of autonomous road vehicle guidance is by far the most developed one: A 5 ton van 'VaMoRs' of our University as well as a 10 t bus and a 7.5 t van 'VITA' of the Daimler-Benz AG have been equipped with our vision system. In experiments ranging over six years by now, the following capabilities have been demonstrated:

- Lane following at high speed: 100 km/h have been achieved limited only by engine performance of VaMoRs. On well marked empty freeways much higher speeds could be handled by the method; limitations may first come from camera resolution at large look-ahead ranges. Both horizontal and vertical curvatures can be estimated to sufficient accuracy [Mysliwetz 90; Mysliwetz, Dickmanns 92] to allow velocity control in order not to exceed preset acceleration limits.

- Lane following on unmarked cross-country roads with shadows from trees and buildings on the road. Speeds up to 60 km/h on empty roads have been demonstrated; even driving under light rain fall with wipers operating in front of the cameras has been shown.

- Night driving on well marked dry roads with normal headlights at low speeds has been performed with the Daimler-Benz bus and VITA on test tracks.

- Driving on unsealed country roads at speeds below 20 km/h has been achieved by VaMoRs; however, in order to obtain more robust performance, computing power both for image processing and on the higher levels has to be expanded.

- Recognition of well visible obstacles of more than 0,5 m^2 cross-section (black trash can) in a look-ahead range of 30 to 50 m has been demonstrated at speeds up to 50 km/h on unmarked two-lane roads. The situation assessment level decides whether the vehicle is autonomously stopped at a safe distance in front of the obstacle or whether a lane change and passing maneuver is performed. Similar demonstrations have been performed with the Daimler bus stopping in front of another bus. Passenger cars can be detected at ranges up to 100 m with a 25 mm tele-lens. Monocular distance estimation through motion stereo (an inherent property of the 4D approach exploiting data fusion from odometry) is achieved with sufficient accuracy up to about 50 m; the introduction of inertial gaze stabilization will allow larger focal lengths with correpondingly improved viewing ranges.

- Convoying behind another vehicle has been initially demonstrated in our hardware-in-the-loop simulation facility, lateron with the test vehicles; 'stop-and-go' experiments are a special case of this capability shown in 1990.

- Lane changings to the left and right have been performed in daytime and at night, triggered by the human operator who has to take care for other vehicles in neighboring lanes.

- Driving on public German 'Autobahnen' has been started in 1992 with the transputer system as the latest achievement. Besides lane recognition two other objects may be detected, tracked and interpreted in parallel.

Aircraft landing approach

One of the most crucial maneuvers in autonomous flight is the final approach phase to the landing strip. Under good visual conditions, human pilots are able to land an aircraft safely without any support from the ground by using just visual cues from the airport environment and the runway. In 1982 we started studying this problem in the simulation loop with the goal to develop methods which would allow autonomous unmanned aircraft with the capability of machine vision to do the same. G. Eberl in his dissertation work [Eberl 87] laid the foundation for the solution available now. From 1987 onward, R. Schell continued the development till the first flight experiments successfully performed in 1991.

The initial 9 years of development have been performed in the simulation loop exclusively. Results have been published in [Dickmanns 88; Dickmanns, Schell 89]. Over the years, realism in simulation and the use of real image processing hardware has been steadily increased. Space does not allow to describe the system developed in detail; the interested reader is referred to [Schell 92; Schell, Dickmanns 92].

The achievements may be considered a breakthrough in machine vision application. It has been shown that full

spatial motion in all rotatory and translatory degrees of freedom can be controlled by onboard autonomous dynamic machine vision with a relatively small set of today's microprocessors, using the 4D approach. In simulation, the control loop has been closed and landing approaches have been performed from about 1.5 km distance till touchdown, including wind effects and gusts. Fig. 6 shows a simulated approach situation with the hashed squares indicating the image areas evaluated for information extraction. In both the simulation loop and in the real flight experiments the camera was suspended on a two-axis pan-and-tilt platform for visual runway fixation.

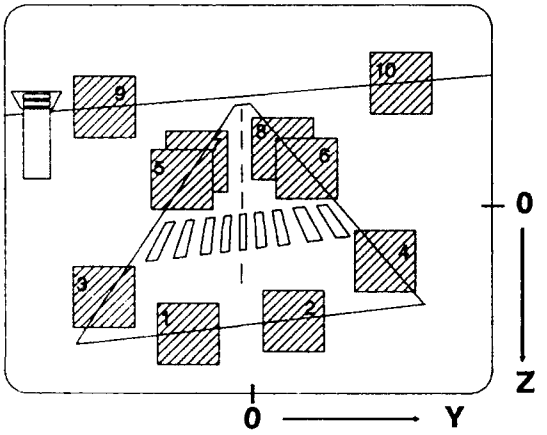


Figure 6. Simulated landing approach with subareas evaluated for information extraction

In the flight experiments, funded by the German Science Foundation (DFG) and performed with the twin turboprop aircraft Dornier Do-128 of the University of Braunschweig (see fig. 7), inertial angular rates and orientations have been measured by gyros and were fed into the interpretation system, with data fusion performed through the two sixth order dynamical models separated for the longitudinal and lateral degrees of freedom.

Since the aircraft was not yet certified for active computer control, only the real-time state estimation part exploiting dynamic vision could be tested. This, however, has been very successful; after only one week of installation work and interface testing, due to the careful preparations performed in the simulation loop with the complete vision system, first trajectory and state estimation results could be achieved. Fig. 8 shows the visually estimated altitude as compared to a radio-altimeter measurements and those from the Global Positioning System (GPS). The landing approaches were abandoned at about 5 m altitude in order to make a fly-around for the next trial. It can be seen that visually estimated and radio-altimeter measurements agree very well in the vicinity of the runway (time > 13 sec); aircraft speed was about 55 m/s (200 km/h). Estimation quality of the longitudinal position was considered sufficiently good whereas lateral position estimation fluctuated with about 2 m amplitude relative to the GPS-results; this will have to be studied further.

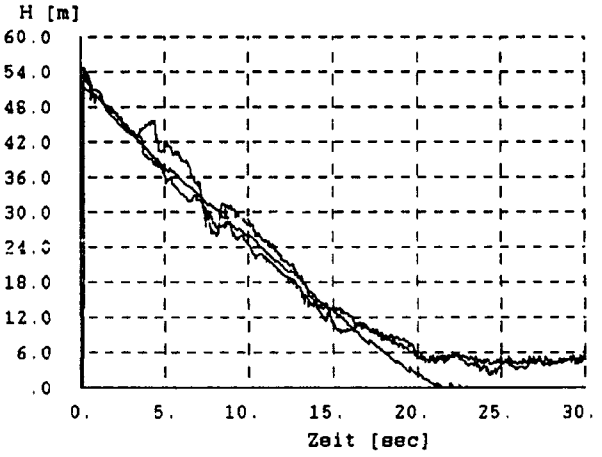


Figure 8. Estimated altitude time history

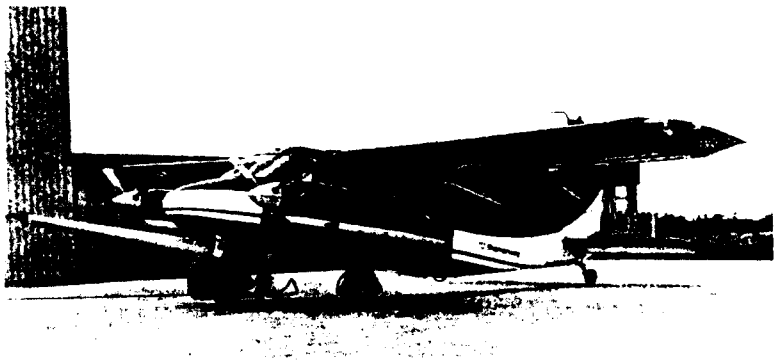


Figure 7. Test aircraft Do-128 of TU-Braunschweig

CONCLUSIONS

Machine perception and vision-based intelligent motion control should take advantage of the recursive state estimation techniques developed in control engineering. The '4D approach' developed at UniBwM over the last decade generalizes the extended Kalman filter to image sequence processing. In its sequential formulation it is well suited for solving major parts of the problem of dynamic scene understanding even under the condition of occlusion. The dynamical models are well suited for knowledge representation in the spatio-temporal domain.

The 4D approach has been developed with the goal in mind to achieve dynamic vision performance similar to the human one, at least in motion control. Introducing

time as an independent variable right from the beginning as the basis for integral spatio-temporal object models, allows to develop very efficient data processing schemes. Unlimited image sequences may be processed without the need for storing previous images; the effects of historical development are accumulated in the state of physical objects, internally represented in 3D space and time.

It has been shown in several application areas, that microprocessors available today, already allow surprising performance levels when exploiting this method as compared to quasi-steady approaches usually studied in Artificial Intelligence. For high level performance in complex scenes, these engineering-based methods need to be complemented with ones well suited for explicit knowledge representation and decision making.

It has been sketched how machine intelligence can possibly be developed based on the feedback scheme for motion control exploiting the high-level spatio-temporal world models which are at the core of recursive state estimation. In human history of science, dynamical models (i.e. differential eqs.) have been a rather late but very consequential achievement in understanding the world we happen to live in. This powerful insight in basic properties of processes in the real world should be exploited for making machine perception more effective.

LITERATURE

Behringer, R.; v. Holt, V.; Dickmanns, D.: 'Road and Relative Ego-State Recognition'. In 'Intelligent Vehicles', IEEE, SAE, Detroit, June 1992.

Bierman, G.J.: 'Measurement Updating Using the U-D Factorization'. Proc. IEEE Control and Decision Conf., Houston, Tx., 1975, pp 337-346.

Bierman, G.J.: 'Factorization Methods for Discrete Sequenctial Estimation'. Acad. Press, New York, 1977.

Bryson, A.E. jr.; Ho, Y.: 'Applied optimal control'. Hemisphere Publishing Corp., Washington D.C., 1975

Christians, T.; Wuensche H.J.; E.D. Dickmanns: 'Schätzung der räumlichen Relativlage von Objekten durch Rechnersehen'. UniBwM/LRT/WE 13/IB/89-6, 1989.

Dickmanns, E.D.: 'Computer Vision forFlight Vehicles'. Zeitschrift für Flugwissenschaft und Weltraumforschung (ZFW), Jan. 1988.

Dickmanns, E.D.: 'Subject-Object Discrimination in 4D-Dynamic Scene Interpretation for Machine Vision'. Proceedings IEEE-Workshop on Visual Motion, Newport Beach, March 1989, pp 298-304.

Dickmanns, E.D.: Simulation for the Development of a Visual Autopilot-System for Road Vehicles'. In M.R. Heller (ed.): Automotive Simulation. Springer-Verlag Berlin, May 1989, pp 11-22.

Dickmanns, E.D.: Dynamic Vision for Intelligent Motion Control'. IEEE-Int. Workshop on Intelligent Motion Control, Istanbul, Aug. 1990.

Dickmanns, E.D.: '4D Dynamic Vision for Intelligent Motion Control'. In C. Harris (ed): Special issue of the Int. Journal for Engineering Applications of AI (IJEAAI) on 'Intelligent Autonomous Vehicles Research' 1991.

Dickmanns, E.D.; Schell, R.: 'Visual Autonomous Automatic Landing of Airplanes'. AGARD-GCP: Integrated and Multi-Function Navigation, Ottawa, Canada, May 1992.

Dickmanns, E.D.; Zapp, A.: 'A Curvature-based Scheme for Improving Road VehicleGuidance by Computer Vision'. In: 'Mobile Robots', SPIE-Proc. Vol. 727, Cambridge, Mass., Oct. 1986, pp 161-168.

Dickmanns, E.D.; Zapp, A.: 'Autonomous High Speed Road Vehicle Guidance by Computer Vision'. 10th IFAC World Congress, Munich, July 1987, Preprint Vol. 4, pp 232-237.

Dickmanns, E.D.; Graefe, V.: a) 'Dynamic monocular machine vision' b) 'Application of dynamic monocular machine vision'. J. Machine Vision & Application, Springer-Int., Nov. 1988, pp 223-261.

Dickmanns, E.D.; Christians, T.: 'Relative 3D-state estimation for autonomous visual guidance of road vehicles'. In: Kanade, T. e.a. (ed.): 'Intelligent Autonomous Systems 2', Amsterdam, Dec. 1989, Vol.2 pp. 683-693.

Dickmanns, E.D.; Mysliwetz, B.; Christians, T.: 'Spatio-temporal guidance of autonomous vehicles by computer vision'. IEEE-Transactions on Systems, Man and Cybernetics, Vol. 20, No. 6, Nov/Dec 1990 ,Special Issue on Unmanned Vehicles and Intelligent Systems, pp 1273-1284.

Eberl, G.: 'Automatischer Landeanflug durch Rechnersehen'. Dissertation, Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München, 1987.

Hock, C.: 'Landmark Navigation with ATHENE'. Proceedings of Int. Conference on Advanced Robotics (ICAR), Pisa, 1991.

Kailath, T.: 'Linear systems'. Englewood Cliffs, N.J. Prentice Hall, 1980.

Kalman, R.E.: 'A new approach to linear filtering and prediction problems'. J. Basic Engineering, 1960, pp 35-45.

Kant, I.: Werke in zehn Bänden, Hrsg. W. Weischedel, Wiss. Buchges. Darmstadt, 1983.

Kuhnert, K.D.: 'Zur Echtzeit-Bildfolgenanalyse mit Vorwissen'. Dissertation, Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München, 1988.

Lindberg, D.C.: 'Theories of Vision from Al-Kindi to Kepler'. University of Chicago Press, Chicago and London, 1976.

Maybeck, P.S.: 'Stochastic models, estimation and control'. Vol. 1, Acad. Press, 1979.

Mysliwetz, B.: 'Parallelrechner-basierte Bildfolgen-Interpretation zur autonomen Fahrzeugsteuerung'. Dissertation, Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München, 1990.

Mysliwetz, B.; Dickmanns, E.D.: 'A Vision System With Active Gaze Control for Real-Time Interpretation of Well Structured Scenes'. In: Hertzberger, L.O. (ed.) Proc. of 1-st Conference on Intelligent Autonomous Systems (IAS), Amsterdam, Dec. 1986, pp 477-483.

Mysliwetz, B.; Dickmanns, E.D.: 'Distributed Scene Analysis for Autonomous Road Vehicle Guidance'. Proc. SPIE Conf. on Mobile Robots, Vol. 852, Cambridge, USA, 1987, pp 72-79.

Mysliwetz, B.; Dickmanns, E.D.: 'Recursive 3D Road and Relative Ego-State Recognition'. IEEE-Trans. PAMI, Special Issue on 'Interpretation of 3D Scenes', Febr. 1992.

Nagel, H.H.: 'Overview on image sequence analysis'. In T.S.Huang (ed.): Image Sequence Processing and Dynamic Scene Analysis, NATO-ASI Series F2 Springer-Verlag, Heidelberg, 1983, pp 2-39.

Nagel, H.H.: 'From image sequences towards conceptual descriptions'. Image and Vision Computing, Vol. 6, No. 2, May 1988, pp 59-74.

Pontryagin, L.S.; Boltyansky, V.G.; Gamkrelidse, R.V.; Miscenko, E.F.: 'Mathematical theory of optimal processes'. New York, Wiley, 1962.

Schell, R.: 'Bordautonomer automatischer Landeanflug aufgrund bildhafter und inertialer Meßdatenauswertung'. Dissertation, UniBw München, Fakultät LRT, 1992.

Schell, R.; Dickmanns, E.D.: 'Autonomous Automatic Landing through Computer Vision'. In AGARD Conf. Proc. No. CP-455: 'Advances in Techniques and Technologies for Air Vehicle Navigation and Guidance'. 1989, pp 24.1-24.9.

Schick, J.; Dickmanns, E.D.: Simultaneous Estimation of 3D Shape and Motion of Objects by Computer Vision'. IEEE-Second Workshop on Visual Motion, Princeton, Oct. 1991.

Schopenhauer, A.: 'Die Welt als Wille und Vorstellung'. In: Löhneysen, W. (ed) Arthur Schopenhauer, Sämtliche Werke. Suhrkamp, Frankfurt a.M., 1819 (Nachdruck 1986).

Thomanek, F.; Dickmanns, D.: 'Obstacle Detection, Tracking and State Estimation for Autonomous Road Vehicle Guidance'. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Raleigh NC, 7-10 July 1992.

Vollmer, G.: Was können wir wissen? Bd. 1: Die Natur der Erkenntnis; Bd. 2: Die Erkenntnis der Natur. S. Hirzel Verlag, Stuttgart,1985.

Wuensche, H.J.: 'Bewegungssteuerung durch Rechnersehen'. Fachberichte Messen, Steuern, Regeln Bd. 20, Springer-Verlag, Berlin, 1988.

Zapp, A.: 'Automatische Straßenfahrzeugführung durch Rechnersehen'. Dissertation, Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München, 1988.

3D Computer Vision Techniques for Object Following and Obstacle Avoidance

Richard Evans
Roke Manor Research Limited,
Roke Manor Romsey,
Hampshire SO51 0ZN,
United Kingdom

1. SUMMARY

Imaging sensors are powerful tools enabling remote control, by tele-operation, of numerous tasks where the operator requires an appreciation of the three-dimensional structure of the viewed scene. Passive video sensors also lend themselves to tasks where covert operation or electromagnetic compatibility is required. A commonly mooted tele-operational task is that of driving a known vehicle through an unknown terrain - or keeping station on a known object moving through an unknown terrain. The computer vision aspects of automating this task are divided into two separate vision functions, which are the subjects of this paper:

- Analysis of image sequences of a general scene to extract its three dimensional (3D) structure without any prior information,
- Analysis of images of a well defined object, to extract its 3D position and orientation relative to the sensor.

For both these functions, the paper provides a brief introduction to possible techniques followed by further description of particular systems, DROID and RAPiD, developed by Roke Manor Research Limited. DROID is a general, feature-based 3D vision system using the structure-from-motion principle. That is, it uses the apparent image-plane movement of localised features viewed by a moving sensor to extract the three-dimensional structure of the scene. RAPiD is a model-based real-time tracker which extracts the position (X, Y, Z) and orientation (roll, pitch, yaw) of a known object from image data. The system operates iteratively, using a prediction of object pose (position and orientation) to cue the search for selected edge features in subsequent imagery. This approach results in minimal processing of image pixels, so that the system can be implemented at full video rate using modest hardware.

2. INTRODUCTION

A video image, as displayed on a TV monitor, is intrinsically a two dimensional object, yet a human operator can remotely control a wide range of tasks in the three-dimensional world by use of a video link. In such cases it tempting to ask if such tasks can be automated as the raw data used by the operator - the video data - has already been captured electronically.

The task of following or keeping station, or performing some manoeuvre with respect to a known object, is a commonly hypothesised example. If the application is to keep in formation with a nearby aircraft, dock a satellite module, or even to follow a cooperating vehicle over the uncluttered desert sands, we are generally concerned with known objects which can be defined in some detail in advance. More generally we may wish to manoeuvre a vehicle in a cluttered scene. In such cases the possibility of obstructions of an unknown shape will be a major concern, and the system will need to estimate the sensor platform's path relative to any obstacles.

Work at Roke Manor Research Limited has been directed towards both of the vision tasks implied above. This work has resulted in two systems, DROID and RAPiD, for estimating structure from image sequences and model-based tracking respectively. These systems enable 3D structure and relationships to be established. While some interpretation of 3D measurements is performed by DROID, interpretation of the 3D structure is largely beyond its scope, as are the functions of path planning or control of the movement of the sensor platform. DROID and RAPiD have now reached some maturity, but the methods have not been integrated into a single demonstration, so it must be admitted that the vision task described above is a focus of attention and the two systems will largely be described separately in what follows.

This introduction continues with a non-mathematical overview of the algorithms developed by Roke Manor for extracting scene structure from image sequences and for tracking the position and orientation of a modelled object. A more detailed mathematical description of the algorithms then follows in sections 3 and 4; the reader may wish to omit that description and skip to section 5, which illustrates the techniques in the context of a typical office corridor scene. The remaining sections of this paper describe the development status of the work (including real-time implementation), and provide a brief critical discussion and concluding remarks.

2.1 Structure from Motion

A human controller in a tele-operated system can employ a wide range of depth cues. Given a single static image he may use his general knowledge of the scene's domain to perform scene understanding, and this may be very precise in providing a 3D interpretation in certain domains. He may also use more general cues such as perceived surface shading or shadows. There are many such *shape-from-X* cues (where X stands for shading, shadows, reflectance, texture, perspective, etc.), though for computer vision these approaches currently seem applicable only to simple constrained scenes. In contrast, given a sequence of images, the assumption of scene rigidity and the invariance of 3D geometry with changing viewpoint provides a powerful lever which can be used to automatically extract quantified stuctural information by triangulation. This, the structure from motion approach, is of course only applicable if the correspondence between (image) features observed from differing view-points can be established, and if the movement of the sensor can be estimated between images.

Solutions to the image correspondence problem could be sought in a spatially continuous form as an optical flow field, defining for every point in one image of a sequence, the image coordinates of the corresponding point in the subsequent image. Images frequently contain large bland regions, however, and in such areas a flow field is ill-defined. Alternatively images could be analysed for discrete image tokens, or features, that are likely to correspond to objective 3D scene elements. The attraction of using features, as compared to a spatially continuous method (such as the gradient optical-flow technique [1]), is that appropriately chosen features encapsulate the highest quality information, forming "seeds of perception" [2], and processing effort is not wasted on low quality regions of the image. This is of considerable interest in a real-time application, as an image contains a very large amount of data.

A further attraction of discrete features, is that they can be developed directly into high-level 3D scene descriptors. These provide a convenient mechanism for passing information across a potentially unlimited number of images, so the geometric accuracy of feature-point measurements can be refined over increasingly long triangulation base-lines. A number of algorithms have been proposed for the detection of point-features, sometimes referred to as 'interest' points or 'corners'. DROID uses a proprietary method (described in section 3) which proves to be robust both as a feature detector and in providing reliably matched features between image frames.

Following feature or corner extraction on the first two frames of a sequence, DROID's function is to estimate sensor and feature positions. The processing of these two frames constitutes DROID's *boot* phase. Thereafter, in DROID's *run* mode, the system functions on an iterated cycle updating sensor and feature positions (and instantiating positions of newly detected features). It would be desirable if DROID could optimally update its state-vector of sensor pose (position and orientation) and feature positions. There are typically many tens - possibly hundreds - of 3D features being processed at any time, however, and it is impracticable to consider a treatment of all correlations between ego-motion errors and feature-point position errors (and between one feature-point and another), and consequently the update is performed in two passes:

- calculation of sensor platform motion, i.e. ego-motion,
- optimal instantiation and update of feature 3D positions, assuming the ego-motion calculation is correct.

This simplification leads to a viable system whose overall cycle of algorithm steps is shown in Figure 1. Steps of particular interest are:

2D-2D feature matching: This concerns the matching of uninstantiated features (i.e. those extracted from a previous image frame but which are yet to be projected into 3D) to newly extracted features. The process is based on a combination of spatial constraints (in the image plane) and feature attributes, which describe the characteristics of a feature point. Spatial search regions are bands centred on epi-polar lines. These lines are the projections onto a later image frame of rays passing from the pinhole of the camera through the feature positions seen in an earlier frame. (This projection requires a prior estimate of ego-motion.)

Ego-motion calculation: Ego-motion is estimated by minimising the discrepancy between the observed and predicted positions of matched features. In the boot case, a feature can only be predicted to lie at some point on an epi polar line, so that the measured discrepancy is based on the perpendicular distance to epi-polars as shown in Figure 2. In run mode, i.e. from frame 3, the discrepancy is based on projection of 3D points; see Figure 3. At boot some prior estimate of motion is required: thereafter the system can be free running or use constraints based on past motion to ensure a smooth track estimate.

2D-3D feature matching: Matching of already instantiated 3D features to newly extracted 2D features is similar to the 2D-2D process, but, with an estimate of feature position now available, spatial search constraints are based on a projection of estimated positional error into the image plane.

Kalman filter instantiation/update: feature point positions are estimated and updated in an optimal weighting of new observations and previously estimated (3D) positions. The process can be visualised as in Figure 4, where the uncertainty in feature position is depicted by an elliptical error surface. The new observation constitutes a cylindrical error surface centred on the ray to the observed feature position. Intersection of these error surfaces results in a new smaller error ellipse, which is gradually refined by subsequent observations.

2.2 Model Based Tracking

Three-dimensional (3D) model-based vision is concerned with finding the occurrence of a known 3D object within an image, and obtaining a quantitative measure of the

object's location in three-dimensional space. The location of the object can then be used for tasks such as robotic manipulation, process monitoring, vehicular control, etc. As only certain aspects of the object are utilised, these aspects are said to form a *model* of the object: it is the occurrence of the model that is sought. A geometric model is attractive to work with, because the 3D geometry of an object is invariant to changes in view-point and so can provide reliability and computational simplicity. Additionally, the results from a geometric model will be quantitative. Non-geometric models, utilising such attributes as colour and texture, may serve to reveal the existence of the object, but not a quantitative measure of its 3D location.

Model-based tracking is model-based vision applied to a sequence of video images. Model-based tracking appears initially to be a much more difficult problem than model-based vision, due to the high data-rate in an image sequence (up to 10 Mbytes/second at video-rate). The continuity between successive images can, however, lead to it being a much easier problem, because the motion of the object can be predicted with some precision. It can thus be advantageous to process at the maximum rate, which is at field rate (50Hz) for standard video cameras. The geometric model features used for tracking must be cheap to extract, computationally, if processing is to proceed at near video-rate. Computationally expensive and unreliable model features, such as closed regions representing surfaces, cannot be afforded. This indicates the use of simple local features such as points (or 'corners') and edges.

The tracking of rigid and jointed objects has been performed by Lowe [3] using straight edge segments extracted over the entire image area. This approach is computationally expensive and slow, and has been demonstrated at about 1 Hz using Datacube image-processing hardware. The strength of the approach is that a prior estimate of object pose is not necessary. Another full-image method is that of Bray [4], who uses the discrepancies of the locations of extracted Canny edgels from the projected model to update the pose, and thus needs a good pose estimate. The approach of Stephens [5] is closest to Roke Manor's RAPiD, his model consisting of control points on high-contrast edges, but determination of the pose change, from frame to frame, is performed using many iterations of a Hough transform. Stephens' system has been demonstrated in real-time (about 10 Hz) using a small Transputer array.

The approach taken in RAPiD is to use a 3D model consisting of selected control points situated on high-contrast object edges, such as surface markings, fold edges (such as edges of a cube), and profile edges (such as the outline of a sphere). The processing cycle is illustrated in Figure 5. Given a prior estimate of object pose, these model points are simple to project onto the image, and the corresponding image edges simple to locate by searching the image pixels perpendicularly to the expected edge direction. The set of measured displacements of these edges is used to refine, or update, the estimate of model pose. Since the estimated model pose must be close to the true model pose for the correct image edges to be associated with the model points, the update equations can be safely linearised. This linearisation, together with the minimal image processing required to locate edges at control points, enables RAPiD to function at full video rate using only modest processing hardware in many cases of interest.

If the target object is moving across the image, the above method of updating the object pose will produce a result that lags behind the true pose. Thus it is desirable to include a predictive element in the tracking loop. This prediction is most simply achieved by using a position and velocity predictor/smoother, such as the so-called alpha-beta tracker [6], but, with more sophistication, a Kalman filter [7] can be used to greater effect. The Kalman filter enables the relative uncertainties in the estimated pose to be weighted appropriately and the expected dynamics of the object and the sensor platform can be included in the smoothing/prediction process. Thus RAPiD can be used for tracking a moving object with a fixed camera, or alternatively if a stationary scene is tracked as the camera moves, the pose of the camera is determined.

A number of RAPiD's features make it very robust in operation. The use of a model defined by selected control points on object edges makes it unnecessary to extract the whole of a edge, thus obviating a step which (for simple techniques at least) is generally prone to error in the form of fragmentation and incomplete termination. As will be apparent from the mathematical description, failure to detect an edge at a control point is not catastrophic, though failure to detect features degrades the accuracy of pose estimates; the measurement error model used in the Kalman filter enables the changed uncertainties in measurements to be taken into account in the smoothing/prediction process.

The required model is a small data structure of typically 20-40 control points. These should be placed on straight edges (edges of low curvature are also acceptable) or certain kinds of profile edge, such as conic sections or, surfaces of revolution. Additional robustness can be provided by specifying the expected image polarity of an edge, which can prevent RAPiD being seduced by background edges in a cluttered scene.

3. THE DROID ALGORITHMS

3.1 Feature Extraction

The primitive features extracted by DROID are *feature-points* or *corners*, which abound in natural and man-made scenes. Feature-points are likely to correspond to real 3D structure, such as corners of objects and surface markings, and also to texture of an appropriate scale. The spatial localisation of feature-points can give good repeatability, even for natural scenes where an image decomposition into straight-line fragments is highly erratic. The extraction of feature-points is a spatially and temporally local operation, and is both repeatable and computationally (comparatively) cheap.

On each image processed by DROID, discrete feature-points are first extracted, with feature extraction performed

independently on each image. Feature-points are detected by use of a local auto-correlation operator [8]. Letting the image intensity (grey-level) be I(x,y), at each point in the image construct the 2x2 matrix

$$M = \begin{pmatrix} <(\partial I/\partial x)^2> & <(\partial I/\partial x).(\partial I/\partial y)> \\ <(\partial I/\partial x).(\partial I/\partial y)> & <(\partial I/\partial y)^2> \end{pmatrix}$$

where angle braces indicate local Gaussian smoothing of the arguments (a smoothing size of 1 to 2 pixels is commonly used), and the first gradients, $\partial I/\partial x$ and $\partial I/\partial y$, are obtained by use of a 5x5 mask. The eigenvalues of M encode the shape (the principal curvatures) of the local auto-correlation function: if both are large, the local grey-level patch cannot be moved in any direction on the image-plane without significant grey-level changes occurring, while an edge or line will have one large and one small eigenvalue. A *corner response function*, R, is formulated to respond to both eigenvalues being large, while not requiring explicit evaluation of the eigenvalues:

$$R = \det(M) - [\,\mathrm{trace}(M)\,]^2 \cdot k / (k+1)^2$$

The subtracted term makes the above formulation to some extent 'edge-phobic', to ensure it does not fire off pixellation on strong edges, a common failing of some corner detectors. The value of the parameter k is the maximum ratio of eigenvalues of M to which the response function is positive. Typically a value of 25 is used. The local (3x3) maxima in the response function form candidate corners, and we select either the *n* strongest, or else all those exceeding a pre-defined threshold. The former selection procedure is better suited to image sequences with a widely varying content, frame-to-frame. The convolutions used in obtaining the response function may cause a feature-point to be slightly mis-positioned, but the mis-positioning will usually be consistent over time and so be of little importance. By performing a local quadratic fit to the response function, the feature-points can be located to sub-pixel accuracy.

The most important property [9] of feature-point extraction is high repeatability; with this algorithm often over 80% of the extracted points are matchable between frames. To each feature-point is associated descriptive grey-level attributes, explicitly the local grey-level (as defined by a Gaussian smoothing mask), and the smoothed first spatial gradients. These attributes are assembled into an *attribute vector*, **a**, which will be used to disambiguate matches.

Feature-points are attractive to work with as they are simple to track over time, and are easy to handle in 3D. Straight edge features are similarly attractive and can be handled by DROID, but they are more suited to man-made environments than natural environments, in which they are scarce [10, 11]. Although curving and squiggly edges are abundant in natural scenes, they can be temporally unstable, and present formidable problems in finding a suitable representation to handle the geometric information they contain.

3.2 Camera Calibration

Since DROID is based on the geometry of image features, it is essential that an accurate interpretation of the location of the features is performed. In particular, it is necessary to know the direction in space towards which each of the pixels in the image is looking; this is called the geometric calibration of the camera. By modelling the camera as a pin-hole camera with specific distortions (eg. radial lens distortions), and using only CCD cameras whose sensing elements form a stable rectangular array, a parametric form for the geometric camera calibration can be devised. This model has been found to be good for many CCD cameras and lenses. Camera calibration is performed using two images of an accurately known planar calibration tile [12], resulting in accurate measurements of the focal length, aspect ratio, location of the optical centre, and up to two terms of radial distortion.

The calibration enables the extracted feature-point locations to be transformed to an 'ideal' distortion-free pin-hole camera of unit focal-length (UFL), whose image-plane is positioned in front of the camera pin-hole to avoid tiresome minus signs. A Cartesian camera coordinate system is defined to have its origin at the pin-hole of the camera and Z axis aligned along the optical axis. The X and Y axes are parallel to the image plane. The image x axis is horizontal and pointing to the right, while the image y axis is vertical and pointing downwards. This gives a right-handed coordinate system, as illustrated in Figure 6. A point positioned at $\mathbf{R} \equiv (X,Y,Z)$ in local camera coordinates will be imaged in UFL camera coordinates at

$$\mathbf{r} \equiv (x,y) = (X/Z , Y/Z)$$

This is the *perspective projection*, and henceforth all image positions will be expressed in UFL coordinates.

It will often be necessary to represent the same 3D point in two different coordinate systems, for example in camera coordinates and global coordinates. Consider a point located at $\mathbf{R}_1$ in a first coordinate system, and at $\mathbf{R}_2$ in a second coordinate system. These point locations will be related by

$$\mathbf{R}_2 = A(\boldsymbol{\theta})^T (\mathbf{R}_1 - \mathbf{t})$$

$$\mathbf{R}_1 = A(\boldsymbol{\theta})\, \mathbf{R}_2 + \mathbf{t}$$

where the rotation matrix, $A(\boldsymbol{\theta})$, and the translation vector, **t**, describe respectively the attitude and the location of the second coordinate system with respect to the first. (The superscript T denotes matrix transpose.)

Rotations are represented by a 3-vector $\boldsymbol{\theta}$, whose direction is the axis of rotation, and whose magnitude is the (right-handed) angle of rotation in radians. The elements of the orthonormal 3x3 rotation matrix, $A(\boldsymbol{\theta})$, are:

$$A_{ij} = \cos\theta\, \delta_{ij} + (1 - \cos\theta)\, \hat{\theta}_i \hat{\theta}_j - \sin\theta \sum_k \varepsilon_{ijk} \hat{\theta}_k \qquad 1 \le i,j \le 3$$

where $\theta = |\boldsymbol{\theta}|$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}/\theta$, and ε_{ijk} is the Levi-Civita symbol. The representation is singular at $\theta = 2\pi$, but this is avoided by working always with $\theta \le \pi$. Note that

rotation vectors are neither commutative nor associative (unless they are parallel), and that successive applications of rotations are best handled using quaternions.

The location and attitude of the camera is generally referred to as its *ego-motion*, expressed as the '6-vector', $\mathbf{q} = (\boldsymbol{\theta},\mathbf{t})$. The ego-motion may be measured from the global origin (as illustrated in Figure 6), or may be in some convenient local coordinates. The location and attitude of a rigid body with respect to a reference coordinate system is called its *pose*. The pose of a body is the rotation, $\boldsymbol{\theta}$, and the translation, $\mathbf{t}$, that must be applied to the body coordinate system so as to correctly position the body.

3.3 Boot-Strap Processing

The task of boot-strap processing is to initiate the 3D representation of the viewed scene from feature-points found in the first images, without assuming any knowledge of the scene content. The 3D representation will be in terms of Kalman filtered points. For a monocular system, the first 2 images of the sequence are used for boot. DROID can be operated in a stereo mode [13], in which case boot consists of a conventional stereo process performed on the 2 or more simultaneously captured images comprising the first frame.

3.3.1 *Boot Matching*

The processing of a monocular image sequence is initiated with the first two images. Using a prior estimate of the camera motion, each extracted feature-point from one image generates on the other image an epi-polar search line near which candidate matches are sought. If the prior ego-motion estimate from frame 1 to frame 2 is $\mathbf{q} = (\boldsymbol{\theta},\mathbf{t})$, and the observed point on frame 2 is at $\mathbf{r}_2 = (x_2,y_2)$, then the epi-polar line on frame 1 will pass through the image points $(t_x,t_y)/t_z$ and $(p_x,p_y)/p_z$, where $\mathbf{p} = A(\boldsymbol{\theta})\ (x_2,y_2,1)^T$. The epi-polar line is broadened out into a band in which match candidates are sought, and this broadening is chosen to reflect both the uncertainty in the prior estimate of the camera motion and errors in feature-point positioning. The length of the epi-polar line may be truncated at minimum and maximum depths, to reduce the number of spurious match candidates. Matching ambiguities are resolved by use of the grey-level attributes. If the attribute vectors for two points are $\mathbf{a}_1$ and $\mathbf{a}_2$, then the attribute mismatch between the points is

$$m_{1,2} = |\mathbf{a}_1 - \mathbf{a}_2| / \sqrt{(|\mathbf{a}_1|.|\mathbf{a}_2|)}$$

For a successful match, the mismatch value must be lower than a set threshold, and if there are several candidates, the one with the lowest mismatch is chosen. Typically over 80% of the feature-points are found to be correctly matchable, and the few incorrect matches are discounted by outlier removal procedures (see below). Unmatched feature-points are kept for possible future matching; they are said to be placed in *limbo*.

3.3.2 *Boot Ego-Motion*

Using the feature-point matches, the camera ego-motion, $\mathbf{q} = (\boldsymbol{\theta},\mathbf{t})$, is next determined. The boot-strap ego-motion is calculated by an iterative multi-dimensional Newton scheme, minimising the image-plane distances between the location of feature-points and the truncated epi-polar lines of their matching features [14]. To cope with mis-matches, a robust minimisation is performed. The starting point of the iterative scheme is the prior estimate of camera motion, and good convergence is usually achieved in 4 to 6 cycles. Prior knowledge about the camera motion may be imposed by a set of soft constraints quadratically linking the 6 ego-motion parameters, $\mathbf{q}$. By varying the constraint coefficients, planar, linear, or curved motion may be imposed. It is essential that a translational constraint is imposed at boot to resolve the speed-scale ambiguity, which is otherwise left entirely unresolved by the visual data. The minimisation scheme and the form of the constraints is described below in section 3.4.2.

Once ego-motion has been determined, the 3D locations of matched points can be estimated by triangulation. The uncertainty in the image-plane position of a feature-point leads to uncertainty in its 3D location. This uncertainty is used to start-up a Kalman filter (KF) for each point, whose variables represent the spatial probability distribution function of the point, and consist explicitly of a 3D mean position and covariance. Strictly, it is extended Kalman filters that are being used, as the time evolution of the filter is only being approximated as linear. The KF enables subsequent observations of the point to be optimally and cheaply combined, and high spatial accuracy achieved. The update and initiation of the KFs is described below in section 3.4.3.

3.4 Run Mode

After the 3D representation has been initiated in the boot-mode, successive frames are processed in the run-mode. The run-mode provides an evolving 3D representation, which increases in accuracy and completeness as more frames are processed. Accuracy is achieved by using Kalman filtering to optimally combine observations of an individual feature-point seen over an extended period of time. The representation evolves by the inclusion of newly seen feature-points, and the exclusion of points that are no longer visible. In this way, an unlimited sequence of images can be processed.

Much of the work of DROID is performed in so-called disparity space, for reasons of speed and numerical stability. A point at $\mathbf{R} = (X,Y,Z)$ in Cartesian camera coordinates has coordinates $\mathbf{S} \equiv (x,y,z) \equiv (X/Z,Y/Z,1/Z)$ in the corresponding disparity space. Thus the first two components of $\mathbf{S}$ are the image coordinates of the perspective projection of $\mathbf{R}$, and the third component is the reciprocal depth. Note that straight lines in Cartesian space are straight in disparity space, and similar relationships hold for both planes and conics. The KF of each feature-point contains in disparity space a mean position (or centroid), $\mathbf{S}_{KF}$, and an estimated error covariance Σ_{KF} (a 3x3 matrix). These can be thought of as defining a normal probability distribution function in disparity space.

3.4.1 *Run Matching*

In the run mode, matches are sougnt between extracted image feature-points and existing KFs by projecting the

KFs down onto the image-plane. First of all, the KFs must be transformed from the previously used disparity space to the disparity space of the current estimate of camera ego-motion. This is straightforward for the centroid (by transforming to and from Cartesian space), but for the covariance, using Cartesian space is inadvisable for distant points because of poor numerical conditioning. To overcome this problem, a direct disparity-to-disparity transform has been devised, which uses a well-conditioned similarity transform. By these means the KFs are brought into the currently used disparity space.

The projection of the KF covariance, Σ_{KF}, onto the image-plane is obtained by pre- and post-multiplying with the projection matrix, $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, and its transpose, which simply serves to extract the upper 2x2 block of Σ_{KF}. By linearly combining the projected KF covariance with the observation covariance, Σ_{obs}, a matching covariance matrix is obtained

$$\Sigma_{match} = k_{obs} \cdot \Sigma_{obs} + k_{proj} \cdot P \, \Sigma_{KF} \, P^T$$

where the two coefficients k govern chosen levels of statistical significance. The observation covariance, Σ_{obs}, is usually taken to be diagonal and equivalent to, say, one pixel. The observation covariance coefficient is chosen to be sufficiently large for it to account for uncertainty (error) in the prior estimate of camera motion. If $\mathbf{r}_{KF}$ is the perspective projection of the KF centroid,

$$\mathbf{r}_{KF} = P \, S_{KF}$$

(trivially, the first two coordinates of S_{KF}), and $\mathbf{r}_{obs}$ is the location of an extracted feature-point, then the feature-point is a match candidate if

$$(\mathbf{r}_{KF} - \mathbf{r}_{obs})^T \, \Sigma_{match}^{-1} \, (\mathbf{r}_{KF} - \mathbf{r}_{obs}) < 1$$

that is, it lies in an ellipse centred on the projected KF centroid. The searching for candidates is accelerated by using a coarse binning scheme for the feature-points, and only examining the bins which the ellipse overlays. Candidate matches are assessed using their grey-level attributes, and irresolvable contentions are discarded to ensure that no multiply-defined KFs are generated.

3.4.2 *Run Ego-Motion*

Once feature-point matches have been obtained, the ego-motion, $\mathbf{q}$, is determined by finding the camera attitude and location that brings projected KF centroids, $\mathbf{r}(\mathbf{q})$, into best alignment with their matching observed feature-points, $\mathbf{r}_{obs}$. If $\mathbf{R}_0$ is a KF centroid location in Cartesian camera coordinates, then a relative ego-motion $\mathbf{q} = (\boldsymbol{\theta}, \mathbf{t})$ of the camera will make the centroid project onto the image at

$$\mathbf{r}(\mathbf{q}) = (X(\mathbf{q}), Y(\mathbf{q})) \, / \, Z(\mathbf{q})$$

where

$$\mathbf{R}(\mathbf{q}) \equiv (X(\mathbf{q}), Y(\mathbf{q}), Z(\mathbf{q})) = A(\boldsymbol{\theta}) \, \mathbf{R}_0 + \mathbf{t}$$

The measure of 'best alignment' used above is given by a matching covariance, Σ_{match}, which is, as before, an appropriate combination of the observation and projected KF covariances. The contribution of the i'th matched point to an objective function to be minimised is thus

$$E_i(\mathbf{q}) = (\mathbf{r}(\mathbf{q}) - \mathbf{r}_{obs})^T \, \Sigma_{match}^{-1} \, (\mathbf{r}(\mathbf{q}) - \mathbf{r}_{obs})$$

The ego-motion determination is performed by minimising a single objective function, $E_{total}(\mathbf{q})$, which is composed of a weighted sum of contributions from each matched point, together with a prior-constraint term producing soft constraints:

$$E_{total}(\mathbf{q}) = \mathbf{q}^T \, \Sigma_{prior}^{-1} \, \mathbf{q} + \sum_{\text{points } i} w_i \, E_i(\mathbf{q})$$

For there to be no bias from the prior-constraint term, the ego-motion $\mathbf{q}$ is taken to be relative to the expected or anticipated camera pose. Global ego-motion is not used because rotation vectors can only be approximated as commutative near $\mathbf{q} = 0$.

The objective function is minimised by using a multi-dimensional Newton minimisation, for which the first and second differentials of the objective function must be calculated. These are constructed analytically by using expressions for the first differentials of the projected KF centroids, $\partial \mathbf{r}(\mathbf{q}) / \partial \mathbf{q}$, and by assuming that there is negligible dependence of the matching covariances on $\mathbf{q}$. Each cycle of the Newton scheme produces a new (and, it is to be hoped, better) estimate of the ego-motion, $\mathbf{q}'$, from a previous estimate, $\mathbf{q}$:

$$\mathbf{q}' = \mathbf{q} - [\, \partial^2 E_{total} / \partial \mathbf{q}^2 \,]^{-1} \, [\partial E_{total} / \partial \mathbf{q}]$$

The starting guess of the minimisation is with the camera at its expected position (ie. $\mathbf{q} = 0$), and usually 4-6 iterations give a good convergence.

The main cause of error in the ego-motion calculation is incorrect matches, which, if uncorrected, significantly bias the result. This problem is overcome both by using robust minimisation techniques to de-weight the effect of the mismatches, and by performing the complete matching/ego-motion cycle twice, with tighter search regions on the second pass. The robust minimisation technique ascribes a weight to each point on each cycle of the Newton minimisation. The weight, w_i, of the i'th point on the current cycle depends exponentially on its contribution, $E_i(\mathbf{q})$, to the objective function of the point on the previous cycle:

$$w_i = \exp - (c \, E_i(\mathbf{q}) \, / \, \overline{E_i(\mathbf{q})})$$

The denominator is the (weighted) average objective function contribution of all the points, and is used to estimate the distribution of the E_i's, and this results in outliers being continuously and strongly de-weighted.

Ego-motion determination is generally very accurate in the short to medium term. An example is quoted by Harris [15] of a short sequence of 10 images taken from a helicopter with a generally forward translation of about 10 feet per frame. The accuracy of the attitude component of the ego-motion, the difference between the DROID analysis and the ground truth data, is better than 0.25°,

though the helicopter undergoes a yaw of 15°. The accuracy of the translational components is less than 0.7 feet, which is less than 0.8% of the total flight distance.

In a long image sequence, long-term drifts can occur, in which both the ego-motion and perceived structure are self-consistently in error. For example, both the camera position and the perceived structure might come to be displaced 1 metre to the right of their true values, and yet the visual observations will be entirely self-consistent. Although there is no feedback mechanism to correct such an error from the imagery alone, external ego-motion measurements (eg. odometry) may be of use in resolving these ambiguities. Drifting can occur in both attitude and translation, and also in the speed-scale factor. Speed-scale drift is where both the speed of the camera and the perceived scale of the structure are in error by the same factor. The speed-scale ambiguity is resolved by using stereo, as the stereo base-line provides a yard-stick for the structure. The problem of drift is exacerbated by the camera turning by an angle greater than the width of its field-of-view, so that previously established structure is lost from sight and no longer acts as a stable reference.

3.4.3 *Kalman Filter Update*

Each time a point is observed and matched, a more precise estimate of its 3D position may be obtained. This is because the new observation provides further information relating to the 3D position of the point. Kalman filtering is a method of combining a number of noisy measurements which is, in certain circumstances, statistically optimum. In DROID, each tracked point has its own filter whose job is to estimate both the point's most likely 3D location, and its positional uncertainty. An alternative approach, that of using a single high dimensionality filter containing the coupled coordinates of *all* the points, permits the imposition of geometric constraints [16], but at a high computational cost, and a danger of irrecoverably coupling unassociated features.

To explain the use of the KF, consider just a single point, as all are treated independently and in a similar fashion. Let the feature-point be observed in the current image at image-plane position, $\mathbf{r}_{obs}$; this is the KF *measurement*. Its estimated positional accuracy is specified by the observation covariance matrix, Σ_{obs}. The *state space* for the KF is the 3D location of the point in disparity space. Let the current estimate for the point's location be $\mathbf{S}_{KF}$ (called the centroid), and the accompanying estimate of its positional accuracy be given by the covariance Σ_{KF}. The covariance and centroid after updating the KF with the current observations are given by

$$\Sigma'_{KF} = [\, \Sigma_{KF}^{-1} + \mathbf{P}^T \Sigma_{obs}^{-1} \mathbf{P} \,]^{-1}$$

$$\mathbf{S}'_{KF} = \Sigma'_{KF} \, [\, \Sigma_{KF}^{-1} \mathbf{S}_{KF} + \mathbf{P}^T \Sigma_{obs}^{-1} \mathbf{r}_{obs} \,]$$

where, as before, P is the projection matrix. (The process noise term, often used in Kalman Filtering, has been omitted from the filter because past observations of a point are considered to be as valid as current observations, and there is no time-evolution because the points are assumed to be stationary in Global coordinates.) As DROID in fact works with the inverse covariance matrix, the former equation reduces to a matrix addition, and the latter to solving a set of 3 simultaneous linear equations. If, after update, the disparity coordinate of the centroid is negative, it is reset to a small positive value to prevent the point subsequently flipping behind the camera.

The KF update process is illustrated in Figure 4, in which surfaces of constant probability density are shown in disparity space. The vertical tube represents the observed feature-point and its covariance, while the larger and smaller ellipsoids represent the KF before and after update respectively.

3.4.4 *Kalman Filter Creation and Destruction*

The feature-points on the current frame that fail to match to existing KFs, may be epi-polar matched (i.e. 2D to 2D matched) to those that remained unmatched from earlier frames and were retained in limbo. This enables KFs for new points to be initiated. The epi-polar matching is the same as in boot (section 3.3.1). The KF initiation, which is also the same as boot, simply makes use of the KF update equations applied to the pair of initial observations.

KFs which repeatedly fail to match are discarded or *purged*, whilst those leaving the field of view are *retired* (matches are no longer sought), but kept on for a while for use in the structural representation. Points that are incorrectly matched at boot will cause KFs to be initiated at locations that in general will not be supported by matches on subsequent frames, and so these erroneous KFs will be purged from the system.

3.5 Surface Interpretation

A 3D geometrical representation should ideally describe all the visible surfaces, seen in the current image or in the past, and should perhaps even infer the existence of unseen surfaces (eg. the continuity of a wall behind a lamp-post). An ideal surface representation would use high-level components, such as planes and conics, to describe the scene, but in unconstrained environments, especially natural scenes, such components may be rare, ill-fitting or ill-conditioned. A more adaptable representation is needed, one which can cope with the inaccurate and spatially non-uniform data that is obtained from real vision systems. Since surfaces cannot be directly measured, and must be inferred from surface markings, bounding edges, etc., a flexible interpolation scheme based on the measured geometric features would be appropriate.

The maintenance of a low-level geometric representation for parts of the scene that have left the field of view for a period of time does not seem worthwhile: it is expensive to maintain (in computer time and space), and even if low-level features are seen again, they are not likely to be recognised as the same ones because of changes of appearance (scale, aspect, reflectance, etc.). Such a 'forgetful' system operates both in people, as the 'persistence of vision', and in DROID. Using the currently visible features to construct surfaces leads to an ego-centric representation, such as a depth-map or the 2.5D sketch [17].

3.5.1 *Planar Facet Representation*

The 3D points from DROID form a sparse depth map, bland regions of the image containing no points. To obtain a surface representation, an interpolation scheme based on the current image is used to construct a full depth map. As only currently visible points on the image are maintained in 3D, a single-valued surface (in range) passing through them should approximate to the depth map. The use of an ego-centric (camera-based) representation avoids the need for multiply-valued surfaces with the associated danger of incorrect point assignment, which could occur, for example, with overhanging structure in a plan-view projection. Working with points that are sufficiently mature to be reliable, the depth map is filled-out by a piece-wise linear interpolation between the image-plane locations of the 3D points. This is performed by using the Delaunay triangulation in the image-plane: each resulting triangle is interpreted as a 3D triangular planar facet passing through three 3D points. The Delaunay triangulation is chosen as it forms compact triangles (long thin triangles are physically implausible), and is cheap to compute (nearly linear in the number of points). The resulting surface is continuous and single-valued in range, but will not fill the entire image-plane unless supported by previously seen points now outside the image. The surface may be relatively coarse as it can be no finer than the separation of the features, and so cover over fine structure in the manner of a draped-sheet. Depth discontinuities in the surface are not currently permitted. As the surface is constructed anew at each new image, it will quickly respond to changes in the structure, but it does suffer from an amount of temporal instability.

3.5.2 *Using Surfaces*

The explicit 3D structural information made available by DROID is intended for open-ended use in a range of high-level tasks, such as obstacle detection, recognition, navigation and path-planning. Such tasks are currently being investigated in relation to performing automatic visual guidance of wheeled or tracked robot vehicles in both indoor and outdoor environments. The most immediate task is to provide safe operation (don't crash!), and this is performed by locating upstanding structural elements in the planar facet surface representation.

For movement in the vicinity of man-made structures, the location of prominent structural elements such as vertical walls and corridors, is of value. Detection of such structures can lead to map registration and on to more sophisticated navigational abilities. The detection of vertical walls around a ground vehicle is being undertaken by considering the plan-view coordinates of DROID points with heights above the floor level. A vertical wall should appear as a straight line in plan-view, and this may be extractable using a Hough transform.

4. THE RAPiD ALGORITHMS

4.1 Single Frame Pose Estimation

The coordinate systems used in RAPiD are shown in Figure 7. Define the Cartesian camera coordinate system, which has its origin at the camera pin-hole, Z-axis aligned along the optical axis of the camera, and X and Y axes aligned along the horizontal (rightward) and vertical (downward) image axes respectively. Imaging of points in 3D will be handled by the introduction of a conceptual image-plane situated at unit distance in front of the camera pin-hole. The conversion to these coordinates from pixels is facilitated by the use of the geometric calibration of the camera, and henceforth all image locations will be expressed in these conceptual image-plane units, and not in pixels. A point at position $\mathbf{R} = (X,Y,Z)^T$ in camera coordinates will project to image position $\mathbf{r} = (x,y)^T = (X/Z,Y/Z)^T$.

Define a model coordinate system, with origin located at **T** in camera coordinates, and with axes aligned with the camera coordinate system. (A different orientation of model axes may be more suitable for the original specification of the control points of the model; in which case assume that the model is pre-rotated from a reference attitude used for specification.) Consider a control point on the model located at **P** in model coordinates, and situated on a prominent 3D edge. This control point will project onto the image at $\mathbf{r} = (T_X+P_X, T_Y+P_Y) / (T_Z+P_Z)$. Let the tangent to the 3D edge on which the control point is located be called the control edge. The orientation of the edge at the control point is defined by specifying a companion control point to **P**, often also located on the same physical edge, and which projects onto the image at **s**. By considering the image displacement between **r** and **s**, the expected orientation of the control edge on the image can be determined. Let this be an angle α from the image x-axis, so that

$$\cos\alpha = \left(\frac{s_x - r_x}{|\mathbf{s} - \mathbf{r}|}\right) \quad , \quad \sin\alpha = \left(\frac{s_y - r_y}{|\mathbf{s} - \mathbf{r}|}\right)$$

As a step towards refining an initial pose estimate, we wish to find the perpendicular distance of projected model control point **r** from the corresponding imaged object edge. Assuming that the orientations of the imaged edge and the projected model edge are nearly the same, a one-dimensional search for the image edge can be conducted by looking perpendicularly to the expected control edge from **r**. To search for the edge along an exact perpendicular would, however, require finding the image intensity at non-pixel positions. To avoid this inconvenience and computational cost, the edge search is performed in one of four directions: horizontally, vertically, or diagonally (that is, by simultaneous unit pixel displacements in both the horizontal and vertical directions). If the pixels are square, the diagonal direction will be at 45°, but with different image aspect ratios, other angles will be traversed. The direction which is closest to perpendicular to the control edge is chosen, and a line of pixel values centred on **r**, the projection of the control point, is read from the image.

Write the orientation of the line of pixels from the x-axis on the image-plane as the angle β, as shown in Figure 8. On the image-plane, let the dimensions of a pixel be k_x and k_y in the x and y directions respectively (thus k_x is the reciprocal of the focal length in pixels). Hence the orientation of the diagonal directions of the row of pixels will be $\beta = \pm\beta^*$, where $\tan\beta^* = k_y/k_x$.

The position of the actual edge brightness step within the extracted line is located by a simple threshold crossing. Suppose the imaged edge is encountered at a displacement from the projected control point **r** of n_x pixels in the x-direction and n_y pixels in the y-direction. (For diagonal directions, $n_x = \pm n_y$, otherwise either n_x or n_y will be zero.) Then the image-plane distance of **r** from the image edge along the row of pixels will be

$$d = \sqrt{n_x^2 k_x^2 + n_y^2 k_y^2}$$

and the perpendicular distance to the edge will be

$$l = d \sin(\beta - \alpha)$$

Let n be the number of pixel steps (horizontal, vertical or diagonal) traversed along the row of pixels before the edge is encountered. For the four permissible orientations of the row of pixels, the above equation for l is explicitly:

Horizontal ($\beta = 0$) $\quad l = -n\, k_x \sin\alpha$

Vertical ($\beta = \frac{\pi}{2}$) $\quad l = n\, k_y \cos\alpha$

Up diag ($\beta = \beta^*$) $\quad l = n\left(k_y \cos\alpha - k_x \sin\alpha\right)$

Down diag ($\beta = -\beta^*$) $\quad l = n\left(k_y \cos\alpha + k_x \sin\alpha\right)$

Each control point will result in a measured perpendicular distance, l, as illustrated in Figure 9. The set of these perpendicular distances will be used to find the small change in the object pose that should minimise the perpendicular distances on the next frame processed.

Consider rotating the model about the model origin by a small angle $\boldsymbol{\Theta}$, and translating it by a small distance $\boldsymbol{\Delta}$. Write these two small displacements as the 'six-vector', **q**. This will move the model point **P**, located in model coordinates at $\mathbf{R} = \mathbf{P} + \mathbf{T}$, to **R'** in camera coordinates

$$\begin{aligned}\mathbf{R}'(\mathbf{q}) &= (X', Y', Z')^T \\ &\approx \mathbf{T} + \boldsymbol{\Delta} + \mathbf{P} + \boldsymbol{\Theta} \times \mathbf{P}\end{aligned}$$

$$= \begin{pmatrix} T_x + \Delta_x + P_x + \theta_y P_z - \theta_z P_y \\ T_y + \Delta_y + P_y + \theta_z P_x - \theta_x P_z \\ T_z + \Delta_z + P_z + \theta_x P_y - \theta_y P_x \end{pmatrix}$$

This will project onto the image at

$\mathbf{r}'(\mathbf{q}) = (x', y') = (X'/Z',\ Y'/Z')$

Expanding in small $\boldsymbol{\Delta}$ and $\boldsymbol{\Theta}$, and retaining terms up to first order, gives

$$x' = x + [\, \Delta_x + \theta_y P_z - \theta_z P_y - x\,(\Delta_z + \theta_x P_y - \theta_y P_x)\,] \,/\, [T_z + P_z]$$

$$y' = y + [\, \Delta_y + \theta_z P_x - \theta_x P_z - y\,(\Delta_z + \theta_x P_y - \theta_y P_x)\,] \,/\, [T_z + P_z]$$

Thus $\mathbf{r}'(\mathbf{q})$ can be written

$$\mathbf{r}'(\mathbf{q}) = \mathbf{r} + \begin{pmatrix} \mathbf{q}.\mathbf{a} \\ \mathbf{q}.\mathbf{b} \end{pmatrix}$$

where

$\mathbf{a} = (-xP_y,\ xP_x + P_z,\ -P_y,\ 1,\ 0,\ -x)^T \,/\, (T_z + P_z)$

$\mathbf{b} = (-yP_y - P_z,\ yP_x,\ P_x,\ 0,\ 1,\ -y)^T \,/\, (T_z + P_z)$

Hence the perpendicular distance of the image edge from the control point is

$$\begin{aligned} l'(\mathbf{q}) &= l + \mathbf{q}.\mathbf{a} \sin\alpha - \mathbf{q}.\mathbf{b} \cos\alpha \\ &= l + \mathbf{q}.\mathbf{c} \end{aligned}$$

where

$\mathbf{c} = \mathbf{a} \sin\alpha - \mathbf{b} \cos\alpha$

and l is the measured distance to the edge.

Consider now not just one control point, but N control points, labelled i = 1..N. The perpendicular distance of the i'th control point to its image edge is

$l'_i(\mathbf{q}) = l_i + \mathbf{q}.\mathbf{c}_i$

We would like to find the small change of pose, **q**, that aligns the model edges precisely with the observed image edges, that is to make all $l'_i(\mathbf{q})$ zero. If the number of control points, N, is greater than 6, then this is not in general mathematically possible as the system is over-determined. Instead, we choose to minimise an objective function, E, the sum of squares of the perpendicular distances

$$E(\mathbf{q}) = \sum_{i=1}^{N} \left[l_i + \mathbf{q}.\mathbf{c}_i \right]^2 .$$

By setting to zero the differentials of E with respect to **q**, the following equations are obtained

$$\left(\sum_{i=1}^{N} \mathbf{c}_i \mathbf{c}_i^T \right) \mathbf{q} = - \sum_{i=1}^{N} l_i \mathbf{c}_i$$

This is a set of 6 simultaneous linear equations, and so can be solved using standard linear algebra.

The pose change, $\mathbf{q} = (\boldsymbol{\Theta}, \boldsymbol{\Delta})$, in the model pose specified by the above algorithm must now be applied to the model. Applying the change in model position is straightforward

$\mathbf{T} := \mathbf{T} + \boldsymbol{\Delta}$

The change in object attitude, however, causes some practical difficulties. Conceptually, the positions of the control points on the model should be updated thus

$\mathbf{P}_i := \mathbf{P}_i + \boldsymbol{\Theta} \times \mathbf{P}_i$

After thousands of cycles of the algorithm, finite numerical precision and the approximation to rotation represented by the above equation, results in the control points no longer being correctly positioned with respect to each other, and thus the model distorts. To overcome this problem, the attitude of the model is represented by the rotation vector $\boldsymbol{\phi}$ (a 3-vector whose direction is the axis of rotation and whose magnitude is the angle of rotation

about this axis), which rotates the model from its reference attitude, in which the model has its axes aligned with the camera coordinate axes. From the rotation vector ϕ can be constructed the orthonormal rotation matrix $A(\phi)$, which appropriately rotates any vector to which it is applied. Conceptually, the rotation matrix, $A(\phi)$, should be updated by the model attitude change, Θ, thus

$$A(\phi) := A(\Theta)\, A(\phi)$$

but by doing this, the orthonormality of the rotation matrix may be lost in time due to rounding errors, since, even allowing for the symmetry of the rotation matrix, it is still redundantly specified. Instead, the rotation vector, ϕ, is updated directly by use of quaternions. If $A(\phi)$ is the rotation matrix after the rotation vector has been updated, and the i'th model point is located in some reference coordinates at $\mathbf{P}_i^{(ref)}$, then the position of this point in model coordinates at the beginning of the next cycle will be

$$\mathbf{P}_i = A(\phi)\, \mathbf{P}_i^{(ref)}.$$

4.2 Kalman Filter

When applying the RAPiD technique to a practical case of a moving object, it is possible, in principle, to use the pose estimate, calculated by processing one video frame, as the initial estimate of the object's pose in the next video frame. This approach to tracking a moving object has the disadvantage that the object's motion would be limited to small movements between frames since RAPiD searches for model edges in a limited region about the predicted position. This problem can be overcome by using a simple predictor, such as an α, β tracker which also has the advantage of performing a temporal smoothing of pose estimates. In practice however, it has been found difficult to set the tracker parameters as the measurement noise depends on the number and position of edges found, and also on the current pose of the object. In some extreme cases, the edges detected in a particular frame may not define all the object's degrees of freedom; clearly a more sophisticated predictor/filter is required.

4.2.1 *Kalman Filter Outline*

This section repeats the formulation of a standard Kalman filter [19]. A good description of the Kalman filter and associated techniques is given by Bar Shalom [20].

Let $\hat{x}_t$ be a vector that represents the estimated state of a system at time t. Given a new measurement, y_t, made at that same instant, the state vector estimate is updated to $\hat{x}'_t$, given by

$$\hat{x}'_t = \hat{x}_t + K(y_t - H\hat{x}_t),$$

where K is the Kalman gain matrix and H is a matrix which maps the estimated state to the corresponding expected observation. Between observations it is assumed that the true state of the system evolves according to

$$x_{t+1} = Ax_t + \varepsilon_t,$$

where the process noise, ε_t, is a random variable of zero mean and covariance defined by the matrix Q_t. Thus given $\hat{x}'_t$, $\hat{x}_{t+1} = A\,\hat{x}'_t$. If the error in the observation y_t has zero mean and covariance R_t, and the error in $\hat{x}_t$ has zero mean and covariance P_t, then the optimal choice of K (that which minimises the trace of P'_t, the covariance of $\hat{x}'_t$) is

$$K = P_t H^T [HP_t H^T + R_t]^{-1}, \text{ and}$$
$$P'_t = P_t - KHP_t.$$

In the time to the next observation, however, confidence in the state vector estimate worsens because of the uncertainty in evolution, thus

$$P_{t+1} = AP'_t A^T + Q_t.$$

4.2.2 *The Object Motion Model*

In this application of Kalman filtering, the RAPiD pose estimate, y_t, is the 6-vector change in pose found by the minimisation of $E(\mathbf{q})$. In the simplest moving object case we assume uniform motion, so the state vector contains both position and velocity terms. In particular we write,

$$x = (\mathbf{r}, \theta, \dot{\mathbf{r}}, \dot{\theta})^T,$$

where $\mathbf{r}$ is the object's position 3-vector (relative to the camera), and θ is a rotation 3-vector defining its orientation;

$$A = \begin{bmatrix} I_6 & I_6 \\ 0_6 & I_6 \end{bmatrix} \text{ and}$$

$$H = \begin{bmatrix} I_6 & 0_6 \end{bmatrix},$$

where I_6 and 0_6 are the 6-by-6 identity and zero matrices. We assume that the above motion model is accurate apart from a random fluctuation in velocities due to forces acting on the model making it accelerate, so that the state covariance is of the form

$$Q = \begin{bmatrix} 0_6 & 0_6 \\ 0_6 & Q_6 \end{bmatrix}$$

The form of Q_6 will depend on the the dynamics of both the camera and the tracked object and their relative position [7].

4.2.3 *The Measurement Model*

If the object pose is in error by $\mathbf{q}$, then the probability of getting the set of measurements $\{l_i\}$ is

$$P(\{l_i\} \mid \mathbf{q}) \propto \prod_i \exp - \frac{1}{2\sigma^2} [\, l_i + \mathbf{q}.\mathbf{c}_i \,]^2$$

where the measurement accuracies in determining an individual edge position are assumed to be uncorrelated and of size σ. Using Bayes theorem, the probability of the pose being in error by an amount $\mathbf{q}$ is

$$P(\mathbf{q} \mid \{l_i\}) \propto \exp - \frac{1}{2\sigma^2} \sum_i [\, l_i + \mathbf{q}.\mathbf{c}_i \,]^2$$

We can re-write this equation in the usual form of a multivariate normal distribution as follows

$$P(\mathbf{q} \mid \{l_i\}) \propto \exp - \frac{1}{2} [\, \mathbf{q} - \mathbf{q}_0 \,]^T R^{-1} [\, \mathbf{q} - \mathbf{q}_0 \,]$$

where $\mathbf{q}_0$ is the best estimate for the pose error, and the observation error covariance, R, is given by

$$R = \sigma^2 \left[\sum_i \mathbf{c}_i \mathbf{c}_i^T \right]^{-1}$$

Unfortunately, when fewer than 6 control points are detected, the matrix inverse cannot be calculated because of

rank deficiency. This is also true in certain situations when the detected control points do not fully define the pose of the object. The formula defining the Kalman filter gain can be re-arranged, however, to avoid the need to compute the inverse, thus

$$K = PH^T R^{-1}[\, HPH^T R^{-1} + I \,]^{-1}$$

With this formulation for K, the filter gain can be calculated robustly for each filter cycle, weighting each measurement according to its expected accuracy.

5. ILLUSTRATIVE EXAMPLES

The operation of DROID is illustrated in Figures 10 to 13 for the application of DROID to an image sequence recorded in a typical corridor of an office building. Figure 10 shows two consecutive frames of the sequence, which is processed at an image resolution of 256 by 256 pixels over a field of view of about 50 degrees. The distance moved between processed frames in this sequence is about 3-5cm, depending on the speed of the sensor platform.

Superimposed on the grey levels of Figure 10 are the positions of extracted point features; these are the points which are tracked from frame to frame. While a few of these features are not detected in every frame the majority are sufficiently stable to be tracked over several frames. Such persistent features are shown in Figure 11; these are the points at which 3D information is available.

Though range estimates are only generated for the tracked feature points, ranges to other points can be obtained by assuming some model of an interpolating surface. DROID assumes the surface can be described by planar triangular facets, the triangles themselves being drawn by a Delauney triangulation process with results shown in Figure 12. This triangulation method tries to avoid long thin triangles and it is seen to be successful near the centre of the image. Near the boundaries of the described structure, triangles tend to be less good natured and an erroneous depth estimate for a particular feature can have an unwanted effect over a large part of the scene.

Once the triangulation is determined, contours can be drawn on the interpolated surface as in Figure 13. 'Contours' here are drawn 20cm apart down-range and cross-range. (Imagine a net of 20cm squares projected onto the scene from above.) We see that the general structure of the scene has been captured - a flat floor with vertical walls to the left, right and in front. The system does not quite have sufficient resolution, however, to clearly distinguish the presence of the pile of rubbish stacked in the right-hand corner. An interesting feature of these results is the cluster of erroneous feature depths on the door to the left of the framed certificate on the wall. These arise from structure seen in reflections on the shiny door surface! 3D edge processing in a scene such as this would have considerable advantages, with the crisp man-made skirting boards and wall panels.

The operation of RAPiD is illustrated in Figures 14 to 18. These show RAPiD tracking a 'bat' symbol. The scenario is shown in Figure 14, with the camera on a remotely controllable platform, though in this demonstration the target is to be moved relative to a stationary camera. The particular target here is a planar object, which is convenient for laboratory trials, but RAPiD is not limited to this class of target. The definition of the corresponding target model is given in Figure 15. Figure 16 shows two views of the target as seen by the tracking camera, with graphics generated by RAPiD superimposed. These mark selected parts of the target outline and show estimates of the target's position and attitude relative to the camera. Note the outline segments shown are not generated by 2D edge extraction, but are the result of projecting the model, in its estimated pose, onto the image plane. The close alignment, of the modelled target edges with the real ones, indicates the accuracy of the estimated track. (The superimposed outline is difficult to see in monochrome imagery.) The white spots around the bat mark the control points at which RAPiD is searching for edge information.

Figure 17 shows a plot of track parameters for the portion of movement between the above images. Using a planar target and a single image, RAPiD is unable to determine very accurately the direction of the perpendicular to the model surface (pitch and yaw) when the orientation is very near fronto-parallel, but with Kalman filtering, the orientation of the target and its position in camera coordinates are generally stable. RAPiD can be applied to a range of objects, with non-planar models. In such cases the relative accuracy of the different pose components is improved.

In addition to the example illustrated here, DROID has been demonstrated in other domains:

- a hypothetical robot work-cell [18]
- country lane and DRA laboratory grounds [21]
- pot plant foliage! [22]
- laboratory and office scenes [13]
- a circular vehicle test track [23]
- an airfield laid out with parked vehicles, viewed from a low flying helicopter [15]

Similarly RAPiD has a wide range of applicability. See for example Figure 18. Other reported applications include:

- laboratory demonstrations with, a floppy disc box, painted cone, and an egg! [6]
- an airfield runway viewed from a descending aircraft [7]
- airborne object release monitoring, and following a Land Rover along a test track [24].

6. DEVELOPMENT STATUS

DROID has been developed as an off-line process using general purpose hardware. In this form DROID has been applied to a range of domains. The initial development was in the context of a laboratory robot work-cell, but DROID has performed well in other indoor and outdoor contexts, including scenes dominated by natural vegetation, and others structured with human artefacts.

In a software implementation, feature detection is the slowest component in DROID, taking 2 seconds on a Sparc 2 workstation for a 256x256 pixel image, while the

subsequent geometric processing takes 0.2 - 0.3 seconds per frame.

Dedicated video-rate hardware (25Hz) will shortly be available from Roke Manor to perform feature extraction for either 512x512 pixel imagery, or up to 4 camera stereo imagery at 256x256 pixels. (Note that use of 512x512 pixel imagery would indicate the use of a frame-capture camera, since the two fields produced by conventional cameras are captured at 1/50'th second intervals and would be torn apart even by moderate camera motion.) DROID systems, based on this front-end hardware, are currently in development; these are expected to perform overall at near video rate.

Given the modest hardware requirements of RAPiD, development has been based on real-time assessment from the beginning. Near real-time performance was originally achieved with a multi-user VAX 3400! Current development and applications work is generally for the analysis of video recorded trials, such as the analysis of released-store trajectories and the landing path of unmanned aircraft. For convenience of software development, and ancillary facilities, RAPiD has been implemented on workstations supplemented by a video capture/display card. In a dedicated application, a two-card solution is readily feasible.

7. A CRITICAL DISCUSSION

DROID and RAPiD might be considered to lie at opposite ends of the range of computer vision tasks, with DROID extracting the 3D structure of unknown scenes and RAPiD plotting the position of a known object. The two systems have developed in this fashion, but it is possible to imagine a unified DROID-RAPiD system. Instead of fully known models we may imagine partially known models in which either (a) newly observed features - specified by DROID-like processing - are added to an existing model, or (b) known yet approximately specified features of a model are refined. Similarly RAPiD processing of a modelled component in a scene may generate ego-motion estimates for use in instantiating previously unknown features.

Returning to the original focus of attention for this paper, (i.e. the following of a known object through unknown terrain), it would be appropriate to consider some apparent deficiencies with the DROID-RAPiD approach. The greatest limitation would seem to lie at the outset with the feature-based approach. While DROID can be demonstrated to provide measurements with at times surprising accuracy, the concentration on high quality features leads to a sparse representation of the viewed structure; the sparseness can be catastrophic in very bland scenes. This underlines the power of the human brain in using a wide range of depth cues, general scene understanding, shape from shading and the other shape-from-X methods. Work is in progress to enrich DROID's structural representation by the use of edge features which should be beneficial in man-made environments particularly. It seems apparent, however, that DROID should be regarded as a measurement system and some applications may require a further tier of image interpretation to achieve a complex objective.

A second weakness expected in the DROID philosophy lies in DROID's use of structure to derive ego-motion and *vice versa*. This is particularly important in the transition from boot to run-mode processing as errors in structure made at boot may be frozen into the system at an early stage, leading to future errors in ego-motion and subsequent structure errors in future structure. In practical cases, however, this does not appear to be a problem, with initial errors decaying over the first few processed frames of a sequence. The resulting structure may well be erroneous with respect to an initial global coordinate frame, but it seems to be generally accurate with respect to local coordinates.

An observed weakness in DROID has been a long term drift in the estimated ego motion, though short term performance is believed to be generally good. This drift is important if it is required to relate currently viewed structure to features which have long ago left the camera's field of view. (This effect is more pronounced with cameras of a narrow field of view, and when the features of the viewed scene are concentrated in a small range of depths.) A particularly common drift has been observed in the estimated speed of estimated sensor motion, which results in a corresponding drift in the estimated scale of the viewed scene. This speed-scale drift does not apply to the use of DROID in a stereo mode [13, 23], which has a generally stabilising effect, particularly at boot. Drifts in the ego-motion estimates may also be stabilised by use of external odometry; other motion constraints, such as constant forward speed may be appropriate in particular circumstances.

Turning to the use of RAPiD to follow known objects, a major weakness here is the reliance on a specific geometric model. This may not be a problem with cooperating targets, especially as the complexity of the required model is not onerous, though the readiness of new models may limit the system's flexibility. With non-cooperating targets, there is a system requirement to identify the object to be tracked so that the appropriate model can be applied. It is feasible that RAPiD can be extended to include estimation of a small number of model parameters, and perhaps a model might be defined to minimise reliance on variable components, but it remains that RAPiD, as currently formulated, is not applicable to the problem of tracking a freely moving generic object.

8. CONCLUDING SUMMARY

It has been demonstrated that DROID can extract sensor ego motion and scene structure to some accuracy, and RAPiD with suitable models can track known objects to high precision. DROID has been applied successfully in a range of indoor and outdoor scenes, and RAPiD too has been used in a range of applications. Together these systems make a considerable contribution to the task of obstacle avoidance and object following.

This paper has described the basic structure-from-motion algorithms used by DROID to generate a description of scene structure and sensor motion from a mono image sequence. The resulting scene structure is represented by the estimated 3D positions of localised point features.

This paper has also described the basic algorithms of the RAPiD tracker. RAPiD is eminently suited to real-time processing with modest hardware, and real-time processor implementations of DROID are now in development.

In addition to the techniques detailed here, DROID has been extended to stereo operation and use of edge features is being researched. Stereo generally enhances the stability of the system and edges are expected to enrich the available 3D structural representation, though this will be of most utility in man-made environments.

This paper has also mentioned possible weakness in the DROID/RAPiD approach, in particular the sparseness of output in bland scenery and the need for target-specific models. To perform complex tasks, we may need to use these methods as measurement subsystems within a larger processing and interpretation framework. It is clear however that DROID and RAPiD are powerful tools in their own right, as shown by the range of environments in which they have been demonstrated.

9. ACKNOWLEDGEMENTS

The author is greatly indebted to all the members of the Computer Vision team at Roke Manor Research and in particular to Dr Chris Harris who has been responsible for much of the technical material included in this presentation.

The DROID 3D vision system has been developed at Roke Manor Research under a combination of privately funded and jointly funded projects under the UK's Alvey Initiative and the EC ESPRIT programme, the most recent being the VOILA project (ESPRIT P2502). RAPiD is a private development of Roke Manor Research Limited.

Development of particular applications and implementations of DROID and RAPiD, and some extensions to the basic algorithms are being performed under contracts to the UK Defence Research Agency

10. REFERENCES

1. B K P Horn & B G Schunck, *Determining Optical Flow*, Artificial intelligence, 17, pp. 185-203, 1981.

2. M Brady, *Seeds of Perception*, Proceedings of 3rd Alvey Vision Conference, Cambridge, September 1987.

3. Lowe, D. G. *Stabilized Solution for 3D Model Parameters* Proceedings of the First European Conference on Computer Vision, ECCV90, 1990.

4. A J Bray *Tracking objects using image disparities* Proceedings of the fifth Alvey Vision Conference, AVC89, Reading, 1989.

5. Stephens, R. S. *Real-Time 3D Object Tracking* Proceedings of the fifth Alvey Vision Conference, ACV89, Reading, 1989.

6. Harris, C. G. & Stennett, C. *3D Object Tracking at Video Rate - RAPiD* Proceedings of the first British Machine Vision Conference, BMVC90, Oxford 1990.

7. R J Evans *Filtering of Pose Estimates Generated by the RAPiD Tracker in Applications* Proceedings of the first British Machine Vision Conference BMVC90, Oxford, 1990.

8. C G Harris & M J Stephens, *A Combined Corner and Edge Detector*, Proceedings of 4th Alvey Vision Conference, Manchester, August 1988.

9. L Kitchen & A Rosenfeld, *Grey Level Corner Detection*, Pattern Recognition Letters, 1, pp. 95-102, 1982.

10. M J Stephens, *Matching Features from Edge-processed Image Sequences*, Proceedings of 3rd Alvey Vision Conference, Cambridge, September 1987.

11. N Ayache & O D Faugeras, *Building, Registrating and Fusing Noisy Visual Maps*, Proceedings of 1st ICCV, London, June 1987.

12. C G Harris, *Camera Calibration*, submitted to Third European Conference on Computer Vision, ECCV92, Genoa, 1992.

13. E P Sparks & M J Stephens *Integration of Stereo and Motion*, Proceedings of 1st British Machine Vision Converence (BMVC90), Oxford, 1990

14. C G Harris, *Determination of Ego-motion from Matched Points*, Proceedings of 3rd Alvey Vision Conference, Cambridge, September 1987.

15. C G Harris *DROID Analysis of the NASA Helicopter Images* Proceedings of the IEEE Special Workshop on Passive Ranging, 10 October 1991.

16. J Porrill, S B Pollard, T P Pridmore, J B Bowen, J E W Mayhew & J P Frisby,*TINA: A 3D Vision System for Pick and Place*, Proceedings of 3rd Alvey Vision Conference, Cambridge, September 1987.

17. D Marr,*Vision*, W H Freeman & Co., San Francisco, 1982.

18. C G Harris & J M Pike, *3D Positional Integration from Image Sequences*, Proceedings of 3rd Alvey Vision Conference, Cambridge, September 1987.

19. Kalman, R. E. *A new approach to linear filtering and prediction problems* Trans. ASME, J. of Basic Engineering, March 1960.

20. Y Bar Shalom & T E Fortmann *Tracking and Data Association* Academic Press, San Diego, California, 1988.

21. D Charnley & R J Blissett *Surface Reconstruction from Outdoor Image Sequences* Proceedings of the 4th Alvey Vision Conference, Manchester, August 1988.

22. M J Stephens, E P Sparks & R J Blissett *Surface Perception and Localisation using Passive Vision* Proceedings of the 9'th International Conference on Automated Inspection and Product Control, Stuttgart, May 1989.

23. E P Sparks & M J Stephens *Stereo DROID* Proceedings of the IEE Colloquium on Active & Passive Techniques for 3D Vision, London, February 1991, p6/1.

24 Roke Manor Research *Information Sheet: RAPiD Real-Time Video Motion Analysis*, 1991.

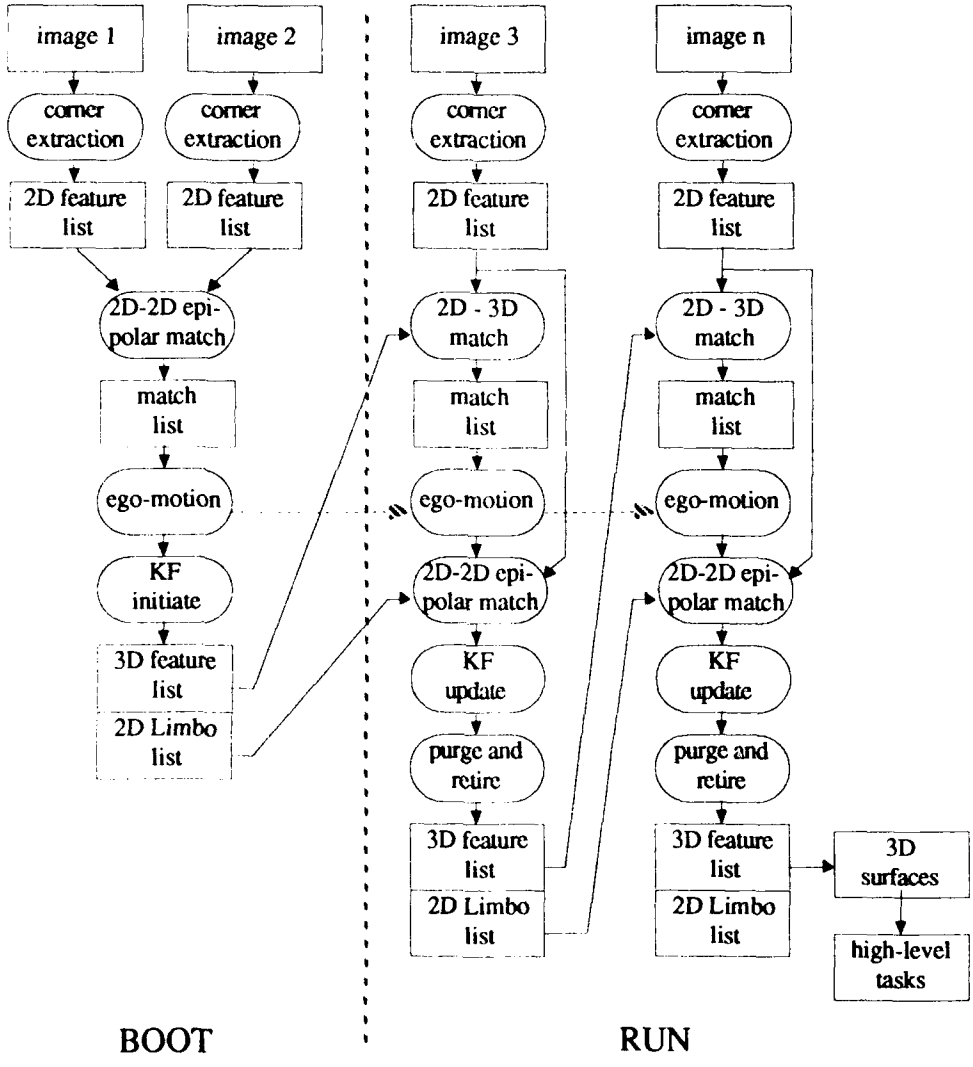


Figure 1. DROID process flowchart.

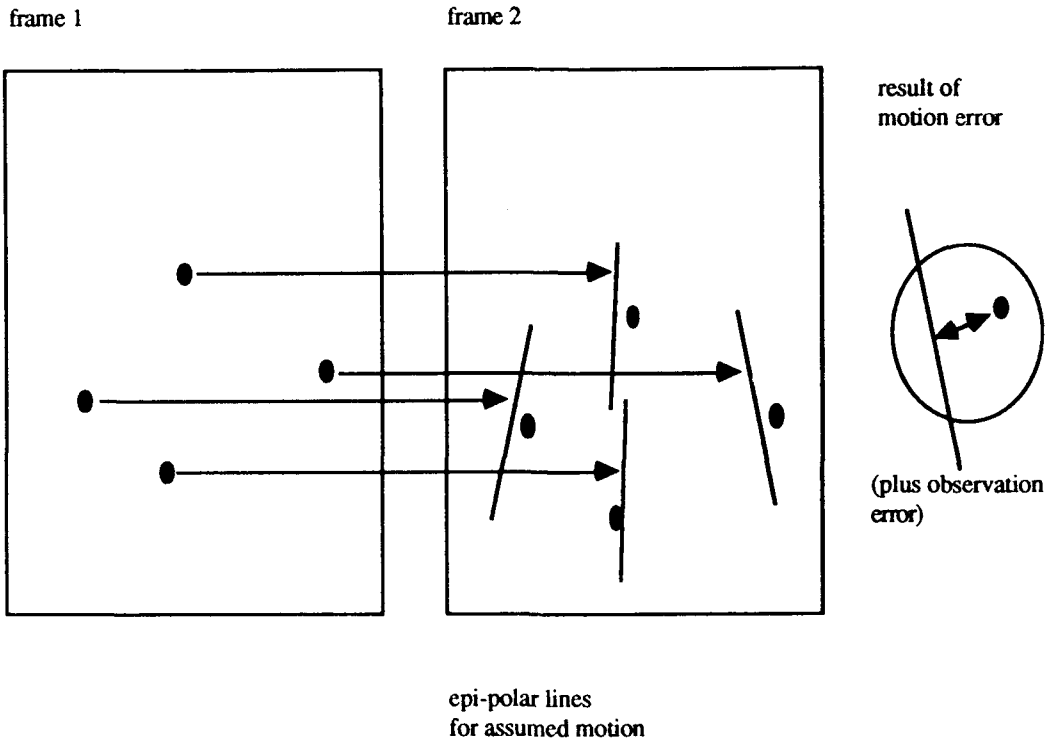


Figure 2. Estimation of DROID ego-motion for boot phase

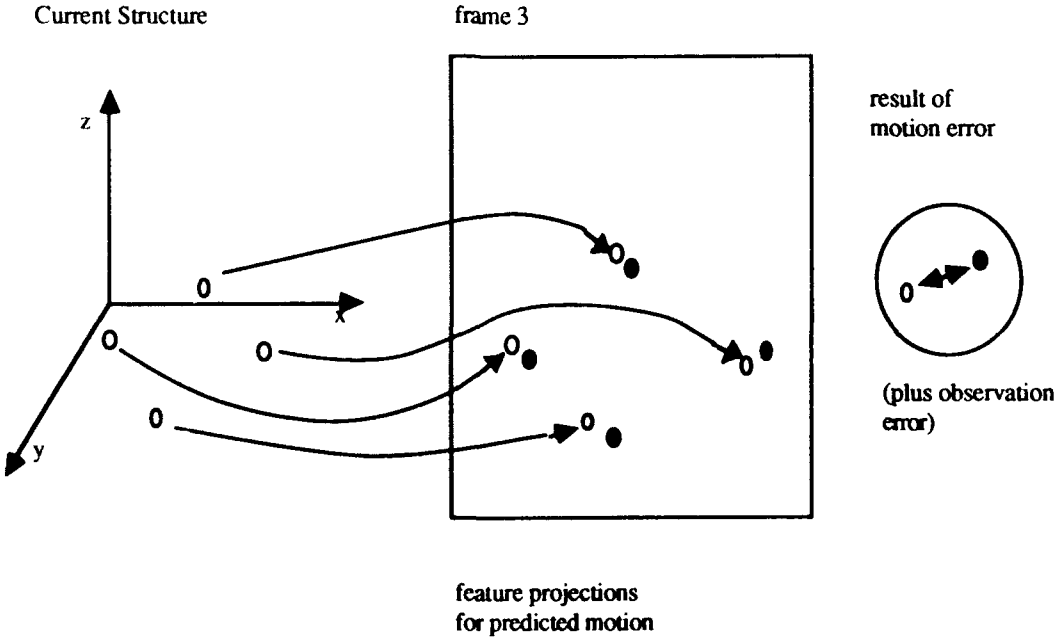


Figure 3. Estimation of DROID ego-motion in run-mode

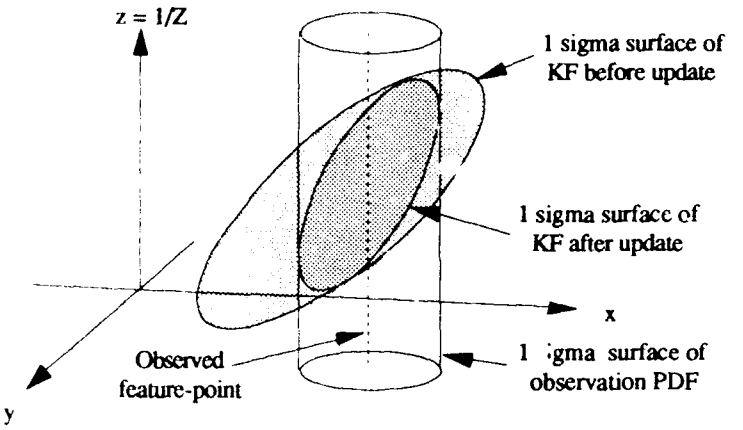


Figure 4. Updating of the Kalman filter in disparity space.

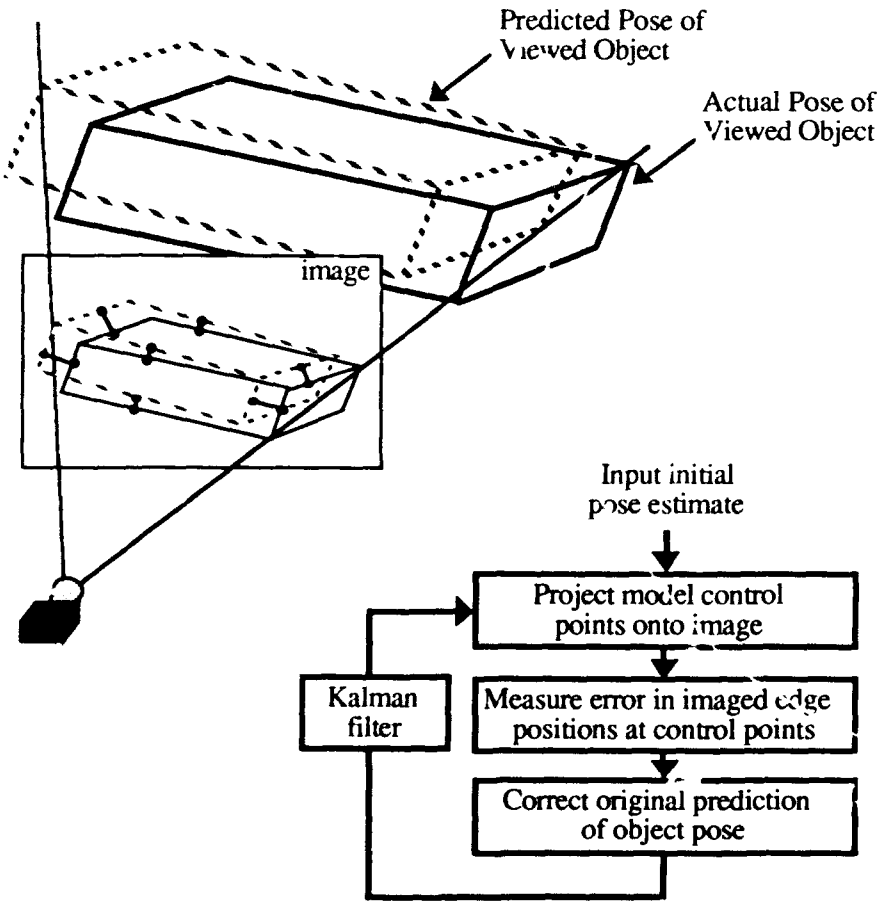


Figure 5. RAPID overview

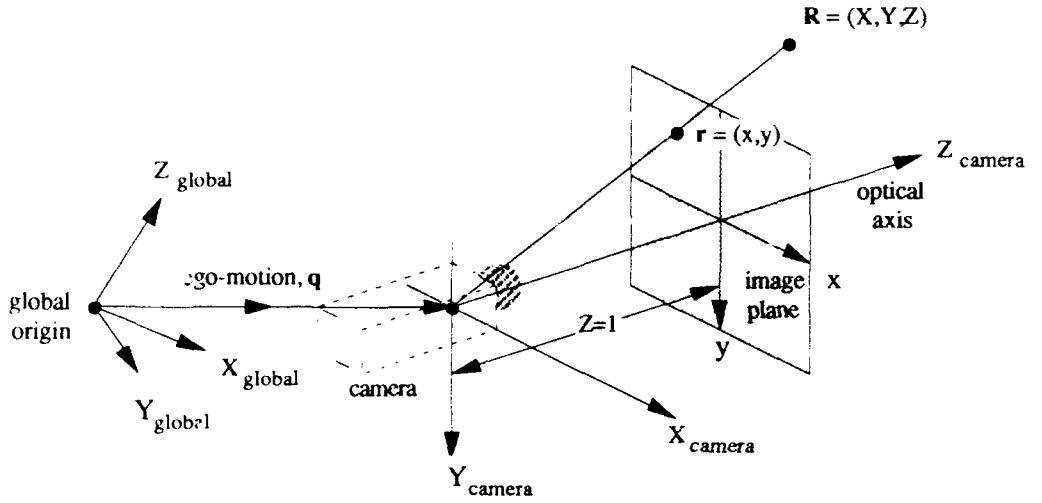


Figure 6. DROID camera coordinates and global coordinates.

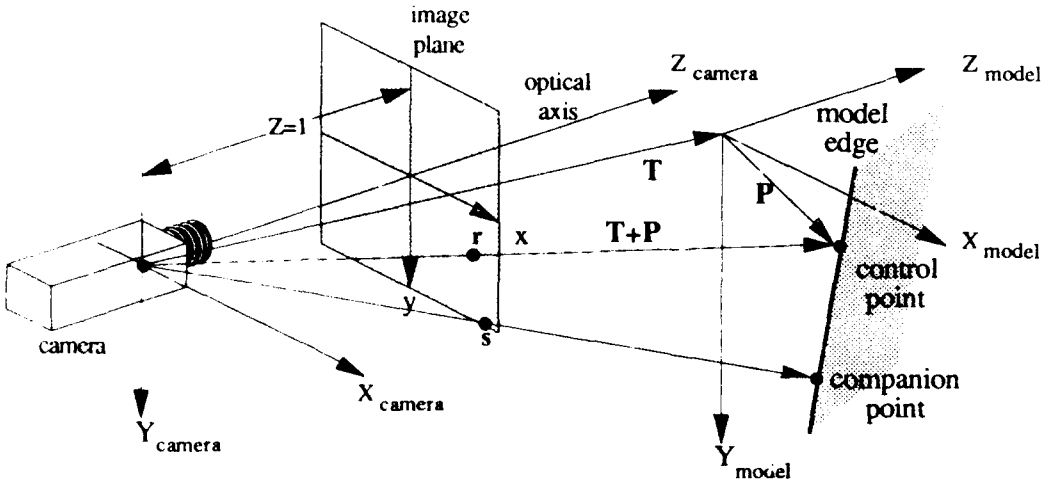


Figure 7. RAPID camera and model coordinate systems.

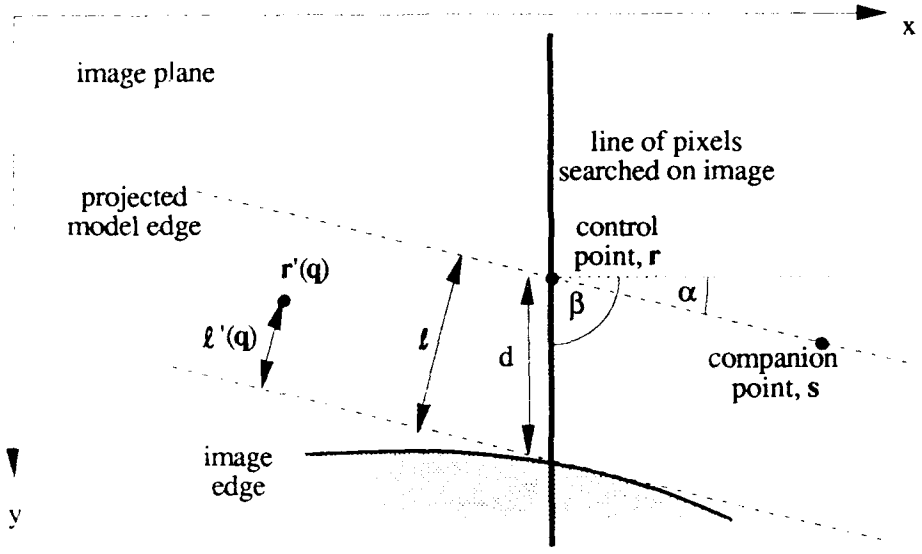


Figure 8. The perpendicular distance, l, of a RAPID control point from its image edge.

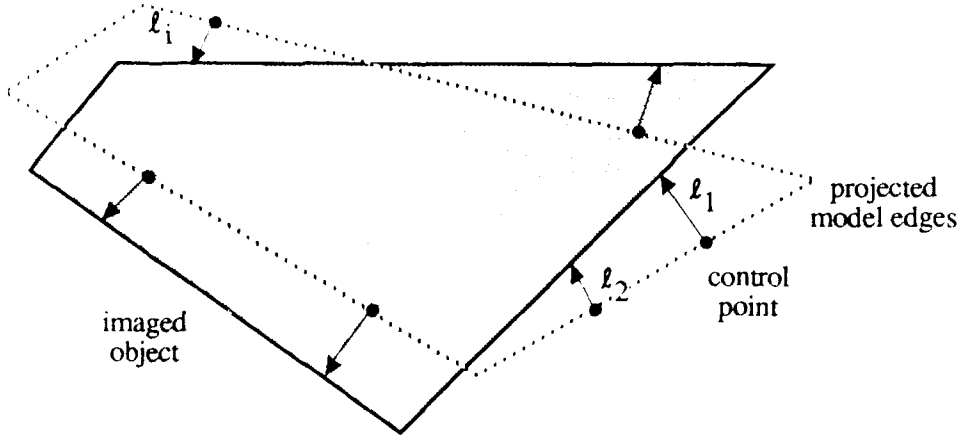


Figure 9. The set of perpendicular distances, $\{l_i\}$, used by RAPID to estimate the model pose.

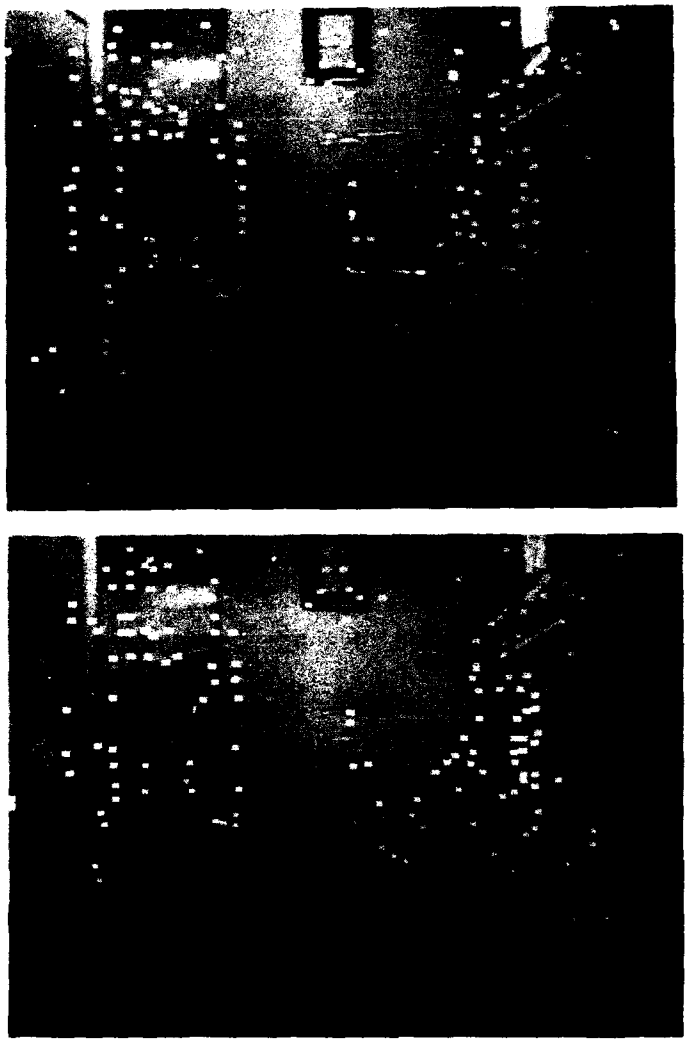


Figure 10. Two consecutive frames from corridor sequence with DROID extracted feature-points marked by white spots .

Figure 11. Reliably tracked DROID feature-points.

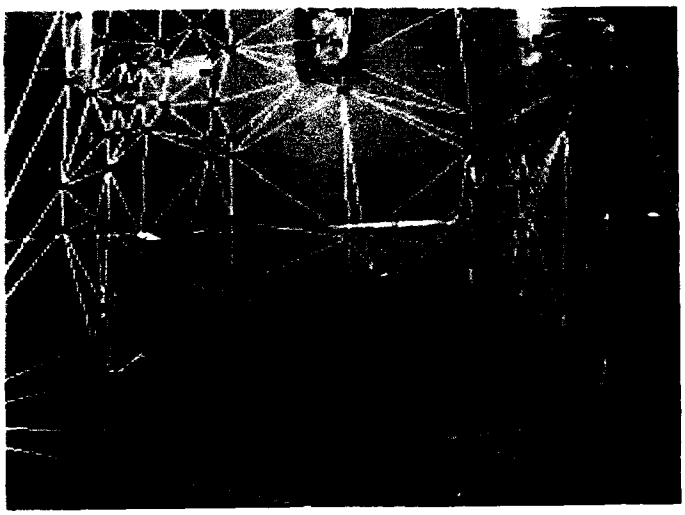


Figure 12. Delauney triangulation of image plane using tracked features.

Figure 13. Contour map of scene derived by interpolation between feature-points using triangulated surface.

Figure 14. General view of RAPID demonstration scenario.

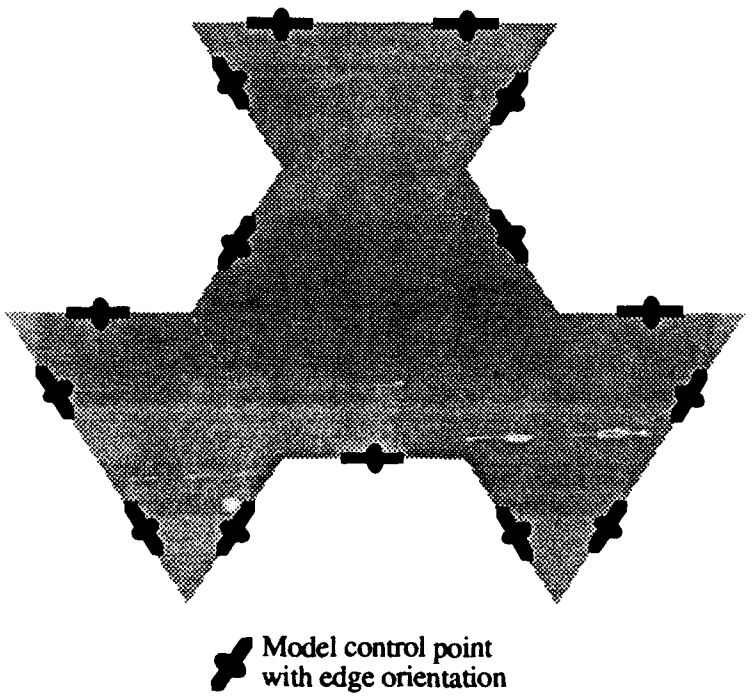


Figure 15. 'Bat' target model used by RAPID.

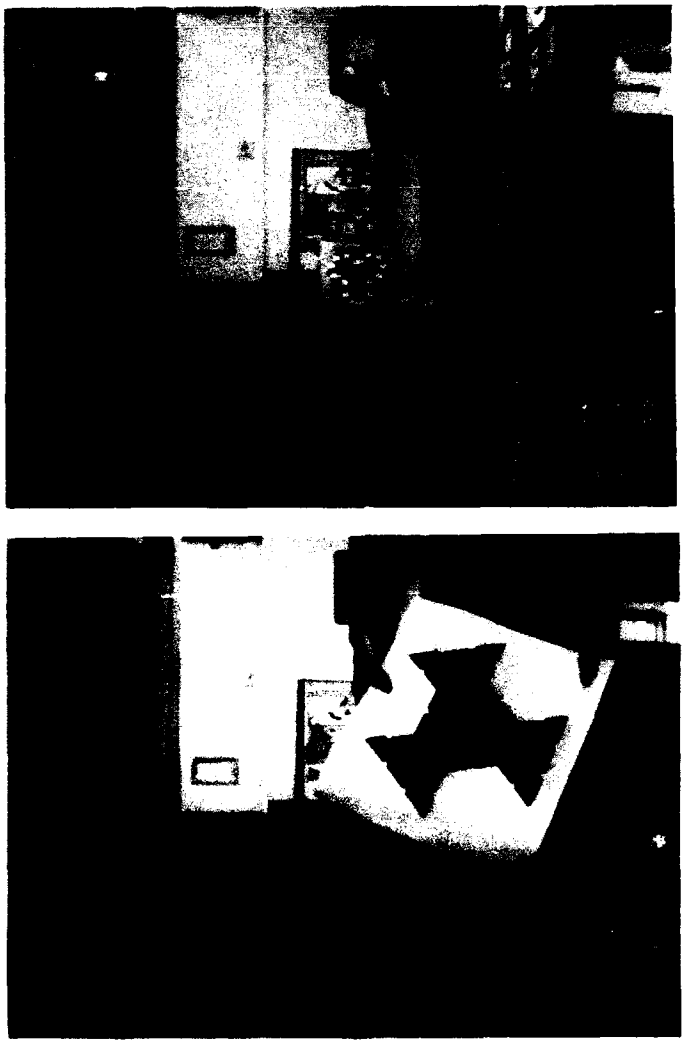


Figure 16. Two images of target as seen by the RAPID camera with target outlines and pose data superimposed.

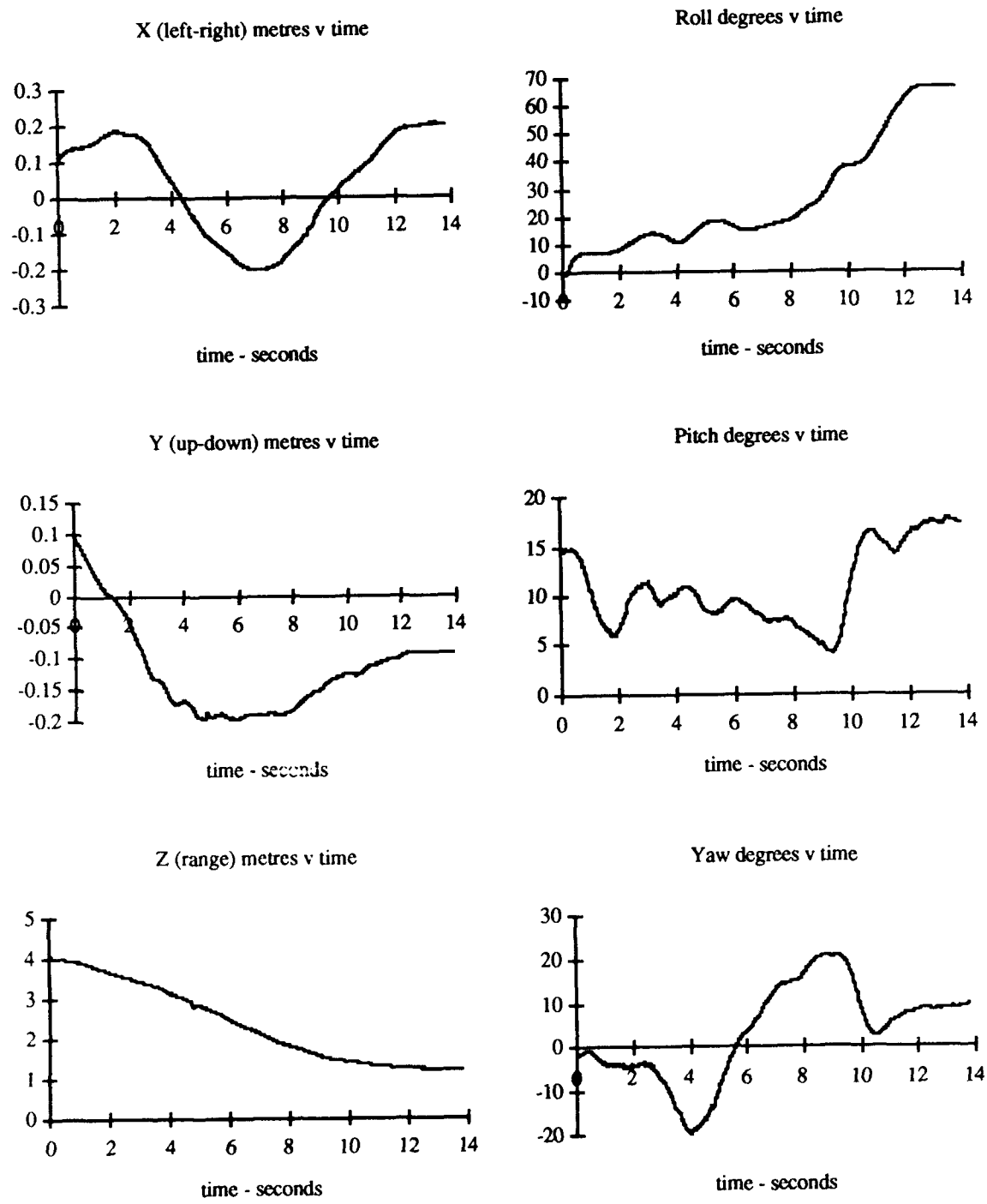


Figure 17. Plots of pose parameters generated by RAPID.

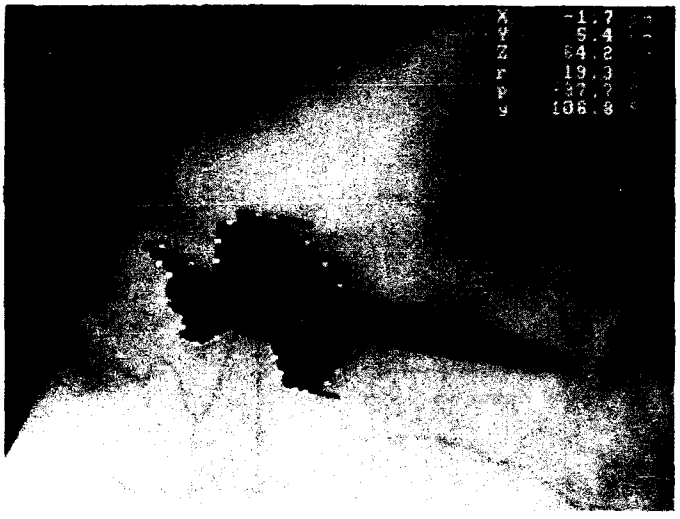


Figure 18. RAPID applied to a model aircraft.

BIBLIOGRAPHY

This bibliography was prepared by the Fachinformationszentrum, Karlsruhe, Germany in support of this Lecture Series in association with the Director of the Lecture Series, Prof. Dr R. Onken.

TYPE 1/4/1
Quest Accession Number : 91A47146
91A47146 NASA IAA Journal Article Issue 20
Teaching network connectivity using simulated annealing on a massively parallel processor
(AA)WILSON, STEPHEN S.
(AA)(Applied Intelligent Systems, Inc., Ann Arbor, MI)
IEEE, Proceedings (ISSN 0018-9219), vol. 79, April 1991, p. 559-566.
910400 p. 8 refs 8 In: EN (English) p.3511

A simulated annealing technique for automatically training a machine vision system to recognize and locate complex objects is described. In this method, the training is used to find an optimum connectivity pattern of a fixed number of inputs that have fixed weights, rather than the usual technique of finding the optimum weights for a fixed connectivity. The recognition model uses a two-layer artificial neural network, where the first layer consists of image edge vectors in four directions. Each neuron in the second layer has a fixed number of connections that connect only to those first layer edges that are best for distinguishing the object from a confusing background. Simulated annealing is used to find the best parameters for defining edges in the first layer, as well as the pattern of connections from the first to the second layer. Weights of the connections are either plus or minus one, so that multiplications are avoided, and the system speed is considerably enhanced. In industrial applications on a low-cost parallel SIMD (single instruction multiple data) architecture, objects can be trained by an unskilled user in less than 1 min, and after training, parts can be located in about 100 ms. This method has been found to work very well on integrated circuit patterns.
I.E.

TYPE 1/4/2
Quest Accession Number : 91A44332
91A44332# NASA IAA Preprint Issue 18
Computer vision of the Martian rover - Hardware/software technique
(AA)SHAMIS, V.; (AB)AVANESOV, G.; (AC)KOGAN, A.; (AD)LANGE, M.; (AE)SHAMANOV, I.
(AE)(AN SSSR, Institut Kosmicheskikh Issledovanii, Moscow, USSR)
AIAA PAPER 88-5012 AIAA and NASA, International Symposium on Space Automation and Robotics, 1st, Arlington, VA, Nov. 29, 30, 1988. 8 p.
881100 p. 8 In: EN (English) p.3073

The present study examines principles of computer vision design for autonomous planetary rovers. Some optional computer vision system (CVS) techniques used to measure environment parameters of the Martian rover are compared, with due account for its diminished payload. Expert estimates of the main design parameters for every feasible option of the rover's CVS are adduced. Attention is given to the CVS optical range finder, stereo system with linear source, stereo system with matrix source (active systems), and stereo system with edge detection, multistereo system, and stereo system with mapped search (passive systems). Consideration is given to CVS detection of obstacles within a viewing angle. The algorithm used to detect local obstacles is described.
P.D.

TYPE 1/4/3
Quest Accession Number : 91A35147
91A35147 NASA IAA Conference Paper Issue 14
Environment learning using a distributed representation
(AA)MATARIC, MAJA J.
(AA)(MIT, Cambridge, MA)
N00014-86-K-0685 IN: 1990 IEEE International Conference on Robotics and Automation, Cincinnati, OH, May 13-18, 1990, Proceedings. Vol. 1 (A91-35126 14-63). Los Alamitos, CA, IEEE Computer Society Press, 1990, p. 402-406. Hughes Aircraft Co.-supported research. 900000 p. 5 refs 15 In: EN (English) p.2354

A method for robust mobile robot navigation and environmental learning is presented. It was implemented and tested on a physical robot. The method consists of a collection of simple, incrementally designed robot behaviors. The behaviors receive sonar and compass data which they use to dynamically detect landmarks and construct a distributed map of the environment. The map is represented as a graph in which each node is a collection of augmented finite state machines functioning a parallel. The distributed nature of the map allows for localization in constant time. The method utilizes a modified spreading of activation scheme to accomplish robust linear-time path planning. It is capable of generating both topologically and physically shortest paths to the goal. The method uses local information to achieve the global task without having to replan if the robot becomes lost or strays off the desired path.
I.E.

TYPE 1/4/4
Quest Accession Number : 91A35146
91A35146 NASA IAA Conference Paper Issue 14
Robot navigation using an anthropomorphic visual sensor
(AA)TISTARELLI, MASSIMO; (AB)SANDINI, GIULIO
(AB)(Genova, Universita, Genoa, Italy)
IN: 1990 IEEE International Conference on Robotics and Automation, Cincinnati, OH, May 13-18, 1990, Proceedings. Vol. 1 (A91-35126 14-63). Los Alamitos, CA, IEEE Computer Society Press, 1990, p. 374-381. Research supported by CNR and NATO. 900000 p. 8 refs 24 In: EN (English) p. 2354

The use of an anthropomorphic, retinalike visual sensor for navigation tasks is investigated. The main advantage, besides the topological scaling and rotation invariance, stems from the considerable data reduction obtained with nonuniform sampling, in conjunction with high resolution in the part of the field of view corresponding to the focus of attention. Active movements are also considered to be a beneficial feature, solving the depth-from-motion problem and maintaining a three-dimensional representation of the viewed scene. For short-range navigation, a tracking egomotion strategy is adopted which greatly simplifies the motion equations and complements the characteristics of the retinal sensor (the displacement is smaller wherever the image resolution is higher). An algorithm for the computation of depth from motion is developed for image sequences acquired with the retinal sensor, and an error analysis is carried out to determine the uncertainty of range measurements. An experiment is presented in which depth maps are computed from a sequence of images sampled with the retinalike sensor, building a volumetric representation of the scene.
I.E.

TYPE 1/4/5
Quest Accession Number : 91A30851
91A30851 NASA IAA Meeting Paper Issue 12
NAECON 90; Proceedings of the IEEE National Aerospace and Electronics Conference, Dayton, OH, May 21-25, 1990. Vols. 1-3
(AA)PALAZZO, FRANK L.
(AA)ED.
(AA)(Questech, Inc., Dayton, OH)
Conference sponsored by IEEE. New York, Institute of Electrical and Electronics Engineers, Inc., 1990, p. Vol. 1, 466 p.; vol. 2, 456 p.; vol. 3, 424 p. For individual items see A91-30852 to A91-31031. 900000 p. 1346 In: EN (English) Price of three vols., members, $70.; nonmembers, $140 p.1899

The present conference discusses advancements in VLSI components/packaging, signal processing, airborne computers, data transmission, advanced avionics architectures. optical applications, data control and display, airborne image processing, target acquisition and recognition, airborne radar and fire control, navigation, weapons guidance and interfaces, Kalman filtering, power generation and control, and command control and communications. Also discussed are flight control reconfiguration, multivariable control theory, flight management, Ada language applications, object-oriented Ada simulations, software management and quality assurance, visual system software, voice-interaction applications, human/machine interfaces, pilot acceleration protection, electronic combat analysis, modular avionics, expert systems, machine vision/optical image processing, adaptive networks, logistics readiness, automated testing, and total quality management.
O.C.

TYPE 1/4/6
Quest Accession Number : 91N30843
91N30843# NASA STAR Thesis Issue 22
Application of Gestalt theory concepts for image interpretation for robot movement navigation / M.S. Thesis - 14 Feb. 1990
UMA APLICACAO DE CONCEITOS DA TEORIA DE GESTALT NA INTERPRETACAO DE IMAGENS PARA A NAVEGACAO DE ROBOS MOVEIS
(AA)ODASHIMA, EUNICE KINUYO
Instituto de Pesquisas Espaciais, Sao Jose dos Campos (Brazil). (IO601891)
INPE-5225-TDL/438 910300 p. 144 In PORTUGUESE; ENGLISH summary In: AA (Mixed) Avail: NTIS HC/MF A07 p.3741

Research involved the development of machine vision for a vehicle capable of moving from one place to another while employing collision avoidance capabilities. The specific objective of the study was the use of image segmentation of the interior space and the obstacles therein to construct a cognitive map of the robot's movements. The paradigm is based on Gestalt psychology and geometry.
Author

TYPE 1/4/7
Quest Accession Number : 91N29801
91N29801# NASA STAR Conference Proceedings Issue 21
Workshop on Automation and Robotics: Proceedings
Lawrence Livermore National Lab., CA. (LH075075)
DE91-015175; CONF-910274 W-7405-ENG-48 910200 p. 243 Workshop held in Livermore, CA, 6 Feb. 1991 In: EN (English) Avail: NTIS HC/MF A11 p.3562

This workshop provided a forum in which Lawrence Livermore National Laboratory scientists and engineers exchanged ideas and information on the latest internal developments in the field of robotic and automation technologies. The material presented constitutes most of the presentations given during the workshop. Presentations were given on the following session topics: robotics and automation in hazardous environments; laboratory and machine tool automation; neural networks, machine vision, and sensors; applied real time control; future technologies and applications; intelligent man-machine interaction issues. Individual papers have been cataloged separately.
DOE

TYPE 1/4/8
Quest Accession Number : 91A29762
91A29762# NASA IAA Journal Article Issue 11
Star pattern identification aboard an inertially stabilized aircraft
(AA)KOSIK, JEAN CLAUDE
(AA)(CNES, Toulouse, France)
Journal of Guidance, Control, and Dynamics (ISSN 0731-5090), vol. 14, Mar.-Apr. 1991, p. 230-235. 910400 p. 6 refs 6 In: EN (English) p. 1713

Comparative statistical analyses are conducted for several star-identification algorithms applicable to inertially stabilized spacecraft: polygon-matching, the pole technique, polygon angular-matching, and orientation-angle-magnitude. While the pole technique was both the most complex and least efficient, so that the polygon-match algorithm was superior even without any a priori information on attitude, the possession of crude attitude data allowed the polygon angular-matching algorithm to yield the best results; its code was nearly as simple as that for the polygon-match, and its efficiency was shown by the present probabilistic approach to be greatly improved over the alternatives.
O.C.

TYPE 1/4/9
Quest Accession Number : 91A28855
91A28855 NASA IAA Journal Article Issue 11
Background characterization techniques for target detection using scene metrics and pattern recognition
(AA)NOAH, PAUL V.; (AB)NOAH, MEG A.; (AC)SCHROEDER, JOHN; (AD)CHERNICK, JULIAN
(AC)(Ontar Corp., Brookline, MA); (AD)(U.S. Army, Material Systems Analysis Activity, Aberdeen Proving Ground, MD)
DAAA15-88-C-0021 Optical Engineering (ISSN 0091-3286), vol. 30, Jan. 1991, p. 254-258. 910100 p. 5 refs 11 In: EN (English) p.1827

Autonomous homing munitions (AHM) using infrared, visible, millimeter wave and other sensors have been investigated in order to develop ground target detection and identificaton systems in a clutter enviroment. Pattern recognition and artificial intelligence techniques combined with multisensor data fusion have been used to evaluate a set of image metrics applied to infrared terrain clutter scenes. The application of discriminant function analysis to target detection and identification is demonstrated.
O.G.

TYPE 1/4/10
Quest Accession Number : 91N27411
91N27411# NASA STAR Technical Report Issue 19
The effects of user's training on the performance of an automatic speech recognizer for a self-paced task / Final Report
(AA)SMYTH, CHRISTOPHER C.
Human Engineering Labs., Aberdeen Proving Ground, MD. (H6521544)
AD-A235844; HEL-TM-10-91 DA PROJ. 1L1-62716-AH-70 910400 p. 84 In: EN (English) Avail: NTIS HC/MF A05 p.3150

The results of a recent experiment concerning the effects of training on the performance of subjects using the automatic speech recognizer are reported. Over a 5-day period, 20 military enlisted grade male subjects were trained and tested in using a connected speech (speaker-dependent) machine automatic speech recognizer in a self-paced task controlling a generic tactical display by voice command. Experimental results show that a majority of the subjects had little difficulty with the automatic speech recognizer and that for these subjects training produced only a slight improvement in recognizer performance. These subjects performed at a high machine recognition rate. However, during the first session, a large minority (35 percent) of the subjects had difficulty training their speech to be machine recognizable. These subjects required at least two training sessions to perform the task at their best ability, and even after they were trained, their performance never reached the performance level of other subjects.
GRA

TYPE 1/4/11
Quest Accession Number : 91N26815
91N26815# NASA STAR Technical Report Issue 18
Northeast Artificial Intelligence Consortium (NAIC). Volume 1: Executive summary / Final Report, Sep. 1984 - Dec. 1989
(AA)WEISS, VOLKER; (AB)BRULE, JAMES F.
Northeast Artificial Intelligence Consortium, Syracuse, NY. (N4144152)
AD-A234880; RADC-TR-90-404-VOL-1 F30602-85-C-0008 901200 p. 71 In: EN (English) Avail: NTIS HC/MF A04 p.3045

The Northeast Artificial Intelligence Consortium (NAIC) was created by the Air Force Systems Command, Rome Air Development Center, and the Office of Scientific Research. Its purpose was to conduct pertinent research in artificial intelligence and to perform activities ancillary to this research. This report describes progress during the existence of the NAIC on the technical research tasks undertaken at the member universities. The topics covered in general are: (1) versatile expert system for equipment maintenance; (2) distributed AI for communications systems control; (3) automatic photointerpretation; (4) time-oriented problem solving; (5) speech understanding systems; (6) knowledge-base reasoning and planning; and (7) a knowledge acquisition, assistance, and explanation system. This volume provides the executive summary of the NAIC.
GRA

TYPE 1/4/12
Quest Accession Number : 91N26792
91N26792# NASA STAR Technical Report Issue 18
Using genetic algorithms to select and create features for pattern classification
(AA)CHANG, E. I.; (AB)LIPPMANN, RICHARD P.
Massachusetts Inst. of Tech., Lexington. (MJ728827) Lincoln Lab.
AD-A235165; TR-892; ESD-TR-90-144 F19628-90-C-0002 910311 p. 90 In: EN (English) Avail: NTIS HC/MF A05 p.3042

Genetic algorithms were used to select and create features and to select reference exemplar patterns for machine vision and speech pattern classification tasks. On a 15-feature machine-vision inspection task, it was found that genetic algorithms performed no better than conventional approaches to feature selection but required much more computation. For a speech recognition task, genetic algorithms required no more computation time than traditional approaches but reduced the number of features required by a factor of five (from 153 to 33 features). On a difficult artificial machine-vision task, genetic algorithms were able to create new features (polynomial functions of the original features) that reduced classification error rates from 10 to almost 0 percent. Neural net and nearest-neighbor classifiers were unable to provide such low error rates using only the original features. Genetic algorithms were also used to reduce the number of reference exemplar patterns and to select the value of k for a k-nearest-neighbor classifier. On a 338 training pattern vowel recognition problem with 10 classes, genetic algorithms simultaneously reduced the number of stored exemplars from 338 to 63 and selected k without significantly decreasing classification accuracy. In all applications, genetic algorithms were easy to apply and found good solutions in many fewer trials than would be required by an exhaustive search. Run times were long but not unreasonable. These results suggest that genetic algorithms may soon be practical for pattern classification problems as faster serial and parallel computers are developed.
GRA

TYPE 1/4/13
Quest Accession Number : 91A26612
91A26612* NASA IAA Conference Paper Issue 10
Kalman filter based range estimation for autonomous navigation using imaging sensors
(AA)SRIDHAR, B.; (AB)CHENG, V. H. L.; (AC)PHATAK, A. V.
(AB)(NASA, Ames Research Center, Moffett Field, CA); (AC)(Analytical Mechanics Associates, Mountain View, CA)
National Aeronautics and Space Administration. Ames Research Center, Moffett Field, CA. (NC473657)
IN: Automatic control in aerospace; IFAC Symposium, Tsukuba, Japan, July 17-21, 1989, Selected Papers (A91-26606 10-12). Oxford, England and New York, Pergamon Press, 1990, p. 45-50. 900000 p. 6 refs 12 In: EN (English) p.1553

The ability to detect and locate obstacles using on-board sensors and modify the nominal trajectory is necessary for safe landing of an autonomous lander on Mars. This paper examines some of the issues in the location of objects using a sequence of images from a passive sensor, and describes a Kalman filter approach to improve the range estimation to obstacles. The filter is also used to track features in the images leading to a significant reduction of search effort in the feature extraction step of the algorithm. The lack of suitable flight imagery data presents a problem in the verification of concepts for obstacle detection. An experiment is designed to acquire a sequence of images along with sensor motion parameters and the range estimation results using this imagery are presented.
Author

TYPE 1/4/14
Quest Accession Number : 91À?6349
91A26349 NASA IAA Book/Monograph Issue 09
Intelligent robotics (Book)
(AA)LEE, MARK H.
(AA)(University College of Wales, Aberystwyth)
Research supported by University of Auckland and SERC. New York/Milton Keynes, England, John Wiley & Sons/Open University Press, 1989, 224 p. 890000 p. 224 refs 55 In: EN (English) $61.95 p.1454

The fundamental principles of intelligent-robot design and application are discussed in a general introduction for engineering students and practicing engineers. Chapters are devoted to the current status of robotics technology, sensor technology, artificial sight, the problem of perception, building a knowledge base, and machinery for thinking about actions. Also considered are the emulation of an expert; errors, failures and disasters; a robotic assembly system; and proposals for a science of physical manipulation. Extensive diagrams, drawings, and graphs are provided.
T.K.

TYPE 1/4/15
Quest Accession Number : 91N24046
91N24046*# NASA STAR Conference Paper Issue 15
Intelligent vision system for autonomous vehicle operations
(AA)SCHOLL, MARIJA S.
Jet Propulsion Lab., California Inst. of Tech., Pasadena. (JJ574450)
In NASA, Washington, Technology 2000, Volume 2 p 34-43 (SEE N91-24041 15-99) 910000 p. 10 In: EN (English) Avail: NTIS HC/MF A16 p.2536

A complex optical system consisting of a 4f optical correlator with programmatic filters under the control of a digital on-board computer that operates at video rates for filter generation, storage, and management is described.
Author

TYPE 1/4/16
Quest Accession Number : 91N23766
91N23766# NASA STAR Technical Report Issue 15
Synergetic multisensor fusion / Final Report, 1 Jul. 1987 - 30 Sep. 1990
(AA)AGGARWAL, J. K.
Texas Univ., Austin. (TT636128) Computer and Vision Research Center.
AD-A232089; ARO-25021.5-PH DAAL03-87-K-0089 901130 p. 60 In: EN (English) Avail: NTIS HC/MF A04 p.2486

Synergetic multisensor fusion is the process of integrating information obtained from different sensing modalities in order to extract additional information that cannot be obtained by separately processing the signals from the different sensors. The development of a computer vision system using synergetic multisensor fusion is a complex task which encompasses: sensor modeling; environment modeling; determining the analytic models used to interrelate the different sensing mechanisms; determining the models used to interrelate the sensed parameters of imaged objects (such as thermal emissivity, visual reflectance, and radar reflectance); and devising algorithms to exploit the derived models. We have developed powerful and robust algorithms for computer vision tasks based upon synergetic multisensor fusion. Our approach is suitable for applications such as object recognition, tracking, surveillance, and autonomous guidance.
GRA

TYPE 1/4/17
Quest Accession Number : 91A23123
91A23123 NASA IAA Journal Article Issue 08
The DARPA Image Understanding Benchmark for parallel computers
(AA)WEEMS, CHARLES; (AB)RISEMAN, EDWARD; (AC)HANSON, ALLEN; (AD)ROSENFELD, AZRIEL
(AC)(Massachusetts, University, Amherst); (AD)(Maryland, University, College Park)
DACA76-86-C-0015 Journal of Parallel and Distributed Computing (ISSN 0743-7315), vol. 11, Jan. 1991, p. 1-24. Research supported by DARPA. 910100 p. 24 refs 15 In: EN (English) p.1258

DARPA has undertaken an evaluation of parallel architectures applicable to knowledge-based machine vision, with a view to the formulation of a benchmark capable of addressing the issue of system performance on an integrated set of tasks. This Integrated Image Understanding Benchmark encompasses a model-based object-recognition problem, two sources of sensor-input and intensity and range data, and a data base of candidate models consisting of rectangular surface configurations in orthographic projection in the presence of both noise and spurious nonmodel surfaces. The benchmark can be used to gain insight into processor strengths and weaknesses, thereby guiding the development of next-generation parallel-vision architectures.
O.C.

TYPE 1/4/18
Quest Accession Number : 91N22769
91N22769*# NASA STAR Conference Proceedings Issue 14
The 1991 Goddard Conference on Space Applications of Artificial Intelligence
(AA)RASH, JAMES L.
(AA)ed.
National Aeronautics and Space Administration. Goddard Space Flight Center, Greenbelt, MD. (NC999967)
NASA-CP-3110; REPT-91B00064; NAS 1.55:3110 Washington 910500 p. 361 Conference held in Greenbelt, MD, 13-15 May 1991 In: EN (English) Avail: NTIS HC/MF A16 p.2312

The purpose of this annual conference is to provide a forum in which current research and development directed at space applications of artificial intelligence can be presented and discussed. The papers in this proceeding fall into the following areas: Planning and scheduling, fault monitoring/diagnosis/recovery, machine vision, robotics, system development, information management, knowledge acquisition and representation, distributed systems, tools, neural networks, and miscellaneous applications. For individual titles, see N91-22770 through N91-22797.

TYPE 1/4/19
Quest Accession Number : 91A20480
91A20480 NASA IAA Meeting Paper Issue 06
Intelligent robots and computer vision VIII: Systems and applications; Proceedings of the Meeting, Philadelphia, PA, Nov. 9, 10, 1989
(AA)BATCHELOR, BRUCE G.
(AA)ED.
(AA)(Cardiff, University College, Wales)
SPIE-1193 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 1193), 1990, 356 p. For individual items see A91-20481 to A91-20484. 900000 p. 356 In: EN (English) Members, $51.; nonmembers, $64 p.918

Recent advances in robot optical sensors and their applications are discussed in reviews and reports. Sections are devoted to planning schemes, intelligent robots, industrial robots, and sensors and processing. Particular attention is given to planning based on multisensor input, an object-oriented approach to simulation of perception and navigation for mobile robots, fast visual foothold finding for an autonomous bipedal robot, hierarchical modeling of mobile seeing robots, a robot tactile sensor for peghole assembling, incorporating ultrasound into robot vision, the use of projection to extract a range map, the tracking of partially occluded two-dimensional shapes, and corner detection from thinned-edge images using a Kalman filter.
T.K.

TYPE 1/4/20
Quest Accession Number : 91A20226
91A20226 NASA IAA Meeting Paper Issue 06
Mobile robots IV; Proceedings of the Meeting, Philadelphia, PA, Nov. 6, 7, 1989
(AA)WOLFE, WILLIAM J.; (AB)CHUN, WENDELL H.
(AA)ED.; (AB)ED.
(AA)(Colorado, University, Denver); (AB)(Martin Marietta Space Systems Co., Denver, CO)
SPIE-1195 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 1195), 1990, 420 p. For individual items see A91-20227 to A91-20231. 900000 p. 420 In: EN (English) Members, $45.; nonmembers, $56 p.918

The present conference on mobile robot systems discusses high-speed machine perception based on passive sensing, wide-angle optical ranging, three-dimensional path planning for flying/crawling robots, navigation of autonomous mobile intelligence in an unstructured natural environment, mechanical models for the locomotion of a four-articulated-track robot, a rule-based command language for a semiautonomous Mars rover, and a computer model of the structured light vision system for a Mars rover. Also discussed are optical flow and three-dimensional information for navigation, feature-based reasoning trail detection, a symbolic neural-net production system for obstacle avoidance and navigation, intelligent path planning for robot navigation in an unknown environment, behaviors from a hierarchical control system, stereoscopic TV systems, the REACT language for autonomous robots, and a man-amplifying exoskeleton.
O.C.

TYPE 1/4/21
Quest Accession Number : 91A19827
91A19827 NASA IAA Journal Article Issue 06
Estimating 3-D egomotion from perspective image sequences
(AA)BURGER, WILHELM; (AB)BHANU, BIR
(AA)(Linz, Universitaet, Austria); (AB)(Honeywell Systems and Research Center, Minneapolis, MN)
DACA76-86-C-0017 IEEE Transactions on Pattern Analysis and Machine Intelligence (ISSN 0162-8828), vol. 12, Nov. 1990, p. 1040-1058. Research supported by DARPA. 901100 p. 19 refs 33 In: EN (English) p.916

Computing sensor motion from sets of displacement vectors obtained from consecutive pairs of images is discussed. The problem is investigated with emphasis on its application to autonomous robots and land vehicles. The effects of 3-D camera rotation and translation upon the observed image are discussed, particularly the concept of the focus of expansion (FOE). It is shown that locating the FOE precisely is difficult when displacement vectors are corrupted by noise and errors. A more robust performance can be achieved by computing a 2-D region of possible FOE locations (termed the fuzzy FOE) instead of looking for a single-point FOE. The shape of this FOE region is an explicit indicator of the accuracy of the result. It has been shown elsewhere that given the fuzzy FOE, a number of powerful inferences about the 3-D sense structure and motion become possible. The aspects of computing the fuzzy FOE are presently emphasized, and the performance of a particular algorithm on real motion sequences taken from a moving autonomous land vehicle is shown.
I.E.

TYPE 1/4/22
Quest Accession Number : 91A19501
91A19501 NASA IAA Meeting Paper Issue 06
Intelligent robots and computer vision VIII: Algorithms and techniques; Proceedings of the Meeting, Philadelphia, PA, Nov. 6-10, 1989. Parts 1 & 2
(AA)CASASENT, DAVID P.
(AA)ED.
(AA)(Carnegie-Mellon University, Pittsburgh, PA)
SPIE-1192 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 1192), 1990, p. Pt. 1, 512 p.; pt. 2, 382 p. For individual items see A91-19502 to A91-19509. 900000 p. 894 In: EN (English) Price of two parts, members, $73.; nonmembers, $91 p.928

Theoretical and practical aspects of computer-vision systems for robotics applications are discussed in reviews and reports. Sections are devoted to pattern recognition for intelligent robots and computer vision; segmentation, image processing, and feature extraction; three-dimensional shape determination and representation; color and range image processing; and neural networks and associative processors for advanced vision processing. Also considered are the biological basis for machine vision, fuzzy logic in intelligent systems and computer vision, image understanding and analysis, time-sequential image processing, and polar exponential grid processing for synthetic vision systems. Extensive diagrams, graphs, and sample images are provided.
T.K.

TYPE 1/4/23
Quest Accession Number : 91A16419
91A16419 NASA IAA Meeting Paper Issue 04
Optics, illumination, and image sensing for machine vision IV; Proceedings of the Meeting, Philadelphia, PA, Nov. 8-10, 1989
(AA)SVETKOFF, DONALD J.
(AA)ED.
(AA)(Synthetic Vision Systems, Inc., Ann Arbor, MI)
SPIE-1194 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 1194), 1990, 317 p. No individual items are abstracted in this volume. 900000 p. 317 In: EN (English) Members, $45.; nonmembers, $56 p.514

Various papers on optics, illumination, and image sensing for machine vision are presented. Individual topics addressed include: extraction of the 'time to contact' from real visual data, position-decoupled optical inspection relay system, TDI imaging in industrial inspection, time delay and integration camera for machine vision, special scanning modes in CCD cameras, scale-invariant processing multiple wavelengths, incoherent optical correlators, light-source models for machine vision, design and testing of a microscopic reflectometer, prediction scheme for a verification vision system, accurate calibration technique for 3-D laser strip sensors, triangulation-based camera calibration for machine-vision system. Also discussed are: 3-D gradient and curvature measurement using local image information, depth from defocus of structured light, range sensing by projecting multiple slits with random cuts, use of linear arrays in electronic speckle pattern interferometry, new 3-D vision sensor for shape-measurement applications, 3-D imager with wide area and high dynamic range, integration of stereo camera geometries, surface orientation from two-camera stereo with polarizers, application-oriented overview of stereoscopic vision.
C.D.

TYPE 1/4/24
Quest Accession Number : 91N13941
91N13941*# NASA STAR Technical Report Issue 05
A discrepancy within primate spatial vision and its bearing on the definition of edge detection processes in machine vision
(AA)JOBSON, DANIEL J.
National Aeronautics and Space Administration. Langley Research Center, Hampton, VA. (ND210491)
NASA-TM-102739; NAS 1.15:102739 307-51-10 900900 p. 31 In: EN (English) Avail: NTIS HC/MF A03 p.707

The visual perception of form information is considered to be based on the functioning of simple and complex neurons in the primate striate cortex. However, a review of the physiological data on these brain cells cannot be harmonized with either the perceptual spatial frequency performance of primates or the performance which is necessary for form perception in humans. This discrepancy together with recent interest in cortical-like and perceptual-like processing in image coding and machine vision prompted a series of image processing experiments intended to provide some definition of the selection of image operators. The experiments were aimed at determining operators which could be used to detect edges in a computational manner consistent with the visual perception of structure in images. Fundamental issues were the selection of size (peak spatial frequency) and circular versus oriented operators (or some combination). In a previous study, circular difference-of-Gaussian (DOG) operators, with peak spatial frequency responses at about 11 and 33 cyc/deg were found to capture the primary structural information in images. Here larger scale circular DOG operators were explored and led to severe loss of image structure and introduced spatial dislocations (due to blur) in structure which is not consistent with visual perception. Orientation sensitive operators (akin to one class of simple cortical neurons) introduced ambiguities of edge extent regardless of the scale of the operator. For machine vision schemes which are functionally similar to natural vision form perception, two circularly symmetric very high spatial frequency channels appear to be necessary and sufficient for a wide range of natural images. Such a machine vision scheme is most similar to the physiological performance of the primate lateral geniculate nucleus rather than the striate cortex.
Author

TYPE 1/4/25
Quest Accession Number : 90A37407
90A37407 NASA IAA Conference Paper Issue 16
Background characterization techniques for pattern recognition applications
(AA)NOAH, MEG A.; (AB)NOAH, PAUL V.; (AC)SCHROEDER, JOHN; (AD)KESSLER, B. V.; (AE)CHERNICK, JULIAN
(AC)(Ontar Corp., Brookline, MA); (AD)(U.S. Navy, Naval Surface Warfare Center, White Oak, MD); (AE)(U.S. Army, Army Material Systems Analysis Activity, Aberdeen Proving Ground, MD)
N60921-87-C-0044; DAAA15-88-C-0021 IN: Aerospace pattern recognition; Proceedings of the Meeting, Orlando, FL, Mar. 30, 31, 1989 (A90-37401 16-63). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 55-70. 890000 p. 16 refs 14 In: EN (English) p. 2594

The development of such sensor hardware as that of large IR and mm-wave detector arrays for air and ground vehicle detection in a cluttered battlefield environment has outpaced the development of signal processing techniques. Attention is presently given to a novel methodology for background clutter characterization, target detection, and target identification, employing multivariate statistical analysis to evaluate a set of image metrics applied to IR cloud imagery and terrain clutter scenes. This methodology is here applied to (1) the characterization of atmospheric water vapor cloud scenes for the U.S. Navy's IR Search and Track system, and (2) the detection of ground vehicles for the U.S. Army's Autonomous Homing Munitions problem.
O.C.

TYPE 1/4/26
Quest Accession Number : 90A32156
90A32156 NASA IAA Conference Paper Issue 13
An update on strategic computing computer vision - Taking image understanding to the next plateau
(AA)SIMPSON, ROBERT L., JR.
(AA)(DARPA, Information Science and Technology Office, Arlington, VA)
IN: Image understanding and the man-machine interface II; Proceedings of the Meeting, Los Angeles, CA, Jan. 17, 18, 1989 (A90-32151 13-61). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 52-58. 890000 p. 7 In: EN (English) p.2064

Development of knowledge-based technology enabling the construction of complete robust high-performance image understanding systems is addressed. A new-generation system, visual modeling and recognition, dynamic scene and motion analysis, obstacle detection and avoidance, parallel computing environment for vision, and technology transfer are covered among important accomplishments achieved in the first phase of the research, and the project summaries of the above developments are outlined. Integration of the component technologies into a new-generation system and demonstration of the utility of emerging vision software for autonomous navigation tasks are emphasized. The integration task represents a major research itself, since it addresses the architectural problems of sensor fusion and communication between the sensing and reasoning modules.
V.T.

TYPE 1/4/27
Quest Accession Number : 90A32152
90A32152 NASA IAA Conference Paper Issue 13
Neural networks for computer vision - A framework for specifications of a general purpose vision system
(AA)SKRZYPEK, JOSEF; (AB)MESROBIAN, EDMOND; (AC)GUNGNER, DAVID
(AC)(California, University, Los Angeles)
N00014-86-K-0395 IN: Image understanding and the man-machine interface II; Proceedings of the Meeting, Los Angeles, CA, Jan. 17, 18, 1989 (A90-32151 13-63). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 16-29. Research supported by IBM Corp., Hewlett Packard Co., and University of California. 890000 p. 14 refs 42 In: EN (English) p.2063

A general-purpose machine vision system capable of perceiving and understanding images in an unconstrained environment is considered. Fifteen systems built during the last ten years are analyzed along five dimensions - image attributes, perceptual primitives, knowledge base, object representation, and control strategy. The human visual system is analyzed as an underlying mechanism necessary for the development of general purpose vision. An interdisciplinary approach to vision research based on the combination of computational neuroscience with computer science and electrical engineering is proposed. A methodology for synthesizing a framework for a general-purpose machine vision system is addressed, and visual tasks such as edge detection and texture discrimination are covered, along with complex pattern analysis and the formation of visual categories.
V.T.

TYPE 1/4/28
Quest Accession Number : 90N27406
90N27406# NASA STAR Preprint Issue 21
Dynamic monocular machine vision and applications of dynamic monocular machine vision
(AA)DICKMANNS, ERNST DIETER; (AB)GRAEFE, VOLKER
Universitaet der Bundeswehr Muenchen, Neubiberg (Germany, F.R.). (U1005765) Inst. fuer Systemdynamic und Flugmechanik.
LRT-WE-13-FB-88-3; ETN-90-97334 Sponsored in part by BMFT; DFG; Daimler Benz A.G.; and MBB 880700 p. 99 In: EN (English) Avail: NTIS HC A05/MF A01 p.3061

A new approach to realtime machine vision in dynamic scenes is presented. It is based on special hardware and methods for feature extraction and information processing. Using integral spatio-temporal models, it bypasses the nonunique inversion of the perspective projection by applying recursive least squares filtering. By prediction error feedback methods, all spatial states variables including the velocity components are estimated. Only the last image of the sequence needs to be evaluated. Two applications in the field of robotics are given.
ESA

TYPE 1/4/29
Quest Accession Number : 90N27394
90N27394# NASA STAR Technical Report Issue 21
Parallel algorithms for computer vision / Final Report, 31 Aug. 1988 - 31 Jan. 1990
(AA)POGGIO, TOMASO
Massachusetts Inst. of Tech., Cambridge. (MJ700802) Artificial Intelligence Lab.
AD-A221871; ETL-0564 DACA76-85-C-0010 900400 p. 64 In: EN (English) Avail: NTIS HC A04/MF A01 p.3059

An integrated vision system, (the Vision Machine) based on a parallel supercomputer, is examined. The core of the Vision Machine is in fact a set of parallel algorithms for visual recognition and navigation in an unstructured environment. The present version of the Vision Machine was demonstrated to process images in close to real time by: (1) computing first several low level cues, such as edges, stereo disparity, optical flow, color and texture, (2) integrating them to extract a cartoon-like description of the scene in terms of the physical discontinuities of surfaces, and (3) using this cartoon in a recognition stage, based on parallel model matching. In addition to the development of the parallel algorithms, their implementation and testing, work was performed in several areas that are very closely related. These include: (1) design and fabrication of VLSI circuits to transfer to potentially cheap and fast hardware some of the software algorithms; (2) initial development of techniques to synthesize by learning vision algorithms; and (3) several projects involving autonomous navigation of small robots.
GRA

TYPE 1/4/30
Quest Accession Number : 90N22242
90N22242*# NASA STAR Conference Paper Issue 15
Ames vision group research overview / Abstract Only
(AA)WATSON, ANDREW B.
National Aeronautics and Space Administration. Ames Research Center, Moffett Field, CA. (NC473657)
In its Vision Science and Technology at NASA: Results of a Workshop p 52 (SEE N90-22216 15-54) 900200 p. 1 In: EN (English) Avail: NTIS HC A04/MF A01 p.2143

A major goal of the reseach group is to develop mathematical and computational models of early human vision. These models are valuable in the prediction of human performance, in the design of visual coding schemes and displays, and in robotic vision. To date researchers have models of retinal sampling, spatial processing in visual cortex, contrast sensitivity, and motion processing. Based on their models of early human vision, researchers developed several schemes for efficient coding and compression of monochrome and color images. These are pyramid schemes that decompose the image into features that vary in location, size, orientation, and phase. To determine the perceptual fidelity of these codes, researchers developed novel human testing methods that have received considerable attention in the research community. Researchers constructed models of human visual motion processing based on physiological and psychophysical data, and have tested these models through simulation and human experiments. They also explored the application of these biological algorithms to applications in automated guidance of rotorcraft and autonomous landing of spacecraft. Researchers developed networks for inhomogeneous image sampling, for pyramid coding of images, for automatic geometrical correction of disordered samples, and for removal of motion artifacts from unstable cameras.
Author

TYPE 1/4/31
Quest Accession Number : 90N22237
90N22237*# NASA STAR Conference Paper Issue 15
Computer vision techniques for rotorcraft low altitude flight
(AA)SRIDHAR, BANAVAR
National Aeronautics and Space Administration. Ames Research Center, Moffett Field, CA. (NC473657)
In its Vision Science and Technology at NASA: Results of a Workshop p 45-46 (SEE N90-22216 15-54) 900200 p. 2 In: EN (English) Avail: NTIS HC A04/MF A01 p.2142

Rotorcraft operating in high-threat environments fly close to the earth's surface to utilize surrounding terrain, vegetation, or manmade objects to minimize the risk of being detected by an enemy. Increasing levels of concealment are achieved by adopting different tactics during low-altitude flight. Rotorcraft employ three tactics during low-altitude flight: low-level, contour, and nap-of-the-earth (NOE). The key feature distinguishing the NOE mode from the other two modes is that the whole rotorcraft, including the main rotor, is below tree-top whenever possible. This leads to the use of lateral maneuvers for avoiding obstacles, which in fact constitutes the means for concealment. The piloting of the rotorcraft is at best a very demanding task and the pilot will need help from onboard automation tools in order to devote more time to mission-related activities. The development of an automation tool which has the potential to detect obstacles in the rotorcraft flight path, warn the crew, and interact with the guidance system to avoid detected obstacles, presents challenging problems. Research is described which applies techniques from computer vision to automation of rotorcraft navigtion. The effort emphasizes the development of a methodology for detecting the ranges to obstacles in the region of interest based on the maximum utilization of passive sensors. The range map derived from the obstacle-detection approach can be used as obstacle data for the obstacle avoidance in an automatic guidance system and as advisory display to the pilot. The lack of suitable flight imagery data presents a problem in the verification of concepts for obstacle detection. This problem is being addressed by the development of an adequate flight database and by preprocessing of currently available flight imagery. The presentation concludes with some comments on future work and how research in this area relates to the guidance of other autonomous vehicles.
Author

TYPE 1/4/32
Quest Accession Number : 90N18188
90N18188# NASA STAR Conference Proceedings Issue 10
High-Level Vision and Planning Workshop Proceedings / Final Report
(AA)BLOOM, MICHAEL I.
(AA)ed.
Institute for Defense Analyses, Alexandria, VA. (IJ564258)
AD-A215982; AD-E501178; IDA-D-649; IDA/HQ-89-034738 MDA903-89-C-0003 890800 p. 256 Workshop held in Rehovot, Israel, 25 Apr. 1988; sponsored by DARPA, US-Israel Binational Science Foundation and Institute for Defense Analyses In: EN (English) Avail: NTIS HC A12/MF A01 p.1420

The slides, papers, and graphic illustrations presented at the joint U.S.-Israeli workshop on artificial intelligence are provided in this Institute for Defense Analyses document. This document is based on a broad exchange of ideas about current approaches and research issues in the areas of design automation and autonomous robotic systems. A list of participants is provided along with applicable references for individual papers.
GRA

TYPE 1/4/33
Quest Accession Number : 90N16734
90N16734# NASA STAR Conference Paper Issue 09
Autonomous automatic landing through computer vision
(AA)SCHELL, R.; (AB)DICKMANNS, E. D.
Hochschule der Bundeswehr, Munich (Germany, F.R.). (HV212637) Dept. of Aerospace Technology.
In AGARD, Advances in Techniques and Technologies for Air Vehicle Navigation and Guidance 9 p (SEE N90-16731 09-04) 891200 p. 9 In: EN (English) Avail: NTIS HC A09/MF A02; Non-NATO Nationals requests available only from AGARD/Scientific Publications Executive p.1163

The automatic autonomous landing approach through computer vision was investigated in a simulation loop with real image sequence processing hardware and software. The use of integral spatio-temporal world models is the presupposition to achieve real time performance with the microprocessors currently available. Results achieved for a business-jet aircraft demonstrate that this set up is powerful enough to solve the problem of autonomous unmanned landing approach.
Author

TYPE 1/4/34
Quest Accession Number : 90N15453
90N15453# NASA STAR Technical Report Issue 07
Research in knowledge-based vision techniques for the Autonomous Land Vehicle Program / Final Annual Report, 1 Jun. 1988 - 31 May 1989
(AA)NEVATIA, R.; (AB)PRICE, K.; (AC)FRANZEN, W.; (AD)GAZIT, S.; (AE)MEDIONI, G.; (AF)PENG, S.; (AG)SAINT-MARC, P.
(AA)ed.; (AB)ed.
University of Southern California, Los Angeles. (U6203125) Inst. for Robotics and Intelligent Systems.
AD-A213440; IRIS-255; ETL-0545 DACA76-85-C-0009 890800 p. 59 In: EN (English) Avail: NTIS HC A04/MF A01 p.942

The authors' basic approach to detecting and tracking motion is to extract and match features, such as lines and regions, from a sequence and to generate motion estimates from these. They present one report on spatio-temporal analysis for tracking edges through very closely spaced sequences. They also present a report on matching edge-based contours using edges from multiple scales with low resolution guiding high resolution matches. They also present an analysis of estimating 3-D motion and structure of moving object with uniform acceleration.
GRA

TYPE 1/4/35
Quest Accession Number : 90A14975
90A14975 NASA IAA Conference Paper Issue 04
Image understanding techniques in geophysical data interpretation
(AA)ROBERTO, V.; (AB)PERON, A.; (AC)FUMIS, P. L.
(AC)(Udine, Universita, Italy)
IN: Issues on Machine Vision, Course, Udine, Italy, July 1988, Proceedings (A90-14971 04-63). Vienna and New York, Springer-Verlag, 1989, p. 263-274. 890000 p. 12 refs 9 In: EN (English) p.0

This paper covers some topics in geophysical signal interpretation, by means of Artificial Intelligence (Machine Vision) techniques. In particular, the low-level processing modules of a Knowledge-Based System for seismic reflection image understanding are presented, as well as an explanation of their structural and functional characteristics. Preliminary results are also given and discussed.
Author

TYPE 1/4/36
Quest Accession Number : 90A14974
90A14974 NASA IAA Conference Paper Issue 04
Neural networks, supercomputers and computer vision
(AA)JOHNSON, O.; (AB)PIERONI, G.; (AC)RAKOTOMALALA, M.
(AA)(Houston, University, TX); (AB)(Udine, Universita, Italy; Houston, University, TX); (AC)(HARC, Woodlands, TX)
IN: Issues on Machine Vision, Course, Udine, Italy, July 1988, Proceedings (A90-14971 04-63). Vienna and New York, Springer-Verlag, 1989, p. 163-175. 890000 p. 13 refs 16 In: EN (English) p.0

A PDP program for simulating neural networks ia applied to problems in machine vision. The PDP program avoids explicit pattern matching with reference model segments as well as the creation of hypotheses in order to utilize the neural networks' ability to perform pattern matching with distorted and incomplete data. The problem of recognizing simple four-sided polygons in a two-dimensional scene of straight lines is considered. Supercomputers which use neural network software are discussed.
C.D.

TYPE 1/4/37
Quest Accession Number : 90A14971
90A14971 NASA IAA Meeting Paper Issue 04
Issues on Machine Vision, Course, Udine, Italy, July 1988, Proceedings
(AA)PIERONI, G. G.
(AA)ED.
(AA)(Udine, Universita, Italy)
Course organized by the International Centre for Mechanical Sciences; Supported by CNR, UNESCO, Centro Ricerche FIAT, et al. Vienna and New York, Springer-Verlag, 1989, 344 p. For individual items see A90-14972 to A90-14975. 890000 p. 344 In: EN (English) $57.20 p.0

Various papers on machine vision are presented. Individual topics addressed include: data processing via associative memory; picture labeling and shape descriptors for machine vision; morphological approach to industrial image inspection of honeycomb composite materials; two-dimensional digital filter design by the adaptive differential correction algorithm; comparison of hierarchical topologies for megamicrocomputers; constrained Delaunay triangulation algorithms for surface representation; medium-level language for pyramid architectures; vision problems in sparse images; machine vision for inspection; neural networks, supercomputers, and computer vision; software issues for machine vision; multiresolution approach for segmenting surfaces; signed Euclidean distance transform applied to shape analysis; image understanding techiques in geophysical data interpretation; knowledge integration for machine vision; motion parameter estimation for robot application; and industrial applications of machine vision.
C.D.

TYPE 1/4/38
Quest Accession Number : 90N13235
90N13235# NASA STAR Technical Report Issue 04
Research in computer vision for autonomous systems / Progress Report, Jun. - Sep. 1988
(AA)KAK, AVI; (AB)YODER, MARK; (AC)ANDRESS, KEITH; (AD)BLASK, STEVE; (AE)UNDERWOOD, TOM
Purdue Univ., West Lafayette, IN. (P9391092) School of Electrical Engineering.
AD-A212420 DAAL01-85-C-0456 880915 p. 532 In: EN (English) Avail: NTIS HC A23/MF A03 p.555

This report addresses FLIR processing, LADAR processing and electronic terrain board modeling. In our discussion on FLIR processing, issues were analyzed for classifiability of FLIR features, computationally efficient algorithms for target segmentation, metrics, etc. The discussion on LADAR includes a comparison of a number of different approaches to the segmentation of target surfaces from range images, extraction of silhouettes at different ranges, and reasoning strategies for the recognition of targets and estimation of their aspects. Regarding electronic terrain board modeling, it was shown how the readily available wire-frame data for strategic targets can be converted into volumetric models utilizing the concepts of constructive solid geometry; then is was shown how from the resulting volumetric models it is possible to generate synthetic range images that are very similar to real LADAR images. Also shown is how sensor noise can be added to these synthetic images to make them even more realistic.
GRA

TYPE 1/4/39
Quest Accession Number : 90A11742
90A11742 NASA IAA Conference Paper Issue 02
Real time imaging rangefinder for autonomous land vehicles
(AA)KERR, J. RICHARD
(AA)(FLIR Systems, Inc., Portland, OR)
IN: Mobile robots III; Proceedings of the Meeting, Cambridge, MA, Nov. 10, 11, 1988 (A90-11726 02-14). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 349-356. 890000 p. 8 In: EN (English) p.190

A three-dimensional sensor that achieves 50 microsteradian resolution over a 90 x 40 degree field of view (FOV) at full video frame rates has been designed for robotic vehicles. A combination of coarse and fine range resolution provides sensing from one to approximately 100 meters with short-range accuracies of less than 10 cm. The system utilizes an eyesafe diode laser configuration along with proprietary mechanical scanning elements, wide-field relay optics, and avalanche photodiode detectors. Range determination is accomplished with dual subcarrier modulation which results in the output of an unambiguous, binary word on a pixel-by-pixel basis. The approach also provides for electronic pitch stailization.
Author

TYPE 1/4/40
Quest Accession Number : 90A11730
90A11730 NASA IAA Conference Paper Issue 02
Terrain classification using texture for the ALV
(AA)MARRA, MARTY; (AB)DUNLAY, R. TERRY; (AC)MATHIS, DON
(AC)(Martin Marietta Information and Communications Systems, Denver, CO)
DACA76-84-C-0005 IN: Mobile robots III; Proceedings of the Meeting, Cambridge, MA, Nov. 10, 11, 1988 (A90-11726 02-14). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 64-70. Research supported by DARPA. 890000 p. 7 refs 13 In: EN (English) p. 237

Off-road navigation is a very demanding visual task in which texture can play an important role. Travel on a smooth road or path can be done with greater speed and safety in general than on rough natural terrain. In addition, recognition of off-road terrain types can aid in finding the fastest and safest route through a given area. Implementations of two texture methods for identifying certain terrain features in video imagery are briefly discussed. The first method uses edge and morphological filters to identify roadways from off-road. The second method uses a neural net to identify several terrain types based on color, directional texture, global variance and location in the image. Plans to integrate the terrain labeled image produced by the latter method into the ALV's perception system are also discussed.
Author

TYPE 1/4/41
Quest Accession Number : 90A11696
90A11696 NASA IAA Conference Paper Issue 02
An intelligent system for autonomous navigation of airborne vehicles
(AA)CAMERON, WILLIAM L.; (AB)FAIN, HOWARD; (AC)BEZDEK, JAMES C.
(AB)(Boeing Aerospace, Seattle, WA); (AC)(Boeing Electronics, Seattle, WA)
IN: Sensor fusion: Spatial reasoning and scene interpretation; Proceedings of the Meeting, Cambridge, MA, Nov. 7-9, 1988 (A90-11676 02-63). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 451-469. 890000 p. 19 refs 8 In: EN (English) p. 143

Autonomous navigation of airborne platforms requires the integration of diverse sources of sensor data and contextual information. This paper describes a system that utilizes polarimetric radar cross-section and range data to generate position estimates based on four kinds of information: area segmentation, ground contours, landmarks, and road networks. Ground truth in the form of terrain feature maps is correlated with each type of data stream. Finally, an arbitrator integrates these inputs with contextual knowledge about the preplanned flight path to resolve conflicts and arrive at a final estimate of current position.
Author

TYPE 1/4/42
Quest Accession Number : 90A11683
90A11683 NASA IAA Conference Paper Issue 02
Neural network model for fusion of visible and infrared sensor outputs
(AA)AJJIMARANGSEE, PONGSAK; (AB)HUNTSBERGER, TERRANCE L.
(AB)(South Carolina, University, Columbia)
IN: Sensor fusion: Spatial reasoning and scene interpretation; Proceedings of the Meeting, Cambridge, MA, Nov. 7-9, 1988 (A90-11676 02-63). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1989, p. 153-160. 890000 p. 8 refs 16 In: EN (English) p. 235

Integration of outputs from multiple sensors has been the subject of much of the recent research in the machine vision field. This paper presents a neural-network model for the fusion of visible and thermal-IR sensor outputs. A model is developed based on six types of bimodal neurons found in the optic tectum of the rattlesnake. These neurons integrate visible and thermal-IR sensory inputs. The neural network model has a series of layers which include a layer for unsupervised clustering in the form of self-organizing feature maps, followed by a layer which has multiple filters that are generated by training a neural net with experimental rattlesnake response data. The final layer performs another unsupervised clustering for integration of the output from the filter layer. The results of a number of experiments are also presented.
Author

TYPE 1/4/43
Quest Accession Number : 90A11032
90A11032 NASA IAA Meeting Paper Issue 01
Optics, illumination, and image sensing for machine vision III; Proceedings of the Meeting, Cambridge, MA, Nov. 8, 9, 1988
(AA)SVETKOFF, DONALD J.
(AA)ED.
(AA)(Synthetic Vision Systems, Inc., Ann Arbor, MI)
SPIE-1005 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 1005), 1989, 271 p. For individual items see A90-11033 to A90-11035. 890000 p. 271 In: EN (English) Members, $41.; nonmembers, $51 p.90

Various papers on optics, illumination, and image sensing for machine vision are presented. Some of the optics discussed include: illumination and imaging of moving objects, strobe illumination systems for machine vision, optical collision timer, new electrooptical coordinate measurement system, flexible and piezoresistive touch sensing array, selection of cameras for machine vision, custom fixed-focal length versus zoom lenses, performance of optimal phase-only filters, minimum variance SDF design using adaptive algorithms, Ho-Kashyap associative processors, component spaces for invariant pattern recognition, grid labeling using a marked grid, illumination-based model of stochastic textures, color-encoded moire contouring, noise measurement and suppression in active 3-D laser-based imaging systems, structural stereo matching of Laplacian-of-Gaussian contour segments for 3D perception, earth surface recovery from remotely sensed images, and shape from Lambertian photometric flow fields.
C.D.

TYPE 1/4/44
Quest Accession Number : 89A41730
89A41730 NASA IAA Journal Article Issue 17
Schemas and neural networks for sixth generation computing
(AA)ARBIB, MICHAEL A.
(AA)(Southern California, University, Los Angeles, CA)
NIH-7-R01-NS-24926 Journal of Parallel and Distributed Computing (ISSN 0743-7315), vol. 6, April 1989, p. 185-216. 890400 p. 32 refs 102 In: EN (English) p.2680

Sixth-generation computer architectures are presently conjectured to profitably involve networks of one or more specialized devices structured as highly-parallel arrays of neuronlike interacting (and perhaps also adaptive) components. Schemas are suggested to be a germane basis for the programming languages that will typify sixth-generation computers; the characteristics of schemas are illustrated for the case of their use in high-level machine vision. An integrated system of investigations, the 'Rana computatrix', demonstrates the fusion of neural-network and schema models of the visuomotor-coordination mechanism in frogs and toads. The 'domain-specific' structure of neural networks is emphasized.
O.C.

TYPE 1/4/45
Quest Accession Number : 89A40426
89A40426 NASA IAA Meeting Paper Issue 17
Applications of digital image processing XI; Proceedings of the Meeting, San Diego, CA, Aug. 15-17, 1988
(AA)TESCHER, ANDREW G.
(AA)ED.
(AA)(Lockheed Research Laboratories, Palo Alto, CA)
SPIE-974 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 974), 1988, 421 p. For individual items see A89-40427 to A89-40452. 880000 p. 421 In: EN (English) Members, $44.; nonmembers, $57 p.2673

Theoretical and applications aspects of digital image processing are discussed in reviews and reports of recent investigations. Topics addressed include enhancement and restoration, transmission and vision, PC-based and graphics applications, architectures and systems, and hybrid and unconventional image-processing methods. Consideration is given to morphology in wrap-around image algebra, maximum-likelihood image restoration with subpixel accuracy, high-resolution digitization of color images, a lighting and optics expert system for machine vision, image-data compression in a PC environment, rule-based processing for string-code identification, digital-image velocimetry, aircraft navigation using IR image analysis, aircraft recognition using a parts-analysis technique, and an image-quality measure based on the human visual system.
T.K.

TYPE 1/4/46
Quest Accession Number : 89N27136
89N27136# NASA STAR Technical Report Issue 21
JTECH (Japanese Technology Evaluation Program) panel report on advanced sensors in Japan
(AA)MILLER, G. L.; (AB)GUCKEL, H.; (AC)HALLER, E.; (AD)KANADE, T.; (AE)KO, W.; (AF)RADEKA, V.
Science Applications International Corp., McLean, VA. (SD708880)
PB89-158760 Sponsored by NSF, Washington, DC; DARPA, Arlington, VA and Department of Commerce, Washington, DC 890100 p. 293 In: EN (English) Avail: NTIS HC A13/MF A01 p.3012

The document provides the results of a detailed evaluation of the current state of Japanese sensor development. The analysis was performed by a panel of technical experts drawn from U.S industry and academia. It covers not only specific technical work, but also covers issues of organization, trends, funding, and methods of organizing work and setting priorities. The topics covered include: Tutorial introduction to sensors, machine vision (charge coupled device (CCD) sensors, vision processing systems, active 3-D range sensors, Research Institution on Machine Vision); sensors for electromagnetic radiation (far infrared, near infrared, visible light, X-rays, gamma-rays); sensors for factory automation and robotics; micromechanical and superconducting sensors; gas sensors; ion sensors; ion selective field effect transistors (ISFET); and biosensors. Also included is an extensive listing of Japanese sensor manufacturers.
GRA

TYPE 1/4/47
Quest Accession Number : 89N23152
89N23152# NASA STAR Conference Paper Issue 16
Combining information in low-level vision
(AA)ALOIMONOS, JOHN; (AB)BASU, ANUP
Maryland Univ., College Park. (MI915766) Computer Vision Lab.
DAAB07-86-K-F073 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 2 p 862-906 (SEE N89-23115 16-61) 880400 p. 45 In: EN (English) Avail: NTIS HC A99/MF E03 p.2320

Low level modern computer vision is not domain dependent, but concentrates on problems that correspond to identifiable modules in the human visual system. Several theories have been proposed in the literature for the computation of shape from shading, shape from texture, retinal motion from spatiotemporal derivatives of the image intensity function and the like. The basic problems with some of the existing approaches if several available cues are combined, disappear in most cases; the resulting algorithms compute robustly and uniquely the intrinsic parameters (shape, depth, motion, etc.). The problem of machine vision is explored here from its basics. A low level mathematical theory is presented for the unique and robust computation of intrinsic parameters. The computational aspect of the theory envisages a cooperative highly parallel implementation, bringing in information from five different sources (shading, texture, motion, contour and stereo), to resolve ambiguities and ensure uniqueness of the intrinsic parameters.
Author

TYPE 1/4/48
Quest Accession Number : 89N23124
89N23124# NASA STAR Conference Paper Issue 16
Three-dimensional vision for outdoor navigation by an autonomous vehicle
(AA)HEBERT, MARTIAL; (AB)KANADE, TAKEO
Carnegie-Mellon Univ., Pittsburgh, PA. (CH188052) Robotics Inst.
DACA76-85-C-0003; F33615-87-C-1499; NSF DCR-86-04199 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 2 p 593-601 (SEE N89-23115 16-61) 880400 p. 9 In: EN (English) Avail: NTIS HC A99/MF E03 p.2315

Progress in range image analysis for autonomous navigation in outdoor environments is reported. The goal of the work is to use range data from an ERIM laser range finder to build a three-dimensional description of the environment. Techniques are described for building both low-level description, such as obstacle maps or terrain maps, as well as higher level description using model-based object recognition. These techniques have been integrated in the NAVLAB system.
Author

TYPE 1/4/49
Quest Accession Number : 89N23121
89N23121# NASA STAR Conference Paper Issue 16
An operational perception system for cross-country navigation
(AA)DAILY, MICHAEL J.; (AB)HARRIS, JOHN G.; (AC)REISER, KURT
Hughes Research Labs., Calabasas, CA. (H5849026) Artificial Intelligence Center.
DACA87-85-C-0007 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 2 p 568-575 (SEE N89-23115 16-61) 880400 p. 8 In: EN (English) Avail: NTIS HC A99/MF E03 p.2314

An operational perception system for cross-country navigation which has been verified in both simulated and real world environments is presented. Range data from a laser range scanner is transformed into an alternate representation called the Cartesian Elevation Map (CEM). A detailed vehicle model operates on the CEM to produce traversability information along selected trajectories. This information supports a real-time reflexive planning system. The successful demonstration of obstacle detection and avoidance algorithms on board an Autonomous Land Vehicle is discussed.
Author

TYPE 1/4/50
Quest Accession Number : 89N23120
89N23120# NASA STAR Conference Paper Issue 16
Using flow field divergence for obstacle avoidance in visual navigation
(AA)NELSON, RANDAL C.; (AB)ALOIMONOS, JOHN
Maryland Univ., College Park. (MI915766) Computer Vision Lab.
In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 2 p 548-567 (SEE N89-23115 16-61) Sponsored in part by DARPA, Washington, DC 880400 p. 20 In: EN (English) Avail: NTIS HC A99/MF E03 p.2314

The practical recovery of quantitative structural information about the world from visual data has proven to be a very difficult task. In particular, the recovery of motion information which is sufficiently accurate to allow practical application of theoretical shape from motion results has so far been infeasible. Yet a large body of evidence suggests that use of motion is an extremely important process in biological vision systems. It has been suggested that qualitative visual measurements can provide powerful perceptual cues, and that practical operations can be performed on the basis of such clues without the need for a quantitative reconstruction of the world. The use of such information is termed inexact vision. The investigation of one such approach to the analysis of visual motion is described. Specifically, the use of certain measures of flow field divergence was investigated as a qualitative cue for obstacle avoidance during visual navigation. It is shown that a quantity termed the directional divergence of the 2-D motion field can be used as a reliable indicator of the presence of obstacles in the visual field of an observer undergoing generalized rotational and translational motion. Moreover, the necessary measurements can be robustly obtained from real image sequences. A simple differential procedure for robustly extracting divergence information from image sequences which can be performed using a highly parallel, connectionist architecture is described. The procedure is based on the twin principles of directional separation of optical flow components and temporal accumulation of information. Experimental results are presented showing that the system responds as expected to divergence in real world image sequences, and the use of the system to navigate between obstacles is demonstrated.
Author

TYPE 1/4/51
Quest Accession Number : 89N23118
89N23118# NASA STAR Conference Paper Issue 16
Dynamic model matching for target recognition from a mobile platform
(AA)NASR, HATEM; (AB)BHANU, BIR
Honeywell Systems and Research Center, Minneapolis, MN. (HY989092)
DACA76-86-C-0017 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 2 p 527-536 (SEE N89-23115 16-61) 880400 p. 10 In: EN (English) Avail: NTIS HC A99/MF E03 p.2314

A novel technique called dynamic model matching (DMM) is presented for target recognition from a moving platform such as an autonomous combat vehicle. The DMM technique overcomes major limitations in present model-based target recognition techniques that use a single, static target model, and therefore cannot account for continuous changes in the target's appearance caused by varying range and perspective, DMM addresses this problem by combining a moving camera model, 3-D object models, spatial models, and expected range and perspective to generate multiple 2-D image models for matching. DMM also generates recognition strategies that can emphasize different object features at varying ranges. DMM operates within a larger system for landmark recognition based on the perception, reasoning, action, and expectation paradigm called PREACTE. Results are presented on a number of test sites using color video data obtained from the autonomous land vehicle.
Author

TYPE 1/4/52
Quest Accession Number : 89N23115
89N23115# NASA STAR Meeting Paper Issue 16
Proceedings: Image Understanding Workshop, volume 2 / Annual Technical Report, Feb. 1987 - Apr. 1988
(AA)BAUMANN, LEE S.
(AA)ed.
Science Applications International Corp., McLean, VA. (SD708880)
AD-A197559 N00014-86-C-0700; ARPA ORDER 5605 880400 p. 678 Workshop held in Cambridge, MA, 6-8 Apr. 1988; sponsored by DARPA In: EN (English)
Avail: NTIS HC A99/MF E03 p.2313

Annual progress reports and technical papers presented by the participants at the Image Understanding Workshop sponsored by the Information Science and Technology Office, Defense Advanced Research Projects Agency are presented. Also included are copies of invited papers presented at the workshop and additional technical papers which were not presented (volume 2). Topics addressed included: intelligent image understanding, machine vision and robotics, knowledge-based systems, motion detection and tracking, object and target recognition, parallel computation, stereo vision, and image processing. For individual titles, see N89-23116 through N89-23180.

TYPE 1/4/53
Quest Accession Number : 89N23108
89N23108# NASA STAR Conference Paper Issue 16
Integration effort in knowledge-based vision techniques for the autonomous land vehicle program
(AA)PRICE, KEITH; (AB)PAVLIN, IGOR
University of Southern California, Los Angeles. (U6203125) Inst. for Robotics and Intelligent Systems.
DACA76-85-C-0009 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 417-422 (SEE N89-23074 16-61) 880400 p. 6 In: EN (English) Avail: NTIS HC A22/MF A01 p.2312

A methodology is presented and some early results are demonstrated in the integration of knowledge-based image analysis programs. The domain of complete three-dimensional motion analysis in the context of the Autonomous Land Vehicle is specifically addressed. The integrated system exploits the strengths and minimizes the weaknesses of the individual techniques, resulting in performance which is considerably improved over the performance of any of the independently developed programs.
Author

TYPE 1/4/54
Quest Accession Number : 89N23107
89N23107# NASA STAR Conference Paper Issue 16
Autonomous navigation in cross-country terrain
(AA)KEIRSEY, DAVID M.; (AB)PAYTON, DAVID W.; (AC)ROSENBLATT, J. KENNETH
Hughes Research Labs., Calabasas, CA. (H5849026) Artificial Intelligence Center.
DACA76-85-C-0017 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 411-416 (SEE N89-23074 16-61) 880400 p. 6 In: EN (English) Avail: NTIS HC A22/MF A01 p.2312

Progress and experimentation with an autonomous robotic vehicle in cross-country terrain is described. Experiments were performed on the Autonomous Land Vehicle in natural terrain. An overview of the software architecture used for this achievement is discussed; descriptions of experiments and details of planning techniques are presented. Experiments describe the vehicle's avoidance of both known and unknown obstacles in its path.
Author

TYPE 1/4/55
Quest Accession Number : 89N23094
89N23094# NASA STAR Conference Paper Issue 16
Kalman filter-based algorithms for estimating depth from image sequences
(AA)MATTHIES, LARRY; (AB)SZELISKI, RICHARD; (AC)KANADE, TAKEO
Carnegie-Mellon Univ., Pittsburgh, PA. (CH188052) Dept. of Computer Science.
F33615-87-C-1499 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 199-213 (SEE N89-23074 16-61) 880400 p. 15 In: EN (English) Avail: NTIS HC A22/MF A01 p.2309

Using known camera motion to estimate depth from image sequences is an important problem in robot vision. Many applications of depth from motion, including navigation and manipulation, require algorithms that can estimate depth in an on-line, incremental fashion. This requires a representation that records the uncertainty in depth estimates and a mechanism that integrates new measurements with existing depth estimates to reduce the uncertainty over time. Kalman filtering provides this mechanism. Previous applications of Kalman filtering to depth from motion have been limited to estimating depth at the location of a sparse set of features. A pixel-based (iconic) algorithm is introduced which estimates depth and depth uncertainty at each pixel and incrementally refines these estimates over time. The algorithm for translations parallel to the image plane is described and its formulation and performance contrasted to that of a feature-based Kalman filtering algorithm. The performance of the two approaches is compared by analyzing their theoretical convergence rates, by conducting quantitative experiments with images of a flat poster, and by conducting qualitative experiments with images of a realistic outdoor scene model. The results show that the method is an effective way to extract depth from lateral camera translations and suggest that it will play an important role in low-level vision.
Author

TYPE 1/4/56
Quest Accession Number : 89N23093
89N23093# NASA STAR Conference Paper Issue 16
The MIT vision machine
(AA)POGGIO, T.; (AB)LITTLE, J.; (AC)GAMBLE, E.; (AD)GILLETT, W.; (AE)GEIGER, D.; (AF)WEINSHALL, DAPHNA; (AG)VILLALBA, M.; (AH)LARSON, N.; (AI)CASS, TODD ANTHONY; (AJ)BUELTHOFF, H.
Massachusetts Inst. of Tech., Cambridge. (MJ700802) Artificial Intelligence Lab.
In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 177-198 (SEE N89-23074 16-61) 880400 p. 22 In: EN (English) Avail: NTIS HC A22/MF A01 p.2309

The vision Machine, its goals, and achievements to date are described. The Vision Machine is a computer system that attempts to integrate several vision cues to achieve high performance in unstructured environments for the tasks of recognition and navigation. It is also a test-bed for theoretical progress in early vision algorithms, their parallel implementation and their integration. The Vision Machine consists of a movable two-camera Eye-Head system (the input device) and a 16K Connection Machine (the main computational engine). Several parallel early vision algorithms which compute edge detection, stereo, motion, texture and surface color in close to real-time were developed and implemented. The integration stage is based on the technique of coupled Markov Random Field models, and leads to a cartoon-like map of the discontinuities in the scene, with a partial labeling of the brightness edges in terms of their physical origin. Available recognition algorithms will interface with the output of the integration stage and the analog and hybrid Very Large Scale Integration (VLSI) implementations of the Vision Machine main components has begun.
Author

TYPE 1/4/57
Quest Accession Number : 89N23091
89N23091# NASA STAR Conference Paper Issue 16
The Maryland approach to image understanding
(AA)ALOIMONOS, JOHN; (AB)DAVIS, LARRY S.; (AC)ROSENFELD, AZRIEL
Maryland Univ., College Park. (MI915766) Computer Vision Lab.
DAAB07-86-K-F073 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 154-165 (SEE N89-23074 16-61) 880400 p. 12 In: EN (English) Avail: NTIS HC A22/MF A01 p.2309

In an effort to understand images, while still working on initial processes of low and middle level vision, emphasis is being placed on the integration of multiple sources of information for visual reconstruction, on navigation and on object recognition. A methodological paradigm for research in vision is introduced, namely: while research is continuing top-down in the Marr paradigm, work also progresses in a bottom-up fashion in that paradigm. It is suggested that the Marr paradigm (computational theory, algorithms, data structures, and implementation) should be augmented with one more level, that of robustness, that Marr left implicit in his writings.
Author

TYPE 1/4/58
Quest Accession Number : 89N23083
89N23083# NASA STAR Conference Paper Issue 16
Image understanding and robotics research at Columbia University
(AA)KENDER, JOHN R.; (AB)ALLEN, PETER K.; (AC)BOULT, TERRANCE E.; (AD)IBRAHIM, HUSSEIN A. H.
Columbia Univ., New York, NY. (CV146013) Dept. of Computer Science.
DACA76-86-C-0024 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 78-87 (SEE N89-23074 16-61) 880400 p. 10 In: EN (English) Avail: NTIS HC A22/MF A01 p. 2307

Diverse research investigations in vision and robotics are identified and summarized. Since it is difficult to separate those aspects of robotic research that are purely visual from those that are vision-like (for example, tactile sensing) or vision-related (for example, integrated vision-robotic systems), all robotic research that is not purely manipulative is listed. Areas of research that are identified are low-level vision: theories involving stereo, data representations, and applications to graphics; middle-level vision: regularized surface reconstruction and stereo, sensory fusion, shape from dynamic shadowing, and application to range data; spatial relations: representations of objects and space, and theory and practice of navigation; parallel algorithms: low- and middle-level vision theory, research and applications on tree machines, and research and applications on pipelined machines; and, finally, robotics and tactile sensing: system development, and multi-fingered object recognition.
Author

TYPE 1/4/59
Quest Accession Number : 89N23081
89N23081# NASA STAR Conference Paper Issue 16
Summary of image understanding research at the University of Massachusetts
(AA)RISEMAN, EDWARD M.; (AB)HANSON, ALLEN R.
Massachusetts Univ., Amherst. (MK149394) Dept. of Computer and Information Science.
DACA76-85-C-0008; DACA76-86-C-0015; F30602-87-C-0140; N00014-82-K-0464; DMA800-85-C-0012; AF-AFOSR-0021-86; NSF DCR-85-00332 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 62-72 (SEE N89-23074 16-61) 880400 p. 11 In: EN (English) Avail: NTIS HC A22/MF A01 p.2307

Several areas of research in the Image Understanding Program are summarized, including: (1) knowledge-based vision; (2) database support for symbolic vision processing; (3) motion processing; (4) perceptual organization (grouping); (5) image understanding architecture; (6) integrated vision benchmark for parallel architectures; and (7) mobile vehicle navigation. A fundamental goal of the computer vision research environment is the integration of a diverse set of research efforts into a system that is ultimately intended to achieve real-time image interpretation. Two major system integration efforts are the VISIONS static interpretation system, which is a knowledge-based computer vision system utilizing parallel modular processes that communicate via a blackboard, and an autonomous mobile vehicle for navigation through a partially known environment.
Author

TYPE 1/4/60
Quest Accession Number : 89N23080
89N23080# NASA STAR Conference Paper Issue 16
Image understanding research at SRI International
(AA)FISCHLER, MARTIN A.; (AB)BOLLES, ROBERT C.
SRI International Corp., Menlo Park, CA. (SY423852) Artificial Intelligence Center.
MDA903-86-C-0084; DACA76-85-C-0004 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 53-61 (SEE N89-23074 16-61) 880400 p. 9 In: EN (English) Avail: NTIS HC A22/MF A01 p.2307

The Image Understanding research program is a broad effort spanning the entire range of machine vision research. The progress in two programs is described: the first is concerned with modeling the earth's surface from aerial photographs; the second is concerned with visual interpretation for land navigation. In particular, the following are described: progress in the design of a core knowledge structure; representing, recognizing, and rendering complex natural and man-made objects; recognizing and modeling terrain features and man-made objects in image sequences; interactive techniques for scene modeling and scene generation; automated detection and delineation of cultural objects in aerial imagery; and automated terrain modeling from aerial imagery.
Author

TYPE 1/4/61
Quest Accession Number : 89N23076
89N23076# NASA STAR Conference Paper Issue 16
USC image understanding research: 1987-1988
(AA)NEVATIA, RAMAKANT
University of Southern California, Los Angeles. (U6203125) Inst. for Robotics and Intelligent Systems.
DACA76-85-C-0009; F33615-87-C-1436 In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 13-16 (SEE N89-23074 16-61) 880400 p. 4 In: EN (English) Avail: NTIS HC A22/MF A01 p.2306

University of Southern California Image Understanding research projects are summarized and references to more detailed projects and papers are provided. The work has focussed on the topics of: mapping from aerial images, robotics vision, motion analysis for autonomous land vehicles (ALV), some general techniques, and parallel processing.
Author

TYPE 1/4/62
Quest Accession Number : 89N23075
89N23075# NASA STAR Conference Paper Issue 16
MIT progress in understanding images
(AA)POGGIO, T.
Massachusetts Inst. of Tech., Cambridge. (MJ700802) Artificial Intelligence Lab.
In Science Applications International Corp., Proceedings: Image Understanding Workshop, Volume 1 p 1-12 (SEE N89-23074 16-61) 880400 p. 12 In: EN (English) Avail: NTIS HC A22/MF A01 p.2306

Work in the past year has concentrated on three main projects, each one representing a complementary aspect of a complete vision system. The first project - a parallel Vision Machine - has the goal of developing a system for integrating early vision modules and computing a robust description of the discontinuities of the surfaces and of their physical properties. Additional goals of the project are the refinement of early vision algorithms and their implementation on a massively parallel architecture such as the Connection Machine System. The second project concerns visual recognition; several schemes for model based recognition were developed and implemented. Finally, work has continued on autonomous navigation. Around these main themes, additional work, at the theoretical and implementation level, has been done in motion analysis, navigation, photogrammetry, visual routines, and learning.
Author

TYPE 1/4/63
Quest Accession Number : 89N23074
89N23074# NASA STAR Meeting Paper Issue 16
Proceedings: Image Understanding Workshop, volume 1 / Annual Technical Report, Feb. 1987 - Apr. 1988
(AA)BAUMANN, LEE S.
(AA)ed.
Science Applications International Corp., McLean, VA. (SD708880)
AD-A197558 N00014-86-C-0700; ARPA ORDER 5605 880400 p. 525 Workshop held in Cambridge, MA, 6-8 Apr. 1988; sponsored by DARPA In: EN (English)
Avail: NTIS HC A22/MF A01 p.2306

This document contains the annual progress reports and technical papers presented on the research activities in image understanding at a workshop conducted on 6 to 8 April 1988, in Cambridge, Massachusetts. Also included are copies of invited papers presented at the workshop and additional technical papers from the research activities which were not presented due to lack of time but are germane to this research field. Topics discussed include: intelligent systems, robotics, knowledge-based vision, algorithms, pattern matching, feedback, tracking, autonomous navigation, parallel processing, target recognition, data integration, motion recognition, and image analysis. For individual titles, see N89-23075 through N89-23114.

TYPE 1/4/64
Quest Accession Number : 89N22597
89N22597# NASA STAR Technical Report Issue 16
Dynamic image interpretation for autonomous vehicle navigation / Annual Report, 26 Feb. 1987 - 25 Feb. 1988
(AA)RISEMAN, EDWARD M.; (AB)HANSON, ALLEN R.
Massachusetts Univ., Amherst. (MK149394) Dept. of Computer and Information Science.
AD-A204167; ETL-0516 DACA76-85-C-0008 880900 p. 33 In: EN (English) Avail: NTIS HC A03/MF A01 p.2222

The results of the project on Dynamic Image Interpretation for Autonomous Land Vehicle (ALV) Navigation is presented for the time period 2/26/87 to 2/25/88. The purpose of the ALV project is to develop algorithms and tools to enable a vehicle to navigate autonomously through realistic landscapes. Contents: Visual Motion Analysis- Computation of the Optical Flow Field; The Recovery of Environmental Motion and Structure from a Mobile Vehicle; Alternatives to General Motion Analysis; Stereoscopic Motion Analysis; Analysis of Constant General Motion; Token-Based Approaches to Motion and Perceptual Organization; Mobile Vehicle Navigation; Perceptual Organization (Grouping)- The Perceptual Organization of Image Curves; Extracting Geometric Structure; Database Support for Symbolic Vision Processing- ISR1, ISR2, Generic Views and Indexing.
GRA

TYPE 1/4/65
Quest Accession Number : 89A21185
89A21185* NASA IAA Journal Article Issue 07
Model-based orientation-independent 3-D machine vision techniques
(AA)DE FIGUEIREDO, R. J. P.; (AB)KEHTARNAVAZ, N.
(AA)(Rice University, Houston, TX); (AB)(Texas A & M University, College Station)
Rice Univ., Houston, TX. (RV347060)
NAG9-192; NAG9-208 (California Institute of Technology, Workshop on Space Telerobotics, Pasadena, Jan. 1987) IEEE Transactions on Aerospace and Electronic Systems (ISSN 0018-9251), vol. 24, Sept. 1988, p. 597-607. Research supported by Texas Instruments, Inc. 880900 p. 11 refs 17 In: EN (English) p.1037

Orientation-dependent techniques for the identification of a three-dimensional object by a machine vision system are represented in parts. In the first part, the data consist of intensity images of polyhedral objects obtained by a single camera, while in the second part, the data consist of range images of curved objects obtained by a laser scanner. In both cases, the attributed graphic representation of the object surface is used to drive the respective algorithm. In this representation, a graph node represents a surface patch and a link represents the adjacency between two patches. The attributes assigned to nodes are moment invariants of the corresponding face for polyhedral objects. For range images, the Gaussian curvature is used as a segmentation criterion for providing symbolic shape attributes. Identification is achieved by an efficient graph-matching algorithm used to match the graph obtained from the data to a subgraph of one of the model graphs stored in the commputer memory.
I.E.

TYPE 1/4/66
Quest Accession Number : 89N19165
89N19165 NASA STAR Conference Paper Issue 11
Automatic shape parametrisation in machine vision
(AA)LEAVERS, V. F.; (AB)BOYCE, J. F.
Kings Coll., London (England). (KV801251) Dept. of Physics.
In Optical Society of America, Topical Meeting on Machine Vision p 93-96 (SEE N89-19145 11-74) 800000 p. 4 In: EN (English) Avail: Issuing Activity p.1591

A fully automatic, computational method is proposed which will allow the extraction of parameters characterising various shape primitives in the image space from their shape indicative distributions in a two dimensional parametric transform space. It is known that the parametric transformation of image data allows space characterising parameters to be determined. The usefulness of such methods is always qualified by the erroneous assumption that its drawbacks are an exponential growth of memory space requirement and computational cost as a function of the number of parameters. A general method is presented which uses the definition of a Radon transform as a means of defining a two dimensional transform space in which information about shape primitives may be simultaneously encoded. Examples are given illustrating how the shape indicative distributions within the transform space may be deduced. The results show that each set of coded information is transparent to any other and that each shape indicative distribution may be located using a convolution mask peculiar to that distribution.
Author

TYPE 1/4/67
Quest Accession Number : 89N17426
89N17426# NASA STAR Technical Report Issue 09
Temporal pattern recognition
(AA)PRIEBE, CAREY E.; (AB)SUNG, CHEN-HAN
(AB)(San Diego State Univ., CA.)
Naval Ocean Systems Center, San Diego, CA. (NR473487) Architecture and Applied Research Branch.
AD-A200090; NOSC/TD-1332 Prepared in cooperation with California Univ., San Diego, La Jolla 880900 p. 7 In: EN (English) Avail: NTIS HC A02/MF A01 p.1285

A self-organizing network architecture for the learning of recognition codes corresponding to temporal patterns is described. The problem presents itself in many real-world situations. In any non-trivial environment in which a proposed system will function the spectre of temporal information (information coming into the system over a period of time) is evident. In many cases it is not sufficient to process the information independent of its relative time-order. Disciplines as diverse as speech recognition, robotics and data fusion/situation analysis require that temporal aspect of the data be considered. In temporal environments such as these the information lost when using a non-temporal approach can be prohibitive. This approach is formulated to make use of this important temporal information. The network described takes as its input individual incoming events. Sequences of these events (letters, phonemes, or, more abstractly, object sightings in a vision system), received by the system over time are categorized as specific sequences by the temporal system. The Temporal system produces Gaussian classifications that represent the statistics of the temporal data, and the system uses a noisy environment, giving as output a Gaussian distance from the stored sequence, thus providing an analog measure of closeness of fit to currently known patterns.
GRA

TYPE 1/4/68
Quest Accession Number : 89N17236
89N17236# NASA STAR Technical Report Issue 09
3-D vision techniques for autonomous vehicles
(AA)HEBERT, MARTIAL; (AB)KANADE, TAKEO; (AC)KWEON, INSO
Carnegie-Mellon Univ., Pittsburgh, PA. (CH188052) Robotics Inst.
AD-A199643; CMU-RI-TR-88-12 DACA76-85-C-0003; NSF DCR-86-04199; ARPA ORDER 5351 880800 p. 68 In: EN (English) Avail: NTIS HC A04/MF A01 p.1252

A mobile robot needs an internal representation of its environment in order to accomplish its mission. Building such a representation involves transforming raw data from sensors into a meaningful geometric representation. In this paper, we introduce techniques for building terrain representations from range data for an outdoor mobile robot. We introduce three levels of representations that correspond to levels of planning: obstacle maps, terrain patches, and high resolution elevation maps. Since terrain representations from individual locations are not sufficient for many navigation tasks, we also introduce techniques for combining multiple maps. Combining maps may be achieved either by using features or the raw elevation data. Finally, we introduce algorithms for combining 3-D descriptions with descriptions from other sensors, such as color cameras. We examine the need for this type of sensor fusion when some semantic information has to be extracted from an observed scene and provide an example application of outdoor scene analysis. Many of the techniques presented in this paper have been tested in the field on three mobile robot systems developed at CMU.
GRA

TYPE 1/4/69
Quest Accession Number : 89A14255
89A14255 NASA IAA Journal Article Issue 03
Parallel architectures for vision
(AA)MARESCA, MASSIMO; (AB)LAVIN, MARK A.; (AC)LI, HUNGWEN
(AA)(Genova, Universita, Genoa, Italy); (AB)(IBM Thomas J. Watson Research Center, Yorktown Heights, NY); (AC)(IBM Almaden Research Center, San Jose, CA)
IEEE, Proceedings (ISSN 0018-9219), vol. 76, Aug. 1988, p. 970-981. IBM-supported research. 880800 p. 12 refs 103 In: EN (English) p.383

Options are examined that drive the design of a vision-oriented computer, beginning with the analysis of the basic vision computation and communication requirements. The classical taxonomy is briefly reviewed for parallel computers, based on the multiplicity of the instruction and data stream. A recently proposed criterion, the degree of autonomy of each processor, is applied to further classify fine-grain SIMD (single-instruction, multiple-data-stream) massively parallel computers. Three types of processor autonomy, namely, operation autonomy, addressing autonomy, and connection autonomy, are identified. For each type, the basic definition is given and some examples shown. The concept of connection autonomy, which is believed to be the key point in the development of massively parallel architectures for vision, is presented. Two examples are shown of parallel computers featuring different types of connection autonomy-the Connection Machine and the Polymorphic-Torus-and their cost and benefits are compared.
I.E.

TYPE 1/4/70
Quest Accession Number : 89N13222
89N13222# NASA STAR Technical Report Issue 04
Adaptive machine vision / Annual Report
(AA)STONER, WILLIAM W.; (AB)BRILL, MICHAEL H.; (AC)BERGERON, DOREEN W.
Science Applications International Corp., Billerica, Mass. (SD705905)
AD-A197039 N00014-86-C-0601 880308 p. 91 In: EN (English) Avail: NTIS HC A05/MF A01 p.552

The mission of the Strategic Defense Initiative is to develop defenses against threatening ballistic missiles. There are four distinct phases to the SDI defense; boost, post boost, midcourse and terminal. In each of these phases, one or more machine vision functions are required, such as pattern recognition, stereo image fusion, clutter rejection and discrimination. In this document the SDI missions of coarse track, stereo track and discrimination are examined from the point of view of a machine vision system.
GRA

TYPE 1/4/71
Quest Accession Number : 88A42656
88A42656 NASA IAA Conference Paper Issue 17
Video road-following for the autonomous land vehicle
(AA)TURK, MATTHEW A.; (AB)MORGENTHALER, DAVID G.; (AC)GREMBAN, KEITH D.; (AD)MARRA, MARTIN
(AD)(Martin Marietta Corp., Denver, CO)
DACA76-84-C-0005 IN: 1987 IEEE International Conference on Robotics and Automation, Raleigh, NC, Mar. 31-Apr. 3, 1987, Proceedings. Volume 1 (A88-42626 17-63). Washington, DC, IEEE Computer Society Press, 1987, p. 273-280. 870000 p. 8 refs 15 In: EN (English) p.2922

A description is given of the vision system for Alvin, the Autonomous Land Vehicle, addressing in particular the task of road-following. The system builds symbolic descriptions of the road and obstacle boundaries using both video and range sensors. Road segmentation methods are described for video-based road-following, along with approaches to boundary extraction and the transformation of boundaries in the image plane into a vehicle-centered three-dimensional scene model. Alvin has performed public road-following demonstrations, traveling distances up to 4.5 km at speeds up to 20 km/hr along a paved road, equipped with an RGB video camera with pan/tilt control and a laser range scanner.
I.E.

TYPE 1/4/72
Quest Accession Number : 88A42649
88A42649 NASA IAA Conference Paper Issue 17
Structure and motion from two noisy perspective views (for mobile robot navigation)
(AA)TOSCANI, G.; (AB)FAUGERAS, O. D.
(AB)(Institut National de Recherche en Informatique et en Automatique, Le Chesnay, France)
IN: 1987 IEEE International Conference on Robotics and Automation, Raleigh, NC, Mar. 31-Apr. 3, 1987, Proceedings. Volume 1 (A88-42626 17-63). Washington, DC, IEEE Computer Society Press, 1987, p. 221-227. 870000 p. 7 refs 26 In: EN (English) p.2922

An acute problem of determining the motion from two perspective views has to be solved in order to make mobile robot navigation work. Structure from motion is needed in many applications including monitoring dynamic industrial processes and image processing. It is known that existing techniques for motion estimation perform poorly on real images, when the image-point feature are noisy. The authors describe robust techniques to recover structure and movement from noisy images. Closed-form solutions are derived for the case of general three-dimensional motion. These solutions are used as initial estimates for another technique, called reconstruction and reprojection. The authors also present a solution for the case of planar motion, which is the case of a mobile robot moving over a flat surface. These techniques have been tested on synthetic as well as real images and the test results are described and compared with an improved version of the Longuet-Higgins technique.
I.E.

TYPE 1/4/73
Quest Accession Number : 88A36311
88A36311* NASA IAA Conference Paper Issue 14
Real-time model-based vision system for object acquisition and tracking
(AA)WILCOX, BRIAN; (AB)GENNERY, DONALD B.; (AC)BON, BRUCE; (AD)LITWIN, TODD
(AD)(California Institute of Technology, Jet Propulsion Laboratory, Pasadena)
Jet Propulsion Lab., California Inst. of Tech., Pasadena. (JJ574450)
IN: Optical and digital pattern recognition; Proceedings of the Meeting, Los Angeles, CA, Jan. 13-15, 1987 (A88-36301 14-63). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1987, p. 276-281. 870000 p. 6 refs 9 In: EN (English) p.2278

A machine vision system is described which is designed to acquire and track polyhedral objects moving and rotating in space by means of two or more cameras, programmable image-processing hardware, and a general-purpose computer for high-level functions. The image-processing hardware is capable of performing a large variety of operations on images and on image-like arrays of data. Acquisition utilizes image locations and velocities of the features extracted by the image-processing hardware to determine the three-dimensional position, orientation, velocity, and angular velocity of the object. Tracking correlates edges detected in the current image with edge locations predicted from an internal model of the object and its motion, continually updating velocity information to predict where edges should appear in future frames. With some 10 frames processed per second, real-time tracking is possible.
V.L.

TYPE 1/4/74
Quest Accession Number : 88A35988
88A35988 NASA IAA Meeting Paper Issue 14
Image understanding and the man-machine interface; Proceedings of the Meeting, Los Angeles, CA, Jan. 15, 16, 1987
(AA)PEARSON, JAMES J.; (AB)BARRETT, EAMON
(AA)ED.; (AB)ED.
(AB)(Lockheed Missiles and Space Co., Inc., Sunnyvale, CA)
SPIE-758 Meeting sponsored by SPIE. Bellingham, WA, Society of Photo-Optical Instrumentation Engineers (SPIE Proceedings. Volume 758), 1987, 191 p. For individual items see A88-35989 to A88-35993. 870000 p. 191 In: EN (English) Members, $33.; nonmembers, $43 p.2329

Various papers concerning image understanding concepts and models, image understanding systems and applications, advanced digital processors and software tools, and advanced man-machine interfaces are presented. Individual topics addressed include: prospects for artificial neural systems in vision computations, optical bidirectional associative memories, model-based approaches for some image understanding problems, strategic computing computer vision, organizing the landscape for image understanding purposes, issues in image registration, and smoothing splines with discontinuities for image analysis. Also considered are: connection machine vision applications, parallel processor for dynamic image processing, LISP-based PC vision workstation, separation of form perception and stereopsis, automating knowledge acquisition for aerial image interpretation, toward an ideal three-dimensional CAD system, and object-oriented image analysis.
C.D.

TYPE 1/4/75
Quest Accession Number : 88A34852
88A34852 NASA IAA Conference Paper Issue 13
Vision-based road following in the autonomous land vehicle
(AA)SEIDA, STEVEN; (AB)MORGENTHALER, DAVID G.; (AC)PODLASECK, MARK; (AD)DOUGLAS, BOB; (AE)MCSWAIN, JON
(AE)(Martin Marietta Corp., Denver, CO)
DACA76-84-C-0005 IN: IEEE Conference on Decision and Control, 26th, Los Angeles, CA, Dec. 9-11, 1987, Proceedings. Volume 3 (A88-34702 13-63). New York, Institute of Electrical and Electronics Engineers, Inc., 1987, p. 1814-1819. 870000 p. 6 In: EN (English) p.2164

The navigation system for Martin Marietta Denver Aerospace's autonomous land vehicle project receives information from the vision system about road boundaries and obstacle locations. This information is used in an optimization equation to create trajectory points on the road. The operation and the algorithms of the vision subsystem are described briefly. The operation and algorithms of the navigation, or reasoning, subsystem is then considered. An obstacle-avoidance navigator is presented.
I.E.

TYPE 1/4/76
Quest Accession Number : 88A29425
88A29425 NASA IAA Book/Monograph Issue 11
Pattern recognition and natural language understanding by a computer (Russian book)
Paspoznavanie obrazov i mashinnoe ponimanie estestvennogo iazyka
(AA)FAIN, VITALII SAMOILOVICH
Moscow, Izdatel'stvo Nauka, 1987, 176 p. In Russian. 870000 p. 176 refs 68 In: RU (Russian) p.0

An approach to the problem of the interaction in the system user-computer-production (or control) environment is presented for the case of a stationary environment. It is shown that problems in a number of areas of computer science, such as artificial intelligence, natural language understanding, and half-tone computer vision, are reduced in the case of stationary environments to pattern recognition problems, which in many cases provides for more efficient solutions. Data on the practical applications of the methods described here are presented.
V.L.

TYPE 1/4/77
Quest Accession Number : 88A22798
88A22798* NASA IAA Conference Paper Issue 07
Applications of artificial intelligence to rotorcraft
(AA)ABBOTT, KATHY H.
(AA)(NASA, Langley Research Center, Hampton, VA)
National Aeronautics and Space Administration. Langley Research Center, Hampton, Va. (ND210491)
IN: AHS, Annual Forum, 43rd, Saint Louis, MO, May 18-20, 1987, Proceedings. Volume 2 (A88-22726 07-01). Alexandria, VA, American Helicopter Society, 1987, p. 1011-1019. 870000 p. 9 refs 17 In: EN (English) p.1084

The application of AI technology may have significant potential payoff for rotorcraft. In the near term, the status of the technology will limit its applicability to decision aids rather than total automation. The specific application areas are categorized into onboard and nonflight aids. The onboard applications include: fault monitoring, diagnosis, and reconfiguration; mission and tactics planning; situation assessment; navigation aids, especially in nap-of-the-earth flight; and adaptive man-machine interfaces. The nonflight applications include training and maintenance diagnostics.
Author

TYPE 1/4/78
Quest Accession Number : 88A20288
88A20288* NASA IAA Journal Article Issue 06
The cortex transform - Rapid computation of simulated neural images
(AA)WATSON, ANDREW B.
(AA)(NASA, Ames Research Center, Moffett Field, CA)
National Aeronautics and Space Administration. Ames Research Center, Moffett Field, Calif. (NC473657)
Computer Vision, Graphics, and Image Processing (ISSN 0734-189X), vol. 39, Sept. 1987, p. 311-327. 870900 p. 17 refs 31 In: EN (English) p. 852

With a goal of providing means for accelerating the image processing, machine vision, and testing of human vision models, an image transform was designed, which makes it possible to map an image into a set of images that vary in resolution and orientation. Each pixel in the output may be regarded as the simulated response of a neuron in human visual cortex. The transform is amenable to a number of shortcuts that greatly reduce the amount of computation.
I.S.

TYPE 1/4/79
Quest Accession Number : 88N15464
88N15464# NASA STAR Technical Report Issue 07
Proceedings of Image Understanding Workshop, volume 2 / Annual Report, Dec. 1985 - Feb. 1987
(AA)BAUMANN, LEE S.
Science Applications International Corp., McLean, Va. (SD708880)
AD-A186104 N00014-86-C-0700; ARPA ORDER 5605 870200 p. 613 Workshop held in Los Angeles, Calif., 23-25 Feb. 1987 In: EN (English) Avail: NTIS HC A99/MF A01 p.902

The partial contents of the Proceedings of the Image Understanding Workshop are as follows: Guiding an Autonomous Land Vehicle Using Knowledge-Based Landmark Recognition; The Image Understanding Architecture; Initial Hypothesis Formation in Image Understanding Using an Automatically Generated Knowledge Base; What Is a Degenerate View; Recognizing Unexpected Objects: A Proposed Approach; Minimization of the Quantization Error in Camera Calibration; Tracing Finite Motions Without Correspondence; The Formation of Partial 3D Models from 2D Projections - An Application of Algebraic Reasoning; Qualitative Information in the Optical Flow; Detecting Blobs as Textons in Natural Images; and Parallel Optical Flow Computation.
GRA

TYPE 1/4/80
Quest Accession Number : 88A13400
88A13400 NASA IAA Conference Paper Issue 03
An emergency command recognizer for voiced system control
(AA)WETTERLIND, P.; (AB)JOHNSTON, WAYMON L.
(AA)(California State University, Bakersfield); (AB)(Texas A & M University, College Station)
IN: SAFE Association, Annual Symposium, 24th, San Antonio, TX, Dec. 11-13, 1986, Proceedings (A88-13376 03-54). Newhall, CA, SAFE Association, 1987, p. 181-184. 870000 p. 4 refs 16 In: EN (English) p.313

An algorithm for accepting speaker-independent voiced input, aimed especially at accommodating emergency acoustic commands, is described. The algorithm is directed toward correctly identifying commands from speaker-independent acoustic input using machine recognition of common, standarized phonemic input, using these recognized sounds to reconstruct entire words and phrases. Speaker-dependent phonemes are not used during the command reconstruction process, so that speaker idiosyncracies are accommodated. Machine recognition extends to voice pitch and emotional tension characteristics.
C.D.

TYPE 1/4/81
Quest Accession Number : 87A42734
87A42734 NASA IAA Journal Article Issue 19
Associative network applications to low-level machine vision
(AA)OYSTER, J. MICHAEL; (AB)VICUNA, FERNANDO; (AC)BROADWELL, WALTER
(AA)(Hughes Image and Signal Processing Laboratory, El Segundo, CA); (AC)(IBM Los Angeles Scientific Center, CA)
Applied Optics (ISSN 0003-6935), vol. 26, May 15, 1987, p. 1919-1926. 870515 p. 8 refs 15 In: EN (English) p.3064

This paper explores the application of a parallel computational model, the associative network, to problems in low-level machine vision. A formal description of the associative network model is presented. Then associative networks are designed for performing Boolean functions, edge detection, and the Hough transform. Associative networks feature very flexible processor interconnections. The flexible processor interconnections allow for parallelism in the algorithm design beyond what is feasible in other parallel computational models. This work demonstrates that image processing transformations, often too slow to be practical on a sequential machine, can be executed rapidly with associative networks.
Author

TYPE 1/4/83
Quest Accession Number : 87A31115
87A31115# NASA IAA Preprint Issue 12
Computational themes in applications of visual perception
(AA)JAIN, RAMESH; (AB)SCHUNCK, BRIAN G.; (AC)WEYMOUTH, TERRY
(AC)(Michigan, University, Ann Arbor)
AIAA PAPER 87-1674 AIAA, NASA, and USAF, Symposium on Automation, Robotics and Advanced Computing for the National Space Program, 2nd, Arlington, VA, Mar. 9-11, 1987. 10 p. 870300 p. 10 refs 47 In: EN (English) p.1842

The paper summarizes the current research in the Computer Vision Research Laboratory at the University of Michigan. The laboratory concentrates on developing generic vision algorithms for industrial applications. Generic vision algorithms can be applied to a wide variety of inspection problems. The paper includes a discussion of the current state of the machine vision industry and provides recommendations for improving the transfer of vision technology from research to practice.
Author

TYPE 1/4/84
Quest Accession Number : 87N24891
87N24891# NASA STAR Technical Report Issue 18
Representation and control in the interpretation of complex scenes / Final Scientific Report, 1 Oct. 1984 - 30 Sep. 1985
(AA)HANSON, ALLEN R.; (AB)RISEMAN, EDWARD M.
Massachusetts Univ., Amherst. (MK149394) Dept. of Computer and Information Science.
AD-A179116; AFOSR-87-0301TR F49620-83-C-0099; AF-AFOSR-0005-85 870000 p. 61 In: EN (English) Avail: NTIS HC A04/MF A01 p.0

The system being developed, called VISIONS, is an investigation into issues of general computer vision. The goal is to provide an analysis of color images of outdoor scenes, from segmentation through symbolic interpretation. The output of the system is intended to be a symbolic representation of the three-dimensional world depicted in the two-dimensional image, including the naming of objects, their placement in three-dimensional space, and the ability to predict from this representation the rough appearance of the scene from other points of view. The emphasis of the research over the past year has been on three issues critical to furthering our understanding of machine vision. The first area addresses the issue of image segmentation and the failure of recent research to provide robust procedures applicable to complex imagery. The second area focusses on the use of domain knowledge in the interpretation task. The third area focusses on techniques for controlling the use of system resources during interpretation and on ways of resolving conflicting partial interpretations.
GRA

TYPE 1/4/85
Quest Accession Number : 87N23017
87N23017# NASA STAR Technical Report Issue 16
Computer vision research and its applications to automated cartography / Final Report, 11 Jun. 1984 - 31 May 1986
(AA)FISCHLER, MARTIN A.
SRI International Corp., Menlo Park, Calif. (SY423852)
AD-A178815 MDA903-83-C-0027; ARPA ORDER 5355 870300 p. 19 In: EN (English) Avail: NTIS HC A02/MF A01 p.0

The SRI Image Understanding program is a broad effort spanning the entire range of machine vision research. Three major concerns are: (1) to develop a computational description of the physics and mathematics of the vision process; (2) to develop a knowledge-based framework for interpreting sensed (imaged) data; and (3) to develop a machine-based environment for effective experimentation, demonstration, and evaluation of our theoretical results, as well as providing a vehicle for technology transfer. This final report summarizes progress in these and related areas.
Author (GRA)

TYPE 1/4/86
Quest Accession Number : 87N20138
87N20138# NASA STAR Technical Report Issue 12
Domain-dependent reasoning for visual navigation of roadways
(AA)LEMOIGNE, JACQUELINE
Maryland Univ., College Park. (MI915766) Center for Automation Research.
AD-A174786; CAR-TR-230; CS-TR-1721; ETL-0445 DACA76-84-C-0004 861000 p. 36 In: EN (English) Avail: NTIS HC A03/MF A01 p.1701

A Visual Navigation System for Autonomous Land Vehicles includes several modules, among them a Knowledge-based Reasoning Module that is described in this report. This module utilizes domain-dependent knowledge (in this case, road knowledge) in order to analyze and label the visual features extracted from the imagery by the Image Processing Module. Knowledge and general hypotheses are given in Section 2. The Reasoning Module itself is described in Section 3 and results are presented in Section 4. Finally, some conclusions are proposed in Section 5.
GRA

TYPE 1/4/87
Quest Accession Number : 86N32751
86N32751# NASA STAR Technical Report Issue 24
Biological visual systems structures for machine vision applied to robotics / Final Report, 15 Sep. 1984 - 31 Jan. 1986
(AA)INIGO, R. M.; (AB)HSIN, C. H.; (AC)NARATHONG, C.; (AD)MCVEY, E. S.; (AE)MINNIX, J. I.
Virginia Univ., Charlottesville. (V3127208) Dept. of Electrical Engineering.
AD-A168521; UVA/525647/EE86/101; AFOSR-86-0282TR AF-AFOSR-0349-84 860200 p. 333 In: EN (English) Avail: NTIS HC A15/MF A01 p.3737

This report describes the research on a biological visual system (BVS) based sensor with possible applications to robotics and automation. The report covers the following subjects: sensor configuration; edge detection modeling for the human visual system and edge detection using the BVS sensor. qualitative motion detection using the BVS; target tracking algorithms for the BVS; and microsaccadic eye movement in the human visual system (HVS).
GRA

TYPE 1/4/88
Quest Accession Number : 86N30333
86N30333# NASA STAR Technical Report Issue 21
Novel architectures for image processing based on computer simulation and psychophysical studies of human visual cortex / Final Report, 15 Apr. 1983 - 15 Apr. 1985
(AA)SCHWARTZ, E. L.
New York Univ. Medical Center. (N0098273)
AD-A166222; AFOSR-86-0059TR F49620-83-C-0108 860102 p. 96 In: EN (English) Avail: NTIS HC A05/MF A01 p.3353

This final report consists of two parts. The first part is a computer simulation of the functional architecture of the visual cortex, and an examination of the possible significance that this architecture may have for understanding both human visual computation and machine vision. The second part of this report is a psychophysical investigation of human shape perception in terms of boundary descriptors of curvature.
GRA

TYPE 1/4/89
Quest Accession Number : 86N29120
86N29120# NASA STAR Technical Report Issue 20
Exploiting sequential phonetic constraints in recognizing spoken words
(AA)HUTTENLOCHER, D. P.
Massachusetts Inst. of Tech., Cambridge. (MJ700802) Artificial Intelligence Lab.
AD-A165913; AI-M-867 N00014-80-C-0505 851000 p. 28 In: EN (English) Avail: NTIS HC A03/MF A01 p.3158

Machine recognition of spoken language requires developing more robust recognition algorithms. A recent study by Shipman and Zue suggest using partial descriptions of speech sounds to eliminate all but a handful of word candidates from a large lexicon. The current paper extends their work by investigating the power of partial phonetic descriptions for developing recognition algorithms. First, we demonstrate that sequences of manner of articulation classes are more reliable and provide more constraint than certain other classes. Alone these results are of limited utility, due to the high degree of variability in natural speech. This variability is not uniform however, as most modifications and deletions occur in unstressed syllables. Comparing the relative constraint provided by sounds in stressed versus unstressed syllables, we discover that the stressed syllables provide substantially more constraint. This indicates that recognition algorithms can be made more robust by exploiting the manner of articulation information in stressed syllables.
GRA

TYPE 1/4/90
Quest Accession Number : 86N24536
86N24536*# NASA STAR Conference Paper Issue 14
Machine vision and the OMV
(AA)MCANULTY, M. A.
Alabama Univ., Birmingham. (AM538929) Dept. of Computer and Information Science.
In NASA. Marshall Space Flight Center Research Reports: 1985 NASA/ASEE Summer Faculty Fellowship Program 24 p (SEE N86-24507 14-80) 860100 p. 24 refs 0 In: EN (English) Avail.: NTIS HC A99/MF E04 p. 2388

The orbital Maneuvering Vehicle (OMV) is intended to close with orbiting targets for relocation or servicing. It will be controlled via video signals and thruster activation based upon Earth or space station directives. A human operator is squarely in the middle of the control loop for close work. Without directly addressing future, more autonomous versions of a remote servicer, several techniques that will doubtless be important in a future increase of autonomy also have some direct application to the current situation, particularly in the area of image enhancement and predictive analysis. Several techniques are presentet, and some few have been implemented, which support a machine vision capability proposed to be adequate for detection, recognition, and tracking. Once feasibly implemented, they must then be further modified to operate together in real time. This may be achieved by two courses, the use of an array processor and some initial steps toward data reduction. The methodology or adapting to a vector architecture is discussed in preliminary form, and a highly tentative rationale for data reduction at the front end is also discussed. As a by-product, a working implementation of the most advanced graphic display technique, ray-casting, is described. Author

TYPE 1/4/91
Quest Accession Number : 86N20008
86N20008# NASA STAR Technical Report Issue 10
Hierarchical multisensor image understanding / Final Report, Oct. 1983 - Aug. 1985
(AA)AGGARWAL, R. K.; (AB)BAZAKOS, M.; (AC)BUDENSKE, J.; (AD)KIM, Y.; (AE)MADER, S.
Honeywell Systems and Research Center, Minneapolis, Minn. (HY989092)
AD-A160324; AFOSR-85-0801TR F49620-83-C-0134 850800 p. 129 In: EN (English) Avail.: NTIS HC A07/MF A01 p.1651

This report describes the research results on Honeywell's Hierarchical Multisensor Image Understanding program. Honeywell is developing a unified framework for the different hierarchical levels of image processing such as segmentation, detection, classification, and identification of outdoor scenes and across different sensor modalities such as millimeter wave, infrared, and visible. Current activities on the project are reviewed under the following headings: (1) A Survey of Multisource Information Fusion Systems; (2) The Role of Structure in Human and Machine Perception; (3) A Knowledge Based Image Segmentation System; (4) The Use of Optical Flow as a Depth Cue in Scene Analysis; and (5) Belief Maintenance for A Fuzzy Reasoning System.
GRA

TYPE 1/4/92
Quest Accession Number : 86N19085
86N19085# NASA STAR Technical Report Issue 09
Computing visible-surface representations
(AA)TERZOPOULOS, D.
Massachusetts Inst. of Tech., Cambridge. (MJ700802) Artificial Intelligence Lab.
AD-A160602; AI-M-800 N00014-75-C-0643 850300 p. 64 In: EN (English) Avail.: NTIS HC A04/MF A01 p.1494

The computational framework offered in this paper addresses, in a unified way, certain visual information processing tasks involved in the representation of visible surfaces. Particular emphasis is placed on utilizing highly parallel, cooperative processing to integrate surface shape information over multiple visual sources, to fuse it across a multiplicity of spatial resolutions, and to maintain the global consistency of the resulting distributed shape representations. The issues are first investigated in terms of a surface reconstruction model rooted in mathematical physics. This formal analysis is augmented by an empirical study of the resulting algorithms, which feature multiresolution iterative processing within hierarchical surface shape representations. The approach is guided by current knowledge of how humans perceive visible surfaces, while applications in machine vision provide a testbed for the algorithms. GRA

TYPE 1/4/93
Quest Accession Number : 86A18651
86A18651 NASA IAA Journal Article Issue 06
Machine perception of visual motion
(AA)BUXTON, B. F.; (AB)MURRAY, D. W.; (AC)BUXTON, H.; (AD)WILLIAMS, N. S.
(AB)(General Electric Co., PLC, Research Laboratories, Wembley, England) ; (AD)(Queen Mary College, London, England)
GEC Journal of Research (ISSN 0264-9187), vol. 3, no. 3, 1985, p. 145-161. Research supported by the Ministry of Defence (Procurement Executive). 850000 p. 17 refs 66 In: EN (English) p.0

An attempt at devising a system for using visual motion to obtain three-dimensional information at the level of Marr's (1982) two-and-one-half-dimensional sketch is described. The algorithm proposed can be implemented efficiently on an SIMD processor array and in the ideal case of a direct 1:1 mapping of the image pixels onto the processor array run at speeds approaching real-time video frame rates. The processing scheme has a potential for performing a multiple regression by introducing new surface and motion parameters to explain variations in the visual motion data and thus can be adapted for a segmentation procedure based on the description of the visible surfaces.
I.S.

TYPE 1/4/94
Quest Accession Number : 86A17019
86A17019 NASA IAA Meeting Paper Issue 05
Pattern recognition and artificial intelligence; French Congress, 4th, Paris, France, January 25-27, 1984, Lectures. Volumes 1 & 2
Reconnaissance des formes et intelligence artificielle; Congres Francais, 4th, Paris, France, January 25-27, 1984, Conferences. Volumes 1 & 2
Congress sponsored by the Ministere de l'Industrie et de la Recherche, Association Nationale du Logiciel, and International Association for Pattern Recognition. Le Chesnay, France, Institut National de Recherche en Informatique et en Automatique, 1984. Vol. 1, 579 p.; vol. 2, 524 p. In French. For individual items see A86-17020 to A86-17024. 840000 p. 1103 In: FR (French) p.0

Two broad topics are addressed: (1) the processing, analysis, and understanding of images; and (2) the analysis and understanding of words. Particular consideration is given to image segmentation; scene analysis; the representation and analysis of two- and three-dimensional forms; industrial vision; and special architectures. Attention is also given to the understanding of natural languages, programming languages, learning theory, and expert systems.
B.J.

TYPE 1/4/95
Quest Accession Number : 85N35634
85N35634# NASA STAR Issue 24
Selected publications in image understanding and computer vision from 1974 to 1983
(AA)VERLY, J. G.
Lincoln Lab., Mass. Inst. of Tech., Lexington. (LQ054005)
AD-A156196; TR-716; ESD-TR-85-180 F19628-85-C-0002; ARPA ORDER 4881 850418 p. 100 In: EN (English) Avail.: NTIS HC A05/MF A01 p.4136

A list of selected publications in image understanding and computer vision is presented. The list was compiled as part of work for the DARPA-sponsored Autonomous IR Sensor Technology program, and the choice of references was directly influenced by the needs of that program. Therefore, emphasis was placed on theories, techniques, and systems for interpreting complex imagery; the more classical fields of image processing, e.g., filtering, enhancement, restoration, coding, and reconstruction, were not included. The topics of edge detection and region segmentation as well as the well-known scene analysis problems of shape recognition from stereo, shading, texture, and motion were also excluded. The bibliography covers the last decade (1974-1983) and is based on the yearly surveys published by A. Rosenfeld in the Journal initially called Computer Graphics and Image Processing (CGIP) and now Computer Vision, Graphics, and Image Processing (CVGIP).
GRA

TYPE 1/4/96
Quest Accession Number : 85A24997
85A24997 NASA IAA Conference Paper Issue 10
Optics for machine vision
(AA)STRAND, T. C.
(AA)(IBM Research Laboratory, San Jose, CA)
IN: Optical computing; Proceedings of the Meeting, Los Angeles, CA, January 24, 25, 1984 (A85-24990 10-60). Bellingham, WA, SPIE - The International Society for Optical Engineering, 1984, p. 86-93. 840000 p. 8 refs 23 In: EN (English) p.0

Current developments in manufacturing technologies have caused a demand for automated inspection and assembly tools. A key requirement regarding such tools is related to machine vision. The term 'machine vision', as used in this discussion, includes any automated acquisition of information via optical sensors. The primary information to be sought with vision systems is spatial information. The normal detection scheme provides all but one of the generally desired variables. The variable not provided is the longitudinal position variable. Information regarding this variable is called 'range information'. The present investigation is mainly concerned with the means of acquiring the range variable. Attention is given to geometric range measurement techniques, time-of-flight range measurement techniques, interferometric techniques, and diffraction range measurement techniques.
G.R.

TYPE 1/4/97
Quest Accession Number : 84A44308
84A44308 NASA IAA Journal Article Issue 21
Parallel processing in machine vision
(AA)STERNBERG, S. R.
(AA)(Machine Vision International, Ann Arbor, MI)
Robotica (ISSN 0263-5747), vol. 2, Jan. 1984, p. 33-40. 840100 p. 8 refs 21 In: EN (English) p.3102

Machine vision systems incorporating highly parallel processor architectures are reviewed. A new processor architecture, the image flow computer, is presented in detail. An interactive image processing programming language based on mathematical morphology is then presented. A detailed example of the use of the system for the inspection of a particular industrial part concludes the presentation.
Author

TYPE 1/4/98
Quest Accession Number : 84N23123
84N23123# NASA STAR Technical Report Issue 13
Machine vision: Three generations of commercial systems / Interim Report
(AA)CROWLEY, J. L.
Carnegie-Mellon Univ., Pittsburgh, Pa. (CH188052) Robotics Inst.
AD-A139037; CMU-RI-TR-84-1 840125 p. 40 In: EN (English) Avail.: NTIS HC A03/MF A01 p.2024

Since 1980, machine vision systems for industrial application have enjoyed a rapidly expanding market. The first generation machines are two-dimensional binary vision systems, patterned after the SRI Vision Module. These systems will soon be joined by a second generation, based on edges description techniques. Both the first and second generation systems are pattern recognition machines. Research in machine vision is leading towards vision systems that will be able to dynamically model the three-dimensional (3-D) surfaces in a scene. This research will lead to a third generation of vision systems which will provide a dramatic increase in capabilities over the first two generations. This article describes these three generations of vision systems. The algorithms, data structures, and hardware architecture are presented for binary vision systems and edge-based systems. A framework is presented for the research problems which must be solved before a commercial vision system can be produced based on dynamic 3-D Scene analysis techniques.
Author (GRA)

TYPE 1/4/99
Quest Accession Number : 83A44078
83A44078 NASA IAA Journal Article Issue 21
Machine vision for robotics
(AA)CORBY, N. R., JR.
(AA)(GE Corporate Research and Development Center, Schenectady, NY)
IEEE Transactions on Industrial Electronics (ISSN 0278-0046), vol. IE-30, Aug. 1983, p. 282-291. 830800 p. 10 refs 14 In: EN (English) p.3135

When applied to robotic tasks, computer or machine vision involves time and space interactions among manipulators, tools, and objects in the work space. Such vision must ultimately be three-dimensional. Attention is given to fundamental characteristics of machine vision processing for binary, grey, and fully three-dimensional cases, and the architectures and control structures for several different vision processing approaches are explored.
O.C.

TYPE 1/4/100
Quest Accession Number : 83A13450
83A13450 NASA IAA Meeting Paper Issue 03
Perceptual capabilities, ambiguities, and artifacts in man and machine
(AA)GINSBURG, A. P.
(AA)(USAF, Aviation Vision Laboratory, Wright-Patterson AFB, OH)
AD-A109864; AFAMRL-TR-81-142 In: 3-D machine perception; Proceedings of the Conference, Washington, DC, April 23, 24, 1981. (A83-13444 03-35) Bellingham, WA, SPIE - The International Society for Optical Engineering, 1981, p. 78-82. 810000 p. 5 refs 11 In: EN (English) p.383

Certain advances in visual science suggesting that perception may be structured from a hierarchy of filtered images are summarized. It is shown that a small numbered set of images created from filters based on biological data can provide a rich array of information about any object: contrast, general form, identification, textures and edges. It is contended that machine perception will require similar parallel processing of an array of filtered images if human-like visual performance is required. Such visual problems as certain visual illusion, multistable objects, and masking are analyzed in terms of the limitations of biological filtering. Machine solutions to these problems are then discussed.
C.R.

TYPE 1/4/101
Quest Accession Number : 83A13444
83A13444 NASA IAA Meeting Paper Issue 03
3-D machine perception; Proceedings of the Conference, Washington, DC, April 23, 24, 1981
(AA)ALTSCHULER, B. R.
(AA)(ED.)
(AA)(USAF, School of Aerospace Medicine, Brooks AFB, TX)
Conference sponsored by SPIE - The International Society for Optical Engineering. Bellingham, WA, SPIE - The International Society for Optical Engineering (SPIE Proceedings. Volume 283), 1981. 145 p. (For individual items see A83-13445 to A83-13450) 810000 p. 145 In: EN (English) MEMBERS, $31.; NONMEMBERS, $37 p.324

Topics discussed include three-dimensional surface mapping and analysis, applications and interfacing, and the three-dimensional display of internal structures. Papers are presented on coherent optical methods for applications in robot visual sensing; real-time three-dimensional vision for parts acquisition; perceptual capabilities, ambiguities, and artifacts in man and machine; and a computerized anatomy atlas of the human brain. Attention is also given to noncontact visual three-dimensional ranging devices, to the application of digital image acquisition in anthropometry, to an overview of data acquisition and processing for three-dimensional displays of internal structures, and to a three-dimensional viewing device for examining internal structure.
C.R.

TYPE 1/4/102
Quest Accession Number : 83A13353
83A13353* NASA IAA Journal Article Issue 03
Feature Identification and Location Experiment
(AA)SIVERTSON, W. E., JR.; (AB)WILSON, R. G.; (AC)BULLOCK, G. F.; (AD)SCHAPPELL, R. T.
(AC)(NASA, Langley Research Center, Hampton, VA); (AD)(Martin Marietta Aerospace, Denver, CO)
National Aeronautics and Space Administration. Langley Research Center, Hampton, Va. (ND210491)
Science, vol. 218, Dec. 3, 1982, p. 1031-1033. NASA-supported research.
821203 p. 3 refs 5 In: EN (English) p.357

The Feature Identification and Location Experiment (FILE), which was flown on the second Space Shuttle flight to test a technique for real-time, autonomous classification of water, vegetation and bare land as well as clouds, snow and ice, senses earth radiation in spectral bands centered at 0.65 and 0.85 microns. The radiance ratio classification algorithm has successfully made automatic data selection decisions. A classification image obtained on the mission is providing data needed to evaluate the FILE algorithm and overall system performance.
O.C.

TYPE 1/4/103
Quest Accession Number : 83A12880
83A12880 NASA IAA Meeting Paper Issue 02
Fast adaptive algorithms for low-level scene analysis - Applications of polar exponential grid /PEG/ representation to high-speed, scale-and-rotation invariant target segmentation
(AA)SCHENKER, P. S.; (AB)WONG, K. M.; (AC)CANDE, E. G.
(AC)(Brown University, Providence, RI)
In: Techniques and applications of image understanding; Proceedings of the Meeting, Washington, DC, April 21-23, 1981. (A83-12875 02-35) Bellingham, WA, SPIE - The International Society for Optical Engineering, 1981, p. 47-57. 810000 p. 11 refs 18 In: EN (English) p.181

This paper presents results of experimental studies in image understanding. Two experiments are discussed, one on image correlation and another on target boundary estimation. The experiments are demonstrative of polar exponential grid (PEG) representation, an approach to sensory data coding which the authors believe will facilitate problems in three-dimensional machine perception. The discussion of the image correlation experiment is largely an exposition of the PEG-representation concept and approaches to its computer implementation. The presentation of the boundary finding experiment introduces a new robust stochastic, parallel computation segmentation algorithm, the PEG-Parallel Hierarchical Ripple Filter (PEG-PHRF).
(Author)

TYPE 1/4/104
Quest Accession Number : 83A12878
83A12878 NASA IAA Meeting Paper Issue 02
Application of image understanding to automatic tactical target acquisition
(AA)HELLAND, A. R.; (AB)WILLETT, T. J.; (AC)TISDALE, G. E.
(AC)(Westinghouse Electric Corp., Systems Development Div., Baltimore, MD)
In: Techniques and applications of image understanding; Proceedings of the Meeting, Washington, DC, April 21-23, 1981. (A83-12875 02-35) Bellingham, WA, SPIE - The International Society for Optical Engineering, 1981, p. 26-31. 810000 p. 6 refs 15 In: EN (English) p.133

Real-time equipment has been developed and is now being tested for automatic recognition of targets on an individual basis. The recent use of frame-to-frame integration techniques has significantly improved the classification performance with this equipment to the point where the human interpreter can sometimes be surpassed. For some imagery, however, initial target segmentation remains unsatisfactory, causing targets to be missed, and the level of false alarms may be too high. As a result, more sophisticated image processing techniques are now being addressed which could provide a comprehensive understanding of overall image content. These include the use of such scene analysis operations as the derivation of motion vectors for passive ranging, false alarm discrimination, and detection of target motion. Additional areas of interest lie in the 'intelligent' tracking of multiple targets, and the autonomous handoff of targets between sensors. The paper discusses the evolution of these areas, and their probable impact on the target acquisition process. It also addresses their impact on hardware implementation.
(Author)

TYPE 1/4/105
Quest Accession Number : 83A11460
83A11460 NASA IAA Meeting Paper Issue 01
Symbolic pattern matching for target acquisition
(AA)NARENDRA, P. M.; (AB)GRABAU, J. J.; (AC)WESTOVER, B. L.
(AC)(Honeywell Systems and Research Center, Minneapolis, MN)
DAAK70-79-C-0114 In: Conference on Pattern Recognition and Image Processing, Dallas, TX, August 3-5, 1981, Proceedings. (A83-11409 01-63) New York, Institute of Electrical and Electronics Engineers, Inc., 1981, p. 481-486. 810000 p. 6 refs 16 In: EN (English) p.8

This paper describes a symbolic pattern matching system for autonomous target acquisition, which requires matching widely disparate views of a scene. The pattern matching system exploits both the object-to-object similarities in the two images and the consistency of configurations of candidate matches. The consistency is evaluated under a general transformation which accounts for a large difference in the sensor positions between the two views. The matching of the symbolic features between the two images is cast in a combinatorial framework. An efficient branch and bound algorithm is developed to find the best match optimizing the criterion function, which measures the goodness of a candidate match. The result of applying the pattern matching system simulation to several pairs of real infrared images are presented both to illustrate the approach and to quantify its performance.
(Author)

TYPE 1/4/106
Quest Accession Number : 82N31312
82N31312# NASA STAR Technical Report Issue 22
Flight plan filing by speech recognition / Final Report
(AA)SHOCHET, E.; (AB)QUICK, P.; (AC)DELEMARRE, L.
Federal Aviation Administration, Atlantic City, N.J. (FI751336) Technical Center.
DOT/FAA/RD-82/39; DOT/FAA/CT-81/64 FAA PROJ. 131-402-540 820700 p. 67
In: EN (English) Avail.: NTIS HC A04/MF A01 p.3080

Automatic flight plan filing by machine recognition is discusssed. The utterance recognition device (URD) was upgraded in preparation for testing the capabilities of voice input for automatic flight plan filing. The URD was modified to include more reliable components, where advisable, and a larger memory to handle the expanded vocabulary. In addition, a dialect study was conducted to determine the locations for collecting a nationally representative voice sample in order to create reference patterns capable of performing well on all American dialects. Subsequently, over 5,000 voices from 24 cities throughout the United States were collected and processed. Initial tests were conducted in which subjects filed simulated flight plans directly into the URD over the telephone. The results indicated that the prototype system, as demonstrated using the adaptation strategy for flight plan filing, has definite potential for application in Model two of the flight service automation program. Moreover, a comparison between the old and new recognition algorithms indicates that the improvement in accuracy with the new data base raises the performance of the mass weather dissemination program to a level quite satisfactory for tne general pilot population.
S.L.

TYPE 1/4/107
Quest Accession Number : 81A44700
81A44700 NASA IAA Meeting Paper Issue 21
Image processing design for autonomous acquisition of targets
(AA)BOYD, W. W.; (AB)MACPHERSON, C. A.; (AC)TAYLOR, J. L.; (AD)TASKETT, J. M.; (AE)LINEBERRY, M. C.
(AE)(Texas Instruments, Inc., Dallas, TX)
In: SOUTHEASTCON '81; Proceedings of the Region 3 Conference and Exhibit, Huntsville, AL, April 5-8, 1981. (A81-44676 21-31) Piscataway, NJ, Institute of Electrical and Electronics Engineers, Inc., 1981, p. 285-290. 810000 p. 6 In: EN (English) p.3617

Primary considerations in designing an image-processing system that can autonomcusly acquire high-value tactical targets are discussed. Attention is given to establishing requirements, and the implications of these requirements on the image-processing algorithms are analyzed. It is pointed out that through these steps, detection and acquisition times can be estimated and, hence, algorithm processing times established. The results of certain candidate algorithms that show promise of meeting mission goals are presented. The design process described takes account of the geographical and climatological features of the area of intended use. Aircraft maneuverability and human factor limits are also considered in establishing system requirements. Analysis shows the feasibility and desirability of employing the seeker and terrain features to cue the aircraft to the target.
C.R.

TYPE 1/4/108
Quest Accession Number : 81A39349
81A39349 NASA IAA Meeting Paper Issue 18
Model-based scene matching
(AA)TSENG, D. Y.; (AB)CONTI, D. K.; (AC)ECKHARDT, W. O.; (AD)OLIN, K. E. ; (AE)MCCULLOH, T. A.; (AF)NEVATIA, R.
(AD)(Hughes Research Laboratories, Malibu, CA); (AE)(Hughes Aircraft Co., Culver City, CA)
F33615-77-C-1227 In: Image processing for missile guidance; Proceedings of the Seminar, San Diego, CA, July 29-August 1, 1980. (A81-39326 18-04) Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1980, p. 225-231. 800000 p. 7 refs 5 In: EN (English) p.3068

Advanced pattern matching techniques were developed that are capable of matching complex terrain scenes for use in midcourse navigational updating of aircraft and missiles. This method utilizes key features in an image to represent scene content. The key features are converted into a line-based model, which is then used in the actual matching process. The pattern-matching approach is more tolerant of scene diversities than are correlation techniques, and it can match scenes containing severe contrast reversal, small prominent features, or scale and orientation differences. Both high- and low-altitude flight profiles are considered, with matches performed for each case. Comparisons with conventional correlation are made for a variety of scenes.
(Author)

TYPE 1/4/109
Quest Accession Number : 81A39342
81A39342 NASA IAA Meeting Paper Issue 18
Application of exact area registration to scene matching
(AA)MERCHANT, J.
(AA)(Honeywell Electro-Optics Center, Lexington, MA)
DAAK40-78-C-0144 In: Image processing for missile guidance; Proceedings of the Seminar, San Diego, CA, July 29-August 1, 1980. (A81-39326 18-04) Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1980, p. 166-177. 800000 p. 12 In: EN (English) p.3128

A description is given of the Exact Area Registration process, which can be used to remove all geometric distortions in autonomous scene-matching systems. It is shown that match noise statistics can be approximated by a set of functions, each one corresponding to an a priori designated region of the reference image. These functions define the confidence level of the scene model as depicted in the reference image within the corresponding image. It is suggested that, for autonomous scene matching under a wide range of conditions, an autonomous smart sensor needs a 'knowledgeable' reference which will not only predict the expected conditions of the sensed image but also define the confidence levels of the prediction. In this way, the autonomous device can make match judgements in a way analogous to that of a human scene matcher.
O.C.

TYPE 1/4/110
Quest Accession Number : 80N17755
80N17755# NASA STAR Thesis Issue 08
Studies in image segmentation algorithms based on histogram clustering and relaxation / Ph.D. Thesis
(AA)NAGIN, P. A.
Massachusetts Univ., Amherst. (MK149394) Dept. of Computer and Information Science.
AD-A076576; COINS-TR-79-15 N00014-75-C-0459 790900 p. 183 refs 0 In: EN (English) Avail.: NTIS HC A09/MF A01 p.1052

The research in this thesis has focussed upon the algorithms and structures that are sufficient to generate an accurate description of the information contained in a relatively complex class of digitized images. This aspect of machine vision is often referred to as 'low-level' vision or segmentation, and usually includes those processes which function close to the sensory data. The bulk of this thesis devotes itself to the exploration of some of the problems typically encountered in segmentation. In addition, a new and robust algorithm is presented that avoids most of these problems. The analysis is carried out through the use of a series of computer-generated tests images with known characteristics. Segmentation algorithms of varying degrees of complexity are applied to each image and their performance is carefully evaluated. It will be shown that even the most sophisticated algorithms that are currently in use often perform poorly when confronted with certain apparently simple images. In particular, it is shown that techniques which rely on histogram clustering often generate gross segmentation errors due to overlap in the distributions of the individual objects in a scene. Moreover, the relaxation processes used to correct these errors are themselves prone to errors, but of a different kind. Both techniques, clustering and relaxation, fail because they are based on information which is too global to be effective in complex scenes.
GRA

TYPE 1/4/111
Quest Accession Number : 80N14303
80N14303# NASA STAR Technical Report Issue 05
Vocabulary specification for automatic speech recognition in aircraft cockpits / Final Report, Sep. 1978 - Jun. 1979
(AA)PETERSEN, R. J.; (AB)LEE, N.; (AC)MEYN, C.; (AD)REGELSON, E.; (AE)SATZER, W.
Logicon, Inc., San Diego, Calif. (L3152614) Tactical and Training Systems Div.
AD-A073703 N00014-78-C-0692 790831 p. 92 refs 0 In: EN (English) Avail.: NTIS HC A05/MF A01 p.592

The general focus of this research was to design a communication media (a vocabulary) that is advantageous to both machine recognition and human production of speech events. The problem was analyzed from a human factors perspective that centered upon the man-computer dialogue (interaction) required for cockpit application of ASR. The results indicated that phrase familiarity and stimulus familiarity had major impact on the learning and utilization of the phrases in the paired-associate task. Phrase length and meaningfulness did not appear to differentially affect either the learning or utilization of the paired associate. In addition, pretraining of stimulus familiarity did not seem to result in improved performance. Acoustic lexical confusability also was discussed in general methodological terms. The results of the study were interpreted in terms of a contextualist viewpoint with the necessity of a broader contextual manipulation being pointed out as a requirement for further research.
GRA

TYPE 1/4/112
Quest Accession Number : 75N20033
75N20033 NASA STAR Issue 11
An environment and system for machine understanding of connected speech / Ph.D. Thesis
(AA)ERMAN, L. D.
Stanford Univ., Calif. (S0380476)
740000 p. 172 In: EN (English) Avail: Univ. Microfilms Order No. 74-27012 p.1301

A description is given of part of the research which led to the development of the first demonstrable live system for machine understanding of connected speech: the HEARSAY system. This system uses syntactic, semantic, and contextual information, as well as the more traditional domains of acoustic-phonetic, phonological, and lexical knowledge, in order to recognize and understand utterances. The efforts involved fall into two classes: (1) the design and implementation of the HEARSAY system itself and (2) the careful construction of an environment within which research in machine perception of speech may be pursued by a number of researchers over a period of years. This consideration for an evolving experimental environment is a prime motivation and direction of the work. Thus, the system itself is viewed as a tool for on-going experimentation.
Dissert. Abstr.

TYPE 1/4/114
Quest Accession Number : 73N26187
73N26187# NASA STAR Technical Report Issue 17
Eyes and ears for computers (Machine perception of speech and vision)
(AA)REDDY, D. R.
Carnegie-Mellon Univ., Pittsburgh, Pa. (CH188052) Dept. of Computer Science.
AD-760153; AFOSR-73-0742TR F44620-70-C-0107; NSF GJ-32784; AF PROJ. 9769 730300 p. 34 refs 0 In: EN (English) Avail.: NTIS p.2002

The paper presents a unified view of the research in machine perception of speech and vision in the hope that a clear appreciation of similarities and differences may lead to better information-processing models of perception. Various factors that affect the feasibility and performance of perception systems are discussed. To illustrate the current state of the art in machine perception, examples are chosen from the HEARSAY speech understanding system and the image processing portion of the SYNAPS neural modelling system. Some unsolved problems in a few key areas are presented.
Author (GRA)

TYPE 1/4/115
Quest Accession Number : 73N23147
73N23147 NASA STAR Conference Paper Issue 14
A procedure for the machine recognition of speech (Computer program for machine recognition of distinctive features in words and sentences)
(AA)MEDRESS, M.
Sperry Rand Corp., St. Paul, Minn. (SX655732)
In IEEE The 1972 Conf. on Speech Commun. and Process. p 113-116 (SEE N73-23119 14-07) 720222 p. 4 refs 0 In: EN (English) p.1623

A hierarchical and fundamental procedure for the machine recognition of words and sentences is proposed, and a preliminary implementation of that procedure is described. The computer program attempts to estimate distinctive features information about some stops, fricatives, and vowels in multi-syllabic words and short sentences without reference to a lexicon, and independent of a speaker. Average correct recognition scores of 92% to 95% were obtained for five adult male speakers and three different vocabularies ranging from 60 short sentences to 100 multi-syllabic words. Only one of the five speakers was used to develop the recognition program; the other four were completely new to the system. Author

TYPE 1/4/116
Quest Accession Number : 73N22127
73N22127 NASA STAR Issue 13
Speech generation and recognition under hybrid computer control / Ph.D. Thesis (Synthetic speech generation and recognition under hybrid computer control, using one set of linguistic rules)
(AA)DOUBLIER, R. M.
University of Southern California, Los Angeles. (U6203125)
720000 p. 239 In: EN (English) Avail: Univ. Microfilms Order No. 72-26009 p.1496

This research was concerned with the design, development and testing of the hardware/software systems necessary to produce synthetic speech, using a set of linguistic rules as its only input data. Evaluation of the quality of the artifically-produced speech is made not only from a spectral analysis standpoint, but also through carefully constructed and administered intelligibility tests. The set of linguistic rules developed as a basis for the generation of artificial speech can be adapted to the initial phases of research into machine recognition of human speech, and several fundamental considerations towards the eventual solution of this problem are presented.
Dissert. Abstr.

TYPE 1/4/117
Quest Accession Number : 70N23733
70N23733# NASA STAR Technical Report Issue 10
Study of acoustic properties of speech 2, and some remarks on the use of acoustic data in schemes for machine recognition of speech (Acoustic properties of different speech sounds and use of acoustic data in schemes for machine recognition of speech)
(AA)STEVENS, K. N.
Bolt, Beranek, and Newman, Inc., Cambridge, Mass. (BS628995) AH710313
AD-698352; AFCRL-69-0339; SR-12; REPT-1871 ARPA ORDER 627; F19628-68-C-0125 690815 p. 53 refs 0 In: EN (English) Avail.: NTIS p.1812

TYPE 1/4/118
Quest Accession Number : 69A34119
69A34119 NASA TAA Issue 17
Continuous speech recognition and synthesis. (Machine recognition of continuous speech at acoustic level, noting low bit rate speech communication system)
(AA)FALTER, J. W.
(AA)/USAF, AVIONICS LAB., WRIGHT- PATTERSON AFB, OHIO/.
INST. OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC., NEW YORK, USAF-SUPPORTED RESEARCH. 690000 p. 6 refs 11 IN- '69 NAECON, INST. OF ELECTRICAL AND ELECTRONICS ENGINEERS, NATIONAL AEROSPACE ELECTRONICS CONFERENCE, 21ST, DAYTON, OHIO, MAY 19-21, 1969, PROCEEDINGS. P. 435-440. A69- 34056 17-09< In: EN (English) p.2932

TYPE 1/4/119
Quest Accession Number : 68N16343
68N16343*# NASA STAR Technical Report Issue 07
A program of research directed toward the efficient and accurate machine
recognition of human speech. A theory of speech perception Final report
(Efficient and accurate machine recognition of human speech - theory of
speech perception)
(AA)YILMAZ, H.
Little (Arthur D.), Inc., Cambridge, Mass. (LW086419)

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-LS-185	ISBN 92-835-0684-7	UNCLASSIFIED

5. Originator Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly sur Seine, France

6. Title MACHINE PERCEPTION

7. Presented on 3rd - 4th September 1992 in Hampton, VA, United States, 14th - 15th September 1992 in Neubiberg, Germany and 17th - 18th September 1992 in Madrid, Spain

8. Author(s) Editor(s)	9. Date
Various	August 1992

10. Author's Editor's Address	11. Pages
Various	206

12. Distribution Statement This document is distributed in accordance with AGARD policies and regulations, which are outlined on the back covers of all AGARD publications.

13. Keywords Descriptors

Pattern recognition
Speech recognition
Guidance and control
Visual perception
Artificial intelligence
Real time operations
Situational awareness
Optical processing
Autonomous vehicles

14. Abstract

Human perceptual capabilities involve the extraction of task-oriented information from environmental stimuli through physical sensing and the use of background knowledge.

There are many activities underway aimed at providing similar capabilities of artificial machine perception. Some success is achieved by exploiting what is known of corresponding human cognitive processes and by making use of the increasing power of information processing techniques. For this purpose, the recognition of sharply contrasted as well as fuzzy patterns (stationary or dynamically changing) plays an important role along with other aspects of processing of complex information structures.

These techniques are beginning to be applied in guidance and control, in particular with regard to artificial visual perception and speech understanding. This application promises major benefits with the advent of autonomous vehicle and mission control, and of intelligent systems for situation awareness support of human operators.

This Lecture Series covers the following subjects:

- Pattern recognition techniques
- Real time visual machine perception, principles and applications in G&C
- Real time speech recognition and understanding in the G&C domain.

This Lecture Series, sponsored by the Guidance and Control Panel of AGARD, has been implemented by the Consultant and Exchange Programme.

AGARD Lecture Series 185
Advisory Group for Aerospace Research and Development, NATO
MACHINE PERCEPTION
Published August 1992
206 pages

Human perceptual capabilities involve the extraction of task-oriented information from environmental stimuli through physical sensing and the use of background knowledge.

There are many activities underway aimed at providing similar capabilities of artificial machine perception. Some success is achieved by exploiting what is known of corresponding human cognitive processes and by making use of the increasing power of information processing techniques. For this purpose, the recognition of sharply

P.T.O.

AGARD-LS-185

Pattern recognition
Speech recognition
Guidance and control
Visual perception
Artificial intelligence
Real time operations
Situational awareness
Optical processing
Autonomous vehicles

AGARD Lecture Series 185
Advisory Group for Aerospace Research and Development, NATO
MACHINE PERCEPTION
Published August 1992
206 pages

Human perceptual capabilities involve the extraction of task-oriented information from environmental stimuli through physical sensing and the use of background knowledge.

There are many activities underway aimed at providing similar capabilities of artificial machine perception. Some success is achieved by exploiting what is known of corresponding human cognitive processes and by making use of the increasing power of information processing techniques. For this purpose, the recognition of sharply

P.T.O.

AGARD-LS-185

Pattern recognition
Speech recognition
Guidance and control
Visual perception
Artificial intelligence
Real time operations
Situational awareness
Optical processing
Autonomous vehicles

AGARD Lecture Series 185
Advisory Group for Aerospace Research and Development, NATO
MACHINE PERCEPTION
Published August 1992
206 pages

Human perceptual capabilities involve the extraction of task-oriented information from environmental stimuli through physical sensing and the use of background knowledge.

There are many activities underway aimed at providing similar capabilities of artificial machine perception. Some success is achieved by exploiting what is known of corresponding human cognitive processes and by making use of the increasing power of information processing techniques. For this purpose, the recognition of sharply

P.T.O.

AGARD-LS-185

Pattern recognition
Speech recognition
Guidance and control
Visual perception
Artificial intelligence
Real time operations
Situational awareness
Optical processing
Autonomous vehicles

AGARD Lecture Series 185
Advisory Group for Aerospace Research and Development, NATO
MACHINE PERCEPTION
Published August 1992
206 pages

Human perceptual capabilities involve the extraction of task-oriented information from environmental stimuli through physical sensing and the use of background knowledge.

There are many activities underway aimed at providing similar capabilities of artificial machine perception. Some success is achieved by exploiting what is known of corresponding human cognitive processes and by making use of the increasing power of information processing techniques. For this purpose, the recognition of sharply

P.T.O.

AGARD-LS-185

Pattern recognition
Speech recognition
Guidance and control
Visual perception
Artificial intelligence
Real time operations
Situational awareness
Optical processing
Autonomous vehicles

contrasted as well as fuzzy patterns (stationary or dynamically changing) plays an important role along with other aspects of processing of complex information structures.

These techniques are beginning to be applied in guidance and control, in particular with regard to artificial visual perception and speech understanding. This application promises major benefits with the advent of autonomous vehicle and mission control, and of intelligent systems for situation awareness support of human operators.

This Lecture Series covers the following subjects:

- Pattern recognition techniques
- Real time visual machine perception, principles and applications in G&C
- Real time speech recognition and understanding in the G&C domain.

This Lecture Series, sponsored by the Guidance and Control Panel of AGARD, and implemented by the Consultant and Exchange Programme of AGARD, presented on 3rd—4th September 1992 in Hampton, VA, United States, 14th—15th September 1992 in Neubiberg, Germany and 17th—18th September 1992 in Madrid, Spain.

ISBN 92-835-0684-7

contrasted as well as fuzzy patterns (stationary or dynamically changing) plays an important role along with other aspects of processing of complex information structures.

These techniques are beginning to be applied in guidance and control, in particular with regard to artificial visual perception and speech understanding. This application promises major benefits with the advent of autonomous vehicle and mission control, and of intelligent systems for situation awareness support of human operators.

This Lecture Series covers the following subjects:

- Pattern recognition techniques
- Real time visual machine perception, principles and applications in G&C
- Real time speech recognition and understanding in the G&C domain.

This Lecture Series, sponsored by the Guidance and Control Panel of AGARD, and implemented by the Consultant and Exchange Programme of AGARD, presented on 3rd—4th September 1992 in Hampton, VA, United States, 14th—15th September 1992 in Neubiberg, Germany and 17th—18th September 1992 in Madrid, Spain.

ISBN 92-835-0684-7

contrasted as well as fuzzy patterns (stationary or dynamically changing) plays an important role along with other aspects of processing of complex information structures.

These techniques are beginning to be applied in guidance and control, in particular with regard to artificial visual perception and speech understanding. This application promises major benefits with the advent of autonomous vehicle and mission control, and of intelligent systems for situation awareness support of human operators.

This Lecture Series covers the following subjects:

- Pattern recognition techniques
- Real time visual machine perception, principles and applications in G&C
- Real time speech recognition and understanding in the G&C domain.

This Lecture Series, sponsored by the Guidance and Control Panel of AGARD, and implemented by the Consultant and Exchange Programme of AGARD, presented on 3rd—4th September 1992 in Hampton, VA, United States, 14th—15th September 1992 in Neubiberg, Germany and 17th—18th September 1992 in Madrid, Spain.

ISBN 92-835-0684-7

contrasted as well as fuzzy patterns (stationary or dynamically changing) plays an important role along with other aspects of processing of complex information structures.

These techniques are beginning to be applied in guidance and control, in particular with regard to artificial visual perception and speech understanding. This application promises major benefits with the advent of autonomous vehicle and mission control, and of intelligent systems for situation awareness support of human operators.

This Lecture Series covers the following subjects:

- Pattern recognition techniques
- Real time visual machine perception, principles and applications in G&C
- Real time speech recognition and understanding in the G&C domain.

This Lecture Series, sponsored by the Guidance and Control Panel of AGARD, and implemented by the Consultant and Exchange Programme of AGARD, presented on 3rd—4th September 1992 in Hampton, VA, United States, 14th—15th September 1992 in Neubiberg, Germany and 17th—18th September 1992 in Madrid, Spain.

ISBN 92-835-0684-7

AGARD

NATO OTAN

7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE

FRANCE

Téléphone (1)47 38.57.00 · Télex 610 176
Télécopie (1)47.38.57.99

DIFFUSION DES PUBLICATIONS
AGARD NON CLASSIFIEES

L'AGARD ne détient pas de stocks de ses publications, dans un but de distribution générale à l'adresse ci-dessus. La diffusion initiale des publications de l'AGARD est effectuée auprès des pays membres de cette organisation par l'intermédiaire des Centres Nationaux de Distribution suivants. A l'exception des Etats-Unis, ces centres disposent parfois d'exemplaires additionnels; dans les cas contraire, on peut se procurer ces exemplaires sous forme de microfiches ou de microcopies auprès de Agences de Vente dont la liste suite.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Fachinformationszentrum,
Karlsruhe
D-7514 Eggenstein-Leopoldshafen 2

BELGIQUE
Coordonnateur AGARD-VSL
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Directeur du Service des Renseignements Scientifiques
Ministère de la Défense Nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Board
Ved Idraetsparken 4
2100 Copenhagen Ø

ESPAGNE
INTA (AGARD Publications)
Pintor Rosales 34
28008 Madrid

ETATS-UNIS
National Aeronautics and Space Administration
Langley Research Center
M/S 180
Hampton, Virginia 23665

FRANCE
O.N.E.R.A. (Direction)
29, Avenue de la Division Leclerc
92322 Châtillon Cedex

GRECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
Aeroporto Pratica di Mare
00040 Pomezia (Roma)

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PAYS-BAS
Netherlands Delegation to AGARD
National Aerospace Laboratory NLR
Kluyverweg 1
2629 HS Delft

PORTUGAL
Portuguese National Coordinator to AGARD
Gabinete de Estudos e Programas
CLAFA
Base de Alfragide
Alfragide
2700 Amadora

ROYAUME UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Milli Savunma Başkanlığı (MSB)
ARGE Daire Başkanlığı (ARGE)
Ankara

LE CENTRE NATIONAL DE DISTRIBUTION DES ETATS-UNIS (NASA) NE DETIENT PAS DE STOCKS DES PUBLICATIONS AGARD ET LES DEMANDES D'EXEMPLAIRES DOIVENT ETRE ADRESSEES DIRECTEMENT AU SERVICE NATIONAL TECHNIQUE DE L'INFORMATION (NTIS) DONT L'ADRESSE SUIT.

AGENCES DE VENTE

National Technical Information Service
(NTIS)
5285 Port Royal Road
Springfield, Virginia 22161
Etats-Unis

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Division
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume Uni

Les demandes de microfiches ou de photocopies de documents AGARD (y compris les demandes faites auprès du NTIS) doivent comporter la dénomination AGARD, ainsi que le numéro de série de l'AGARD (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Veuiller noter qu'il y a lieu de spécifier AGARD-R-nnn et AGARD-AR-nnn lors de la commande de rapports AGARD et des rapports consultatifs AGARD respectivement. Des références bibliographiques complètes ainsi que des résumés des publications AGARD figurent dans les journaux suivants:

Scientifique and Technical Aerospace Reports (STAR)
publié par la NASA Scientific and Technical
Information Division
NASA Headquarters (NTT)
Washington D.C. 20546
Etats-Unis

Government Reports Announcements and Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 22161
Etats-Unis

(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

AGARD

NATO OTAN

7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE

FRANCE

Telephone (1)47.38.57.00 · Telex 610 176
Telefax (1)47.38.57.99

DISTRIBUTION OF UNCLASSIFIED AGARD PUBLICATIONS

AGARD does NOT hold stocks of AGARD publications at the above address for general distribution. Initial distribution of AGARD publications is made to AGARD Member Nations through the following National Distribution Centres. Further copies are sometimes available from these Centres (except in the United States), but if not may be purchased in Microfiche or Photocopy form from the Sales Agencies listed below.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordo
Etat-Ma
Quartie
Rue d'E

CANADA
Directe
Dept o
Ottawa

DENMARK
Danish
Ved Id
2100 (

FRANCE
O.N.E
29 Av
9232

GERMAN
Fachi
Karlsruhe
D-7514 Eggenstein-Leopoldshafen 2

GREECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
Aeroporto Pratica di Mare
00040 Pomezia (Roma)

Pintor Rosales 34
28008 Madrid

TURKEY
Milli Savunma Başkanlığı (MSB)
ARGE Daire Başkanlığı (ARGE)
Ankara

UNITED KINGDOM
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
National Aeronautics and Space Administration (NASA)
Langley Research Center
M/S 180
Hampton, Virginia 23665

THE UNITED STATES NATIONAL DISTRIBUTION CENTRE (NASA) DOES NOT HOLD STOCKS OF AGARD PUBLICATIONS, AND APPLICATIONS FOR COPIES SHOULD BE MADE DIRECT TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS) AT THE ADDRESS BELOW.

SALES AGENCIES

National Technical
Information Service (NTIS)
5285 Port Royal Road
Springfield, Virginia 22161
United States

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Requests for microfiches or photocopies of AGARD documents (including requests to NTIS) should include the word 'AGARD' and the AGARD serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Note that AGARD Reports and Advisory Reports should be specified as AGARD-R-nnn and AGARD-AR-nnn, respectively. Full bibliographical references and abstracts of AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
published by NASA Scientific and Technical
Information Division
NASA Headquarters (NTT)
Washington D.C. 20546
United States

Government Reports Announcements and Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic Database or on CD-ROM)

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

ISBN 92-835-0684-7