

# RECOGNITION OF HAND RAISING GESTURES FOR A REMOTE LEARNING APPLICATION

Bill Kapralos<sup>1</sup>, Andrew Hogue<sup>2</sup>, and Hamed Sabri<sup>1</sup>

<sup>1</sup>Faculty of Business and Information Technology,  
University of Ontario Institute of Technology. Oshawa, Ontario, Canada. L1H 7K4

<sup>2</sup>Department of Computer Science and Engineering, Centre for Vision Research,  
York University, Toronto, Ontario, Canada. M3J 1P3

bill.kapralos@uoit.ca

hogue@cse.yorku.ca

## ABSTRACT

*A central technical issue in developing synchronous distance learning technology is enabling the remote class and the instructor to interact with each other. Issues such as “how does a student capture the instructor’s attention?”, “how can the instructor select one student to converse with?”, and “how can the instructor attend to the student once (s)he has been selected?” are complex problems that must be addressed if the class and instructor are to interact in an effective manner. This paper describes the use of Hidden Markov Models for the recognition of students signaling their intent to interact with the instructor using “traditional” classroom hand gestures such as raising and waving hand motions. Hand raising gestures are detected using motion cues over a sequence of omni-directional images using a set of pre-defined Hidden Markov Models.*

## 1. INTRODUCTION

A central technical issue in developing synchronous distance learning technology is enabling the remote class and the instructor to interact with each other. There are really two parts to this problem: i) how to present the instructor to the remote classroom, and ii) how to present the remote classroom to the instructor. The first part of this problem is perhaps the easiest to solve as there is only one person (the instructor) who must be attended too. Attending to students in the remote classrooms is more difficult. Issues such as “how does a student capture the instructor’s attention?”, “how can the instructor select one student to converse with?”, and “how can the instructor attend to the student once (s)he has been selected?” are complex problems that must be addressed if the class and instructor are to interact in an effective manner. Providing human facilitators at each site is not cost effective and the option of physically wiring each seat with buttons for students to draw the instructor’s attention would require significant modifications to existing classroom spaces. An alternative would be to deploy a sensor system within the classroom that enables student interaction with the instructor at a remote location. But how should the sensor attend to the person who wishes to ask a question? From a practical point of view, how should a sensor be constructed that has a wide enough field of view so that it can capture the entire class at once and be able to attend to the person who wants to speak or ask a question? In addition, once a speaker has been selected, how should the sensor continue to track, localize, and focus on the selected speaker?

This ongoing research project investigates issues related to the development of a remote learning system that permits a remote class-

room to interact with the instructor. This includes issues related to attending to (in both the audio and visual domains) individual students, finding students who wish to speak, permitting the instructor to view the entire remote class, and to attend to audio and visual events within the class. To this end, a novel sensor that combines directional audio and an omni-directional video sensor (Paracamera) to locate students in the classroom who wish to interact with the instructor has been developed (see Fig. 1). Students who wish to interact with the remote instructor may signal their intent via voice (e.g., by speaking aloud) or using hand raising gestures as done in “traditional” classrooms. Hand raising gestures are detected using Hidden Markov Models over a sequence of omni-directional images while sound localization techniques with a steerable microphone array allows for detection of auditory (voice) signals. The system identifies “attention seeking” actions in the audio and video domain and then presents potential speakers to the remote instructor. The instructor can then select (via a touch-screen-based user interface) one of the potential speakers (including speakers who have not sought attention overtly) and the sensor will then attend to that speaker. A high resolution view of the speaker and a beamformed audio signal is then presented to the remote instructor. This paper describes one aspect of the system; the recognition of hand raising gestures over a sequence of omni-directional images using Hidden Markov Models.

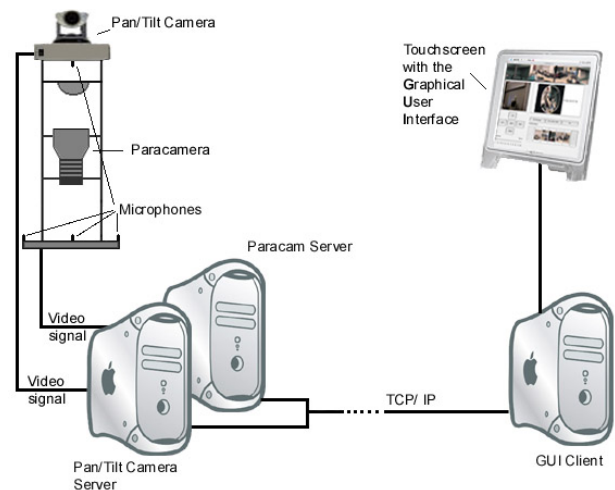


Fig. 1. System overview (see [5] for greater details).

## 2. BACKGROUND

### 2.0.1. Introduction to Hidden Markov Models

Hidden Markov Models are a statistical model used to model a wide variety of temporal (time varying) one-dimensional data [11]. An HMM consists of a number of states, each of which is assigned a probability of transition from one state to another [14]. With time, state transitions occur stochastically and the probability of transitioning from one state to another is dependent on the current and previous states only. HMMs are a “double layer” stochastic process where the first layer is a Markov chain and the second layer is a set of output observations for each state in the Markov chain. HMM states are not directly observable but are rather observed through a sequence of observable symbols. HMMs have successfully been used for pattern recognition and in particular, speech recognition tasks [9]. HMMs have recently been applied to a wide variety of pattern recognition tasks in computer vision applications including gesture recognition [14] where gestures are treated as a parametric random process whose parameters can be determined precisely [7].

The use of HMMs can be divided into two phases: i) learning, and ii) recognition (classification). In the learning phase each HMM is trained such that it will most likely generate the symbol patterns for this category. In other words, the HMM is presented with examples in order to allow it to estimate its parameters,  $\lambda = (A, B, \Pi)$  ( $A$  is the probability transition matrix for the Markov chain of layer one,  $B$  is the , and  $\Pi$  is the set of initial probabilities) and maximize the probability  $P = (O|\lambda)$  for a given set of observation symbols  $O = O_1, O_2, \dots, O_t$  [13]. There is no known analytical solution to solve the training problem and hence maximize  $P = (O|\lambda)$  for a particular observation sequence. However, a model can be chosen such that  $P = (O|\lambda)$  is locally maximized using the Baum-Welch iterative procedure [10] (equivalent to the *expectation maximization* algorithm [3]). At each step, the Baum-Welch procedure adjusts the parameters of the HMM based on probabilities computed in the previous step and tries to maximize the log-likelihood of the model with respect to the data [2].

The observation symbols are essentially derived from a “feature vector” that is extracted from input (e.g., image sequence). The feature vector is specific to the application and may include any feature of interest from the “raw” image data or after the image data has been processed in some manner. In order to recognize a set of observed sequences (as trained in the learning phase), one HMM is created for each possible category of sequences (e.g., when considering gesture recognition, a category corresponds to a distinct gesture). Essentially, when presented with a sequence whose category is unknown, given a sequence  $O$  of observation symbols  $O_1, O_2, \dots, O_t$ , the probability  $Pr(\lambda_i|O)$  for each of the  $C$  HMMs  $\lambda_i$  (where  $i$  ranges from  $0 \dots C - 1$ ) is calculated using the *forward algorithm* [10]. The HMM resulting in the highest probability is chosen and thus defines the model. A complete discussion of the theory of HMMs is beyond the scope of this paper. A detailed mathematical description of HMMs is given by Rabiner [10].

### 2.0.2. Related Work

Yamato *et al.* [14], demonstrated the first application of HMMs in human motion recognition. They were able to classify six tennis strokes performed by three people with an accuracy of over 90% when the training and test data were from the same subject and 78.5% otherwise. Since the work of Yamato *et al.* many other motion recognition applications have employed HMMs including many hand gesture recognition systems. In particular, Starner *et al* [12]

used HMMs to recognize a subset of American Sign Language (ASL) consisting of forty words allowing for 494 randomly constructed five word sentences. In this system, the user was required to wear a distinct colored glove on each hand (to separate the hands from the remaining scene) and sit on a chair very close to the camera. The eight element feature vector extracted from each image in a sequence consisted of the x,y position, angle of axis of least inertia, and the eccentricity of the bounding ellipse of each hand. This system was able to achieve an accuracy of 97% with a “strong” grammar and 91% accuracy without the strong grammar. Tanguay *et al.* [8] developed a similar system for hand gesture recognition. However, they removed the dependence on a “strong” grammar in order to avoid dependencies in any domain. Finally, Hossain and Jenkin [4] describe two approaches to the recognition of attention seeking gestures; one approach models the temporal information using structured constraints in the feature space while the other approach models the temporal information explicitly in the HMM using structured forms in the state transitions within the HMM [4].

## 3. PROPOSED SYSTEM

### 3.1. Video System

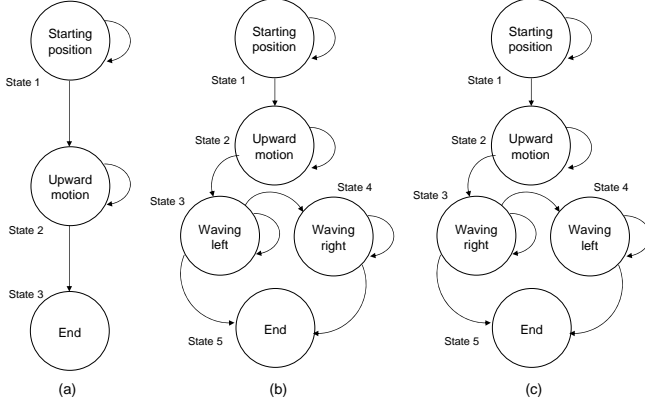
Cyclovision’s Paracamera omni-directional camera system [1] is used to capture a sequence of omni-directional images. The Paracamera consists of a high precision paraboloidal mirror and a combination of special purpose lenses. By aiming a camera to the face of the paraboloidal mirror, the combination of these optics permit the Paracamera to capture a  $360^\circ$  (*hemispherical*) view of potential speakers from a single viewpoint. Once the hemispherical view has been obtained, it may be easily un-warped to produce a panoramic view. From this panoramic, perspective views of any size corresponding to different portions of the scene may be easily extracted. The Paracamera is used to provide a view of the entire visual hemisphere thereby providing multiple dynamic views of the participants within a single image. Once a gesture has been detected, a high-resolution pan-tilt camera and a microphone array are focused in the direction of the individual making the gesture. Details regarding other aspects of the system will not be addressed here but greater details can be found in [5, 6].

### 3.2. Hand-Raising Gesture Recognition

Although the raising of a hand appears to be a fairly simple gesture (when compared to some of the complex gestures used in ASL for example), there is a great amount of variation possible making the task difficult to automate. From its starting point (lowered position) until it reaches its final raised position, the hand (arm) may move straight upwards, diagonally to the right or left, etc. In addition, the direction of motion will generally change continuously throughout the trajectory. Furthermore, once the hand is raised, it may then move (“wave”) from left to right (or right to left) and may also be lowered or raised slightly while doing so or even placed on the head. A description of the hand raising gestures defined in this work is provided below.

**Starting Point** Prior to raising the hand, the hand is at some position below the head (“lowered position”). Usually this position will be to the side of the student or on the desk in front of them.

**Gesture A** The hand follows a straight trajectory from the starting point to its final raised position above the head.



**Fig. 3.** HMM state diagrams. (a) Upward motion (straight upwards, leftward, and rightward). (b) Upward motion with left-to-right waving. (c) Upward motion with right-to-left waving.

**Gesture B** Rather than moving the hand straight upwards, the hand is raised upwards diagonally to the left from the starting point to the final position above the head.

**Gesture C** The hand is raised upwards diagonally to the right from the initial starting position to the final position above the head.

**Gesture D - Waving Hand Gesture** After raising the hand above the head using either of the three gestures (gestures A, B or C), the hand may be waved from left to right (or right to left) any number of times. Once again, the motion may not move left to right (or right to left) in a perfectly straight manner but rather, may deviate slightly.

When these “traditional” hand-raising gestures are captured over a sequence of Paracamera images several observations can be made. In particular, regardless of the position of the participant relative to the camera, as the hand is raised, its distance relative to the center of the image  $r_c$  always increases (see Fig. 2(a)-(f) for an example). This increase in distance occurs regardless of whether the hand is raised straight upwards, towards the left or towards the right (e.g., gestures A, B, and C). Furthermore, the azimuthal angle  $\theta$  of the hand in the Paracamera image is constant while the hand is raised. When the hand is waved (gesture D) from left to right (or right to left),  $r_c$  remains constant and  $\theta$  changes in time either increasing as the hand is waved from left to right or decreasing when the hand is waved from right to left. These observations are unique to omni-directional images and this uniqueness is exploited in this work. Taking into account the observed hand-raising gestures, three HMMs (HMMs A, B, and C) are defined. HMM A is specific to the hand raising gestures that do not contain any waving motions. HMM B is specific to the hand waving gesture whereby after the hand is raised, is waved from left to right while HMM C is specific to the hand waving gesture whereby after the hand is raised, is waved from right to left. A state diagram for each of the HMMs is illustrated in Fig. 3 (currently, only the HMM to recognize gestures A and B has been implemented). To focus on and isolate the effectiveness of the HMMs to recognize a particular hand raising gesture, the position of the hand within each Paracamera image is manually determined (this can be automated using a skin-color segmentation process for example (see [6])). The distance  $r_c$  between the hand and the center of the Paracamera image is given as

$$r_c = \sqrt{(x_p - x_c)^2 + (y_p - y_c)^2} \quad (1)$$

| Gesture                      | Log likelihood |
|------------------------------|----------------|
| B2 (in the training set)     | -33.9          |
| C1 (in the training set)     | -33.9          |
| A4 (not in the training set) | -1626.6        |
| B4 (not in the training set) | -Inf           |
| C4 (not in the training set) | -1982.6        |

**Table 1.** Test one results.

where  $(x_p, y_p)$  and  $(x_c, y_c)$  are the coordinates of the hand and the center of the Paracamera image respectively. The angle  $\theta$  of the region in the image corresponding to the hand is given as

$$\theta = \tan^{-1}((x_p - x_c), (y_p - y_c)) \quad (2)$$

Once  $r_c$  and  $\theta$  have been calculated, the feature vector is determined. The features of interest for the defined HMMs include i) whether  $r_c$  is increasing, decreasing or constant, and ii) whether  $\theta$  is increasing, decreasing or constant. These features are computed over two Paracamera images in sequence and are vector quantized. The quantized feature vector is then input to each HMM and each HMM then computes a probability. The HMM that maximizes the probability defines the particular gesture present in the image sequence.

#### 4. EXPERIMENTAL RESULTS

Testing of the described approach was performed on a series of hand raising gestures obtained with a Paracamera (resolution of  $640 \times 480$ ). Three subjects (subjects A, B, and C) performed the following four hand raising gestures: 1) straight upward motion (see the sequence of images illustrated in Fig. 2 for an example), 2) leftward upward motion, 3) rightward upward motion, and 4) straight upward motion with hand-waving from left to right). Although various other combinations are possible (e.g., rightward or leftward upward motion with the hand waving from left-to-right, right-to-left, etc.), they were not explicitly tested for here but will be tested in the future. Several tests were performed examining the recognition of gestures using various combinations of training gestures. The output for each test is the log likelihood. The log likelihood of the correct category is highest.

In the first test, the following gesture sequences were used to train the HMMs: A1, A2, A3, B1, B2, B3, C1, C2, and C3. The HMMs were then applied to several gesture sequences. Several of the sequences were included in the training set (B2 and C1) while others were not (A4, B4, and C4). A summary of the results is provided in Table 1. In the second test, the third gesture sequence of each of the subjects used in the previous test was removed (e.g., A3, B3, and C3 were removed). A summary of the results is provided in Table 2. In the third test “leave-one-sequence-out testing” was performed by removing the three gesture sequences from subject A (e.g., gestures A1, A2, and A3) from the training set of the first test. The HMM was then applied to the first three gestures of the first subject (A1, A2, and A3). A summary of the results is provided in Table 3. Finally, in the fourth test only one sequence from each subject was used for training (e.g., gestures A1, B2, and C3 were used). The HMM was then applied to gestures included in the training set but from other subjects (e.g., applied to gestures B1, A2, and C1). A summary of the results is provided in Table 4. By examining the results of each test scenario, it is clear that generally, the log likelihood for the gesture sequences not included in the training set are much smaller than the gesture sequences that were included indicating that the HMM was able to correctly detect gestures it has been

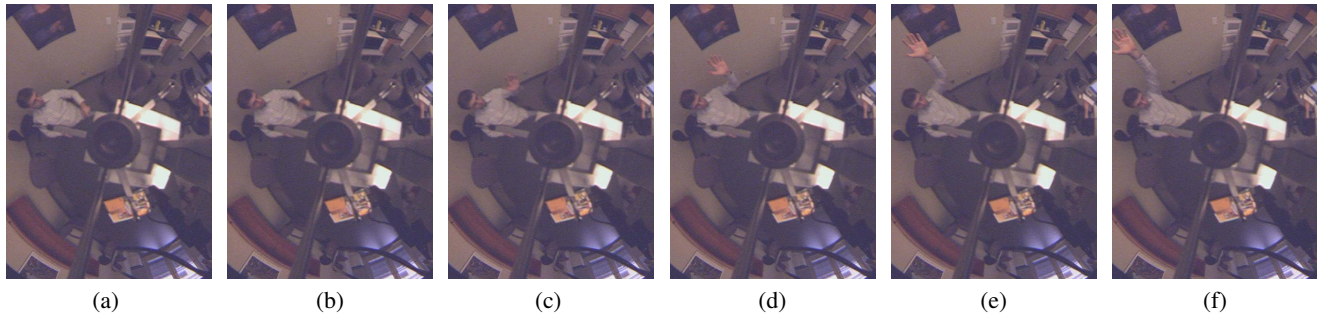


Fig. 2. Sample Paracamera images of a hand-raising gesture sequence.

| Gesture                      | Log likelihood |
|------------------------------|----------------|
| A3 (not in the training set) | -113.2         |
| B3 (not in the training set) | -160.6         |
| C3 (not in the training set) | -Inf           |

Table 2. Test two results.

| Gesture                      | Log likelihood |
|------------------------------|----------------|
| A1 (not in the training set) | -43.0          |
| A2 (not in the training set) | -45.0          |
| A3 (not in the training set) | -133.6         |

Table 3. Test three results.

trained to detect. However, in the third test, gesture A3 was not detected despite having trained the HMM on this gesture thus resulting in a false negative. In the fourth test, the resulting log likelihoods for each of the three gestures indicate the gesture was successfully detected. Finally, the fourth gesture of the first test (gesture B4) and the third gesture of the second test (C3) resulted in a value of “-Inf” indicating more training data is needed.

## 5. CONCLUSIONS

This paper presented an HMM model for hand raising gesture recognition in a sequence of omni-directional (Paracamera) images. This work is part of an ongoing project whose goal is to automatically localize and focus on (in both the audio and visual domains) participants of a remote learning (or remote meeting) application who wish to interact with one another. Although the work described here is currently in the initial stages (work in progress), preliminary results suggest that the system is capable of detecting a hand raising gestures observed in typical classroom scenarios where participants wish to interact. Being in the initial stages, there are various extensions that will follow. Most importantly, the HMM to recognize the two hand waving gestures (gestures B and C of Fig. 3) will be implemented and incorporated into the model. Furthermore, more extensive testing of the presented HMM model to determine its effectiveness under a variety of real-world scenarios (e.g., images obtained within a “real” classroom or videoconference meeting scenario) will be performed. This also includes incorporating a larger training set in the learning phase (e.g., hand raising gestures of many more subjects) and conducting experiments with “non-attention seeking” hand raising gestures (e.g., a participant raising their hand to rub their eyes or scratch their head).

## 6. REFERENCES

[1] S. Baker and S. Nayar. A theory of single viewpoint catadioptric image formation. *IJCV*, 35(2):1–22, 1999.

| Gesture                      | Log likelihood |
|------------------------------|----------------|
| B1 (not in the training set) | -29.6          |
| A2 (not in the training set) | -27.6          |
| C1 (not in the training set) | -32.1          |

Table 4. Test four results.

- [2] M. Bicego and V. Murino. 2D shape recognition by Hidden Markov Models. In *IEEE Proceedings of 11th International Conference on Image Analysis and Processing*, pages 20–24, 2001.
- [3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.
- [4] M. Hossain and M. Jenkin. Recognizing hand-raising gestures using hmm. In *Proc. Canadian Conf. on Comp. and Robot Vision*, Victoria, BC. Canada, May 8-11 2005.
- [5] B. Kapralos, A. Barth, J. Ma, and M. Jenkin. A system for synchronous distance learning. In *Proc. 16th Int. Conf. on Vision Interface*, pages 134–140, Halifax, NS. Canada, 2003.
- [6] B. Kapralos, M. Jenkin, and E. Milios. Audio-visual localization of multiple speakers in a video teleconferencing setting. *IJIST*, 13(1):95–105, 2003.
- [7] R. H. Liang and M. Ouhyoung. A sign language recognition system using hidden markov model and context sensitive search. In *ACM Symposium on Virtual Reality and Software Technology*, volume 6, pages 59–66, Hong Kong, China, July 1996.
- [8] D. O. Tanguay, Jr. Hidden Markov Models for gesture recognition. Master’s thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, August 1995.
- [9] A. Nefian. *A Hidden Markov Model-Based Approach for Face Detection and Recognition*. PhD thesis, Dept. of Electrical Engineering, GIT., Atlanta, Georgia, USA, 1999.
- [10] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.
- [11] F. Samaria and S. Young. HMM based architecture for face identification. *Image and Vision Computing*, 12:537–543, 1994.
- [12] T. E. Starner. Visual recognition of american sign language using Hidden Markov Models. Master’s thesis, School of Architecture and Planning, MIT., Cambridge MA, USA, 1995.
- [13] P. A. Stoll and J. Ohya. Applications of HMM modeling to recognizing human gestures in image sequences for a man-machine interface. In *Proc. 4th IEEE Int. Workshop on Robot and Human Communication*, pages 129–134, Tokyo, Japan, 1995.
- [14] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using Hidden Markov Model. In *Proc. IEEE CVPR*, 1992.