Visual Human Intent Analysis Survey

Michael S. Del Rose, Christian Wagner

Abstract— Visual Human Intent Analysis (VHIA) is a growing field of research devoted to algorithms that categorize human behavior through input of visual image sequences. In VHIA, it is difficult to make an implementation framework that is robust enough to handle the many non-deterministic ways of human physical actions. This paper will survey several techniques currently used in VHIA including non-traditional artificial intelligence techniques, visual languages, statistical algorithms, and others. It will give the reader the background and understanding of how open this domain is to new algorithms; improving on the overall research in VHIA or similarly related problems.

Keywords— Activity recognition, artificial intelligence, visual human intent analysis, visual understanding.

.

I. Introduction

As the technologies for developing Visual Understanding systems (VU) moves toward full commercial possibility, the desire to take VU algorithms out of the research labs and into real-world applications is growing. One sub research area of VU is Visual Human Intent Analysis (VHIA), also referred to as *visual human behavior identification*, *visual action or activity recognition*, and *understanding human actions from visual cues*. In static self security systems, *visual human behavior identification* systems will aide or replace security guards monitoring CCTV feeds. Television stations will require *activity recognition* systems to automatically categorize and store video scenes in a database for quick search and retrieval. The military is pushing robotics to replace the soldier, thus requiring the need to *understand human actions from visual cues* to determine hostile actions from people so the robot can take appropriate actions to secure itself and its mission. These are just a few areas where VHIA will increase current state of the art in the development and use of future systems. This paper surveys current research in the area of Visual Human Intent Analysis.

II. NON-TRADITIONAL TECHNIQUES

A good deal of the work in VHIA uses visual cues of the human actions without any traditional computational intelligent techniques [10, 12, 15, 17, 23, 25, 27, 31, 33, 34, 37, 43, 46, 47, 49, 50, 52, 55, 61, 62, 63]. These algorithms rely on simplicity at the cost of fusing input data or they are using less than typical data inputs. They rely almost exclusively on the pre-processing of the data while using statistical or non-traditional AI algorithms to determine the behavior. For example, M. Cristani et. al. [15] uses both audio and visual data to determine simple events in an office. First, they remove foreground objects and segment the images in the sequence. This output is coupled with the audio data and a threshold detection process is used to identify unusual events. These event sequences are put into an audio visual concurrence matrix (AVC) to compare with known AVC events.

Many systems use their own form of plotting space-time data from the image sequences to calculating the closest distance to pre-determined or automatically determine events to decide what action the human is performing [10, 12, 16, 17, 19, 23, 25, 27, 30, 31, 34, 35, 43, 50, 55, 59]. For example, M. Dimitrijevic et. al. [17] developed a template database of actions based on five male and three female individuals. Each human action is represented by a set of three frames of their 2D silhouette: the frame when the person first touches the ground with one of his/her foot, the frame at the midstride of the step, and the end frame when the person finishes touching the ground with the same foot. The three frame sets were taken from seven camera positions. When determining the event, they use a modified Chamfer's distance calculation to match to the template sequences in the database.

In another example, D. Weiland et. al. [10] uses motion history volumes to determine human gestures by extending the 2D pixel representation with time to make a 3D representation. This is accomplished by using multiple cameras around the person and subtracting out any background. Classes are created manually for each action or gesture. Mahalanobis distance with principle component analysis is used to identify action from the appropriate class.

The work from A. Mokhber et. al. [23] uses volumetric models to determine human actions. Features representative of the action sequence are extracted from the binary images and used to compute the space-time volumes. This is so people are seen globally. The volumes are formed from all the binary images in the action and are concatenated together in chronological order. These volumes make up the feature vector of moments. Comparison of testing data into classes formed from the training data is by Mahalanobis distances.

Eigenspace is used to help categorize actions for distance computations to identify events by M. Rahman et. al. [25]. They believe that these should be used since they are highly mathematical and require less image processing. Motion from a camera is captured and manually placed into classes and then finally into a covariance matrix. This makes up the universal image set for that action. Eigenvalues and

M. S. Del Rose is with the U.S. Army Tank Automotive Research Development and Engineering Center (TARDEC), Warren, MI 49397 USA (Desk: 596.306.4059; email: mike.delrose@us.army.mil).

C. C. Wagner is with Oakland University, Rochester MI, 48309, USA (Desk: 248.370.2215; wagner@oakland.edu).

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **01 MAY 2011**	2. REPORT TYPE **Journal Article**	3. DATES COVERED **01-05-2011 to 01-05-2011**

4. TITLE AND SUBTITLE
VISUAL HUMAN INTENT ANALYSIS SURVEY

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michael Del Rose; Christian Wagner

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army TARDEC ,6501 E.11 Mile Rd,Warren,MI,48397-5000

8. PERFORMING ORGANIZATION REPORT NUMBER
#21296

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army TARDEC, 6501 E.11 Mile Rd, Warren, MI, 48397-5000

10. SPONSOR/MONITOR’S ACRONYM(S)
TARDEC

11. SPONSOR/MONITOR’S REPORT NUMBER(S)
#21296

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Submitted to International Journal on Computational Intelligence

14. ABSTRACT

Visual Human Intent Analysis (VHIA) is a growing field of research devoted to algorithms that categorize human behavior through input of visual image sequences. In VHIA, it is difficult to make an implementation framework that is robust enough to handle the many non-deterministic ways of human physical actions. This paper will survey several techniques currently used in VHIA including non-traditional artificial intelligence techniques, visual languages, statistical algorithms, and others. It will give the reader the background and understanding of how open this domain is to new algorithms; improving on the overall research in VHIA or similarly related problems.

15. SUBJECT TERMS
Activity recognition, artificial intelligence, visual human intent analysis, visual understanding

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

eigenvectors are calculated and using Karhunen-Loeve technique the best ones that describe the action are kept. This makes up an orthogonal coordinate system. To recognize an unknown image sequence behavior, a distance measure is used for the calculated eigenvalues and eigenvectors onto the coordinate system. This study looked at five different cricket events with fairly good recognition.

III. TRADITIONAL ARTIFICIAL INTELLIEGENT TECHNIQUES

Some traditional artificial intelligence techniques are used for identifying intent, many with a spin towards specific motions [13, 19, 20, 26, 28, 35, 40, 42, 45, 58, 63]. For instance, J.-Y. Yang et al. [19] uses neural networks to determine human actions. They differ from the normal human motion capture of placing tags on body parts for tracking. They strap a tri-axial accelerometer to the person's wrist to detect three degrees of motion on a specific body part. Tri-axial data is captured on set time intervals and processed to feed into a neural network to determine if it is a static event, like standing or sitting, or a dynamic event, like walking and running. Once this is determined, either a static event neural network or a dynamic event neural network is used to determine the action. The results are promising for the limited actions to detect (standing, sitting, walking, running, vacuuming, scrubbing, brushing teeth, and working on a computer).

Rule based and Fuzzy systems are other types of computational intelligent technique used to identify patterns and have been adapted to analyze human events [13, 16, 47]. H. Stern et al. [16] created a prototype fuzzy system for picture understanding of surveillance cameras. His model is split into three parts, pre-processing module, a static object fuzzy system module, and a dynamic temporal fuzzy system module. The static fuzzy system module takes in the pre-processed data and outputs the number of people in the image: single person, two people, three people, many people, or no people. Then the dynamic fuzzy system determines the intent of the person based on the temporal movements.

IV. MARKOV MODELS AND BAYESIAN NETWORKS

Developing Markov or Bayesian networks are a common approach to visual human behavior analysis research. These methods seem to fit the logical approach of having a sequence of images making up an action. There are several ways to develop a network from the input data [1, 8, 14, 18, 19, 30, 36, 38, 42, 46, 48, 57]. In the work of Du, Chen, and Xu [1], they use a Coupled Hierarchal Durational – State Dynamic Bayesian Network (CHDS-DBN) to model human actions. They claim that to understand human actions, frameworks should have both motion corresponding to the interaction as well as details of the motion on different scales. For the most part, research did not include interaction with other people as a determinate in understanding intent of a single person. Most work is on motion characteristics of the individual alone. This approach adds a decision base to normal action understanding.

The work by A. Galata et. al. [8] uses variable length Markov models for human understanding. They claim that using variable length Markov models provide a more efficient way to represent behaviors and more flexibility then common models when using a large temporal scale.

V. GRAMMARS

In constructing networks, many researchers use grammars [5, 22, 37, 41, 44, 50, 52, 56, 57, 58, 59] as describing the sequence of events the body makes to help calculate the actions of the person visually. Grammars are mathematically based and they seem to flow well with understanding actions because several frames are used in a network fashion. A. Ogale et. al. [22] uses probabilistic context free grammars (PCFG) in short action sequences of a person from video. Body poses are stored as silhouettes which are used in the construction of the PCFG. Pairs of frames are constructed based on their time slot: the pose from frame 1 and 2 are paired, the pose from frame 2 and 3 are paired, and so on. These pairs construct the PCFG for the given action. When testing the algorithm, the same procedure is followed. Comparing the testing data with the trained data is accomplished through Bayes.

The work from Starner, Weaver, and Pentland [5] also use grammars in constructing their network. Phrase grammars were used to distinguish the type of action from hand signals that can be networked together to form the meaning of the sentence signed using American Sign Language. In this case, phrase grammars limit the search set of words to improve the accuracy of what is being described. They also speed up the process over not using grammars. A Hidden Markov Model is used to train and test the data.

VI. TRADITIONAL HIDDEN MARKOV MODELS

Of all the VHIA networks constructed, Hidden Markov Models (HMM) are the most widely used [3, 4, 5, 7, 8, 9, 11, 20, 21, 24, 26, 38, 45, 48, 51, 64]. Yamato et. al. [3] used HMMs to recognize six different tennis strokes by creating a feature vector of a 25x25 mess overlaid on the image to help describe body positions in each frame. Campbell, Becker, and Azarbayejani [4] used HMMs to recognize eighteen Tai Chi moves. Each move was represented by a series of vectors formed by the 3D position of the head and the hands. Yu and Ballard [9] use HMMs to distinguish similar action based on head and eye movements. The work of Lee and Kim [7] shows how to use HMMs for gesture recognition. A threshold model is built to provide dynamic threshold values to distinguish between meaningful and meaningless gestures. HMMs are used to train and test the predefined gestures. Gehrig and Schulz [24] used HMMs to recognize ten kitchen actions based on the movement of twenty four points on the upper body. They looked at skeletal data and calculated the correct movements of people to reduce the number of body parts down to thirteen. Gao et. al. [26] used both Optical Flow Tensors (OFT) and HMMs to distinguish basketball shot actions from video. Optical flow fields are modeled from the video frames at several resolutions and a tensor is built, this is the OFT. Reducing the dimensionality of the data is accomplished by first applying a general Tensor Discriminate Analysis Function then a Linear Discriminate Analysis function. An HMM is used to train and test the final features.

VII. NON-TRADITIONAL HIDDEN MARKOV MODELS

Other forms of HMMs have been developed to handle more specific problems with HMM based action recognition systems [2, 6, 11, 14, 18, 20, 21, 28, 39, 51, 54, 60, 66]. Wilson and Bobick [6] use a Parametric Hidden Markov Model (PHMM) to recognize gesture. The PHMM has an additional parameter used to represent meaningful variations of gestures across the set of all gestures. This gives PHMMs the ability to distinguish between gesture meanings with similar hand movements.

Oliver et. al. [11] developed a real time system that detects and classifies interactions between people using a Coupled Hidden Markov Model (CHMM). They used synthetic environments to model person to person interactions and thus creating their CHMM. Data from a static camera was used and moving objects were segmented and tracked. Data about their location, heading, and relative location to other people were inputted into the synthetically created CHMM for analysis of the type of interaction. Results show they outperformed standard HMMs. This is not a far stretch since standard HMMs work on single automatons where CHMMs work on coupled automatons, thus HMMs cannot outperform CHMMs in this environment.

Multi-Observation Hidden Markov Models (MOHMM) are discussed in both [28] and [14] from Xaing and Gong. In [28] they use MOHMMs to create breakpoints in video content of activity. Blobs above a certain threshold in each frame are segmented from the pixel change history. Several functions of these blobs are used in the feature vector to classify the video with the MOHMM. In [14] an MOHMM was used to detect security piggybacking of people off someone else's card to open the security door. Piggybacking is when someone follows another person through a security door without using his/her card. The framework of the system allowed for continual changes based on changes in peoples' movements, thus unsupervised learning is used to continually change the model.

Gong and Xiang [18] developed a dynamic multi-linked HMM (DML-HMM) to recognize group activity from an outdoor scene. The DML-HMM is based on salient dynamic interlinks among multiple temporal events using Dynamic Probabilistic Networks (DPN). Standard HMMs cannot take into account the multi processes needed. The DML-HMM was built to handle the multitude of different object events. The topology is determined by the causality and temporal order, automatically made using the Schwarz Bayesian Information Criterion based factorization. They claim that instead of being fully connected like Coupled HMMs (CHMM), the DHL-HMM aims to only connect a subset of relevant hidden state variables across multiple temporal processes. When comparing between a Multi-Observation HMM (MOHMM), a Parallel HMM (PaHMM), and a CHMM, the DHL-HMM performs better since the CHMM and the MOHMM propagates the noise through the systems and the PaHMM discards correlations between multiple temporal processes.

Continuous HMMs (cHMMs) are used in the work of Antonakaki et. al. [20]. Their work classifies abnormal behavior of people based on both the short term behavior of the people and the trajectory of the people. A short term behavior is a behavior that can be classified in twenty five frames, or one second. A one class support vector machine (SVM) is used to distinguish abnormal behavior from the short term behavior sequence. For trajectory data, a one class cHMM is used to determine if the person's movement is abnormal. Both are used to determine the final results.

Layered Hidden Markov Models (LHMMs) are used in Oliver et. al. [21] to detect people's activity in an office. They employ a two level cascade of HMMs with three processing layers. The first layer captures video, audio and keyboard/mouse activity and creates a feature vector. The middle layer has two HMMs, one for creating an audio feature vector and one for creating a video feature vector. The top layer uses the results of the these HMMs along with keyboard/mouse activity and the derivative of the sound localization component as its feature vector. The results of the top layer determine the activity in the office. They claim the layered HMM makes it feasible to decouple different levels of analysis for training and inferences. By using a single HMM it would need a large parameter space, thus need a large amount of data to train.

Liu and Chua [2] use Observation Decomposed Hidden Markov Model (ODHMM) to model and classify multi people activity. They state that to automatically recognize multi agents (person, extremity, or object) is very challenging due to the complexity of interactions between agents. This complexity stems from large dimensionality of the feature vectors and the complex mapping of agents from input data to pre-defined activity models. To handle this problem, they decompose each feature vector into a set of sub-feature vectors for the ODHMM.

VIII. SUMMARY OF VISUAL NUMAN INTENT ANALYSIS DEVELOPMENT

Across all the previous research, several common requirements stand out.

- Detect and segment objects in each frame
- Determine relative position and orientation of each object
- Identify a meaningful sequence of frames from the visual input
- Store and retrieve past sequence of behaviors for identification of current ones.

To detect and segment out objects in each frame, there needs to be a well developed image processing set of tools. If the goal is to identify the arm/hand movements like in [5] to classify American Sign Language words from visual identification of the hands, then the image processing is very important. Any misprocessing of the input data that goes into the classifier may cause inaccuracy.

Determining the relative position and orientation of the object also requires good image processing techniques. In some instances, just the movement is important and not the exact location from frame to frame, as in [11]. This case requires less analysis of the processed input data into the classifier than, say, [10] where body orientation is important to match with preselected human actions from several angles.

To identify a meaningful sequence of frames, stops should be placed before and after each interesting area. There is a large amount of research just on finding separations in actions. If, for example, HMMs are to be used and codebooks are required to identify common actions then having equal length action sequences for comparison is important. This would require either adding to the processed sequence several inputs or subtracting parts in the middle which require little attention, like slow movements. Either way, the intelligence of the algorithm has greatly increased which requires a lot of work in automating. In [17], they have taken out three frames that describe the action from a small set. To automate this process, it requires a lot of image processing to correctly identify the starting, middle and ending location of a person's stride.

Identification of sequence could also mean sequences that have no visual information past on, like in [27]. They have identified a way of processing the visual information to a point where only a few values represent the sequence. The reader is caution that the further away the data is from its original values, the more errors are introduced in the system.

To store past behaviors, there must be knowledge of the behavior. In intelligent systems, this is usually done through learning; however, it can also be done through human intervention. If human intervention is performed on setting up parameters for known behaviors, then often times patterns that are sometimes found through the learning process is missed, thus causing misclassification of actions. It is suggested that a combination of both computer learning and human interaction is used. This requires heavy analysis on the training and testing data to completely encompass the range of each activity being performed, a large data set to perform several iterations of training and testing of data, and in-depth knowledge of the minor facets of the action. Finally, once the user has concluded that the training data is complete and the baseline action sequences are stored, then to retrieve them takes nothing more than a comparison of a newly processed action to the most likely candidate stored.

IX. Conclusion

Much of the human brain is set aside for processing the visual sense. As computing power has continually increased, and as ever great push is made for efficiencies in business and government, letting automated computers perform heretofore human visual tasks could lead to great efficiencies and improvements. Visual Human Intent Analysis (VHIA) is a wide open field of research with several different methods that relate to individual solutions, like identifying hand gestures, understanding different tennis strokes, identifying office activities, and many others. Each method described above plays on the solution's strengths and weaknesses whether it is simplified classification with heavy pre-processing or a more intelligent decision system. The wide range of methods demonstrates the openness to new and innovative solutions that are catered to one's own problem.

However, with all the different techniques there are four common tasks which every VHIA systems must perform: detect and segment objects in each frame, determine relative position and orientation of each object, identify a meaningful sequence of frames from the visual input, and store and retrieve past sequence of behaviors for identification of current ones. These tasks are performed in different ways depending on the type of classification system. They require a lot of image processing, analysis on the input data, and in-depth knowledge of the actions with respect to the processed data.

References

[1] Du, Y., Chen, F., Xu, W., "Human Interaction Representation and Recognition Through Motion Decomposition," *IEEE Signal Processing Letters*, Vol. 14, No. 12, Dec. 2007, pp. 952-955.

[2] Liu, X., Chua, C.S., "Multi-agent Activity Recognition Using Observation Decomposed Hidden Markov Models," *Image and Vision Computing*, Vol. 24, Feb. 2006, pp. 166-175.

[3] Yamato, J., Ohya, J., Ishii, K., "Recognizing Human Action in Time Sequential Images Using Hidden Markov Models," *Proceedings from IEEE Computer Vision and Pattern Recognition (CVPR)*, 1992, pp. 379-385.

[4] Campbell, L., Becker, D., Azarbayejani, A., "Invariant Features for 3-D Jester Recognition," *Proceedings from IEEE Automatic Face and Gesture Recognition (AFGR)*, 1996, pp. 157-162.

[5] Starner, T., Weaver, J., Pentland, A., "Real Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 20, 1998, pp. 1371-1375.

[6] Wilson, A., Bobick, A., "Parametric Hidden Markov Models for Gesture Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 21, 1999, pp. 884-899.

[7] Lee, H., Kim, J. H. "An HMM Based Threshold Model Approach for Gesture Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 21, 1999, pp. 961-973.

[8] Galata, A., Johnson, N., Hogg, D., "Learning Variable Length Markov Models of Behaviour," *Computer Vision and Image Understanding*, Vol. 81, 2001, pp. 398-413.

[9] Yu, C., Ballard, D., "Learning to Recognize Human Action Sequences," *Proceedings from International Conference on Development and Learning*, 2002, pp. 28-33.

[10] Weinland, D., Ronfard, R., Boyer, E., "Motion History Volumes for Free Viewpoint Action Recognition," *Proceedings from IEEE International Workshop on Modeling People and Human Interaction*, 2005

[11] Oliver, N. M., Rosario, B., Pentland, A. P., "A Bayesian Computer Vision System for Modeling Human Interaction," *IEEE Transactions in Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, Aug. 2000, pp. 831-843.

[12] Jang, W. S., Lee, W. K., Lee, I. K., Lee, J., "Enriching a Motion Database by Analogous Combination of Partial Human Motion," *The Visual Computer*, Vol. 24, No. 4, Springer, Berlin, April 2008, pp. 271-280.

[13] Siebel, N. T., Maybank, S. J. "The ADVISOR Visual Surveillance System," *Proceedings from Applications of Computer Vision*, 2004, pp. 103-111.

[14] Xiang, T., Gong, S., "Incremental Visual Behaviour Modelling," *Proceedings from European Conference on Computer Vision*, 2006, pp. 65-72.

[15] Cristani, M., Bicego, M., Murino, V., “Audio-Visual Event Recognition in Surveillance Video Sequences,” *IEEE Transactions on Multimedia*, Vol. 9, No. 2, Feb. 2007, pp. 257-267.

[16] Stern, H., Kartoun, U., Shmilovici, A., “A Prototype Fuzzy System for Surveillance Picture Understanding,” *Proceedings from Visual Imaging and Image Processing Conference*, Sept 2001, pp. 624-629.

[17] Dimitrijevic, M., Lepetit, V., Fua, P., “Human Body Pose Detection Using Bayesian Spatio-Temporal Templates,” *Computer Vision and Image Understanding*, Vol. 104, No. 2, 2006, pp.127-139.

[18] Gong, S., Xiang, T., “Recognition of Group Activity using Dynamic Probabilistic Networks,” *Proceedings from International Conference in Computer Vision*, 2003, pp. 742–749.

[19] Yang, J. Y., Wang, J. S., Chen, Y. P., “Using Acceleration Measurements for Activity Recognition: an Effective Learning Algorithm for Constructing Neural Classifiers,” *Pattern Recognition Letters*, Vol. 29, 2008, pp. 2213–2220.

[20] Antonakaki, P., Kosmopoulos, D., Perantonis, S. J., “Detecting Abnormal Human Behaviour Using Multiple Cameras,” *Signal Processing Journal*, Vol. 89, 2009, pp. 1723 – 1738

[21] Oliver, N., Horvitz, E., Garg, A., “Layered Representations for Human Activity Recognition,” *Proceedings from IEEE International Conference on Multimodal Inferences (ICMI)*, 2002, pp. 3-8.

[22] Ogale, A., Karapurkar, A., Aloimonos, Y., “View-Invariant Modeling and Recognition of Human Actions Using Grammars,” *Lecture Notes in Computer Science*, Vol. 4358, Springer, Berlin, 2007, pp.115-126, doi:10.1007/978-3-540-70932-9_9

[23] Mokhber, A., Achard, C., Milgram, M., “Recognition of Human Behavior by Space-Time Silhouette Characterization,” *Pattern Recognition Letters*, Vol. 29, 2008, pp 81-89.

[24] Gehrig, D., Schulz, T., “Selecting Relevant Features for Human Motion Recognition,” *Proceedings from International Conference on Pattern Recognition*, 2008, pp. 1-4, doi:10.1109/ICPR.2008.4761290

[25] Rahman, M., Nakamura, K., Ishikawa, S., “Recognizing Human Behavior Using Universal Eigenspace,” *Proceedings from International Conference on Pattern Recognition*, 2002, pp. 295-298, doi:10-1109/ICPR.2002.1044694

[26] Gao, X., Yang, Y., Tao, D., Li, X., “Discriminative Optical Flow Tensor for Video Semantic Analysis,” *Computer Image and Understanding*, Vol. 113, No. 3, 2009, pp. 372-383.

[27] Blackburn, J., Ribeiro, E., “Human Motion Recognition Using Isomap and Dynamic Time Warping,” *Lecture Notes in Pattern Recognition*, Vol. 4814, Springer, Berlin, 2007, pp. 285-298, doi:10-1007/978-3-540-75703-0.

[28] Xiang, T., Gong, S., “Activity Based Video Content Trajectory Representation and Segmentations,” *Proceedings from British Machine Vision Conference*, 2004, pp. 177-186.

[29] Shah, M., “Understanding Human Behavior from Motion Imagery,” *Machine Vision and Application*, Vol. 14, No. 4, Springer, Berlin, Sept. 2003, pp. 210-214, doi:10.1007/s00138.0003-0124-3.

[30] Walter, M., Psarrou, A., Gong, S., “Data Driven Gesture Model Acquisition Using Minimum Description Length,” *Proceeds From British Machine Vision Conference*, 2001, pp. 673-683.

[31] Ben-Arie, J., Wang, Z., Pandit, P., Rajaram, S., “Human Activity Recognition Using Multidimensional Indexing,” *IEEE Transactions on Pattern Analysis and Machine Vision (PAMI)*, Vol. 24, No. 8, Aug. 2002, pp. 1091-1104, doi:10.1109/TPAMI.2002.1023805

[32] Morellas, V., Pavlidis, I., Tsaimyartzis, P., “DETER: Detection of Events for Threat Evaluation and Recognition,” *Machine Vision and Application Journal*, Vol. 15, No. 1, 2003, pp. 29-45.

[33] Masoud, O., Papanikolopoulus, N.P., “A Method for Human Action Recognition,” *Image and Vision Computing*, Vol. 21, No. 8, 2003, pp. 729 – 723.

[34] Parameswaran, V., Chellappa, R., “View Invariance for Human Action Recognition,” *International Journal of Computer Vision*, Vol. 66, No. 1, 2006, pp. 83-101.

[35] Oikonomopoulos, A., Patras, I., Pantic, M., “Kernal-Based Recognition of Human Actions Using Spatiotemporal Salient Points,” *Proceedings from Computer Vision and Pattern Recognition Conference*, 2006, pp.151-161, doi:10.1109/CVPRW.2006.114.

[36] Truyen, T. T., Phung, D. Q., Venkatesh, S., Bui, H. H., “AdaBoost.MRF: Boosted Markov Random Forests and Application to Multilevel Activity Recognition,” *Proceedings from Computer Vision and Pattern Recognition Conference*, 2006, pp.1686-1693, doi:10.1109/CVPR.2006.49

[37] Oikonomopoulus, A., Pantic, M., Patras, I., “B-Spline Polynomial Descriptors for Human Activity Recognition,” *Computer Vision and Pattern Recognition Conference*, 2008, pp. 1-6, doi:10.1109/CVPR.2008.4563175.

[38] Robertson, N., Reid, I. D., “A General Method for Human Activity Recognition in Video,” *Computer Vision and Image Understanding*, Vol. 104, No. 2, 2006, pp. 232-248.

[39] Kawanaka, D., Okatani, T., Deguchi, K., “HHMM Based Recognition of Human Activity,” *Institute of Electronics, Information and Communication Engineers Transactions, Oxford Journals*, Vol. E89-D, No. 7, 2006, pp. 2180-2185.

[40] Schuldt, C., Laptev, I., Caputo, B., “Recognizing Human Actions: a Local SVM Approach,” *Proceedings from International Conference on Pattern Recognition*, Vol. 3, 2004, pp. 32-36, doi:10.1109/ICPR.2004.1334462.

[41] Mikolajczyk, K., Uemura, H., “Action Recognition with Motion-Appearance Vocabulary Forest,” *Proceedings from Computer Vision and Pattern Recognition*, 2008, pp. 1-8, doi:10.1109/CVPR.2008.4587628.

[42] Ikizler, N., Duygulu, P., “Human Action Recognition Using Distribution of Oriented Rectangular Patches,” *Journal of Human Motion*, 2007, pp. 271-284.

[43] Ikizler, N., Cinbis, R. G., Duygulu, P., “Human Action Recognition with Line and Flow Histograms,” *Proceedings from International Conference on Pattern Recognition,* 2008, pp. 1-4, doi:10.1109/ICPR.2008.4671434.

[44] Wang, Y., Mori, G., “Human Action Recognition by Semilatent Topic Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 31, No. 10, 2009, pp. 1762-1774.

[45] Perez, O., Piccardi, M., Garcia, J., Patricio, M. A., Molina, J. M., “Comparison Between Genetic Algorithms and the Baum-Welch Algorithm in Learning HMMs for Human Activity Classification,” *Lecture Notes in Computer Science*, Vol. 4448, Springer, Berlin, 2007, pp. 399-406.

[46] Han, L., Liange, W., Wu, X. X., Jia, Y. D., “Human Action Recognition Using Discriminative Models in the Learned Hierarchical Manifold Space,” *Proceedings from IEEE International Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1-6, doi:10.1109/AFGR.2008.4813416.

[47] Herrera, A., Beck, A., Bell, D., Miller, P., Wu, Q., Yan, W., “Behaviour Analysis and Prediction in Image Sequences Using Rough Sets,” *Proceedings from International Machine Vision and Image Processing Conference*, 2008, pp. 71-76, doi:10.1109/IMVIP.2008.24.

[48] Shi, Q.F., Wang, L., Cheng, L., Smola, A., “Discriminative Human Action Segmentation and Recognition Using Semi-Markov Model,” *Proceedings from Computer Vision and Pattern Recognition Conference*, 2008, pp. 1-8, doi:10.1109/CVPR.2008.4587557.

[49] Rodriguez, M. D., Ahmed, J., Shah, M., “Action MACH a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition,” *Proceedings from Computer Vision and Image Processing Conference*, 2008, pp. 1-8, doi:10.1109/CVPR.2008.4587727.

[50] Liu, J. G., Yang, Y., Shah, M., “Learning Semantic Visual Vocabularies Using Diffusion Distance,” *Proceedings from Computer Vision and Image Processing Conference*, 2009, pp. 461-468, doi:10.1109/CVPRW.2009.5206848.

[51] Chakraborty, B., Rudovic, O., Gonzalez, J., “View Invariant Human Body Detection with Extension to Human Action Recognition Using Component-wise HMM of Body Parts,” *Lecture Notes in Computer*

Science, Vol. 5098, 2008, pp. 208-217, doi: 10.1007/978-3-540-70517-8_20

[52] Colombo, C., Comanducci, D., Bimbo, A., “Compact Representation and Probabilistic Classification of Human Actions in Videos,” *Proceedings from IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 342-346, doi: 10.1109/AVSS.2007.4425334.

[53] Thurau, C., Hlavac, V., “n-Grams of Action Primitives for Recognizing Human Behavior,” Lecture Notes in Computer Science, Vol. 4673, Springer, Berlin, 2007, pp. 93-100.

[54] Herzog, D.L., Kruger, V., “Recognition and Synthesis of Human Movements by Parametric HMMs,” *Lecture Notes in Computer Science*, Vol. 5064, Springer, Berlin, 2009, pp. 148-168, doi: 10.1007/978-3-642-03061-1_8

[55] Wang, Y., Huang, K. Q., Tan, T. N., “Group Activity Recognition Based on ARMA Shape Sequence Modeling,” *Proceedings from International Conference on Image Processing*, Vol. 3, 2007, pp. 209-212, dio:10.1109/ICIP.2007.4379283.

[56] Jenkins, O.C., Gonzalez, G., Loper, M., “Dynamic Motion Vocabularies for Kinematic Tracking and Activity Recognition,” *Proceedings from Computer Vision and Pattern Recognition Conference*, 2006, pp. 147-156, doi:10.1109/CVPRW.2006.67.

[57] Yamamoto, M., Mitomi, H., Fujiwara, F., Sato, T., “Bayesian Classification of Task-Oriented Actions Based on Stochastic Context Free Grammar,” *Proceedings from IEEE International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 317-322, doi:10.1109/FGR.2006.28.

[58] Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A., “Recovering the Basic Structure of Human Activity from a Video-Based Symbol String,” *Proceedings from IEEE Workshop on Motion and Video Computing*, 2007, pp. 1-9, doi:10.1109/WMVC.2007.34.

[59] Batra, D., Chen, T. H., Sukthankar, R., “Space-Time Shapelets for Action Recognition,” *Proceedings from IEEE Workshop on Motion and Video Computing*, 2008, pp. 1-6, doi:10.1109/WMVC.2008.4544051.

[60] Mori, T., Segawa, Y., Shimosaka, M., Sato, T., “Hierarchical Recognition of Daily Human Actions Based on Continuous Hidden Markov Models,” *Proceedings from IEEE Conference on Automatic Face and Gesture Recognition*, 2004, pp.779-784, doi:10.1109/AFGR.2004.1301629.

[61] Kam, A. H., Ann, T. K., Lung, E. H., Yun, Y. W., Wang, J. X., “Automated Recognition of Highly Complex Human Behavior,” *Proceedings from International Conference on Pattern Recognition*, Vol. 4, 2004, pp. 327-330, doi:10.1109/ICPR.2004.1333769.

[62] Gao, J., Collins, R. T., Hauptmann, A. G., Wactlar, H. D., “Articulated Motion Modeling for Activity Analysis,” *Proceedings from International Conference on Image and Video Retrieval*, 2004, pp. 1-19.

[63] Chomat, O., Crowley, J. L., “A Probabilistic Sensor for the Perception of Activities,” *Proceedings from IEEE International Conference on Automatic Face and gesture Recognition*, 2000, pp. 314-319.

[64] Babu, R. V., Anantharaman, B., Ramakrishnan, K. R., Srinivasan, S. H., “Compressed Domain Action Classification Using HMM,” *Pattern Recognition Letters*, Vol. 23, No. 10, Aug. 2002, pp. 1203-1213, doi:10.1016/S01167.8655(02)00067-3.

[65] Ghayoori, A., Hendessi, F., Sheikh, A., “Application of Smooth Ergodic Hidden Markov Model in Text to Speech Systems,” *International Journal of Signal Processing*, Vol. 2, No. 3, 2006, pp. 151-157.

[66] Brand, M., Oliver, N., Pentland, A., “Coupled Hidden Markov Models for Complex Action Recognition,” *Proceedings from Computer Vision and Pattern Recognition Conference (CVPR)*, 1997, pp. 994-999.

[67] Li, X., Parizeau, M., Plamondon, R., “Training Hidden Markov Models with Multiple Observations – A Combinatorial Method,” *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 22, No.4, 2000, pp. 371-177.

First A. Author (M’76–SM’81–F’87) and the other authors may include biographies at the end of regular papers. Biographies are often not included in conference-related papers. This author became a Member (M) of **IEEE** in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state or country, and year degree was earned. The author’s major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests ends the paragraph.

The third paragraph begins with the author’s title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the **IEEE**. Finally, list any awards and work for **IEEE** committees and publications. If a photograph is provided, the biography will be indented around it. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.

Michael S. Del Rose earned his BS in mathematics at the University of Michigan – Dearborn, Dearborn Michigan. He has a MSEE in electrical engineering for the University of Michigan – Ann Arbor, Ann Arbor, Michigan and a Ph.D. in systems engineering at Oakland University, Rochester, Michigan.

He currently works for the U.S. Army Tank Automotive and Engineering Center located in Warren, Michigan as a Research Electrical Engineer in robotics where he is responsible for artificial intelligence and image processing research. Prior to this he has served two years in London, England as European Scientist responsible for funding European universities and companies programs in the field of vehicle design, robotics research, power and energy systems, water filtration technology, armor development, and explosives research.