(12)

AD A114121

Prepared by
Bell-Northern Research Ltd.
Ottawa, Ontario

APRIL 1982

FINAL REPORT

A Multi-Bid Rate Interframe Movement Compensated Multimode Coder For Video Conferencing

PREPARED FOR:

Defense Advanced Research Projects Agency

DARPA ORDER NO.: 4121
Contract No.: MDA903-81-C-0130
BNR Ref. No.: P379/TR6735

PREPARED BY:

Bell-Northern Research Ltd.
P.O. Box 3511, Station C
Ottawa, Ontario
K1Y 4H7

EXECUTIVE SUMMARY

This report describes a multi-bit rate video coder for DARPA video conferencing applications. The coder can operate at any preselected transmission bit rate ranging from 1.5 Mb/s to 64 kb/s.

The proposed National Command Authority Teleconferencing System (NCATS) is designed to connect several conferencing sites. The system provides shared audio, video and graphic spaces. The video conferencing system communicates dynamic images of participants to different conferencing sites. The system is designed to operate under different bandwidth constraints. Under emergency situations communications bandwidth can be drastically reduced to allow only for 64 kb/s to carry out the video conferencing system. Under normal conditions larger channel capacity is available for this service.

In order to accomodate the above requirements, a video codec that can operate at different transmission bit rates is needed. This allows for upgrading of picture quality when there is sufficient bandwidth and a graceful reduction of picture quality under severe bandwidth limitations.

The NTSC colour video signal sampled at 14.3 MHz (4 times the colour subcarrier frequency) and uniformly quantized to 8 bits per picture element, requires a transmission bit rate of 114 Mb/s. Such a high bit rate is economically prohibitive especially for video conferencing applications. In order to reduce the transmission bit rate, redundant information in the signal has to be removed and the specific video conferencing environment has to be exploited.

There are two main sources of redundancy in the video signal, namely: statistical redundancy and perceptual redundancy. The statistical redundancy manifests itself in the form of a high degree of spatial and temporal correlation between adjacent picture elements. This source of redundancy is exploited by interframe coding and variable word-length encoding techniques. Perceptual redundancy is utilized by exploiting some of the properties of the human visual system. This is carried out by allowing modifications to the signal which are irreversible. By utilizing the properties of the eye-brain mechanism, the degradations can be placed in areas of the picture where the human visual sensitivity is low. The fidelity criterion in irreversible coding is dependent on the application. In video conferencing applications, some visible degradations are normally acceptable provided that they are not annoying or interfere with communication of non-verbal cues in the video meeting.

In order to achieve the required transmission bit rate (1.5 Mb/s to 64 kb/s) bandwidth compression ratios of approximately 100:1 to 2000:1 have to be attained. This can be realized using interframe coding techniques which fully exploit the statistical properties of signal, the properties of the human visual system, and the video conference environment.

The specific video conference environment in the NCATS specifies a single participant per conference site. Therefore, the full frame need not be coded and instead a window of approximately one-seventh of the screen size is used. The size of this window is large enough to accomodate a head- and-shoulders view of the participant. The video signal inside the window, which has the full NTSC resolution, is fed to the coder. In addition, conference room lighting, background illumination and colour are assumed to be under the control of the system designer.

The present coder combines several data rate reduction techniques in what is termed as a multimode interframe video coder. Due to the utilization of the statistical properties of the video signal, the data generated from the video coder is at a variable rate dependent on the picture activity. Since the channel rate is fixed, a buffer memory is used to smooth these variations. As the buffer memory content increases due to increased picture activity, parameters of the coder are changed in such a way so as to prevent buffer overflow. Feedback from the buffer memory switches the coder from its normal mode of operation to one of its overload modes. If the buffer memory continues to fill, the coder is switched to a higher overload mode. As a result, degradations are introduced gracefully to the signal. As the buffer memory occupancy falls below certain thresholds, the coder switches back to the lower modes of operation.

A key element to achieving the required bit rate reduction while maintaining acceptable picture quality is the utilization of motion compensation techniques. In standard interframe coders (no motion compensation) a prediction of the current frame picture element (pel) is formed using corresponding previous frame picture element(s). The prediction error, i.e., the difference between the current pel value and the predicted value is quantized, coded and transmitted. Therefore, areas of the picture that have changed from one frame to the next have to be coded and transmitted. In movement compensated coding, the displacement of different objects in the picture, i.e., participants motion from one frame to the next, is estimated. The prediction is formed in the direction of motion, i.e., using the displaced frame element as prediction. In this way the percentage of picture area that is fully predictable (prediction error is below a threshold) is increased. In addition, the magnitude of the prediction error in picture areas which are not fully predictable is significantly reduced. The final result is a significant reduction in the bit rate.

In conjunction with movement compensated predictive coding, the following data rate reduction techniques are utilized in the present coder:

(i) Spatial and temporal subsampling

(ii) Temporal filtering and noise reduction

(iii) Adaptive quantization

(iv) Isolated pel noise suppression and change of thresholds

A BNR proprietary displacement estimation and motion compensation technique has been incorporated in the multi-bit rate coder. This technique operates satisfactorily for all the bit rates under consideration.

The BNR/INRS image processing facility (DVS), which is capable of real time acquisition and display of NTSC colour moving sequences, has been used as the main simulation tool for this work. The full coder operates at several bit rates ranging from 1.5 Mb/s to 64 kb/s. Included in these bit rates is necessary overhead information for framing, synchronization and error protection. For example, for the coder operation at 64 kb/s, a 14 kb/s capacity is reserved for channel overhead and 50 kb/s is used for coding of the video signal. Handling of the sound signals has not been included in the above rates as facilities for this are already available in the voice, data and graphics network. It is assumed that proper synchronization of sound and picture will be carried out.

Simulation of the above coder operating at different rates has been carried out using colour sequences of head-and-shoulder pictures with varying amounts and types of motion. Informal subjective viewing of picture quality indicates that at 1.5 Mb/s excellent picture quality is obtained. Similar results are obtained at 750 kb/s transmission bit rate. At 375 kb/s, a very slight jerkiness is noticeable for large amounts of motion. For 256 kb/s rate, granular noise is slightly visible and some jerkiness is noticeable for large motion. At 64 kb/s aliasing on some edges, and granular noise is visible. For large amounts of motion, jerkiness and blurring in the moving areas are quite noticeable. However, picture quality is judged to be acceptable for the intended application.

Proposed future work on this project involves carrying out a system design for the codec. Special emphasis should be placed on the lower end of the bit rate, i.e., coder operation at 256 kb/s - 64 kb/s. The system design involves identifying implementation alternatives, taking into account state-of-the-art high speed technology and economic considerations.

In addition, investigation of techniques for improved handling of very large amounts of motion at the lower bit rates should be carried out. This will improve picture quality, especially at the 64 kb/s rate. The impact of channel errors on the coder operation and picture quality should be examined and suitable error correction and/or concealment techniques identified.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 SCOPE AND MOTIVATION

This work describes a multi-bit rate video coder for DARPA video conferencing applications. The coder can operate at any preselected transmission bit rate ranging from 1.5 Mb/s to 64 kb/s.

The proposed National Command Authority Teleconferencing System (NCATS) is designed to connect several conferencing sites. The system provides shared audio, visual, and graphic spaces. The video conferencing system communicates images of participants to different conferencing sites. It is assumed that each site will include a single participant and the video conferencing environment (e.g. room set-up and lighting) is under the control of the system designer.

The system is designed to operate under different bandwidth constraints. For example, under emergency situations communications bandwidth can be drastically reduced to allow only for 64 kb/s to carry out the video conferencing service. However, under normal conditions larger channel capacity is available for such a service. In order to accomodate these requirements, a video codec that can operate at different bit rates is needed. This allows for upgrading of picture quality when there is sufficient bandwidth and graceful reduction of picture quality under severe bandwidth limitations.

1.2 THE CODING PROBLEM

The NTSC colour video signal sampled at 14.3 MHz (4 times the colour subcarrier frequency) and uniformly quantized to 8 bits per picture element ('pel'), requires a transmission bit rate of 114 Mb/s. Such a high bit rate is economically prohibitive especially for video conferencing applications. In order to reduce the transmission bit rate, redundant information in the signal has to be removed. There are two main sources of redundancy in the video signal, namely: statistical redundancy and perceptual redundancy.

The statistical redundancy in the video signal manifests itself as a high degree of spatial and temporal correlation between adjacent picture elements. There are several techniques to exploit this source of redundancy. For example, in predictive coding systems, the current picture element is predicted using a combination of previously transmitted picture elements. The prediction error, i.e. difference between predicted and actual value, is quantized and transmitted. Normally picture areas which are fully predictable (prediction error is less than a threshold) are not transmitted, and instead, only some addressing information is sent to the receiver. In addition, a variable word-length code is used to transmit the prediction error signal so that the average number of bits per picture element is reduced.

The transmission bit rate can be further reduced by exploiting the properties of the human visual system. This is carried out by allowing modifications to the signal which are irreversible. By utilizing the properties of the eye-brain mechanism, the degradations can be placed in areas of the picture where the human visual sensitivity is low. For example, noise visibility is much higher in flat areas than in busy areas. Therefore, the phenomenon of masking if properly utilized, will lead to a reduction of the required transmission bit rate with minimal impairments to picture material.

The governing factor on how much the perceptual redundancy can be utilized is the fidelity criterion. In many applications, such as video conferencing, visible impairments are acceptable provided that they are not annoying. However, for broadcast TV applications, visible impairments are not acceptable.

Techniques for data rate reduction can be classified into three main categories, namely:

a) Transform coding approach

b) Interpolative coding approach

c) Predictive coding approach

In the transform coding approach the image is subdivided into 2-dimensional blocks (or 3-D cubes). An orthogonal transformation process is performed on each block. The resulting transform coefficients are quantized and transmitted. At the receiver the inverse transformation is performed and the signal is reconstructed.

In the transform coding approach, the choice of the block size and coding parameters is governed by the sampling frequency used initially. In the multi-bit rate codec, more than one sampling frequency has to be used in order to realize the wide range of bit rate reductions (50:1 up to 2000:1). Therefore, several block sizes may have to be used. This will result in a significant increase in complexity of the codec. In addition, at the very low bit rates under consideration, visible impairments to picture quality are unavoidable. If a transform approach is to be used an objectionable block structure will appear. Therefore, this approach is not suitable for this application.

In the interpolative coding approach, samples of the video signal are dropped and are not transmitted. An interpolation process is carried out at both the receiver and the transmitter. At the transmitter the difference between the interpolated and the actual value is quantized and transmitted together with the retained sample values. At the receiver, the missing samples are interpolated and the quantized interpolation errors are added to the signal. Normally this approach provides modest bandwidth compression and is not suitable for the application under consideration.

In the predictive coding approach a prediction of the current picture element is formed using previously transmitted picture elements. The difference between the current value and the predicted value, i.e., prediction error, is quantized and transmitted. This approach lends itself to the multi-bit rate coder problem as will be seen in the following sections.

The video signal is basically three-dimensional. Intraframe processing exploits its spatial properties while interframe processing exploits both the spatial and temporal properties of the signal. In order to achieve the required transmission bit rates, interframe coding techniques have to be utilized.

The basic principle of interframe coding is to transmit information about the frame-to-frame changes in an image (see Fig. 1.1). In the more general case, the difference between the input sample and a predicted value is quantized and transmitted (see Fig. 1.2). The prediction is formed using a function of previously transmitted pels. At the receiver, the prediction error signal is used to reconstruct the original image: the difference signal is zero or insignificant in the background and fixed parts of the image and non-zero in the moving parts of the image. The prediction error will have a smaller variance than the input (i.e., a smaller dynamic range), with small differences much more probable than large differences. The non-uniform distribution of the quantized difference signal is exploited with a variable word-length encoder, which assigns short code words to the most probable signal values (near zero) and longer code words to the less probable large values. Areas of the picture which are predictable are not transmitted and only the addressing information is transmitted instead.

The data generated from the encoding process is generated at a variable rate dependent on the picture activity. As the channel transmission bit rate is fixed, however, a buffer memory is used to smooth these variations. As the buffer memory content increases due to increased picture activity, parameters of the coder are changed in such a way so as to prevent buffer overflow. Feedback from the buffer switches the coder from its normal mode of operation to overload modes; by so doing, quality is degraded in a graceful fashion. The overload modes will degrade the signal, and must be arranged to give the best available subjective quality as the amount of motion increases. As the buffer memory occupancy falls below a safe level, the coder switches back to the lower modes of operation.

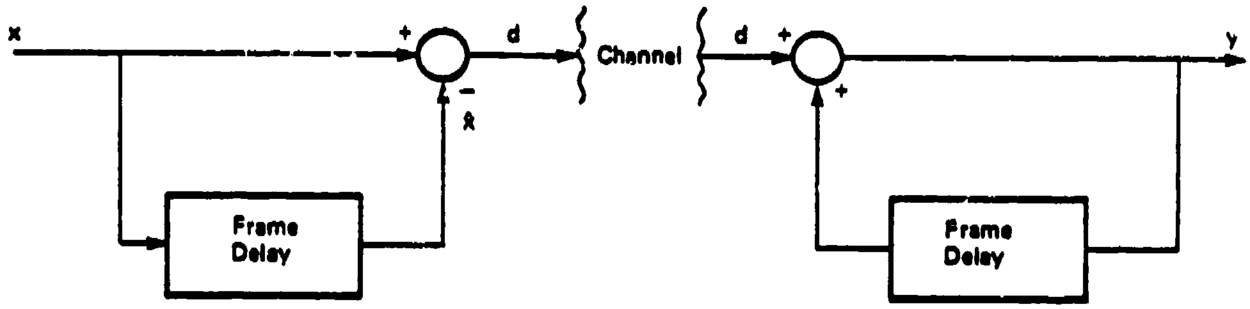


Fig. 1.1 Principle of Interframe Coding

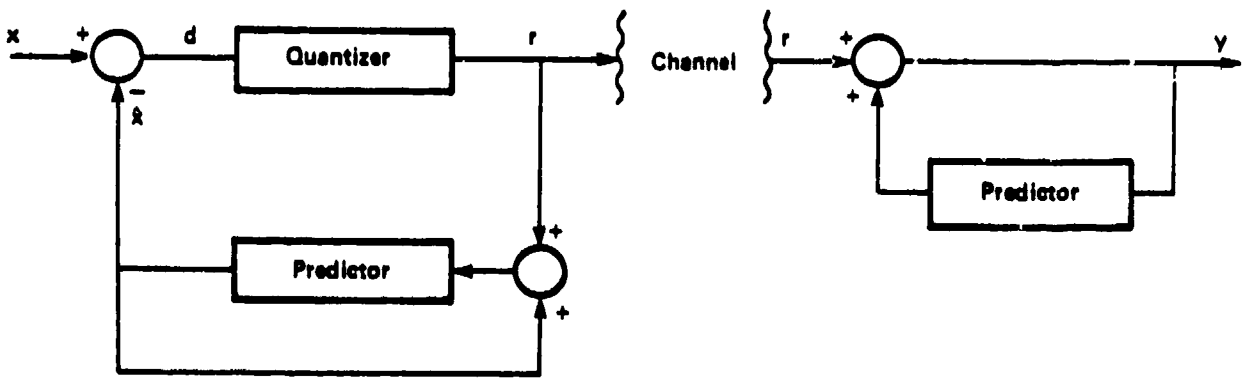
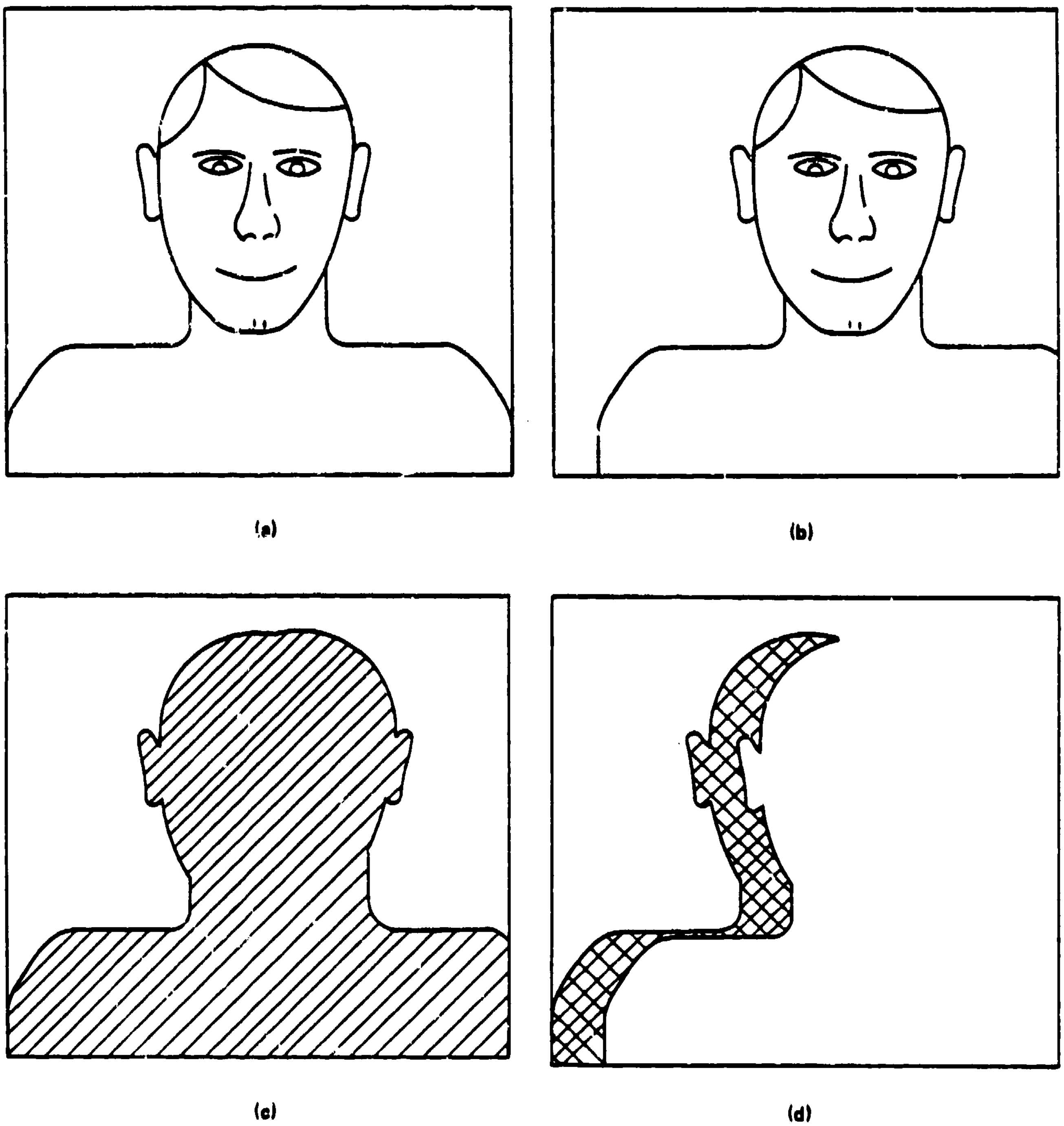


Fig. 1.2 Differential Pulse Code Modulation (DPCM)

In general, it is not possible to achieve acceptable picture quality at very low bit rates with a standard interframe coding technique. In the work described in [6], a scan converter prior to the coder is used to lower the bit rate by reducing the bandwidth and then halving the frame rate. The interframe coding technique known as "conditional replenishment" is used. With this, only the pels that have changed significantly between frames are updated (or replenished) by sending the prediction error signal. The latter is calculated as a linear combination of seven picture elements taken from the actual and of the previous frame. A variation on the basic coder theme is to spatially subsample (drop pels) in the moving area, while linearly interpolating at the receiver. The coder structure can produce black and white video at the bit rates of 128 kb/s and 64 kb/s.

A technique which greatly improves the standard interframe coding is that of movement compensation. A typical video conference scene contains a head-and-shoulders view (Fig. 1.3(a)) of a conference participant (conferee). If the conferee moves to the right after one frame (Fig. 1.3(b)) only the crosshatched region has changed (Fig. 1.3(c)). This represents the area of nonzero prediction error signals which must be transmitted by the standard interframe coder. The concept of movement compensation is understood by realizing that the doubly crosshatched area of Fig. 1.3(d) is not present in the previous frame and represents newly exposed background. If the displacement of the moving area from one frame to the next is calculated simultaneously at the transmitter and receiver (or transmitted to the receiver), then the difference between a present moving area pel and the appropriately displaced pel in the previous frame is zero (zero prediction error signal). Thus, in the ideal case, the only nonzero prediction error signals occur in the newly exposed area which is much smaller than the changing area of a standard interframe coder. In practice, the displacement estimate is not precise so that the prediction errors are not exactly zero in the moving area, but if less than a threshold they are classified as predictable and set to zero; as the prediction errors are small, a comparatively smaller information rate is obtained.



(a) Head and Shoulders View of Conferee
(b) Conferee Moved to the Right
(c) Picture Area Transmitted by Standard Interframe Coder
(d) Picture Area Transmitted by Movement Compensated Interframe Coder

Fig. 1.3

CHAPTER 2.

MULTI-BIT RATE CODER

2.1 INTRODUCTION

In order to achieve the required bit rate reduction, i.e., to the 1.5 Mb/s - 64 kb/s range, the NTSC signal resolution cannot be maintained at all bit rates. As mentioned earlier, the conferencing system is designed to accomodate one participant per conference site. Therefore, the full frame need not be coded and instead a window of approximately 1/7th the screen size is used. The size of the window is large enough to accomodate a head-and-shoulders view of the participant. Inside this window the full resolution is initially maintained. For display purposes the picture material can be enlarged to fill the whole screen by interpolation techniques if needed.

A block diagram of the multi-bit rate coder is shown in Fig. 2.1.a. The composite NTSC signal generated from the camera is sampled at 14.3 MHz (4 times the col ur subcarrier frequency f_{sc}). The inactive portion of the signal as well as the synchronization and blanking intervals are deleted. The active portion of the video signal, i.e., the window containing the picture of the participant, is digitized using 8 bits PCM and then fed to the different parts of the coder.

The composite colour NTSC signal is separated into its three main components: luminance Y and chrominance components I and Q. The luminance and chrominance components are multiplexed and fed to the noise reducer.

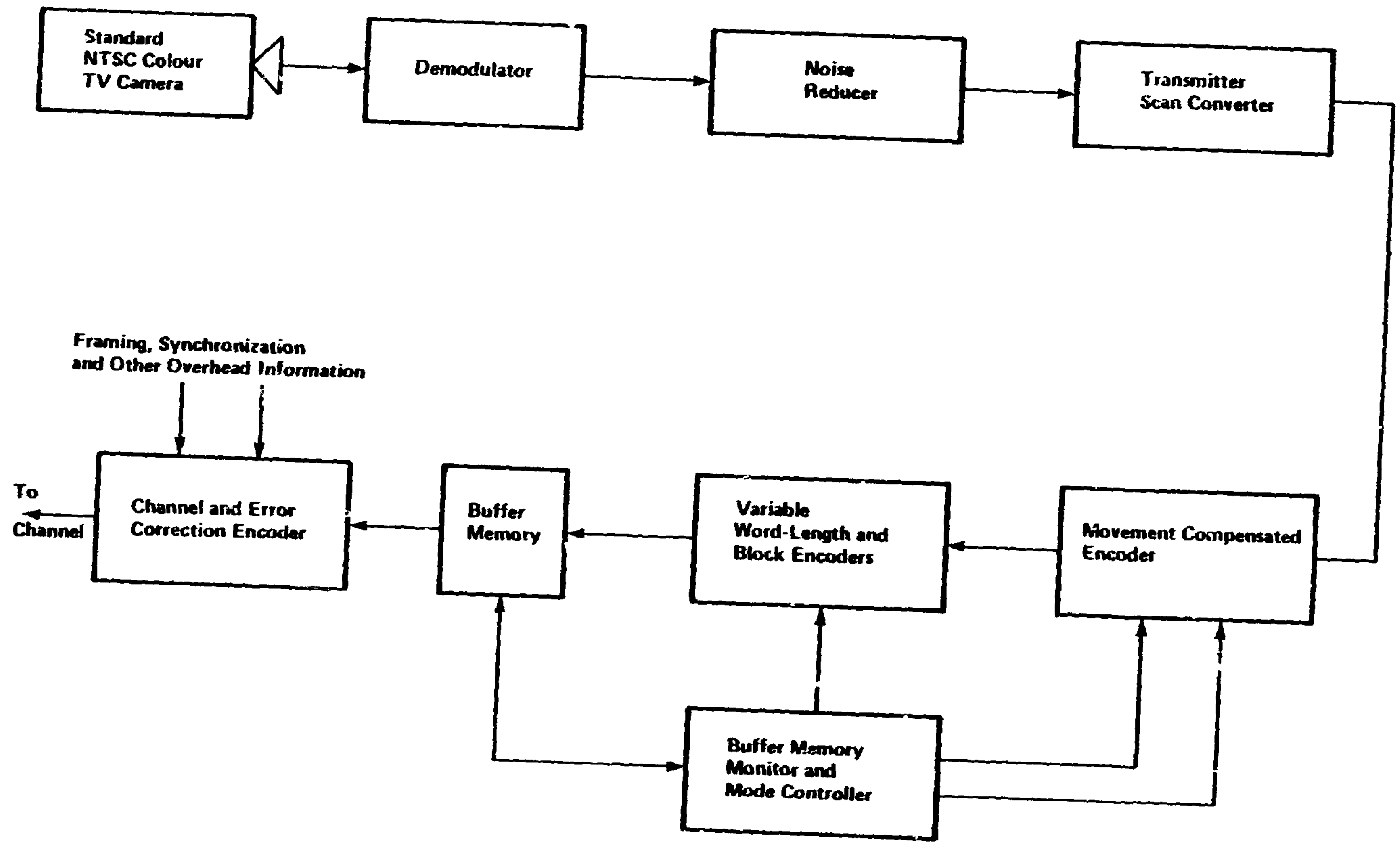


Fig. 2.1a General Block Diagram of the Complete Video Encoder

Noise in the video signal will cause problems to the interframe video coder; i.e., noise will unnecessarily increase the bit rate generated by the coder. In video conferencing applications, two factors contribute to increased noise levels in the signal: the use of inexpensive TV cameras and lighting conditions. Increased lighting in the conference room will reduce the noise levels in the signal. However, it might be uncomfortable for conference participants as it is desirable to operate under normal lighting conditions. Therefore, the use of noise reduction techniques prior to coding of the signal will relax the requirements on the input signal-to-noise ratio (SNR) so that the coder can operate satisfactorily.

After the signal has been processed through the noise reducer, a scan conversion process takes place. Its function is to further reduce the bit rate prior to coding. The issues involved in the design of different elements of the scan conversion process are discussed in the following sections.

Once the scan conversion process is completed, the resulting signal is processed through the movement compensated interframe video coder to reduce the data rate to the desired levels.

The channel encoder adds the supplementary channel data, such as framing, synchronization and error protection bits, prior to transmission over the communication channel.

At the receiver the inverse of the above operations are performed to reconstruct the video signal as shown in Fig. 2.1.b.

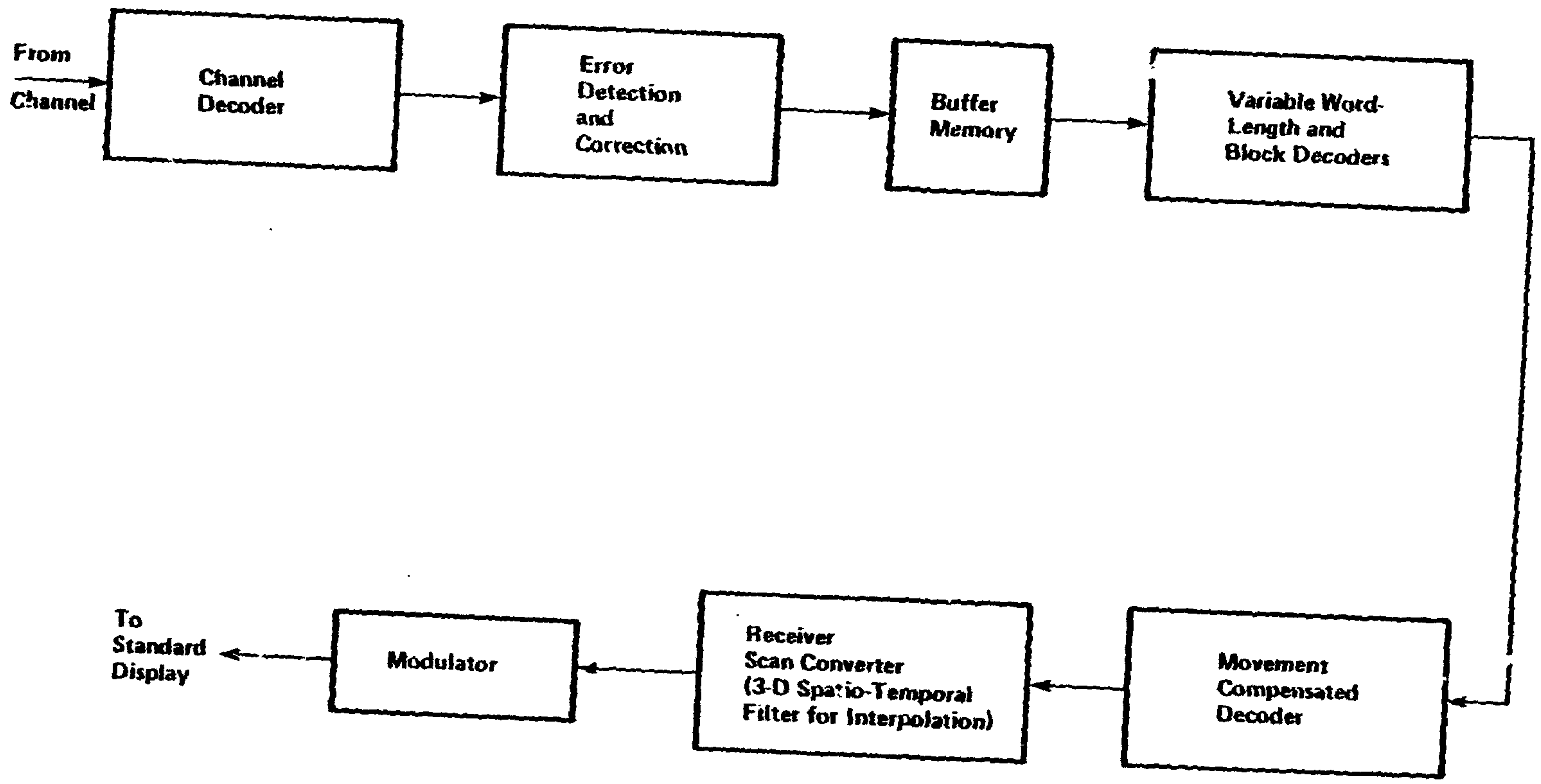


Fig. 2.1b General Block Diagram of the Complete Video Decoder

The following sections describe each system component culminating with the simulation results obtained for the multimode coder.

2.2 DEMODULATION OF NTSC COLOUR SIGNALS

Demodulation of the composite colour NTSC video signal to obtain the luminance signal Y and the two chrominance signals (I and Q) is achieved by 2-D spatial filtering as shown in Fig. 2.2.

The sampling phase is normally selected along the +I axis. Therefore, due to the $4*f_{sc}$ sampling frequency, the I signal is obtained directly by a 2:1 horizontal subsampling and the Q signal is obtained by a one-pel delay followed by a 2:1 horizontal subsampling. Shifting of the I and Q to baseband requires, in this case only, a multiplication by ± 1, i.e., a sign change.

The impulse response of the bandpass filter is:

$$h(n) = (-1,0,6,0,-15,0,20,0,-15,0,6,0,-1)/64$$

whilst the impulse response of the comb filter is:

$$h(n) = (-1,2,-1)/4$$

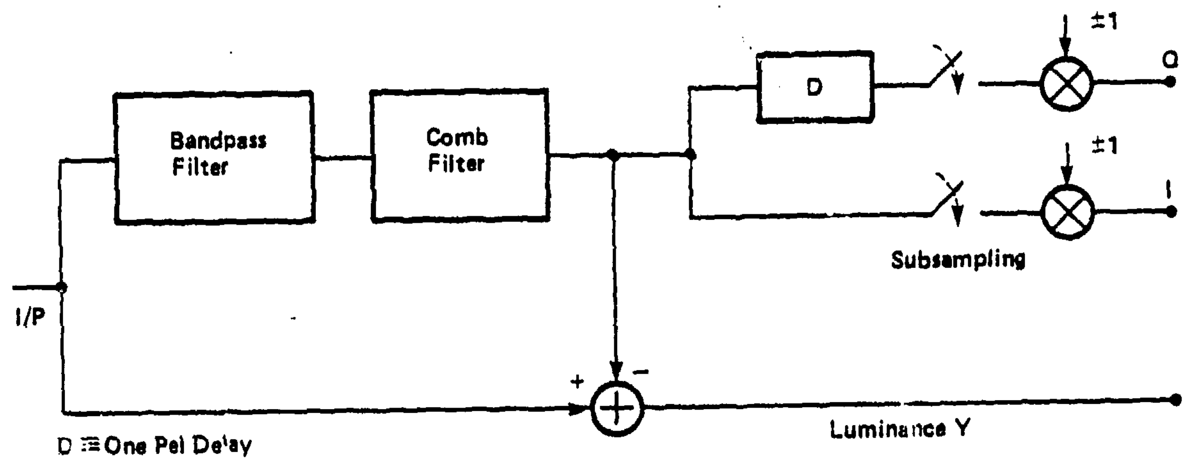


Fig. 2.2 Digital Demodulation of the Composite NTSC Video Signal Sampled at 4^*f_{sc} (14.3MHz)

2.3 NOISE REDUCTION

An essential technique for attainment of low bit rates in this video conferencing application is that of noise reduction. As mentioned earlier, the main noise source is that due to use of inexpensive NTSC colour cameras, as well as the studio set-up (lighting, etc.). The importance of noise reduction can be appreciated by realizing that the coder output represents very little information at low bit rates. In some cases, the noise might represent a comparable or greater portion of the "information" out of the coder. Of course, besides the increase in channel bit rate, the quality of the image is degraded.

Noise reduction in the video signal can be realized by adaptive non-linear temporal filtering [2]. There are two basic structures, namely: FIR (non-recursive) filters or IIR (recursive) filters. The FIR structures have the advantage of having a linear phase response (constant delay response). Hence, impairments in picture quality due to phase nonlinearity, such as the "tailing" of moving objects, do not exist. For a given attenuation, however, recursive structures require fewer frame memories than non-recursive. Practical systems usually use a first-order IIR filter (1 frame memory). The disadvantage of the recursive structures is that their phase responses are nonlinear, and therefore their parameters have to be carefully optimized so as not to introduce visible degradations to the signal.

The configuration of the digital noise reducer is shown in Fig. 2.3. It is composed of three main elements: the predictor, the movement detector and the nonlinear element.

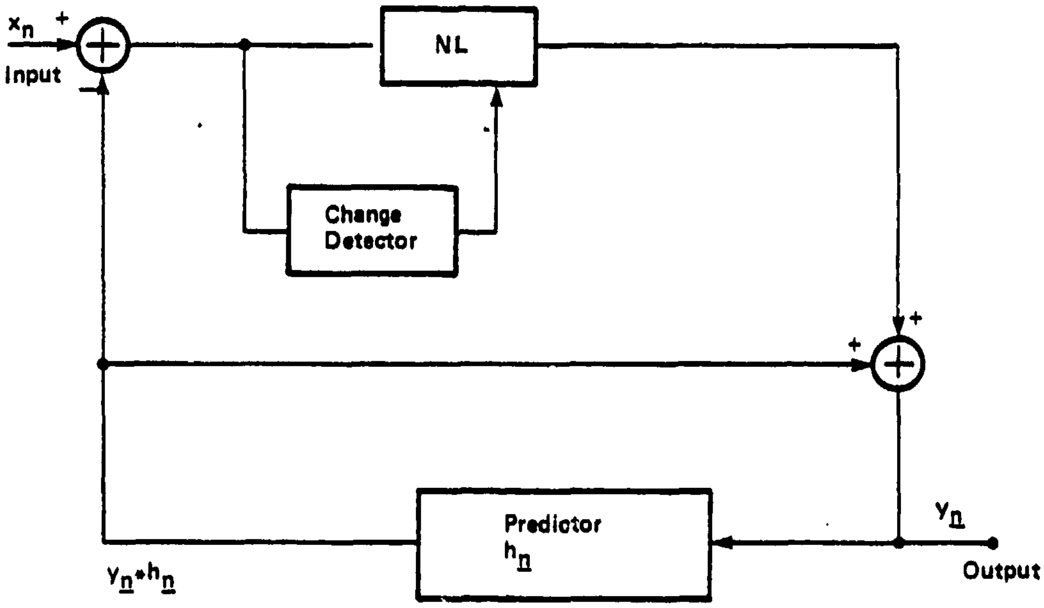


Fig. 2.3 Block Diagram of the Digital Noise Reducer

The function of the movement or changed area detector is to segment the picture into stationary and changing areas. In general, complex segmentation algorithms which usually take the form of 2-D nonlinear filters, are successful when an accurate noise model is known. However, the movement detector used here is based on a pel-by-pel comparison of the frame difference signal with a predetermined threshold.

The prediction error signal (frame difference) is passed through the nonlinear element NL. This element is effectively a multiplier with a varying multiplication coefficient α. The value of α depends on the magnitude of the prediction error as shown in Fig. 2.4. In the stationary area of the picture, small values of α are used as it affects the amount of noise suppression and subsequently the improvement in signal-to-noise ratio (SNR). The value of α in these regions is dependent on the artifacts introduced. This temporal filtering process will modify the temporal spectrum of the noise, but not the noise spatial characteristics. Hence, setting α too low will result in a freezing of the noise, and as α is increased the noise patterns will start to move slowly. To disable the filtering in the moving areas, α is set to unity. In order to avoid introducing artifacts or edge distortions, especially at the boundaries of moving edges, a gradual transition of α is required. This is illustrated in Fig. 2.4 with the nonlinearity defined by (P_e, P_s, α_s).

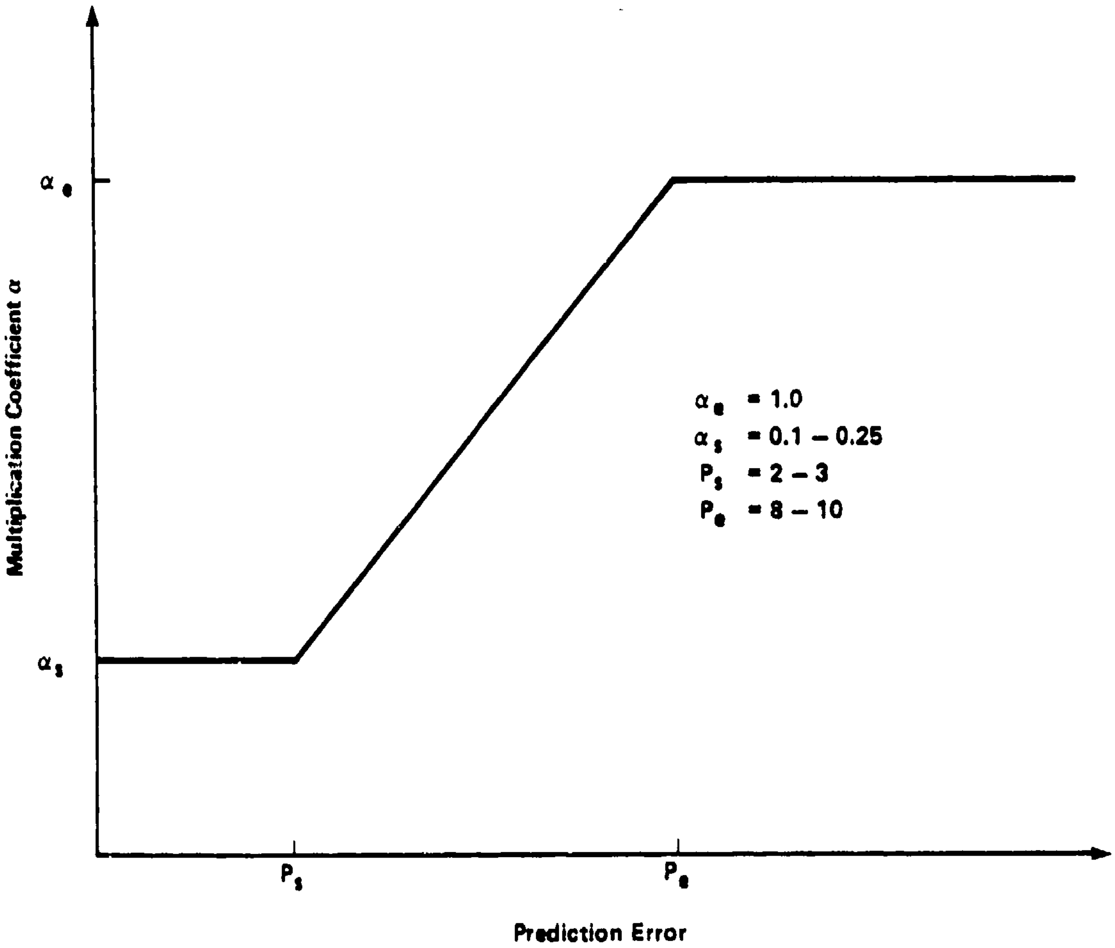


Fig. 2.4 Nonlinear Characteristic (NL)

2.4 SCAN CONVERTER

As mentioned earlier, the full NTSC colour signal resolution cannot be maintained at all bit rates. Therefore, a reduction in the sampling frequencies is needed prior to coding in order to achieve the required bit rates. This process is achieved by the scan converter.

The NTSC signal in analog form is already sampled in the vertical (on a line-by-line) and temporal (field-by-field) dimensions. The vertical and temporal sampling frequencies are specified by the given standards. Therefore the system designer has the flexibility of selecting the horizontal sampling frequency and the three-dimensional sampling pattern.

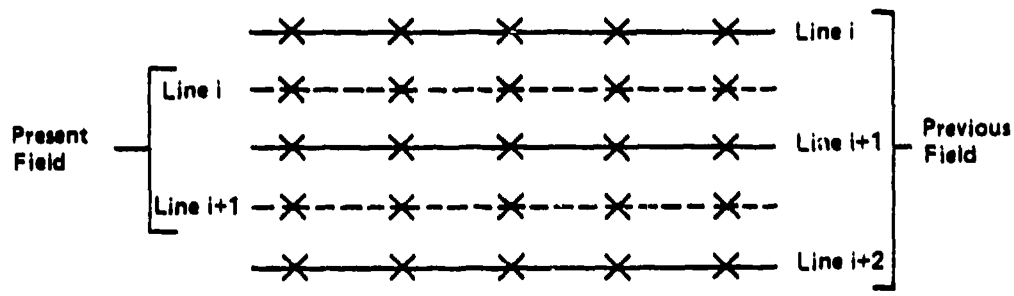
2.4.1 Design Approach

As mentioned earlier, only a window containing the active portion of the image is selected and processed through the coder. Inside this window, a sampling frequency of $4*f_{sc}$ is used. After the demodulation process, i.e., component separation, the resulting sampling frequency for the luminance is $4*f_{sc}$ and that of each of the chrominance components (I or Q) is $2*f_{sc}$.

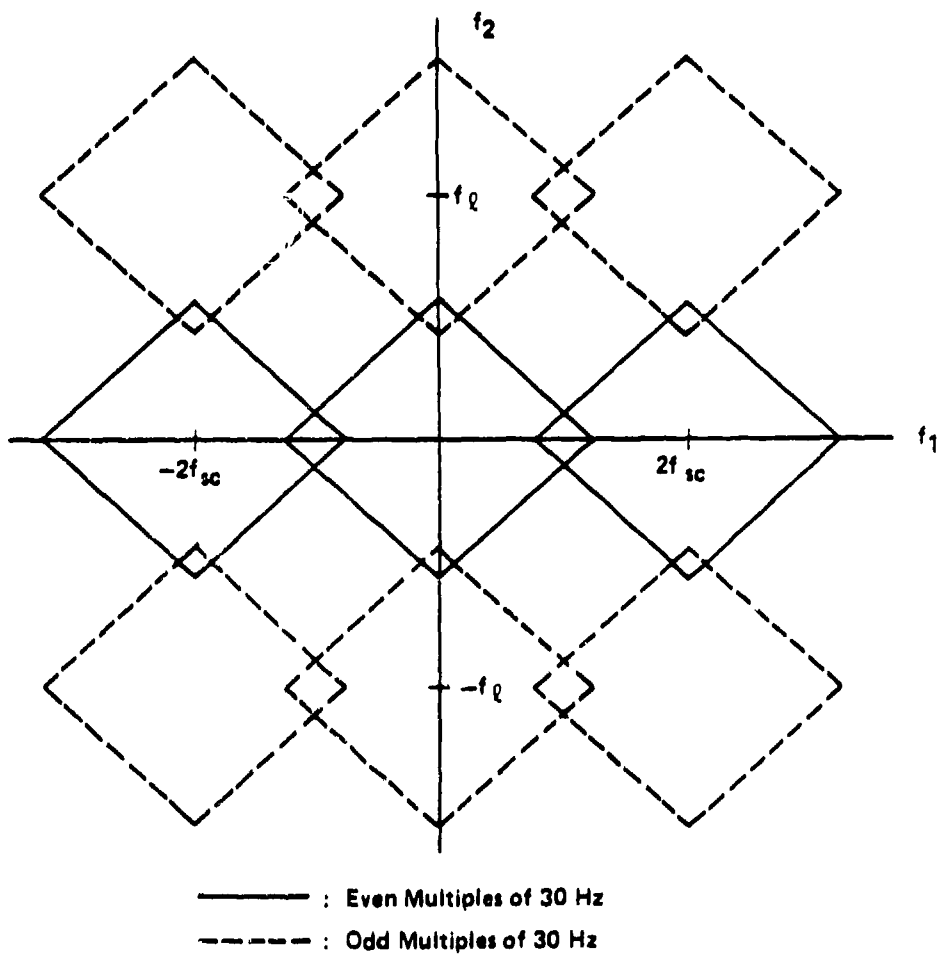
Due to the fact that the chrominance signals I and Q have bandwidths of approximately 1.5 MHz and 0.5 MHz respectively, they may be more severely subsampled than the luminance Y. Therefore, of primary importance is the manner in which the luminance is subsampled. To reduce the amount of aliasing introduced, the signal has to be prefiltered. In so doing, the overlapping of the spectrum replicas is kept to a minimum. The chrominance signals (I and Q) do not have to be prefiltered, due to their low bandwidth.

For the range from 1.5 Mb/s to 256 kb/s, the prefiltered 4^*f_{sc} luminance array is subsampled by a factor of 2 horizontally. Each of the chrominance components is subsampled by a factor of 4 horizontally and a factor of 2 vertically. This results in an effective sampling frequency of 2.5^*f_{sc}. For the low bit rates, in the range of 128 kb/s to 64 kb/s, the luminance is subsampled by a factor of 4 horizontally and each of the chrominance signals is subsampled by a factor of 8 horizontally and a factor of 2 vertically. The effective sampling frequency in this case is 1.25^*f_{sc}. For both ranges, the subsampled Y, I and Q components are multiplexed and fed into the coder.

This subsampling may be accomplished with different sampling patterns among which are: Orthogonal (O) or Field Quincunx (QT). The orthogonal pattern, shown in Fig. 2.5, is rectangular and aligned from field-to-field, leading to easy implementation. The QT pattern, as shown in Fig. 2.6, however, is offset temporally. That is, the grid in the interlaced field is shifted by half of the pel spacing in the previous field. At bit rates lower than 1.5 Mb/s, the coder will actuate field subsampling. Since QT results from offsetting the sampling pattern in alternate fields, dropping this field would defeat the purpose. Of course, this pattern would only be of value when operating at 1.5 Mb/s where, as shown later, field subsampling is rarely utilized. The orthogonal pattern, however, may be used at all bit rates. More complex sampling patterns such as the Line Quincunx and zig-zag patterns have been investigated. However, because of their



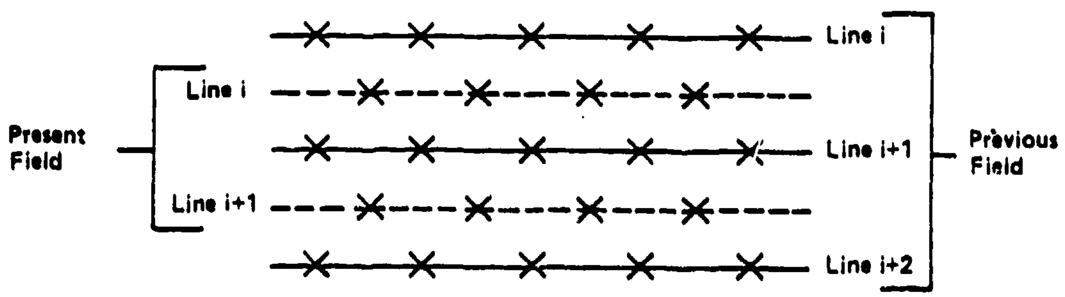
(i)



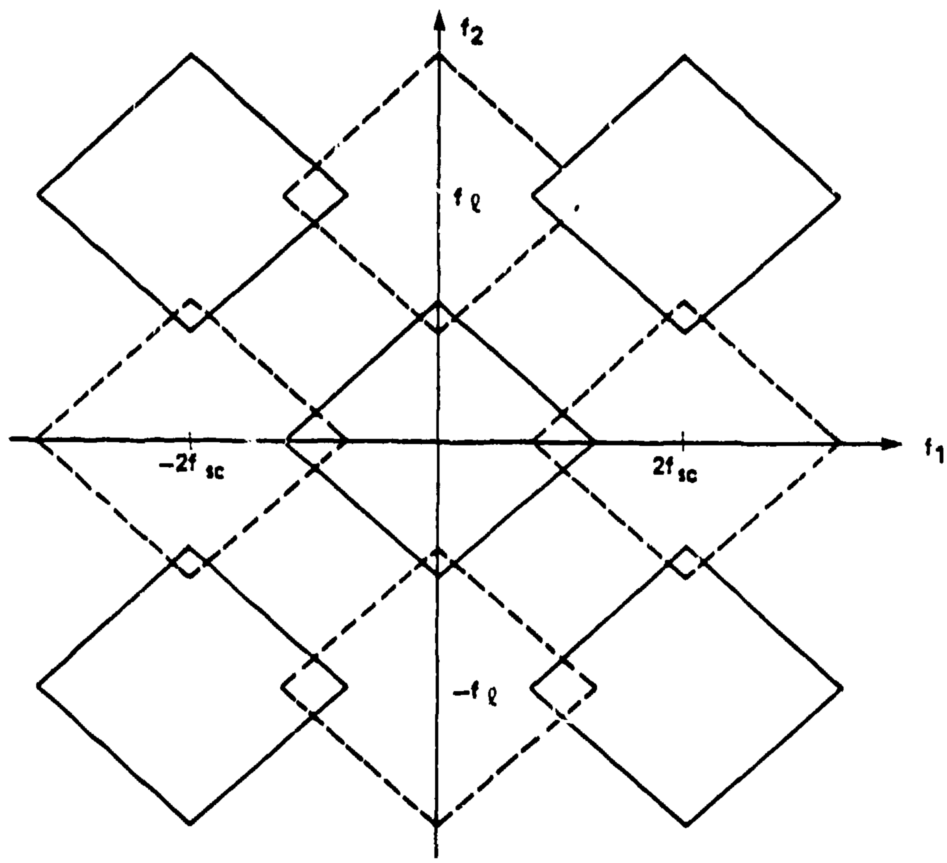
(ii)

(i) Orthogonal Sampling Pattern (O)
(ii) Spectrum of the Digital Luminance Signal Sampled at $2 \cdot f_{sc}$ with the O Structure

Fig. 2.5



(i)



(ii)

——— : Even Multiples of 30 Hz

- - - - : Odd Multiples of 30 Hz

(i) Field Quincunx (QT) Sampling Pattern
(ii) Spectrum of the Digital Luminance Signal Sampled at $2^* f_{sc}$ with the QT Structure

Fig. 2.6

implication on coder complexity they are discarded for this application.

At the receiver all pels which have been dropped spatially (in a field) are restored by interpolation. Any fields that have been dropped by the coder are also restored.

The filters required at the transmitter and receiver are described in the following sections. The design motivation is that, basically, two types of degradation can occur. The first is the attenuation by the filter involved of desired signal components, which usually appears as a loss of resolution. The second is the failure to eliminate unwanted alias components caused by the subsampling. This usually results in spurious patterns i.e. aliasing in the reconstructed signals. For a given sampling pattern, the problem of design for a filter is to achieve a compromise between these types of distortion with a minimum filter complexity.

2.4.2 Prefiltering, Subsampling and Interpolation

The scan converter must reduce the resolution by subsampling; the luminance and chrominance components have different bandwidths, and thus are handled differently.

a) Luminance Prefiltering:

The aliasing produced by the orthogonal and QT sampling patterns, both at $2*f_{sc}$ and $1*f_{sc}$ has to be determined The spatial and vertical-temporal projections of the NTSC signal, sampled at $2*f_{sc}$, have already been shown in Figs. 2.5 and 2.6. At $2*f_{sc}$ (= 7.2 MHz), there is some aliasing in the luminance component.

Most of the information of a video conferencing image is concentrated at the low frequencies. Hence, bandlimiting the signal prior to coding to less than, say, 3 MHz results in negligible loss of resolution. To this end, a study of various digital lowpass filters has been carried out, which would bandlimit to around 2.5 MHz. The filter giving a subjectively pleasing picture is based on a maximally flat design criteria with an impulse response given by:

$$h(n) = (1, 0, -6, 0, 15, 0, 44, 0, 15, 0, -6, 0, 1)/64$$

The output of the filter is 3 dB down at about 2.1 MHz with zeros of transmission at odd multiples of f_{sc}. This filter has been selected for prefiltering of the luminance signal.

If a QT sampling pattern is used, the alias components of the orthogonal structure are offset temporally. Although for this case prefiltering is not needed, a 3-D filter for interpolation is necessary.

For the low bit rate end of the coder, 64 kb/s and 128 kb/s, the signal has to be further subsampled. The luminance signal is subsampled by a factor of 4 horizontally leading to a sampling frequency of $1*f_{sc}$ inside the window.

Consideration of the degree of aliasing, and for other reasons as discussed in the next section, indicates that 4:1 horizontal subsampling is better for both orthogonal and QT patterns. The horizontal prefilter used for both cases is the same as previously described for 2:1 subsampling.

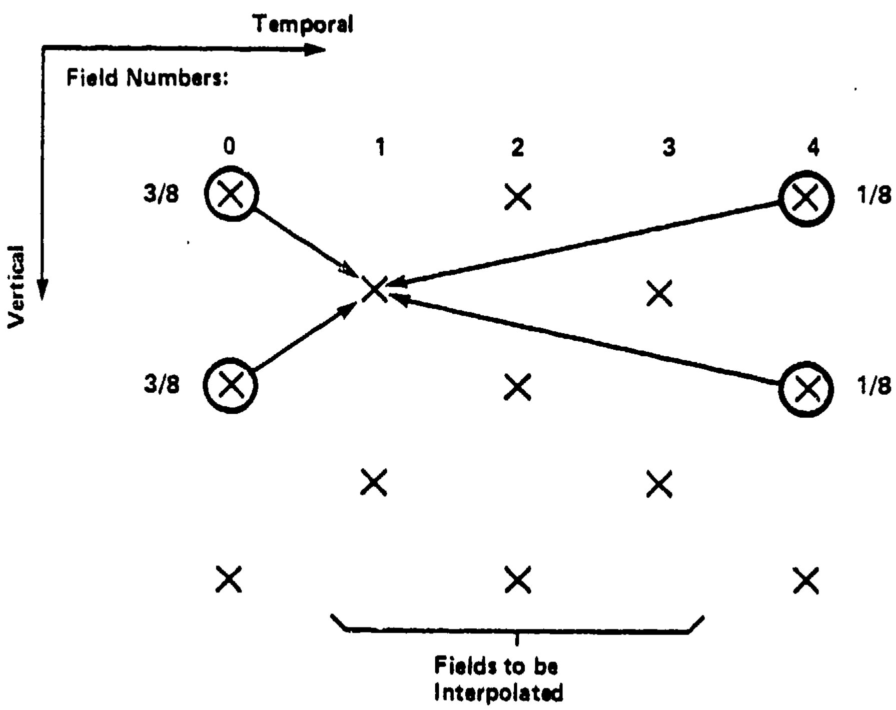
b) Interpolation Filters for the Luminance:

If the coder has performed field subsampling, then, at the receiver the scan converter linearly interpolates the missing fields. For the luminance section of the multiplexed signal, the coefficients are weighted according to the distance to the field to be interpolated and whether the latter is even or odd. This is shown in Fig. 2.7 for a field subsampling ratio of 4:1. The odd fields are not directly aligned with the original (even) fields, so that the weighted average of four pels is taken. The even fields, however, are directly aligned leading to the averaging of two pels.

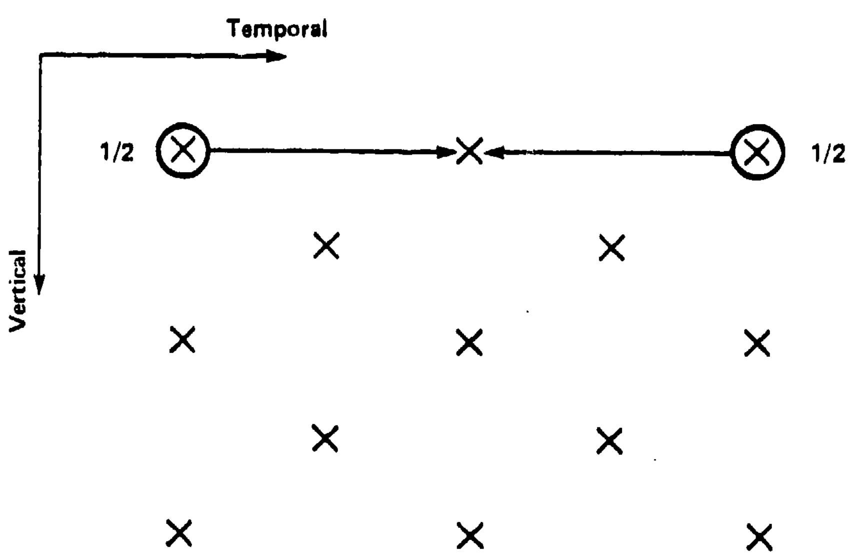
To orthogonally interpolate within a field from 1^*f_{sc} to 4^*f_{sc}, a 15th order "SPLINE" interpolator is used. The impulse response, s(n), is as follows:

$$s(n) = (-3, -8, -9, 0, 19, 40, 57, 64, 57, 40, 19, 0, -9, -8, -3)/64$$

With the SPLINE interpolator the data is interpolated directly from 1^*f_{sc} to 4^*f_{sc}. However, with the QT sampling case, a three-dimensional filter interpolates from 1^*f_{sc} to 2^*f_{sc}, followed by a one-dimensional interpolation from 2^*f_{sc} to 4^*f_{sc}. The operation of the three-dimensional filter is illustrated in Fig. 2.8. For any one pel



(a)



(b)

Fig. 2.7 4:1 Field Interpolation of the Luminance for an (a) Odd Field, and (b) Even Field to be interpolated, two lines in the previous field and one line in current field, are operated on. Interpolation from $2*f_{sc}$ to $4*f_{sc}$ is done with a filter of impulse response:

$$h(n) = (1, 0, 15, 32, 15, 0, 1)/64$$

c) Chrominance Subsampling:

Up to now, only subsampling ratios and patterns for the luminance signal have been investigated. In comparison to the luminance, the chrominance bandwidth is very small: I = 1.5 MHz and Q = 0.5 MHz. As I and Q are each sampled at $2*f_{sc}$ (7.2 MHz), they may be heavily subsampled. Each of the chrominance signals is vertically subsampled by a factor of 2. This leads to the multiplexed format where the Y for the line is followed by either I or Q, the latter two being alternately retained, line-by-line. In addition, for both components, horizontal subsampling by factor of 4 is utilized for the coder operation at 1.5 Mb/s to 256 kb/s. For the coder operation at 128 kb/s and 64 kb/s the chrominance signals are horizontally subsampled by a factor of 8. Therefore the effective sampling frequency inside the window is $2.5*f_{sc}$ and $1.25*f_{sc}$ for the higher and lower ends of the bit rates respectively. The multiplexed version of the data out of the scan converter is shown in Fig. 2.9.

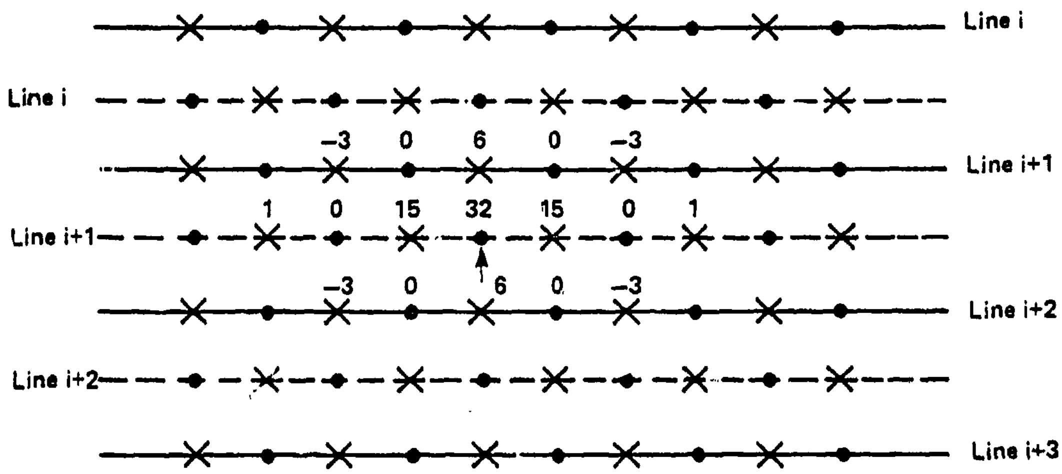


Fig. 2.8 3-D (QT) Interpolation Filter for the Luminance Signal Using One Field Memory

× Pels at 2*f_{sc} Sampling — Current Field

● Missing Pels — Previous Field

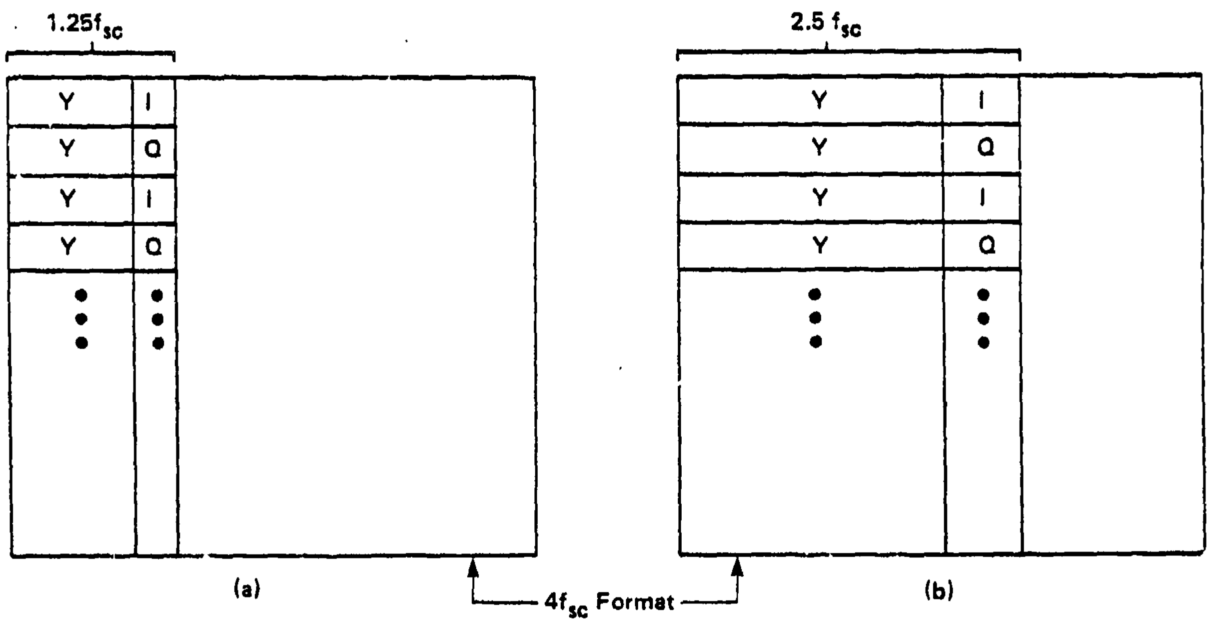


Fig. 2.9 Multiplexed Version of Data From the Scan Converter With Luminance Y and Chrominance (I and Q) at (a) 1 f_{sc} and 0.25 f_{sc}, (b) 2 f_{sc} and 0.5 f_{sc}, Respectively

d) Chrominance Interpolation:

As explained with the luminance, the scan converter at the receiver linearly interpolates the missing field(s) (see Fig. 2.7). The chrominance interpolation is slightly different, however, due to the multiplexed format of I and Q. As the restoration of a component, say I, requires the weighted average of similar I values, pels which are two line intervals apart must be used for the odd fields.

Finally, to reconstruct the full composite video signal, the three components Y, I and Q are necessary. The luminance is interpolated as in the previous section. Each chrominance component (either I or Q on a line) is horizontally interpolated using standard 4:1 (or 8:1) linear interpolators. To reconstruct the composite signal the luminance signal is added, with proper sign change, to the chrominance signals. The missing chrominance signal (I or Q) is repeated from the previous line.

2.4.3 Scan Converter Simulation Results

The operation of the scan converter has been simulated on the BNR/INRS image processing facility. The objective of these simulations has been to determine the best tradeoffs between picture quality and implementation complexity. Some of the issues that have been investigated and the simulation results are summarized in the following:

a) Sampling Patterns:

The Field Quincunx sampling pattern gave excellent results in terms of maintaining high resolution and minimal aliasing problems. However, its use would have resulted in more complex interpolation filters. In addition, since the coder will have to use temporal subsampling in most of the bit rates under consideration, the advantage of this sampling pattern no longer exists. Therefore, it has been decided to use the orthogonal sampling pattern to minimize implementation complexity. Other sampling patterns such as the Line Quincunx pattern have been investigated; their use however will result in increased complexity of the movement compensated coder.

b) Temporal Prefiltering:

Since temporal subsampling is utilized in the system in order to reduce the data rate, temporal aliasing results. Experiments with 3-D FIR temporal prefiltering indicated that the aliasing is reduced. However, this would have required several frame memories for implementation which will significantly add to the complexity. Therefore, it has been determined that 3-D FIR temporal prefiltering not be implemented, as the same function can be realized using the noise reducer as a prefilter if needed. In addition, temporal filtering is also provided inside the interframe coder as one of the techniques for reducing the bit rate.

c) Horizontal Prefiltering:

For effective sampling rates of $2.5*f_{sc}$ and $1.25*f_{sc}$, horizontal prefiltering has been found essential, especially at $1.25*f_{sc}$ sampling. Simulation results have indicated that aliasing errors at $2.5*f_{sc}$ sampling are barely noticeable. However, at $1.25*f_{sc}$ aliasing errors are noticeable but are judged to be acceptable.

d) Vertical Prefiltering and Subsampling:

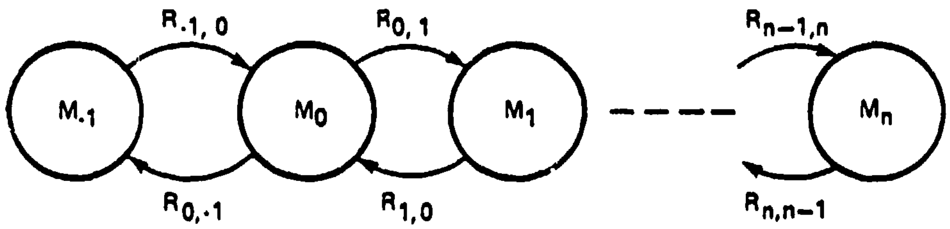
Experiments with vertical subsampling (within the field) indicated that significant loss of vertical resolution results. Therefore, this approach to reducing the data rate has been discarded for this application.

e) Temporal Subsampling:

Several temporal subsampling factors ranging from 2:1 field subsampling to 20:1 field subsampling have been simulated. The results of simulation indicated that temporal subsampling is best realized by both the scan converter and the coder combined. Therefore, for the system operation at 1.5 Mb/s to 256 kb/s, the scan converter does not perform any temporal subsampling. For the 128 kb/s - 64 kb/s range, the scan converter provides an initial 4:1 field subsampling. Additional field subsampling is provided by the coder in order to achieve maximum utilization of the available channel bandwidth.

2.5 MOVEMENT COMPENSATED MULTIMODE CODER

In interframe multimode coders the data is emitted from the variable word-length coder at an irregular rate and hence must be buffered for transmission over a fixed rate channel. As the buffer content increases due to more picture activity, parameters of the coder are altered to prevent buffer overflow. The picture quality is gracefully degraded as feedback from the buffer switches the coder from its normal mode of operation to a set of overload modes. These higher modes will degrade the signal, and must be arranged to give the best subjective quality as the amount of motion increases. The operation of the multimode coder can be represented by a state transition diagram as shown in Fig. 2.10.a. The modes of operation are indicated by M_{-1}, M_0, $M_1, \ldots M_n$; associated with each mode is a set of coding techniques and coding parameters. The mode switching rule from mode M_i to mode M_j is represented by $R_{i,j}$. In fact, $R_{i,j}$ is based on the buffer occupancy; if the buffer occupancy is greater than or equal to a forward threshold $R_{i,i+1}$ the mode changes from mode i to i+1. Similarly, if the occupancy decreases (due to decreasing picture activity) below a backward threshold $R_{i,i-1}$ then the coder will switch to the lower mode of operation M_{i-1}. M_0 is the main mode of operation which is designed to give full available resolution and best picture quality. M_{-1} is the "underflow" mode of operation to insure that the buffer does not underflow. It is also invoked periodically, i.e., used as a refresh mode, to limit the propagation of channel errors by transmitting the 8 bit PCM samples. Modes M_1, $M_2, \ldots, M_n$ are the overflow modes of operation and are invoked successively as the spatio-temporal activities in the picture increase.



M_i = Mode i of Operation
$R_{i,j}$ = Transition Rule From Mode i to Mode j

(a) State Transition Diagram of a Multimode Coder

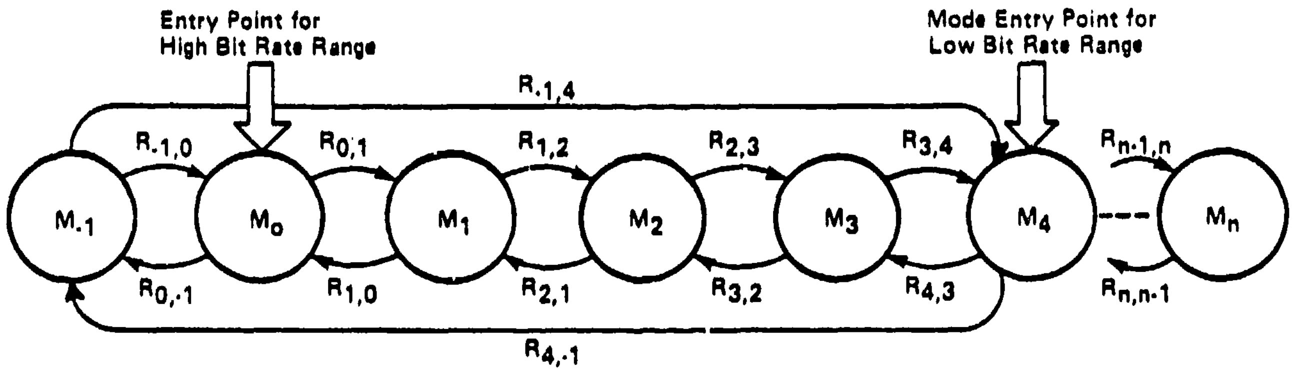


Fig. 2.10 (b) State Transition Diagram Altered to Accomodate a Range of Bit Rates

The multi-bit rate multimode coder, can be thought of as being constructed as an (n+1)-stage transition diagram. However, depending on the desired bit rate, not all stages will be used. For example, when the coder is switched to operate at 64 kb/s, the main mode of operation is mode 4; i.e., the entry point into the state transition diagram is variable as shown in Fig. 2.10.b.

2.5.1 Bit Rate Considerations

For digital transmission of the video information over communication channels, a certain amount of overhead has to be reserved to provide synchronization, framing and error protection. Such overhead has to be provided within the total bit rate allocated for transmission. This will result in a reduction in the number of bits available for coding of the video information.

In many video conferencing applications, it is necessary to carry out voice, data, graphics, and facsimile signals transmission in addition to the video information. In the MCATS a separate system is designed to handle these additional signals. Therefore, no allocation in the current coder is made for such signals. It is assumed that the voice signals will be properly synchronized to the video signal.

The different bit rates allocated for the video and overhead information are given in Table 2.1.

AVAILABLE TRANSMISSION BIT RATE	VIDEO INFORMATION RATE	RESERVED FOR OVERHEAD
64 kb/s	50 kb/s	14 kb/s
128 kb/s	100 kb/s	28 kb/s
256 kb/s	200 kb/s	56 kb/s
448 kb/s	375 kb/s	73 kb/s
832 kb/s	750 kb/s	82 kb/s
1.5 Mb/s	1.35 Mb/s	150 kb/s

TABLE 2.1: Bit rates allocated for video and overhead information.

2.5.2 Techniques Used in Different Modes of Operation

As mentioned before, the multimode coder incorporates several data rate reduction techniques, namely:

1) Movement compensated predictive coding
2) Temporal field subsampling
3) Temporal filtering
4) Isolated pel noise suppression and change of thresholds
5) Switched quantizers
6) Block encoding and variable word-length encoding

In the following sections these techniques are discussed.

2.5.3 Movement Compensated Predictive Coding

In movement compensated predictive coding the displacements of different objects have to be obtained. In moving areas of the picture, the prediction is formed in the direction of the motion. In the following sections displacement estimation techniques are discussed.

2.5.3.1 Estimation of the Translational Displacement

There are several approaches for estimating the displacement of objects. The selection of a suitable technique is governed by several factors such as (i) ability to operate in real-time at high speeds, (ii) implementation complexity, and (iii) performance in the context of the current coder. Based on these factors, the selection can be narrowed down to two approaches.

The first approach is based on a block structured displacement estimation technique [9,10]. In this approach each field is subdivided into rectangular blocks of N pels by M lines. A single displacement is obtained for each block. The resulting displacement estimates for the whole field are then stored to be used in forming the movement compensated prediction, i.e. displaced field (or frame) values, for the next field to be processed.

The image intensity is defined as $u(\underline{x},t)$, expressed as a function of spatial coordinates $\underline{x} = (x,y)$ and time t. Thus with a displacement $\underline{d}$, $u(\underline{x}-\underline{d},t-T)$ represents the pel in the previous frame which has moved to its new position $u(\underline{x},t)$ in the present frame. The displacement $\underline{d}$, which has occurred in one frame interval, may be estimated as follows. Defining the displaced frame difference as:

$$D(\underline{x}, t, \underline{d}) = u(\underline{x}, t) - u(\underline{x}-\underline{d}, t-T) \qquad (2.1)$$

where T is the frame interval.

For each block an estimate $\hat{\underline{d}}$ is obtained as:

$$\hat{\underline{d}} = \tilde{\underline{d}} - \left[\sum_{\underline{x}\,\varepsilon\,MA} \Delta_{\underline{x}}\,\Delta_{\underline{x}}^{T}\right]^{-1} \sum_{\underline{x}\,\varepsilon\,MA} DFD(\underline{x},t,\tilde{\underline{d}})\Delta_{\underline{x}} \qquad (2.2)$$

where $\tilde{\underline{d}}$ is the previously obtained estimate for the same block and $\Delta\underline{x}$ is a finite difference approximation of the spatial gradient. The summations in (2.2) are carried out over the moving area (MA) within the block. Implementation of equation (2.2) requires multiplications. It can be simplified to eliminate the multiplications and reduce the number of additions needed without significantly affecting the performance. The simplified form of (2.2) is given as:

$$d_x^i = d_x^{i-1} - \frac{\sum_{MA} DFD(\underline{x},\ \underline{d}^{i-1}) \cdot Sign\ (G_x)}{\sum_{MA} |G_x|}$$

$$d_y^i = d_x^{i-1} - \frac{\sum_{MA} DFD(\underline{x},\ \underline{d}^{i-1}) \cdot Sign\ (G_y)}{\sum_{MA} |G_y|} \qquad (2.3)$$

where $\underline{d}^i = (d_x^i, d_y^i)$, d_x^i and d_y^i are the displacement estimates in the horizontal and vertical directions respectively. The $\underline{d}^{i-1}$ is the displacement estimate for the same block at frame (or field) i-1. G_x and G_y are the horizontal and vertical gradients respectively. Sign (.) denotes the sign function. Similarly to the DFD, the (standard) Frame Difference may be defined as:

$$FD(\underline{x},\ t) = u(\underline{x},\ t) - u(\underline{x},\ t-T)$$

The block structured approach described above has been previously simulated in the context of the multimode coder. For the higher bit rate range and with picture material containing medium amounts of motion, good results have been obtained. However, for the lower bit rates and when larger temporal subsampling ratios are used, the performance of this approach deteriorated very rapidly. This is largely due to the fact that the basic assumption of a uniform displacement within a block is no longer true when large temporal subsampling is used. In addition, in this approach, the displacement estimates obtained from the previous processed field are used for prediction of the current field. For example, if a temporal subsampling ratio of 8:1 is used, in order to calculate the displaced frame value at current field j, the displacement estimate between field j-8 and j-16 is used. This estimate is potentially inaccurate.

To alleviate the problem with this approach, calculation of the displacement estimates can be taken outside the coder DPCM loop prior to performing the temporal subsampling process. However, such a solution would have required the transmission of the displacement estimates to the receiver. Obviously this will add significant overhead which cannot be accomodated at the lower bit rates. Since it is desirable to have a single displacement estimation technique to be implemented in the coder that can operate satisfactorily over all the bit rates, this approach is discarded because of its inadequacy at the lower bit rates.

The second approach for estimation of the displacement is based on the pel recursive method [3]. This approach has been found suitable for the current coder and is described in the following section.

2.5.3.2 Pel Recursive Estimation of the Displacement

The recursive estimation of displacement is based on calculating the estimate at a given pel based on previously obtained estimates at neighbouring pels. Thus, let $\underline{x}-\underline{i}$, $i \in I$ be a set of pels for which estimates have already been obtained. The new estimate is thus:

$$\underline{d}(\underline{x}, t) = f(\{\underline{d}(\underline{x}-\underline{i}, t), i \in I\}, u) \qquad (2.4)$$

where the function f basically defines the estimator. The set I, proposed in [9] consists of only one previously transmitted pel, either the previous element of the same line, or the same element on the previous line. The $d(\underline{x}-\underline{i},t)$ is modified in order to reduce the displaced frame difference $D(\underline{x},t,\underline{d})$ using the steepest descent algorithm to give:

$$\underline{d}^{i} = \underline{d}^{i-1} - \varepsilon \cdot DFD(\underline{x},\underline{d}^{i-1}) \cdot \nabla I(\underline{x}-\underline{d}^{i-1}, t-T) \qquad (2.5)$$

where $\underline{d}^{i}$ is the displacement estimate at the i^{th} iteration, $\underline{d}^{i-1}$ is the previous displacement estimate, DFD $(\underline{x},\underline{d}^{i-1})$ is the displaced frame difference, ∇I is the spatial gradient, and ε is the convergence control parameter. This estimator is preceded by a segmentation into fixed and moving areas, and the estimate update process is applied in the moving areas only.

Implementation of Eq. (2.5) requires multiplications, and interpolation to evaluate DFD and ∇I. Equation (2.5) can be simplified without seriously affecting performance as follows:

$$d_x^i = d_x^{i-1} - \varepsilon \cdot \mathrm{Sign}\left(\mathrm{DFD}(\underline{x},\underline{d}^{i-1})\right) \cdot \mathrm{Sign}\left(G_x(\underline{x} - [\underline{d}^{i-1}], t-T)\right)$$

$$d_y^i = d_y^{i-1} - \varepsilon \cdot \mathrm{Sign}\left(\mathrm{DFD}(\underline{x},\underline{d}^{i-1})\right) \cdot \mathrm{Sign}\left(G_y(\underline{x} - [\underline{d}^{i-1}], t-T)\right) \tag{2.6}$$

where

$\underline{d}^i = (d_x^i, d_y^i)$, d_x^i and d_y^i being the horizontal and vertical components of the displacement estimate, while $[\underline{d}^i]$ is the integral value of the displacement estimate. G_x and G_y are the horizontal and vertical components of the spatial gradient. The sign function is defined as:

$$\mathrm{Sign}(Z) = \begin{cases} \frac{Z}{|Z|}, & Z \neq 0 \\ 0, & Z = 0 \end{cases} \tag{2.7}$$

In Eq. (2.6) the multiplication has been eliminated and the interpolation process is required only to calculate the displaced frame difference signal. The displacement estimate is updated in the scan direction as shown in Fig. 2.11(a). The update is disabled in the stationary area of the picture, i.e., when the frame difference signal is less than or equal to a certain threshold value K_m. The pels used in the calculation of the displacement estimates are illustrated in Fig. 2.11(b). The horizontal and vertical gradient calculations and interpolation are calculated as:

$$G_x = \frac{I_4 - I_1 + I_3 - I_2}{2}$$

$$G_y = \frac{I_1 - I_2 + I_4 - I_3}{2} \tag{2.8}$$

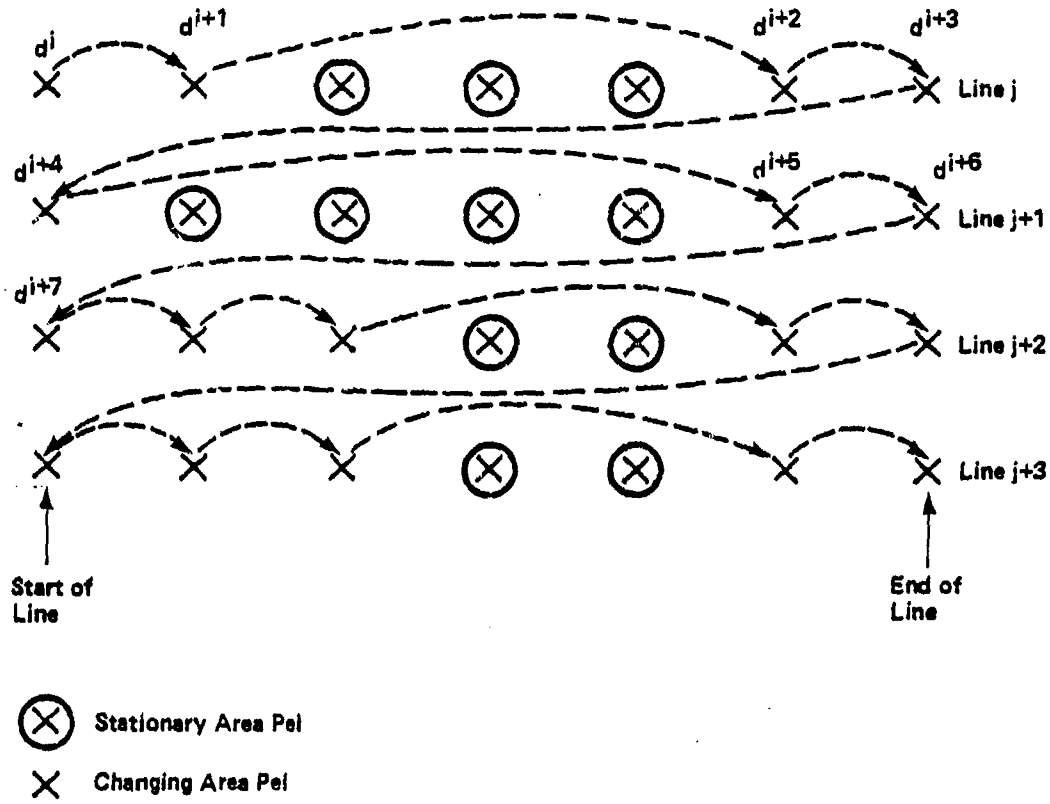
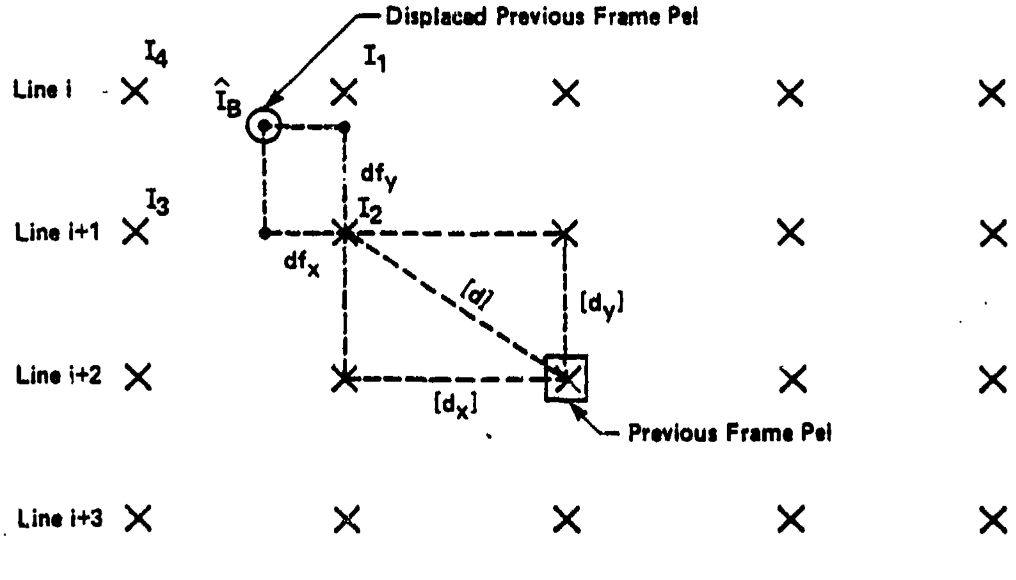
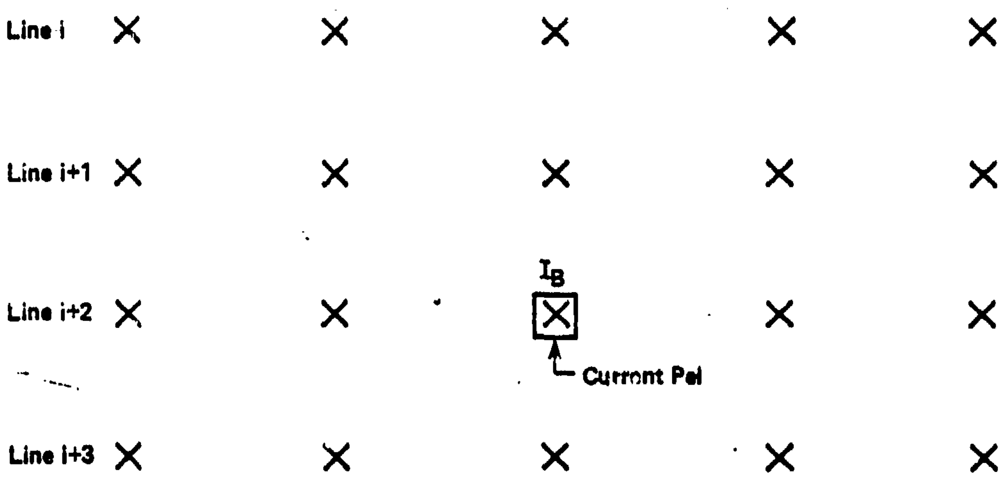


Fig. 2.11(a) Illustration of the Displacement Estimate Updating



(i) Previous Frame (Field n–2)



(ii) Current Frame (Field n)

Fig. 2.11b Illustration of Different Pels Used in the Calculation of the Displacement Estimates

The corresponding displaced frame element I is obtained by interpolation as:

$$\hat{I}_B = (1 - df_y)\cdot[(1 - df_x)\ I_2 + df_y \cdot I_3]$$
$$+ df_y \cdot [(1 - df_x)\ I_1 + df_x \cdot I_4] \qquad (2.9)$$

where df_x and df_y are the fractional parts of the displacement estimates in the horizontal and vertical directions respectively. The displaced frame difference of pel $\underline{x}$ is:

$$DFD(\underline{x}, \underline{d}^{i-1}) = I_B - \hat{I}_B \qquad (2.10)$$

A possible implementation of the recursive algorithm in the context of a movement compensated interframe multimode coder is shown in Figs. 2.12 and 2.13. In this case, the previous frame prediction is a one-frame delay, and the displaced frame difference is obtained by interpolation. It is noted that the displacement estimates need not be transmitted, being imbedded in the data. That is, the displacement estimation is carried out using previously processed picture elements, which are available at both the receiver and transmitter. In Fig. 2.12 two predictions are formed, i.e. previous frame prediction and displaced previous frame prediction. The coder switches between these two predictors depending on which one gives a lower prediction error. In order to avoid transmitting to the receiver information on which predictor has been used at the transmitter, the predictor selection rule

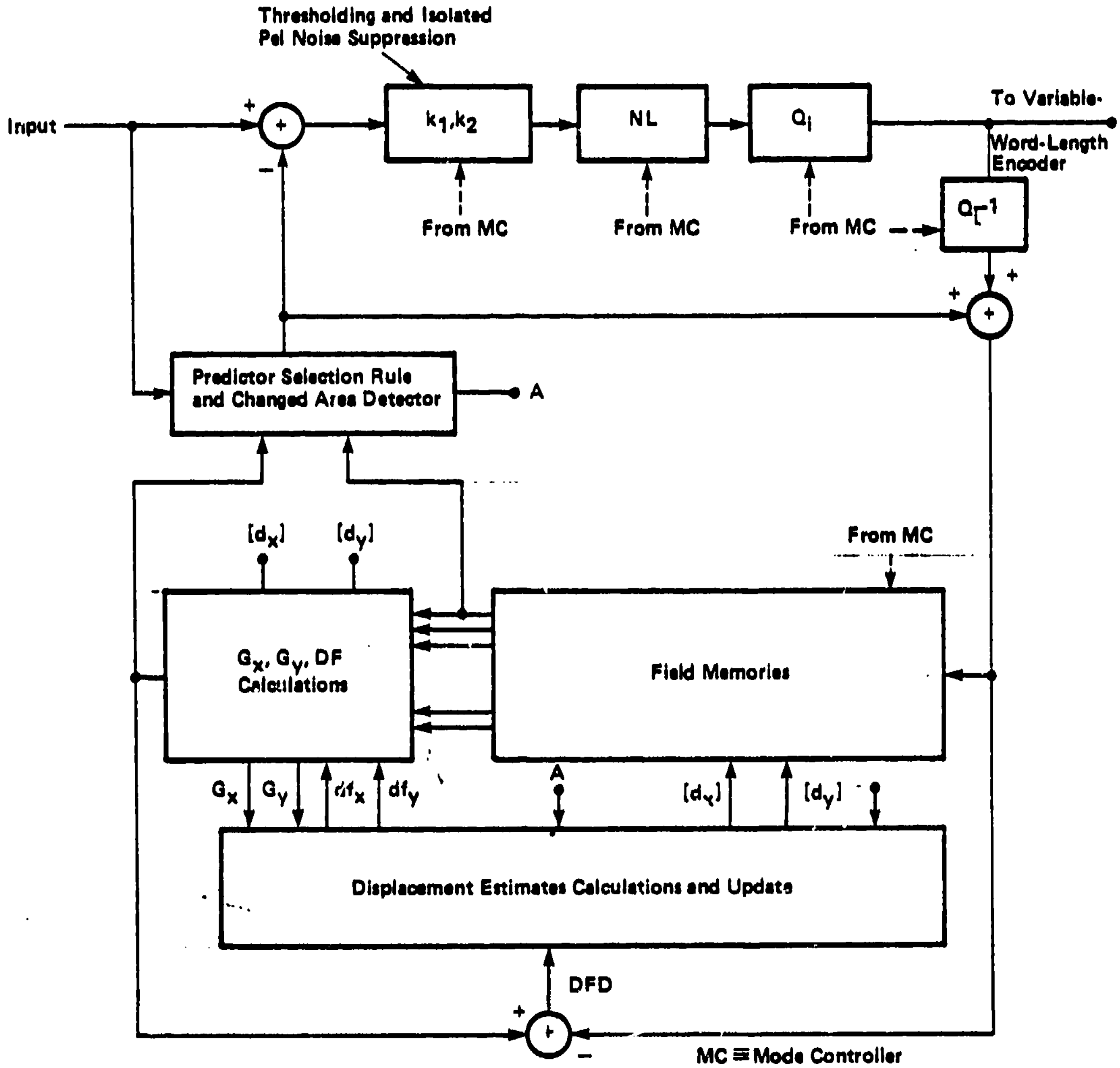
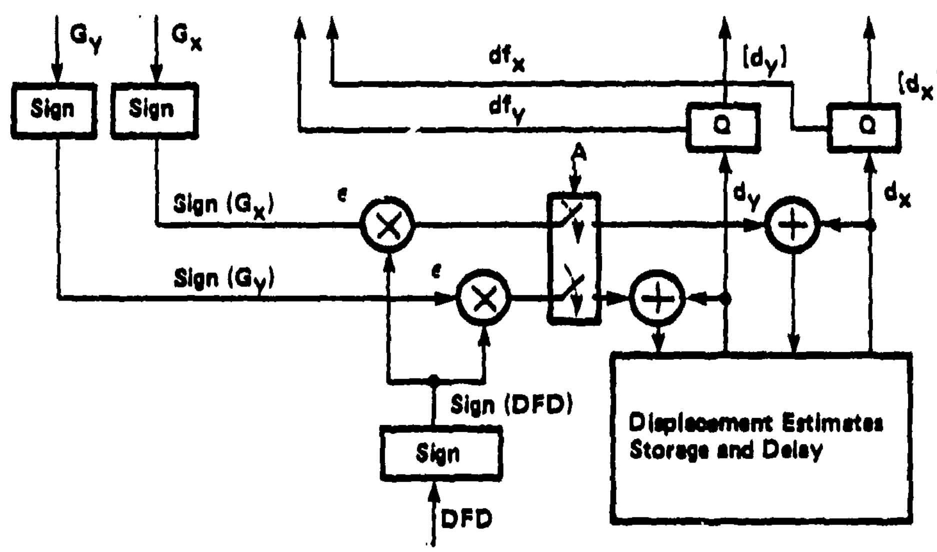


Fig. 2.12 Block Diagram of the Movement Compensated Interframe Video Coder

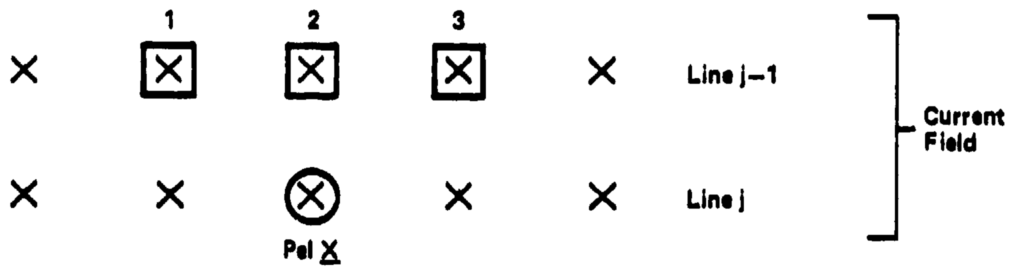


Q ≡ Quantizer DFD ≡ Displaced Frame Difference

Fig. 2.13 Displacement Estimate Calculations and Update for the Movement Compensated Interframe Video Coder has to be based on previously processed data. In Fig. 2.14 the predictor selection rule that has been used is shown. It is based on calculating a function of the frame differences and displaced frame differences at the previous line.

For a particular input pel being coded the predictor to be used is selected as previously described. The displacement from the previous line pel is then used to form the displaced frame pel. The difference signal is then passed through three circuits under control of the mode controller. The first is the Isolated Pel Noise Suppression unit with variable parameters K_1 and K_2. The second is the non-linearity (NL) or temporal filtering unit, the output of which is quantized and passed from the DPCM loop to the Variable Word-length Encoder. Referring to the paths of the transmitted difference signal, it can be seen that the coded pel is reconstructed at both the transmitter and receiver. These values are used to calculate the Frame Difference, which, if greater than or equal to a threshold K_m, causes the displacement estimate for this present pel to be updated. This means effectively, that if a changing or moving area is detected, the output "A" activates the update process shown in Fig. 2.13. To perform equations (2.6), and hence the update, the displacement estimate $\underline{d}^{i-1}$ is decomposed into the integral part ($[d_x]$,$[d_y]$) and fractional part (df_x, df_y) using the quantizer element Q. The integral part is used to locate the 4 pel window used in the calculation of G_x, G_y and DFD. The fractional part (df_x, df_y) is used in the interpolation.

As previously implied, the parameter ε controls the convergence of the algorithm and normally is chosen as large as possible subject to a stability constraint. If ε is large, the convergence rate is increased.



$$\text{Estimated Pel } \hat{u}(\underline{x},t) = \begin{cases} u(\underline{x},t-T) \text{ if } |FD_1| + |FD_2| + |FD_3| \\ \qquad < |DFD_1| + |DFD_2| + |DFD_3| \\ \\ u(\underline{x}-\underline{d},t-T)\text{; Otherwise} \end{cases}$$

DFD_i	$\equiv$	Displaced Frame Difference of Pel i
FD_i	$\equiv$	Previous Frame Difference of Pel i
$u(\cdot)$	$\equiv$	Pel Value in 2nd Previous Field
T	$\equiv$	Frame Time

Fig. 2.14 Predictor Selection Rule for Luminance Part of the Multiplexed Signal

However, doing so can cause serious problems regarding stability of the algorithm and serious oscillations could result. In addition, increasing ε will influence the accuracy of the displacement estimate. This follows as reference to Eq. (2.5) indicates that the displacement estimate can change by only $\pm\varepsilon$, thus limiting the accuracy of interpolation and/or prediction. As a compromise between speed of convergence and accuracy, ε is chosen to be 1/16.

To ensure that the displacement estimation is fairly insensitive to noise, proper segmentation into stationary and moving areas is required. Complex segmentation algorithms are possible; however, in many instances there is little distinction between noise and low contrast fine details in the picture. Thus, it has been found more practical to compare the previous frame prediction error (frame difference) signal to a thresold K_m. Updating of the displacement estimation is disabled whenever the frame difference is less than K_m. If K_m is set too low the noise would be classified as a changing area. Raising this threshold too high would result in moving areas being classified as stationary and would consequently disable the updating of displacement estimates unnecessarily. Ideally K_m should be varied according to the noise level in the signal (if it could be measured). A reasonable choice for this parameter is in the range of $K_m = 3$ to 8, and has been taken as 5.

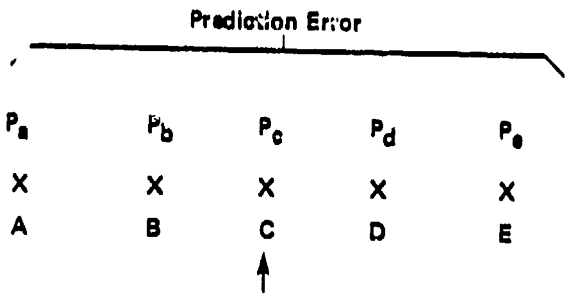
2.5.4 Isolated Pel Noise Suppression and Change of Thresholds

In order to classify a picture element or area of the picture as being predictable, the magnitude of the prediction error (FD or DFD) is compared to a threshold value K_1. Prediction errors with magnitudes less than or equal to K_1 are classified as being predictable and are set to zero. As the value of K_1, is increased, predictable picture areas will increase. However, large values of K_1 will cause parts of the moving area to be classified as background and then repeated from the previous frame. Hence the picture may appear to be moving behind a fixed pattern (the dirty window effect). In addition, low contrast details in the picture will be affected. For the range 50 kb/s to 1.35 Mb/s, K_1 is restricted to the range of 3 to 5 (out of 256). Of course at higher modes the value of K_1 is increased to reduce the amount of output data.

After the predictable and non-predictable pels have been classified, an isolated pel noise suppression is performed. This operation is illustrated in Fig. 2.15. The prediction error at pel C is set to zero if the surrounding pels A,B,D,E are predictable, and the magnitude of the prediction error at pel C is less than a threshold value K_2. The value of K_2 is varied between 5 and 20 depending on the mode of operation.

2.5.5 Temporal Filtering

Once the prediction error has been thresholded and isolated pel noise suppressed, it passes through an operation of temporal filtering. It is simply a multiplier (α) with a value between zero and one that is adaptively altered depending on the mode. Again its effect is a reduction of bit rate. At the lower modes of operation temporal filtering is not invoked (α=1); it is only invoked (α<1) at the higher overflow modes. As the magnitude of the prediction error is reduced,



Prediction Error P_c at Pel C is Set to Zero if:

$P_a, P_b, P_d, P_e < k_1$
and
$P_c < k_2$

Fig. 2.15 One Dimensional Isolated Pel Noise Suppression the inner quantizer levels, which are assigned shorter word-lengths, will be used more often. However, the temporal filtering process results in loss of temporal resolution which manifests itself as a blurring and tailing of moving objects.

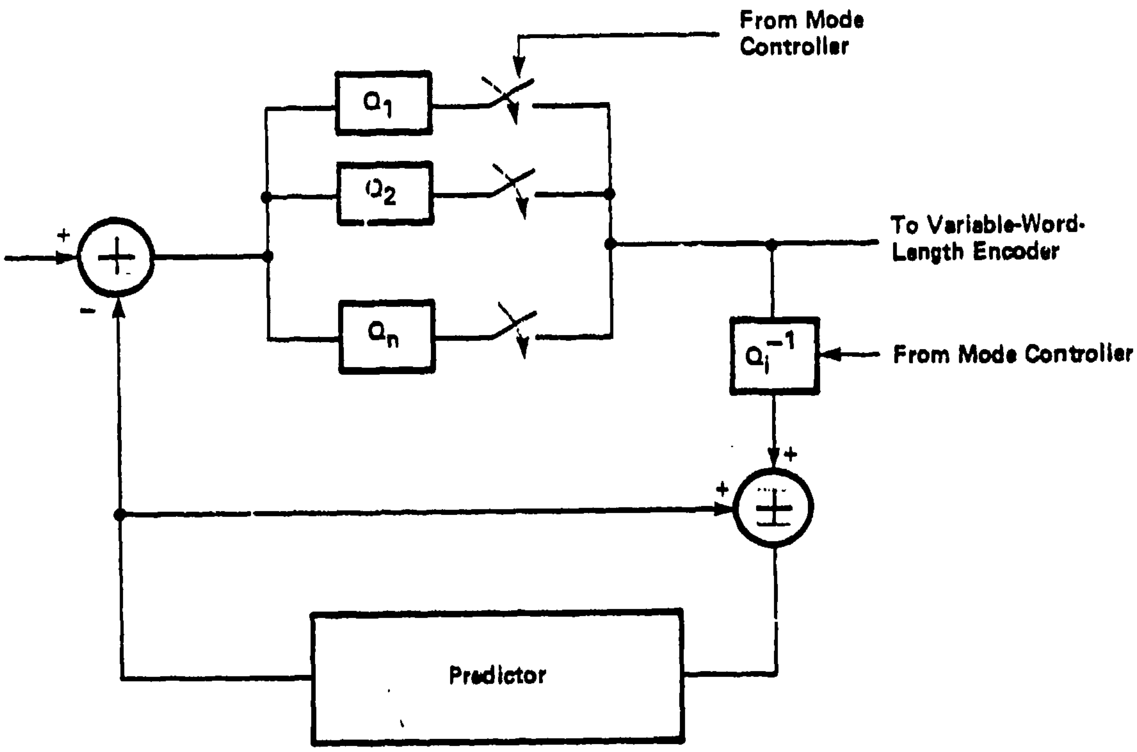
2.5.6 Switched Quantizers

The prediction error signal is quantized to a predetermined number of levels in order to reduce the transmission bit rate. In the main mode of operation the quantizer step size is chosen in the range of 3-5 depending on the transmission bit rate. In overflow modes of operation coarser quantization is invoked to further reduce the amount of information generated. The quantizer step size in these modes varies from 5 to 11.

As variable word-length encoding is used, uniform quantizers perform better than nonuniform quantizers. This is due to the fact that quantizers designed according to a minimum entropy criterion are fairly uniform. This switched quantizer concept is illustrated in Fig. 2.16 using several lookup tables. It can also be realized using a single multiplier for the case of uniform quantizers.

2.5.7 Subsampling

The previous four quantities, namely, thresholds K_1, K_2, K_m and the quantizer step size are permitted to change at the end of each line. A further quantity under buffer occupancy control is the field subsampling ratio; this differs, however, in being permitted to change only at the



$Q_i^{-1} \equiv$ Inverse Quantizer Characteristic

Fig. 2.16 Switched Quantizers

end of a field. To accomodate the range from 50 kb/s to 1.35 Mb/s, field subsampling ratios varying from 0 to 20 are allowed.

In the final overflow mode of operation, frame repeat is used as a last resort to prevent the buffer from overflowing. It is stressed that this mode of operation M will not be used under normal coder operation.

No further vertical or horizontal subsampling, other than that done in the scan converter is performed. As it is, the scan converter reduces the $4*f_{sc}$ horizontal sampling rate to either $2.5*f_{sc}$ or $1.25*f_{sc}$, depending on whether the coder is to be run at a high or a low bit rate, respectively.

2.5.8 Block Encoding and Variable Word-length Encoding

Predictable areas of the picture need not be transmitted, as this information already exists at the receiver and, hence, may be repeated. Therefore only some addressing information indicating if a group of prediction errors are significant or insignificant need be transmitted. A block coding approach is utilized in this coder. Quantized prediction errors are segmented into one dimensional blocks of 8 pels each. Each block is assigned one overhead bit - a "0" or a "1". If all prediction errors inside the block are zero, this bit is set to zero and the prediction errors are not transmitted. If at least one prediction error in the block is not equal to zero, then the overhead bit is set to "1" and all the elements in the block are encoded by the variable word-length coder and transmitted. Therefore, a constant overhead of 0.125 bits per transmitted pel is required.

In order to utilize the statistical properties of the quantized prediction errors, a variable word-length coder is used. For a given probability density distribution, the optimum variable word-length code (Huffman code) could be constructed. However, the Huffman code is sensitive to a variation of the prediction error statistics. In addition, a Huffman decoder is complex to implement and synchronization recovery, in case of channel errors, is difficult to achieve.

The variable word-length code that has been selected is shown in Table 2.2. The level zero is assigned codeword "0". The codeword lengths are symmetrical about 0, and increase by 1 for each level, starting from 3. Each word is bounded by "1" at the beginning and at the end. The number of zeros in-between identifies the level. Implementation of the encoding and decoding in this case is very simple as compared to the Huffman code and synchronization can be easily recovered when channel errors occur.

2.6 MULTIMODE CODER SIMULATION RESULTS

The multi-bit rate coder described in the previous sections has been simulated on the BNR/INRS image processing facility (DVS). A description of this facility is given in Appendix A. The bit rates used in the simulation are shown in Table 2.1. Several video sequences containing head-and-shoulder views have been used as input to the coder. The sequences contained varying amounts of motion in order to evaluate the coder performance under different conditions. Colour photos of one frame of the video sequences, "HARVEY" and "NEWSCASTER", are found in Appendix B.

QUANTIZATION LEVEL	CODE WORD	CODE LENGTH
.	.	.
.	.	.
.	.	.
8	1000000001	10
7	100000001	9
6	10000001	8
5	1000001	7
4	100001	6
3	10001	5
2	1001	4
1	101	3
0	0	1
-1	111	3
-2	1101	4
-3	11001	5
-4	110001	6
-5	1100001	7
-6	11000001	8
-7	110000001	9
-8	1100000001	10
.	.	.
.	.	.
.	.	.

TABLE 2.2: QUANTIZER AND VARIABLE WORD-LENGTH CODE

As explained in Section 2.4, the scan converter provides a digital video signal at either $1.25*f_{sc}$ or $2.5*f_{sc}$ formats. The first is used for the bit rates of 50 kb/s and 100 kb/s while the second is used from 200 kb/s to 1.35 Mb/s. The multimode coder parameters are given in Table 2.3 for the high bit rate range and in Table 2.4 for the low bit rates.

In the following sections, the simulation results obtained for the aforementioned video sequences are discussed. The importance of using the technique of movement compensation is demonstrated, followed by the coder performance over the entire bit rate range. As an objective measure of performance, the buffer occupancy is determined for each sequence at different bit rates. This indicates which modes of operation the coder used in order to achieve the required transmission bit rates. Since higher modes of operations normally introduce degradation gracefully into the pictures, the buffer occupancy gives an indication of picture quality. In addition, picture quality of the processed video sequences at different bit rates has been informally evaluated.

The buffer occupancies for the low bit rate ranges are shown in Figs. 2.18a and 2.19a, while the high bit rate ranges appear in Figs. 2.18b and 2.19b. Note, however, that the entry point for Fig. 2.19b is MODE 2.

	MODE OF OPERATION (HIGH BIT RATES)																	
CODER PARAMETER	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Quantizer Step	1	3	5	5	7	5	7	5	7	5	7	5	7	5	7	9	11	-
Threshold K_1	0	3	3	3	3	5	5	5	5	5	5	5	5	5	5	5	5	255
Threshold K_2	0	5	5	5	5	5	5	5	5	5	5	5	5	6	7	8	9	255
Threshold K_m	-	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	-
ALPHA (α)	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.75	.75	0
Field Subsampling Ratio	-	0*	0	2	2	4	4	6	6	8	8	10	10	12	12	14	20	∞
Frame Repeat	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES
T_i**	- ***	6	13	19	25	31	38	44	50	56	63	69	75	81	88	94	100	-
$\tilde{T}_i$	-	0	3	9	16	22	28	34	41	47	53	59	66	72	78	84	91	-

* No Field Subsampling indicated by 0

** Forward (T_i) and Backward ($\tilde{T}_i$) Thresholds in terms of percentage buffer occupancy $[=(R_{i,j}/32K) \times 100]$

*** Buffer occupancy levels offset

TABLE 2.3: MULTI-BIT RATE MULTIMODE CODER PARAMETERS FOR HIGH BIT RATE RANGE (256 Kb/s +)

	MODE OF OPERATION (LOW BIT RATES)																	
CODER PARAMETER	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Quantizer Step	1	5	7	5	7	5	6	5	7	5	7	5	7	9	11	9	11	-
Threshold K_1	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	255
Threshold K_2	0	5	5	5	5	5	5	5	5	6	7	8	9	10	16	20	20	255
Threshold K_m	-	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	-
ALPHA (α)	-	1	1	1	1	1	1	1	1	1	1	.75	.75	.75	.5	.5	.5	0
Field Subsampling Ratio	-	4	4	6	6	8	8	10	10	12	12	14	14	16	16	16	20	∞
Frame Repeat	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES
T_i*	-	6	13	19	25	31	38	44	50	56	63	69	75	81	88	94	100	-
$\tilde{T}_i$	** -	0	3	9	16	22	28	34	41	47	53	59	66	72	78	84	91	-

* Forward (T_i) and Backward ($\tilde{T}_i$) Thresholds in terms of percentage buffer occupancy $[=(R_{i,j}/32K) \times 100]$

** Buffer occupancy levels offset

TABLE 2.4: MULTI-BIT RATE MULTIMODE CODER PARAMETERS FOR LOW BIT RATE RANGE (64 Kb/s, 128 Kb/s)

2.6.1 Impact of Movement Compensation

A key factor in achieving acceptable picture quality in the multimode coder is the incorporation of movement compensation. Its effectiveness depends on the amount of motion in the image. HARVEY and NEWSCASTER were tested at 50 kb/s using just a previous frame predictor as well as a movement compensated prediction technique. The results are shown in Figs. 2.17a and 2.17b, respectively.

Without movement compensation in HARVEY, the rapid nodding of the head after about field 70 causes the buffer to overflow and initiate the frame repeat mode. Movement compensation, on the other hand, not only prevents buffer overflow, but achieves an overall lowering of the percentage buffer occupancy and a subsequently improved picture quality.

This improvement proves to be even more substantial with NEWSCASTER. Without movement compensation, buffer overflow occurs with the large motion at the beginning of the sequence, while high modes of operation persist for the rest. The effect of movement compensation during this fast motion is to reduce the percentage buffer occupancy to 40 whilst operating at MODE 6 or lower. This is particularly relevant as the type of motion is considered to be highly similar to that of a typical teleconference participant.

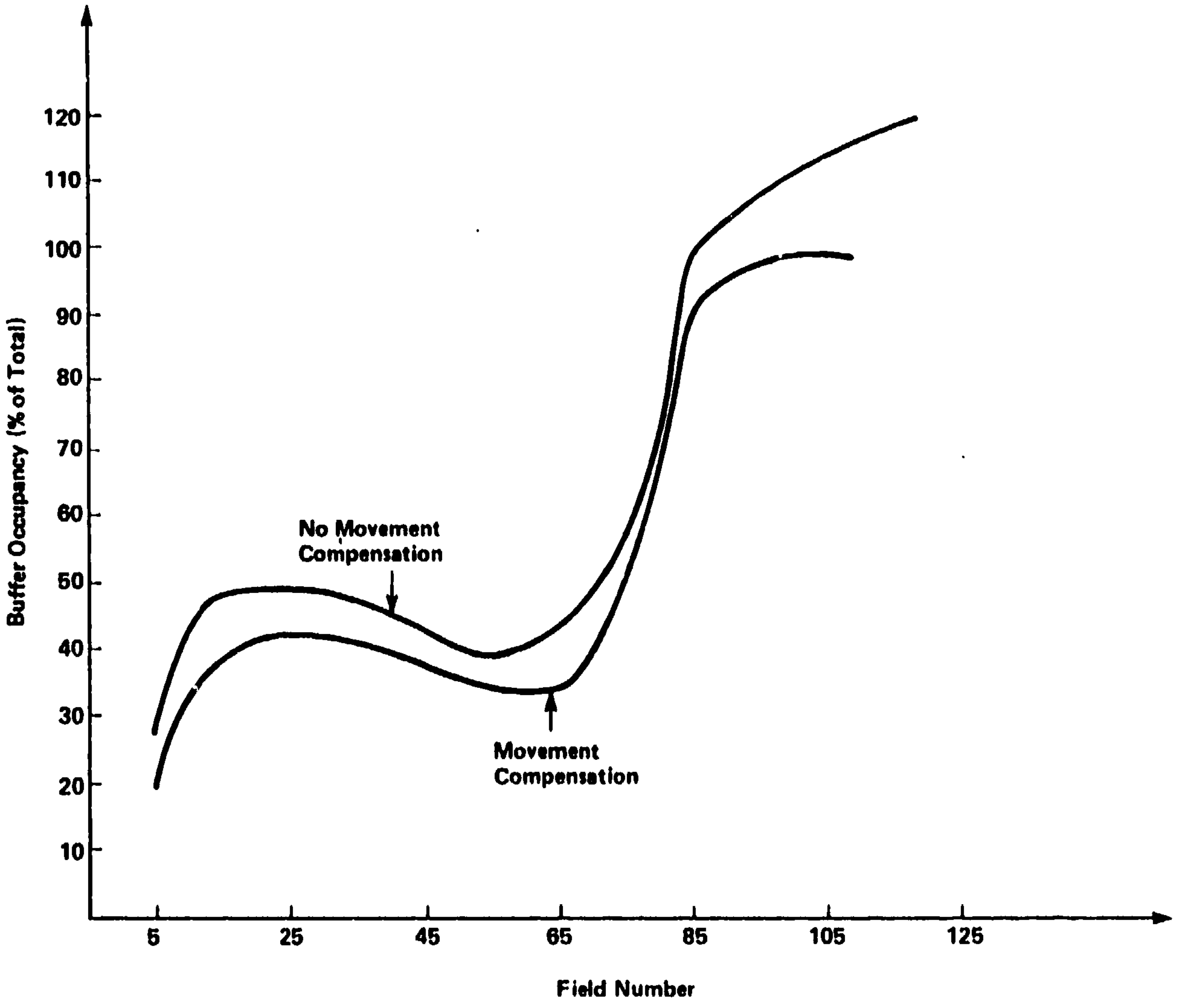


Fig. 2.17a Buffer Occupancy for Video Sequence "HARVEY", With and Without Movement Compensation, at 50 kb/s (1.25 *f_{sc} Format)

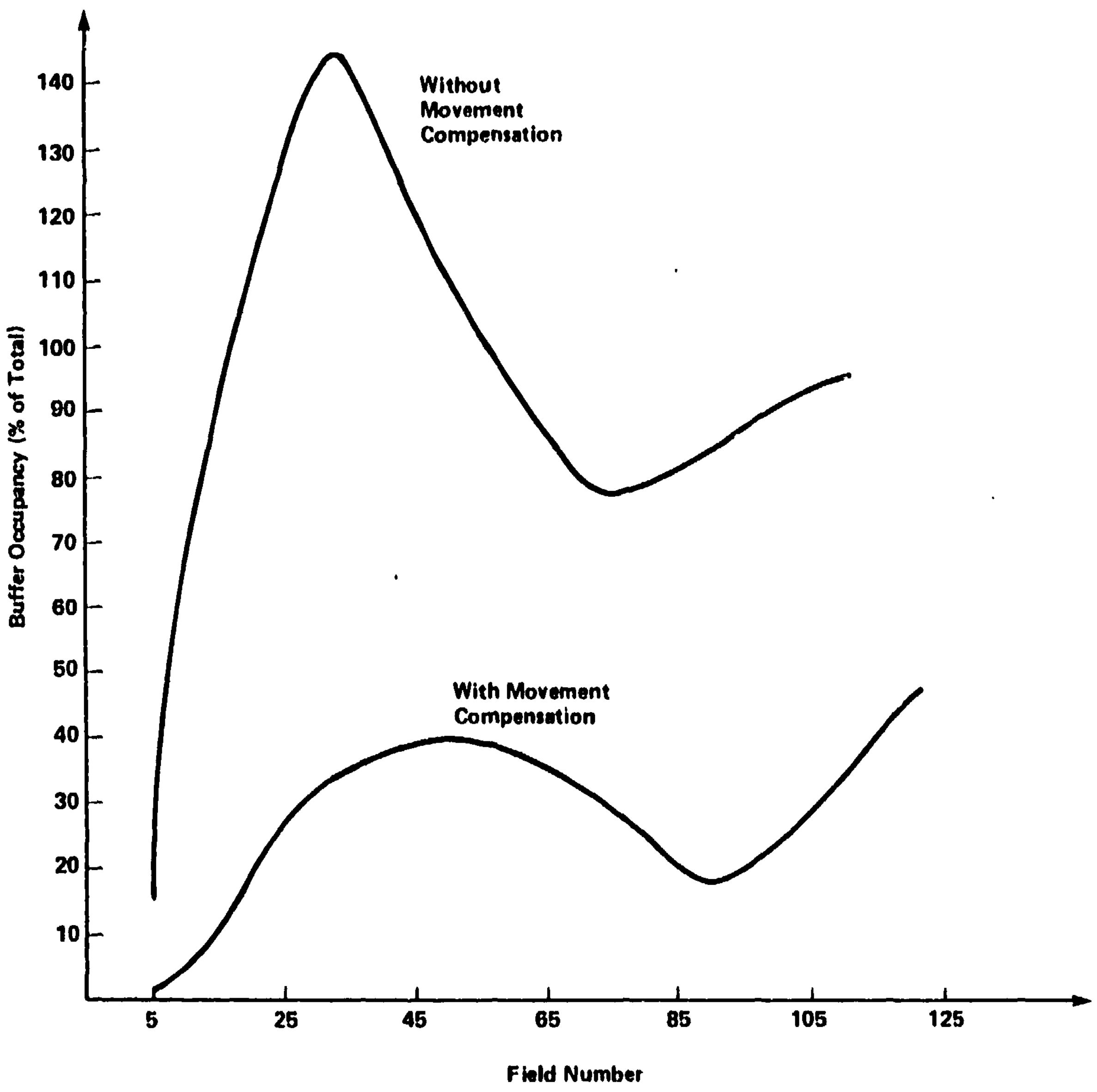


Fig. 2.17b Buffer Occupancy for Video Sequence "NEWSCASTER", With and Without Movement Compensation, at 50 kb/s

2.6.2 Multi-bit Rate Coder Simulation Results

For the two sequences at 1.35 Mb/s, the quality of the coded pictures is excellent. In fact, the coder has utilized the main mode of operation M_0 and quite often the underflow mode of operation M_{-1} is used.

The quality obtained at 750 kb/s for both sequences is close to that at 1.35 Mb/s. The buffer occupancy increases sufficiently to force NEWSCASTER from the underflow mode to operate for 10 percent of the time in MODE 1. With the rapid head motion of HARVEY after field 70, the major mode of operation is MODE 2.

At 375 kb/s, the quality is again very good with a slight amount of jerkiness under fast motion. This is reflected by the fact that NEWSCASTER spends 82 percent of the time in Modes 1 and 2, HARVEY spending 73 percent of the time in modes 2 and 3.

At the bottom end of the high bit rate range, 200 kb/s, quantization noise is slightly visible. In regions of little motion, such as between fields 55 and 90 for NEWSCASTER and up to field 70 for HARVEY, the motion rendition is very good, i.e., not much jerkiness. Outside these regions, the fast head motion in HARVEY causes MODE 14 to be briefly attained (77% Buffer Occupancy) while NEWSCASTER reaches MODE 5 (32% Buffer Occupancy), so that some jerkiness is visible.

Inspection of Figs. 2.18a and 2.19a indicates that a horizontal subsampling factor of 4 (i.e. $1.25*f_{sc}$) is necessary if both NEWSCASTER and HARVEY are to be processed at bit rates less than 200 kb/s with acceptable picture quality. This is evident with HARVEY, where if a horizontal subsampling factor less than 4 were to be used, the buffer would overflow at 50 kb/s. At 50 kb/s and 100 kb/s the spatial aliasing which results is visible but is not annoying. The buffer occupancy curves are as expected. From Fig. 2.18a it can be seen how, at the beginning of NEWSCASTER, motion is fairly large leading to a rapid increase in occupancy. This is followed by a little motion part enabling the coder to gradually empty the buffer. The large motion area of HARVEY initiates high modes of operation and hence high field subsampling ratios. This causes blurring and jerkiness of the moving areas. This effect is not as serious with NEWSCASTER, however, due to a smaller amount of abrupt motion. Nevertheless, for the application at hand, the picture quality at these low bit rates is deemed to be adequate.

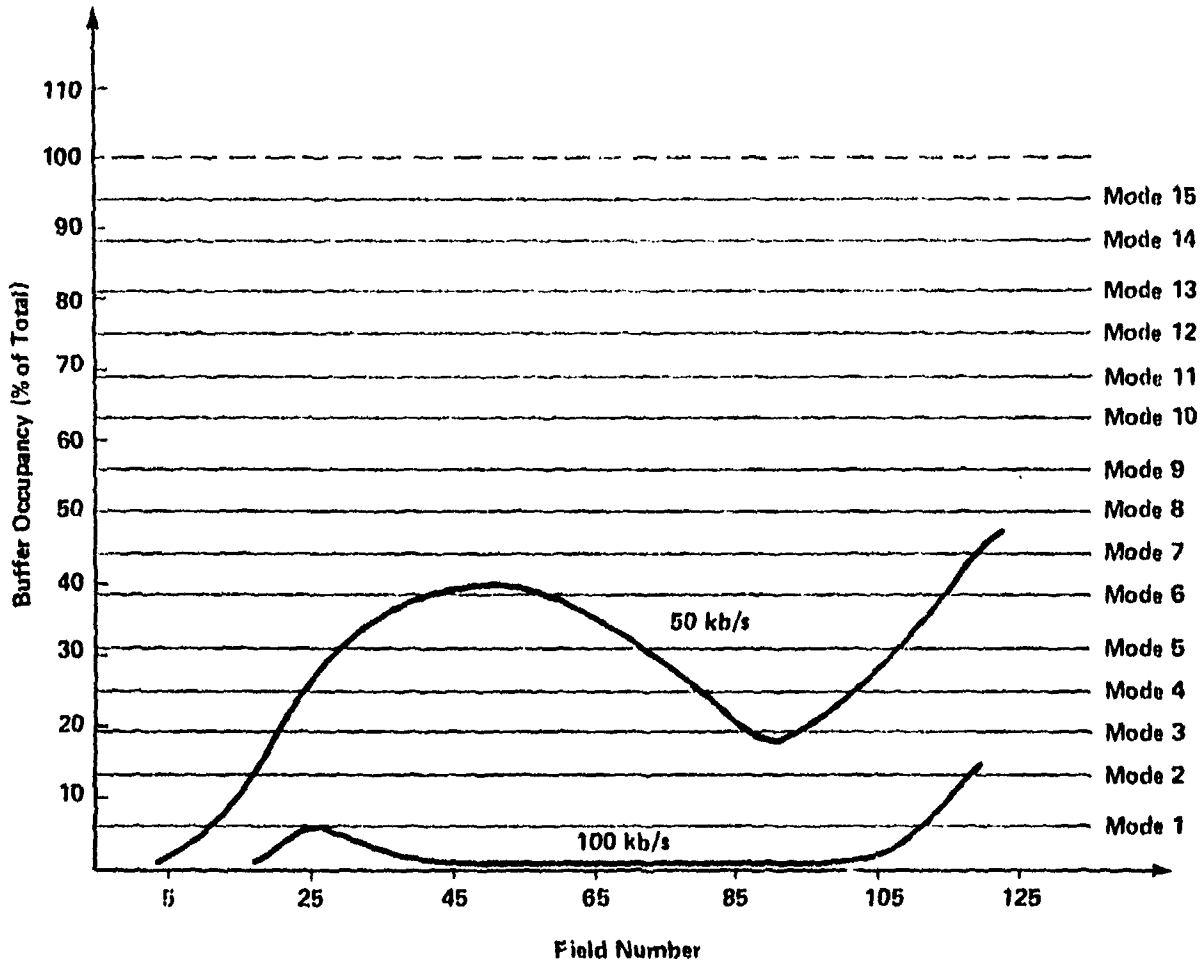


Fig. 2.18a Buffer Occupancy for Video Sequence "NEWSCASTER" Two Bit Rates of 50 kb/s and 100 kb/s are Shown. (Forward Thresholds for Entering Each Mode Shown)

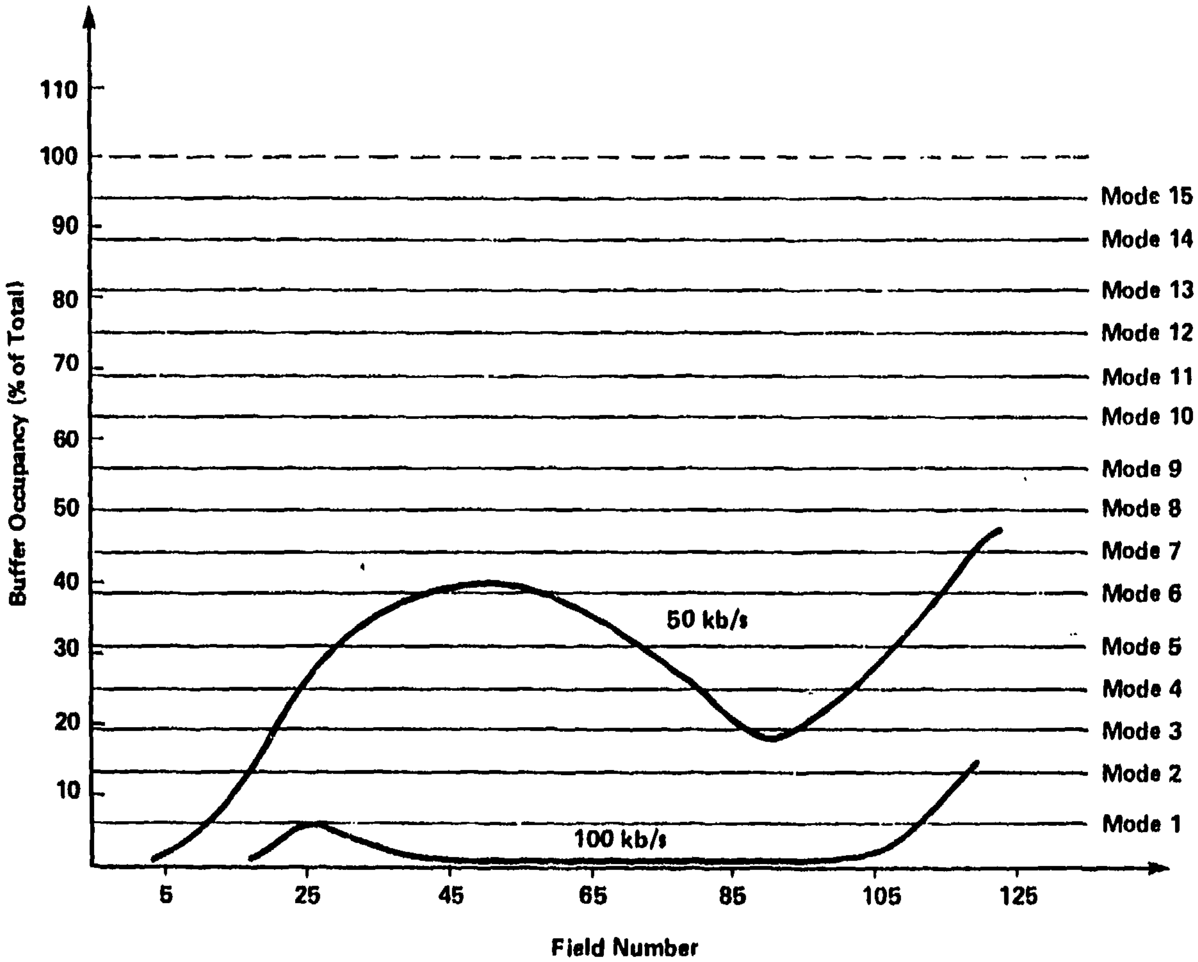


Fig. 2.18a Buffer Occupancy for Video Sequence "NEWSCASTER" Two Bit Rates of 50 kb/s and 100 kb/s are Shown. (Forward Thresholds for Entering Each Mode Shown)

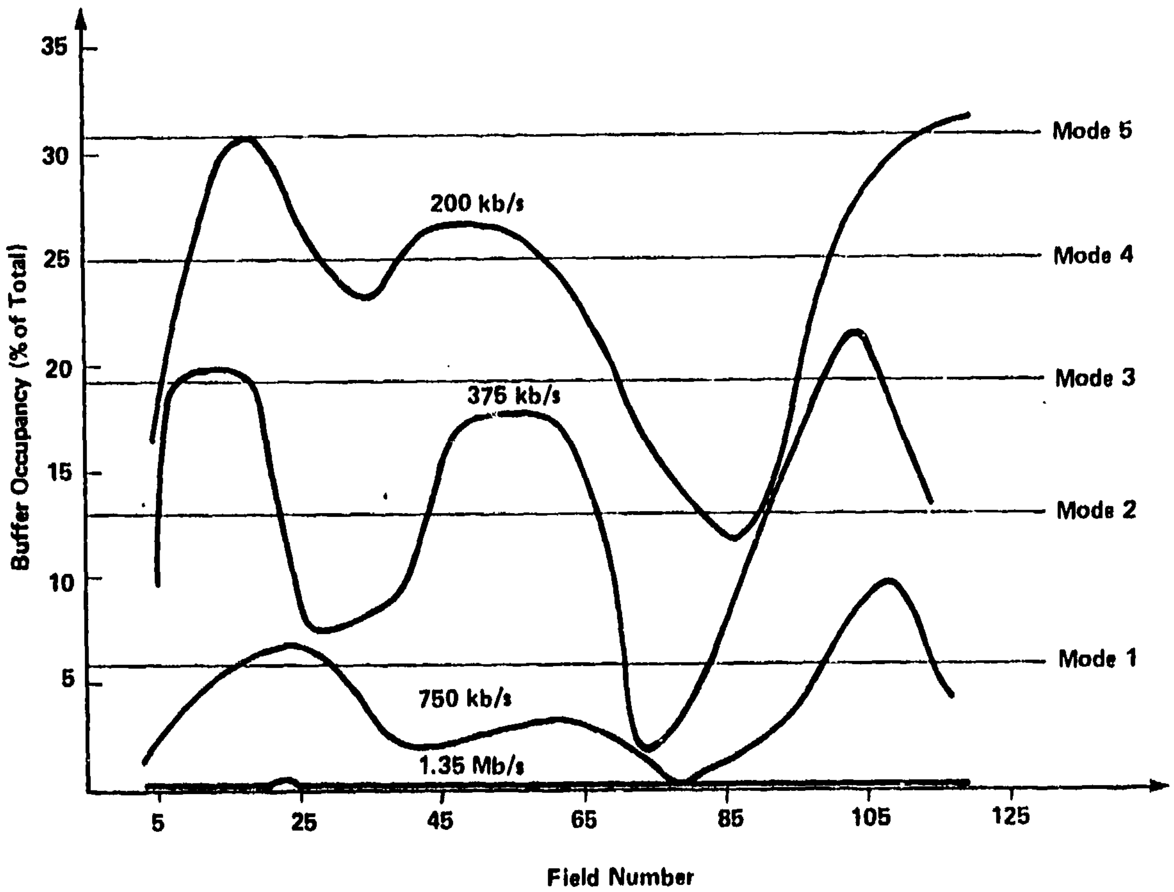


Fig. 2.18b Buffer Occupancy for Video Sequence "NEWSCASTER" at Different Bit Rates; Forward Thresholds for Entering Each Mode Shown

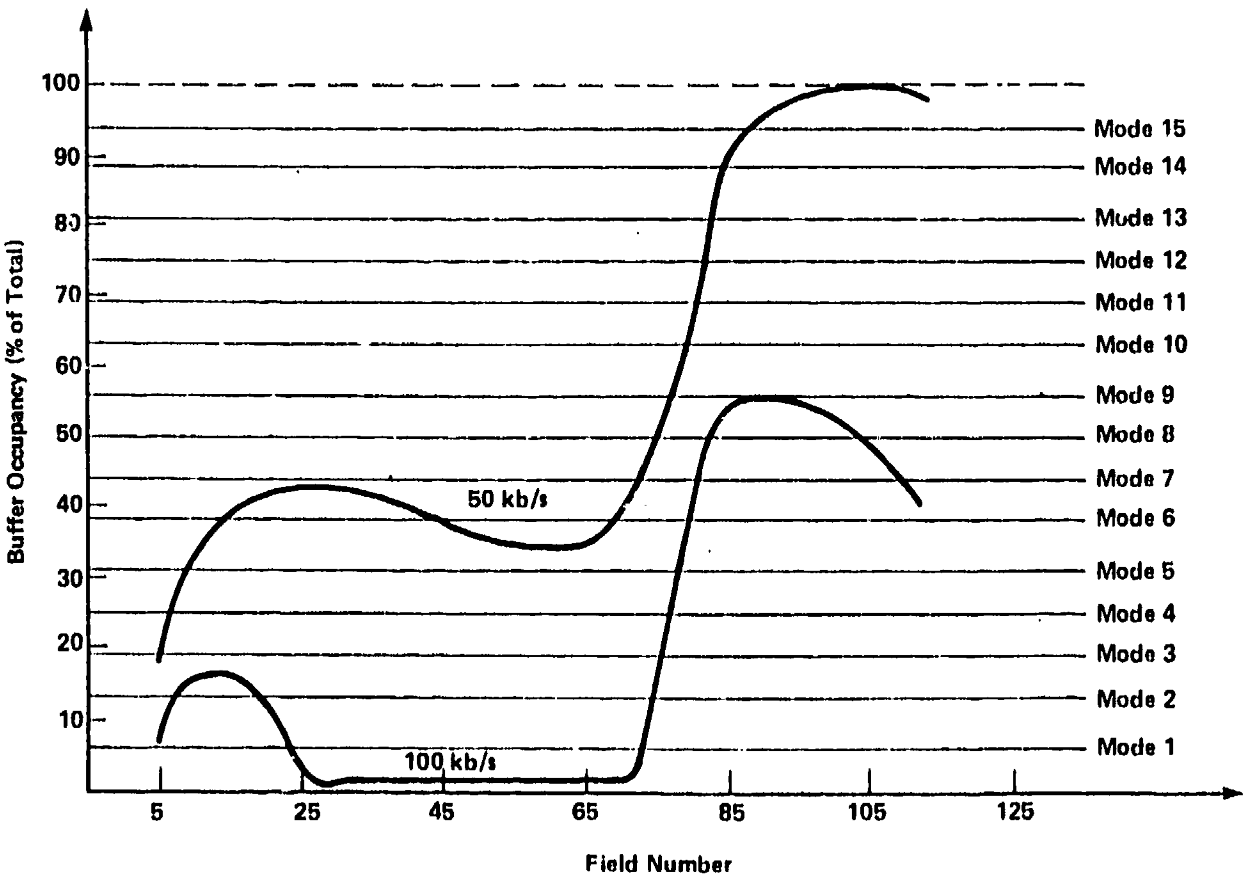


Fig. 2.19a Buffer Occupancy for Video Sequence "HARVEY"; Two Bit Rates of 50 kb/s and 100 kb/s are Shown; Forward Thresholds for Entering Each Mode are Indicated

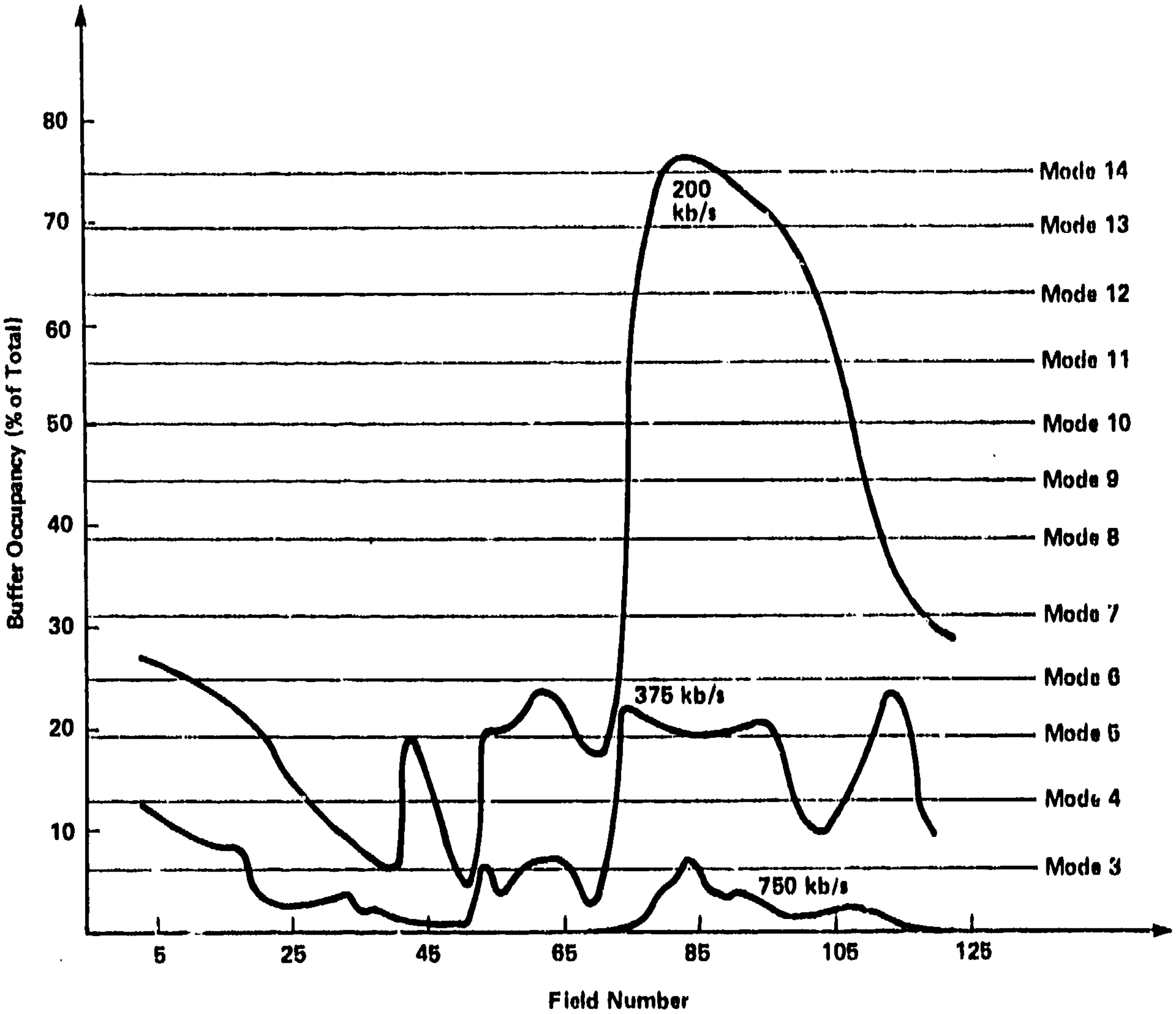


Fig. 2.19b Buffer Occupancy for Video Sequence "HARVEY" at Various Bit Rates

CHAPTER 5

CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

In this study a multi-bit rate multimode movement compensated interframe video coder for video conferencing applications has been presented. Simulation of the coder operation at different bit rates ranging from 1.5 Mb/s to 64 kb/s has been carried out on the BNR/INRS image processing facility DVS. The results obtained are very promising especially at the low bit rate end. Good picture quality is obtained at 256 kb/s and acceptable picture quality is obtained at 64 kb/s. A key element in achieving this goal is the incorporation of a motion compensation technique that can operate at different bit rates in conjunction with other data rate reduction techniques.

Proposed future work on this project will involve carrying out a system design for the codec. Special emphasis should be placed on the lower end of the bit rate, i.e., coder operation at 256 kb/s - 64 kb/s. The system design involves identifying implementation alternatives, taking into account state-of-the-art high speed technology and economic considerations.

In addition, investigation of techniques for improved handling of very large amounts of motion at the lower bit rates should be carried out. This will improve picture quality, especially at the 64 kb/s rate. The impact of channel errors on the coder operation and picture quality should be examined and suitable error correction and/or concealment techniques identified.

Appendix A. A DIGITAL TELEVISION SEQUENCE STORE (DVS)

A.1 INTRODUCTION

The DVS is a general purpose simulation facility for Processing television pictures, especially moving sequences. It provides facilities for real time acquisition and display of digitized colour (NTSC) or black and white sequences. It operates in a non-real time processing environment providing the user random access to the stored data and the flexibility to simulate different processing algorithms. The processed sequences can be displayed in real-time and compared.

The system design concept of DVS [12] involves the use of several moving-head, removable-pack, disk drives operating in parallel to provide the necessary bit rate capability. A semi-conductor buffer memory, which has a high-speed port to accept digitized video and a low speed port to communicate with the disk, is associated with each disk drive. The current implementation of the DVS at the BNR/INRS Signal Processing Laboratory permits a maximum sequence length of 80 seconds and can record/display a 256 x 212 sub-array of the entire frame. However, the DVS has been designed in a modular fashion, allowing expansion in terms of increased data rate (resulting in increased window size of the picture) and/or increased storage capacity (resulting in longer sequences).

The DVS is supported by a VAX 11/780 computer operating under VMS.2.4 in a multi-user environment.

A.2 CAPABILITIES

The NTSC composite signal can be sampled at 2,3 or 4 times the colour sub-carrier frequency (7.16, 10.7, or 14.3 MHz) and can be linearly quantized up to 256 levels (8 bits). The system can easily be enhanced to accomodate other sampling frequencies below 15 MHz.

The DVS has two modes of recording: in one mode, it records a video sequence of a pre-determined length starting at a given time. In the second mode, DVS simulates a recording loop continuously recording the last n seconds of video ($n \leq 80$ s). The recording process can be stopped at any time to preserve the last n seconds of recording. The first mode is useful for automatic sampling of broadcast material, while the second mode is useful in capturing an event after it has occurred. DVS is also capable of recording every m^{th} frame of a sequence.

There are several display modes. The DVS can display a sequence of predefined length either repetitively or in "palindromic" mode. In the latter mode, an arbitrary sub-sequence of a recorded sequence is repetitively displayed, first in the forward direction and then the reverse direction. This makes it possible to present motion without an abrupt discontinuity at the end and the beginning of the sequence. DVS has facilities for slow-motion display as well as for stepping through video frames one by one. An important display feature of DVS is the capability to switch back and forth between several sequences without seeing transient effects on the monitor. For non-real-time processing of video data, the user has random access to the recorded data.

A.3 THE SYSTEM

A block diagram of the DVS is shown in Fig. 2.20. This is a two-disk configuration of DVS. The disks used are CDC 9762-1 80 Mbyte storage module drives (SMD's). These drives have a burst transfer rate of about 1.2 Mbytes/s. Each disk is provided with a high-speed semiconductor buffer memory of 256 kbytes.

The General Video Controller (GVC) includes the digital video switch which connects one buffer memory to either the A/D or D/A converter, the digital time-base generator and the analog television interface. The GVC has an interface to the VAX 11/780 through which it receives control information and transfer timing information.

Each Field Storage Unit (FSU) consists of the semiconductor buffer with high-speed video port, a Computer Bus (Channel) Interface, a disk adapter, a disk controller and a disk drive. The channel interface links the computer to both the buffer and the disk adapter.

The analog video signal is connected to the video ports of the high-speed semiconductor buffer memories via the GVC. The incoming digitized video windows are switched from one buffer memory to another in a cyclical fashion. This "round-robin" mode of operation makes the disks work in parallel and doubles the transfer rate. In the particular implementation, six field windows (1/10 s) are sent before the video input stream is switched to the other buffer. The disk transfer "start up" delay is thus incurred every 1/10 s instead of every 1/60 s, thus increasing the throughput and providing a larger window.

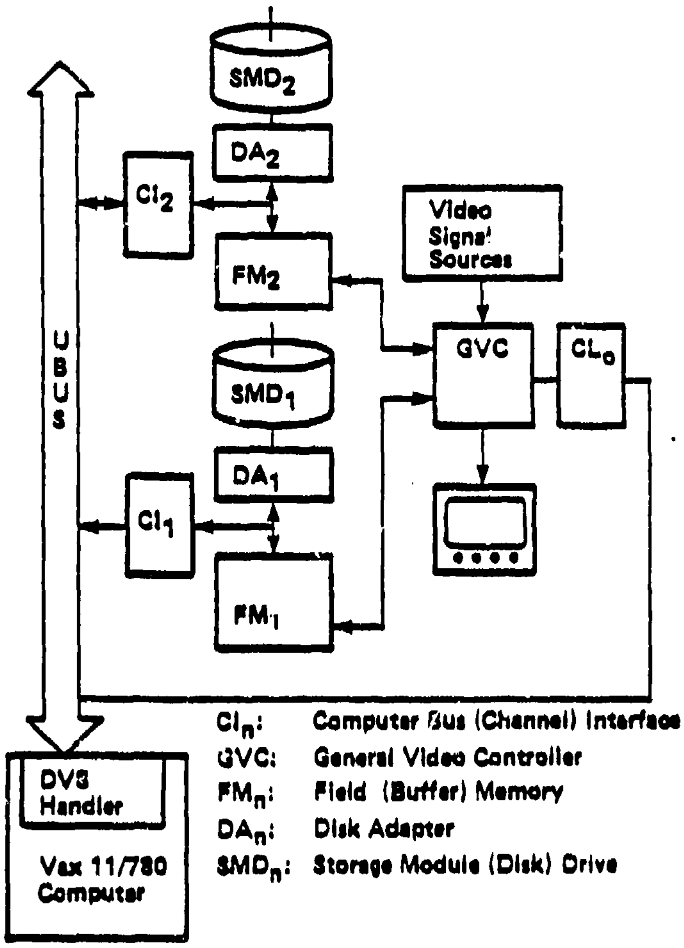


Fig. 2.20 **General Block Diagram of the DVS Image Processing Facility**

A.4 SOFTWARE

DVS has a comprehensive support software package resident in the host computer. DVS has been integrated in the multi-user environment by means of a device handler, which is a special task under VMS-2.2. The handler provides the software interface between a particular hardware device and the application program using that device.

The DVS software support can be sub-divided into four basic functions:

1. Data Base Management,
2. Real-time System Control,
3. Data Processing Support,
4. System Maintenance and Calibration.

The DVS has been designed to support several users. At any time, only one user can access DVS; however, each user can have one or more VISTAs (Visual Information Storage Area) defined on the disk. The data base management system of the DVS provides the users with the following facilities:

a) User Segregation and Data Security,
b) Dynamic Resource Allocation,
c) Archiving Facility,
d) Simple User Interface,
e) Access to Physical Parameter Information.

The software for real-time system control provides the user with convenient facilities for recording video sequences in pre-defined areas in the file system and for subsequent display of sequences/subsequences. The DVS disk data format has been devised to minimize the cylinder-to-cylinder switching time. The fields recorded during the forward motion of disk heads are interleaved with those recorded during the return trip for palindromic display. For "glitch-free" switching between several sequences, it is possible to interleave two or more sequences in the same fashion.

The data processing support software provides convenient facility for reading data form stored sequences and writing processed sequences into file areas. Under the control of the user, video data is transferred between the computer and DVS using direct memory access (DMA) techniques incurring negligible CPU overhead.

The system maintenance and calibration software provides facilities that aid in monitoring the integrity of some of the key components and permit adjustment of system variables to comply with user-defined specifications or with pre-defined default values.

APPENDIX B

PICTURES OF ONE FRAME OF THE MAIN VIDEO SEQUENCES USED IN THE SIMULATION

(a) NEWSCASTER (2 seconds)

(b) HARVEY (2 seconds)

References

1. A. Deczky and M.S. Sabri, "22 Mbit/sec Digital Colour Television CODEC", BNR TR380, January 1977.

2. M.S. Sabri, "Interframe Codecs and Temporal Filtering of Digital Video Signals," BNR, TM 32029, December 1978.

3. E. Dubois, "Movement Compensated Predictive Coding: Displacement Estimation and Coder Simulations," BNR, TM 32059, November 1979.

4. E. Dubois, B. Prasada and M.S. Sabri, "Image Sequence Coding," Ch. 3 of Image Sequence Analysis, T.S. Huang ed., Springer-Verlag, New York 1981.

5. Candy et al., "Transmitting Television as Clusters of Frame-to-Frame Differences," BSTJ, Vol. 50, pp. 1889-1917, July-August 1971.

6. J. Klie, "Encoding of Video Signals for Low Transmission Bit Rates," GERMAN VERSION, TU HANNOVER, 1979.

7. J.L.E. Baldwin, "Sampling Frequencies for Digital Coding of Television Signals," IBA Tech. Review, No. 9, pp. 32-35, September 1976.

8. G.E. Forsythe, M.A. Malcolm, C.B. Moler, "Computer Methods for Mathematical Computations," pp. 63-83, Prentice Hall, Inc., 1977.

9. A.N. Netravali and J.D. Robbins, "Motion-Compensated Television Coding: Part I," BSTJ, Vol. 58, pp. 631-670, March 1979.

10. C. Cafforio and F. Rocca, "Detection and Tracking of Moving Objects in Television Images," in Record of International Workshop on Image Processing: Real-Time Edge and Motion Detection/Estimation, doc. Zich. CCETT: CTN/T/1/80, pp. 19.1-19.10, September 1979.

11. E. Dubois, M.S. Sabri, J.Y. Ouellet, "Three Dimensional Spectrum and Processing of Digital NTSC Colour Signals," SMPTE: Digital Video, R. Marcus, ed., Vol. 3, pp. 72-84, June 1980.

12. R. Johnston et Al., "A Digital Video Sequence Store," IEEE Transactions on Communications, COM 26, p. 594, 1978.