HMM BASED HAND GESTURE RECOGNITION: A REVIEW ON TECHNIQUES AND APPROACHES

M. A. MONI
Pabna University of Science & Technology, Bangladesh
E-mail: moni_cse1@yahoo.com

Abstract: *Many ways of communications are used between human and computer, while using gesture is considered to be one of the most natural ways in a virtual reality system. Hand gesture is one of the typical methods of non-verbal communication for human beings and we naturally use various gestures to express our own intentions in everyday life. Gesture recognizers are supposed to capture and analyze the information transmitted by the hands of a person who communicates in sign language. This is a prerequisite for automatic sign-to-spoken-language translation, which has the potential to support the integration of deaf people into society. This paper present part of literature review on ongoing research and findings on different technique and approaches in gesture recognition using Hidden Markov Models (HMMs) for vision-based approach.*

Keywords*: Gesture Recognition, Sign Language, HMM.*

1. Introduction

The video-based approaches claim to allow the signer to move freely without any instrumentation attached to the body. Trajectory, hand shape and hand locations are tracked and detected by a camera (or an array of cameras). By doing so, the signer is constrained to sign in a closed, some-how controlled environment. The amount of data that has to be processed to extract and track hands in the image also imposes a restriction on memory, speed and complexity on the computer equipment.

A recognition system that we proposed comprises of a framework of two most important entity, which are preprocessing (image processing) and recognition classification (artificial intelligence recognition) as both will have to work relatively in defining the meaning of an input sign language.

Vision-based technique has been proposed as we are using single Matrox camera to collect gesture data. This technique can cover both face and hands signer in which signer does not need to wear data gloves device. All processing tasks are solved by using computer vision techniques which are more flexible and useful than devised based measurement approach. Since sign language is gesticulated fluently and interactively like other spoken languages, a sign language recognizer must able to recognize continuous sign vocabularies in real-time. The authors try to build such a system for the Bahasa Melayu Sign Language. Gesture in this paper will always refer to the hand. Basically, different technique and approaches in recognizing a gesture for sign language will be reviewed.

2. Previous Work

Sensing of human expression is very important for human-computer interactive applications such as virtual reality, gesture recognition, and communication. In recent years, a large number of studies have been made on machine recognition of human expression.

A number of systems have been implemented that recognize various types of sign languages. For example, Starner was able to recognize a distinct forty word lexicon consisting of pronouns, nouns, verbs, and adjectives taken from American Sign Language with accuracies of over 90% [10]. A system developed by Kadous [17] recognized a 95 word lexicon from Australian Sign Language and achieved accuracies of approximately 80%. Murakami and Taguchi were able to adequately recognize 42 finger-alphabet symbols and 10 sign-language words taken from Japanese Sign Language [19]. Lu [21] and Matuso [18] have also developed systems for recognizing a subset of Japanese Sign Language. Finally, Takahashi and Kishino were able to accurately recognize 34 hand gestures used in the Japanese Kana Manual Alphabet [20].

A common approach to describing human motion is to use a state-based model, such as a Hidden Markov Model (HMM), to convert a series of motions into a description of activity. Such systems [1, 2, 3, 4, 5, 6] operate by training a HMM (or some variant thereon) to parse a stream of short-term tracked motions.

Another system [3] recognizes simple human gestures (against a black backdrop), while an earlier system [2] recognizes simple actions (pick up, put down, push, pull, etc.), also based on a trained HMM.

An early effort by Yamato et al. [7] uses discrete HMM's to successfully recognize image sequences of six different tennis strokes among three subjects. This experiment is significant because it used a 25x25 pixel quantized sub sampled camera image as a feature vector that is converted into a discrete label by a vector quantizer. The labels are classified based on discrete HMMs. Even with such low-level information, the model could learn the set of motions to perform respectable recognition rates.

According to Iwai, Shimizu, and Yachida [8], a lot of methods for the recognition of gestures from images have been proposed. Takahasi used spatiotemporal edges for the recognition of gestures based on a DP matching method. Yamato used matrices sampled sparsely from images and Hid-den Markov Models for the recognition. The sparse matrices are quantized by the Categorized VQ method and are used as input for a discrete HMM. Campbell used the 3-D positions and 3-D motion of face, left hand, and right hand extracted from stereo images and HMM for the recognition. HMM method has a better performance than DP matching method for recognition of public gestures because of statistically learning.

Research by Starner and Pentland did not attempt to capture detailed information about the hands [9, 10]. In this recognition system, sentences of the form ('personal pronoun, verb, noun, adjective, (the same) personal pronoun" are to be recognized with a vocabulary of 40 signs. These systems have mostly concentrated on finger signing, where the user spells each word with hand signs corresponding to the letters of the alphabet. They proposed an extensible system which uses a single color camera to track hands in real time and interprets American Sign Language (ASL) using HMM. The hand tracking stage of the system does not attempt to produce a fine grain description of hand shape; studies have shown that such detailed information may not be necessary for humans to interpret sign language [11, 12]. Instead, the tracking process produces only a coarse description of hand shape, orientation, and trajectory as inputs to HMMs. Signs are modeled with four-states-HMM. They use a single camera and plain cotton gloves in two colors or no gloves for a second test. The shape, orientation, and trajectory information is then input to a HMM for recognition of the signed words. They achieve recognition accuracies between 75% and 99% allowing only the simplest syntactical structures.

In [2] Siskind and Morris conjecture that human event perception does not presuppose object recognition. In other words, they think visual event recognition is performed by a visual pathway which is separated from object recognition. To verify the conjecture, they analyze motion profiles of objects that participate in different simple spatial motion events. Their tracker uses a mixture of color based and motion based techniques. Color based techniques are used to track objects defined by set of colored pixels whose saturation and value are above certain thresholds in each frame. These pixels are then clustered into regions using a histogram based on hue. Moving pixels are extracted from frame differences and divided into clusters based on proximity. Next, each region (generated by color or motion) in each frame is abstracted by an ellipse. Finally, feature vector for each frame is generated by computing the absolute and relative ellipse positions, orientations, velocities and accelerations. To classify visual events, they use a set of Hidden Markov Models (HMMs) which are used as generative models and trained on movies of each visual event represented by a set of feature vectors. After training, a new observation is classified as being generated by the model that assigns the highest likelihood. Experiments on a set of 6 simple gestures, "pick up," "put down," "push," "pull," "drop," and "throw," demonstrate that gestures can be classified based on motion profiles.

Sorrentino, Cuzzolin and Frezza [13] present an original technique for hand gesture recognition based on a dynamic shape representation by combining size functions and hidden Markov models (HMM). Each gesture is described by a different probabilistic finite state machine which models a succession of so called canonical postures of the hand. Conceptually, they model a gesture as a sequence of hand postures which project into sets of size functions. The state dynamics describe the transition between canonical postures while the observation equations are maps from the set of canonical postures to size functions. In their proposed technique, a fundamental hypothesis has been made; gestures may be modeled as sequences of a finite number of "canonical" postures of the hand. Each posture is associated to a state of a probabilistic finite state machine, in particular of a Hidden Markov Model. Each gesture is identified with a HMM with

an appropriate number of states and transition probabilities. They come with the result that HMM paradigm deserves to be confirmed, whereas future work should concentrate upon a better shape representation and also the measuring functions proposed in [14] must be improved since they are particularly sensitive to small changes in the position of the center of mass of the images. Further developments on integrating the qualitative modeling philosophy presented in this paper with quantitative techniques which describe the 27 degrees of freedom kinematics of the hand.

Chen, Fu and Huang in [15] have introduced a hand gesture recognition system to recognize continuous gesture before stationary background using 2D video input. Reliable method of image processes has been applied: motion detection, skin color extraction, edge detection, movement justification and background subtraction as a real-time image processing subsystem. The system consists of four modules: a real time hand tracking and extraction, feature extraction, hidden Markov model (HMM) training, and gesture recognition. First, they apply a real-time hand tracking and extraction algorithm to trace the moving hand and extract the hand region, then the Fourier descriptor (FD) has been used to characterize spatial features and the motion analysis to characterize the temporal features. A spatial and temporal feature has been combined of the input image sequence as our feature vector. After the feature vectors have been extracted, we apply HMMs to recognize the input gesture. The gesture to be recognized is separately scored against different HMMs. The model with the highest score indicates the corresponding gesture. In the experiments, we have tested our system to recognize 20 different gestures, and the recognizing rate is above 90%.

Garver [16] has came out with his research to construct interface environment that allowing mouse control through pointing gestures and commands executed by simple arm movements as the goal of having natural gestures as a modality for basic computer control. Hand segmentation system has been performed for individual frames. The hand segmentation system is made up of three image processing algorithms: dynamic background subtraction, statistical skin color segmentation, and depth thresholding. Using the basic filters for segmentation and locating hands using a very simple blob extraction algorithm the tracking system is ready to use. HMM has been proposed for command gesture system to trained gesture classification. After preparing the gesture for executing a command, produced gesture sequence that is then processed by a series of HMMs. Each HMM represents a specific gesture, and the performed gesture is classified as the representative model with the highest probability of correspondence. However only the hand segmentation and tracking follow well. The initial classification of gestures as pointing or command works with reasonable accuracy, need to be improved.

3. Gesture and Hidden Markov Model

3.1. Hand Gesture

Hand and arm gestures receive the most attention among those who study gesture – in fact, many (if not most) references to gesture recognition only consider hand and arm gestures. The vast majority of automatic recognition systems are for deictic gestures (pointing), emblematic gestures (isolated signs) and sign languages (with a limited vocabulary and syntax). Some are components of bimodal systems, integrated with speech recognition. Some produce precise hand and arm configuration while others only coarse motion.

The kinematics of the posture of a human hand can be described by a mechanical system with 27 degrees of freedom [1, 2].

Common experience suggests, however, that a gesture may be characterized by a sequence of only a finite number of hand postures and that, therefore, it is not necessary to describe hand posture as a continuum.

Feature extraction and analysis is a robust way to recognize hand postures and gestures. It can be used not only to recognize simple hand postures and gestures but complex ones as well. Its main drawback is that it can become computationally expensive when a large number of features are being collected, which slows down system response time. This slowdown is extremely significant in virtual environment applications, but should diminish with the advent of faster machines.

The concept of gesture is loosely defined, and depends on the context of the interaction. Recognition of natural, continuous gestures requires temporally segmenting gestures. Automatically segmenting gestures is difficult, and is often finessed or ignored in current systems by requiring a starting position in time and/or space. Similar to this is the problem of distinguishing intentional gestures from other

"random" movements. There is no standard way to do gesture recognition – a variety of representations and classification schemes are used. However, most gesture recognition systems share some common structure.

3.2. Overview of Hidden Markov Model

HMM were introduced in the mid 1990's, and quickly became the recognition method of choice, due to its implicit solution to the segmentation problem.

In describing hidden Markov models it is convenient first to consider Markov chains. Markov chains are simply finite-state automata in which each state transition arc has an associated probability value; the probability values of the arcs leaving a single state sum to one. Markov chains impose the restriction on the finite-state automaton that a state can have only one transition arc with a given output; a restriction that makes Markov chains deterministic. A hidden Markov model (HMM) can be considered a generalization of a Markov chain without this Markov-chain restriction [23]. Since HMMs can have more than one arc with the same output symbol, they are nondeterministic, and it is impossible to directly determine the state sequence for a set of inputs simply by looking at the output (hence the "hidden" in "hidden Markov model").

More formally, a HMM is defined as a set of states of which one state is the initial state, a set of output symbols, and a set of state transitions. Each state transition is represented by the state from which the transition starts, the state to which transition moves, the output symbol generated, and the probability that the transition is taken [23]. In the context of hand gesture recognition, each state could represent a set of possible hand positions. The state transitions represent the probability that a certain hand position transitions into another; the corresponding output symbol represents a specific posture and sequences of output symbols represent a hand gesture. One then uses a group of HMMs, one for each gesture, and runs a sequence of input data through each HMM. The input data, derived from pixels in a vision-based solution or from bend sensor values in a glove-based solution, can be represented in many different ways, the most common by feature vectors [9]. The HMM with the highest forward probbility (described later in this section) determines the users' most likely gesture. An HMM can also be used for hand posture recognition; see Liang and Ouhyoung [24] for details.

Gaolin Fang and Wen Gao [25] stated that there are some limitations of classical HMMs for sign language recognition. Firstly, it is the assumption that the distributions of individual observation parameters can be well represented as a mixture of Gaussian or autoregressive densities. This assumption isn't always consistent with the fact. Secondly, HMMs have the poorer discrimination than neural networks. In the HMMs training, each word model is estimated separately using the corresponding labeled training observation sequences without considering the confused data (other models with similar behavior).

4. Research Methodology

Current phase of our development is still under image processing. Hand region has been successfully detected by implementing five steps image processing as shown in figure 1 under the preprocessing stage.

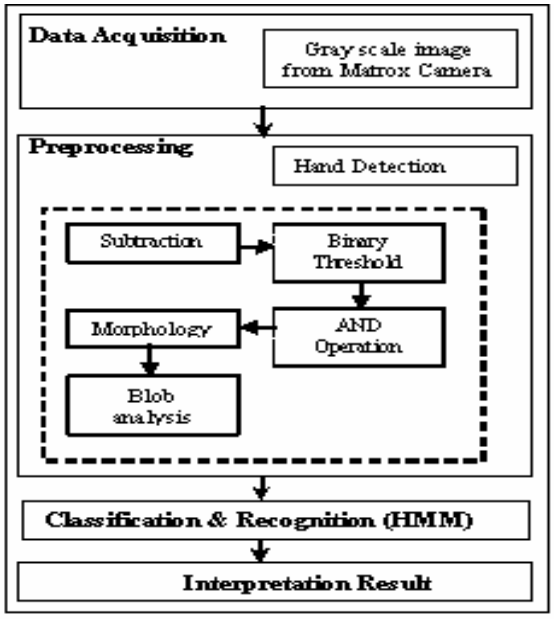

Fig. 1: Basic System Flow of Recognition System

Thus, review on technique for gesture recognition using HMM will contribute the idea on how we are going to model our gesture.

Future plan on the image processing will be discussed on the next part as the output of it will be used for the recognition process. The output is supposed to give much information to model the gesture and perform recognition process.

5. Discussion

The main requirement for the use of gesture recognition in human-computer interfaces is a

continuous online recognition with a minimum time delay of the recognition output, but such a system performance is extremely difficult to achieve, because of the following reasons: The fact that the starting and ending point of the gesture is not known is a very complicated problem.

A frame captured by the frame grabber has to be immediately preprocessed and the features have to be extracted and used for the classification, before the next frame is captured.

Assumption on the lighting, background, signers location need to be done since the Matrox camera that we use capable of capturing images into grayscale image.

6. Conclusion and Future Direction

In this paper, a few literature reviews on technique of gesture recognition using HMM has been reviewed and analyzed. Most of them are using colored images as better result can be achieved.

As the input of gesture recognition, further process need to be done on the detected hand. Edge detection is proposed. Age detection used to mark points at which image intensity changes sharply. So we will get only the hand region.

Hand motion, posture and orientation are other areas that need to be explore. Combination of hand detection with those will be useful input for recognition process.

Referrences

[1] H. Buxton, "Learning and understanding dynamic scene activity: a review," *Image and Vision Computing*, 21(1):125–136, January 2003.

[2] J.M. Siskind and Q. Morris, "A maximum-likelihood approach to visual event classification," *In Proc. 4th European Conference on Computer Vision*, pages II:347–360, 1996.

[3] Y.A. Ivanov and A.F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):852–872, August 2000.

[4] J. Lou, Q. Liu, T. Tan, and W. Hu, "Semantic interpretation of object activities in a surveillance system," *In Proc. International Conference on Pattern Recognition*, pages III: 777–780, 2002.

[5] N.M. Oliver, B. Rosario, and A.P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831–843, August 2000.

[6] V. Nair and J.J. Clark, "Automated visual surveillance using hidden markov models," *In International Conference on Vision Interface*, pages 88–93, 2002.

[7] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov models," *In Proc. 1992 ICCV*, p. 379-385.

[8] Y.Iwai, H.Shimuzu, and M.Yachida, "Real-Time Context-based Gesture Recognition Using HMM and Automation," Department of Systems and Human Science, Osaka University, Toyonaka, Osaka 560-8531, JAPAN, iwai@sys.es.osaka-u.ac.jp

[9] A T. Staner and A. Pentland, "Visual Recognition of American Sign Language using Hidden Markov Models," Technical Report TR-306, Media Lab, MIT, 1995

[10] A T. Staner and A. Pentland, "Real-Time American Sign Language Recognition from Video using Hidden Markov Models," Technical Report TR-306, Media Lab, MIT, 1995

[11] H. Poizner, U. Bellugi, and V. Lutes-Driscoll, "Perception of american sign language in dynamic point-light displays," 7: 430-440, 1981.

[12] G. Sperling, M. Landy, Y. Cohen, and M. Pavel, "Intelligible encoding of ASL image sequences at extremely low information rates," *Comp. Vision, Graphics, and Image Proc.*, 31:335-391, 1985.

[13] A.Sorrentino F.Cuzzolin, and R.Frezza, "Using Hidden Markov Models and Dynamic Size Functions for Gesture Recognition," *British Machine Vision Conference*, 1997.

[14] A. Verri and C. Uras. "Metric-topological approach to shape representation and recognition," *Image Vis. C.*, 14 (3):189–207, 1996.

[15] Sheng Chen, Chih-Ming Chu, Chung-Lin Huang, "Hand Gesture Recognition using Real Time Tracking Method and Hidden Markov models," Institute of Electrical Engineering, National Tsing Hua University, Hsin Chu 300, Taiwan, ROC

[16] R.Garver, "Vision Based Gesture Recognition, Interaction Lab," University of California at Santa Barbara

[17] Kadous, Waleed, "GRASP: Recognition of Australian Sign Language Using Instrumented Gloves," Bachelor's thesis, University of New South Wales, 1995.

[18] Matsuo, Hideaki, Seiji Igi, Shan Lu, Yuji Nagashima, Yuji Takata, and Terutaka Teshima, "The Recognition Algorithm with Non-contact for Japanese Sign Language Using Morphological Analysis," *In Proceedings of the International Gesture Workshop'97* , Berlin, 273-284, 1997.

[19] Murakami, Kouichi, and Hitomi, "Taguchi. Gesture Recognition Using Recurrent Neural Networks," *In Proceedings of CHI'91 Human Factors in Computing Systems*, 237-242, 1991.

[20] Takahashi, Tomoichi, and Fumio Kishino, "Hand Gesture Coding Based on Experiments Using a Hand Gesture Interface Device," SIGCHI Bulletin 23(2):67-73, 1991.

[21] Lu, Shan, Seiji Igi, Hideaki Matsuo, and Yuji Nagashima, "Towards a Dialogue System Based on Recognition and Synthesis of Japanese Sign Language," *In Proceedings of the International Gesture Workshop'97*, Berlin, 259-272, 1997.

[22] J.J. LaViola Jr, A survey of Hand Posture and Gesture Recognition Techniques and Technology, Department

of Computer Science, Brown University, Providence, Rhode Island 02912, June 1999

[23] Charniak, Eugene. Statistical Language Learning. MIT Press, Cambridge, 1993.

[24] Liang, Rung-Huei, and Ming Ouhyoung. A Sign Language Recognition System Using Hidden Markov Model and Context Sensitive Search. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology'96, ACM Press, 59-66, 1996.

[25] Gaolin Fang, Wen Gao, Jiyong Ma. Signer-Independent Sign Language Recognition Based on SOFM/HMM Depatment of Computer Science and Engineering, Harbin Institute of Technology, Harbin, 150001, China