17th International Conference on Pattern Recognition (ICPR'04), Cambridge, United Kingdom, August 23-26, 2004

# Dining Activity Analysis Using a Hidden Markov Model

Jiang Gao[1], Alexander G. Hauptmann[1], Ashok Bharucha[2], and Howard D. Wactlar[1]

[1]*School of Computer Science,Carnegie Mellon University, Pittsburgh, PA 15213*
[2]*University of Pittsburgh Medical Center, Pittsburgh, PA 15213*
[1]{jgao, alex, hdw}@cs.cmu.edu; [2]bharuchaaj@msx.upmc.edu

## Abstract

*We describe an algorithm for dining activity analysis in a nursing home. Based on several features, including motion vectors and distance between moving regions in the subspace of an individual person, a hidden Markov model is proposed to characterize different stages in dining activities with certain temporal order. Using HMM model, we are able to identify the start (and ending) of individual dining events with high accuracy and low false positive rate. This approach could be successful in assisting caregivers in assessments of resident's activity levels over time.*

## 1. Introduction

Activity analysis from video is an active research area in recent years (refer to [1]). Past works have been focused on applying both high-level modeling and low level feature selection strategies to identify events of interest.

Hidden Markov models (Rabiner and Huang, 1993) have been applied successfully as a high level strategy to classify activities. Typically, a HMM model is trained for each activity class. Trained HMM models are then used to compute each model's similarity to a novel input sequence. For example, in Starner (1998), skin color and moments of blobs are used as features, and a 4 state HMM model with one skip transitions is adopted to classify the feature sequences into ASL vocabularies. Yamato (1992) are the first to use HMM in vision to recognize tennis strokes.

In this paper, our goal is to measure feeding difficulties in nursing home residents with severe dementia. "Feeding" for the purpose of this study is defined as, "the act of moving food from a plate to the mouth and, thereafter, swallowing it" (Siebens, 1986). "Difficulty" then is defined as any aspect of feeding which leads to reduced food intake (Watson, 1993). It is estimated that 25% of the total cost of caring for a demented person who is totally dependent could be accounted for by feeding related activities.

We plan to automatically measure number of hand movements to the mouth using computer vision technology. The frequency of feeding difficulties in a small group of severely demented institutionalized residents will be estimated from the digital data that will permit identification of the frequency, persistence, and pattern of food/fluid aversive behaviors that occur in individual subjects.

We have developed a motion segmentation algorithm to provide motion features for our system (Gao, 2004a). The algorithm is able to detect and label natural human motion with normal clothing and in natural environments. Combined with tracking and temporal consistency filtering, we are able to generate a motion feature that is self-initiating, robust to noise, and at the same time sensitive to subtle motions.

Based on motion features, in this paper, we first propose a strategy to analyze the subspace of an individual person, and measure distances between moving regions in the subspace. The distance measure provides an auxiliary feature to identify dining events.

Based on motion and distance features, we further propose a HMM model to identify relevant dining events. We use a single HMM model, with each state represents a different stage in dining activities. By applying the model to input observations, the state labels obtained by Viterbi decoding provide identification of each stage. Two stages: start and end of an eating event, is of interest to us. They provide frequency and duration of a resident's dining activity, and are important measurements to help caregivers in assessment of resident's activity levels over time.

The major contribution of this paper is providing a systematic approach to identify dining events based on motion and distance features. We have obtained results with high detection rate while maintain a low false positive rate. As far as we know, no previous systems were able to identify dining activities from video steams with such an accuracy in a long period of time.

The organization of this paper is as follows: Section 2 briefly reviews our method for obtaining motion features used in our system. Section 3 outlines our approach for modeling dining activities of an individual person and estimating distances between moving regions in the subspace of an individual person. Section 4 proposes a HMM model for identifying dining events based on motion and distance features. Section 5 gives experimental results. Finally, section 6 concludes the paper and discusses future works.

## 2. Motion segmentation and tracking

In this section we describe our motion features obtained by a combination of motion segmentation and tracking. *Motion segmentation* aims to find major motion patterns in image sequences, and segments images into regions corresponding to these motion patterns. We developed a motion segmentation algorithm based on dense optical flow and RANSAC (Fischler and Bolles, 1981).

Based on motion segmentation result, we *track* the segmented regions through a limited time window. At the same time, we perform a consistent motion filtering, and only keep regions with consistent motion directions over a time interval. In this way we further compensate errors in motion segmentation and tracking. This algorithm is tested extensively on nursing home video sequences, and can achieve an average detection rate of over 90 percent for motions of interest over a long term.

Fig. 1 shows an example of our motion features, which include motion vectors and segmented moving regions. A detailed description of our algorithm was presented in Gao (2004b).

**Fig. 1.** Automatically estimated moving regions and motion vectors. The red arrows are computed motion vectors; the blue masks indicate segmented moving regions.

## 3. Analyzing activities of individual person

Based on the motion features, we further analyze the subspace of individual person to characterize their activities. There are three steps in our algorithm:

1) Find individual persons;
2) Detect faces;
3) Characterize activities.

### 3.1. Finding individual person

We find the boundaries of individual persons by accumulating the segmented moving regions. Let $M(x, y, t)$ be a binary mask indicating all regions of motion in frame *t*, i.e.,

$$M(x, y, t) = 1 \qquad (1)$$

indicates the pixel at $(x, y)$ in frame $t$ belong to a moving region. Then the individual person regions are obtained by accumulating $M(x, y, t)$ over time. Let

$$Cluster_{\tau}(x, y, t) = \bigcup_{i=0}^{\tau-1} M(x, y, t-i). \qquad (2)$$

$\tau$ is the temporal duration. If $\tau$ is large enough, $Cluster_{\tau}(x, y, t)$ will approximately indicate the individual person regions at time *t*. Typically, we select $\tau$ to be around 2 minutes. A result of this algorithm is given in Fig. 2.

(a) (b)

**Fig. 2.** Finding individual person. (a) A frame of the video. (b) The individual person regions obtained using our method (duration is 2 minutes).

### 3.2. Face detection

We use a face detection algorithm (Schneiderman and Kanade, 2000) to find the faces of each person in individual regions, and track the faces over time.

### 3.3. Relative motion estimation

Within each individual region and with faces detected, we define a head-hand model as shown in Fig. 3. At the present stage, this model only describes persons sitting frontal or half frontal relative to the camera.

The head position in this model is provided by the face detection algorithm initially and tracked over time. The other two components in the head-hand model, i.e. 2 hands/arms are detected based on major regions in motion found by our algorithm. At each moment, we assign the 2 regions with consistent motions to the two hands/arms. If there are more than two regions in the candidate list, we select the regions overlapping with most other regions. The motion vectors of these two regions are then mapped to the main axes between the head and hands, as shown in Fig. 3. We use the projected distance change on these axes to characterize activities of each person.

After finding regions corresponding to the head and 2 hands/arms, we compute the distances between head region and 2 hands/arms regions. The distances are used as an auxiliary feature, as described in section 4.

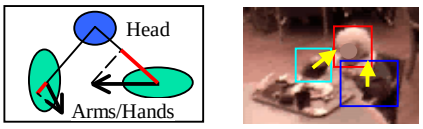


**Fig. 3.** Head-hand model for an individual person. The black arrows are motion vectors of 2 major arm/hand components. They are mapped to the head-hand axes to characterize the dining activity of the person. The red lines indicate normalized motion vectors used to characterize dining activities.

Note that for our task accurate identification of right and left hands is not required, as long as the major hand motions are captured using our approach. The dining events identified from motions of both hands /arms are then merged to obtain the final result.

## 4. Events detection using HMM

Based on motion and distance features, we developed a high level strategy to identify activities in a dining room. We use hidden Markov models for our purpose.

The topology of the HMM model is determined by estimating how many different states are involved in specifying the activities in a dining room. We use a 4 states topology to describe stages in dining activities. 2 states model the 2 movements that can be used to mark the start and end of each instance of eating: state $q_1$ model the movements with hands moving relatively toward head, while state $q_2$ model the relative movement in the opposite direction: hands moving away from the head. In addition, to model instances with no motion detected and tracked, we use a "don't care" state dc.

Unless we have an ideal situation, i.e., people stay motionless until and after instances of eating, the above 3 states cannot reliably model the activities in a dining room. Typically, people would move their hands between dishes, and sometimes talking with other people using gestures. We model motions from these activities using an additional state $q_3$.

The observation vector at each time instance is composed of motion and distance features for each hand. For each state, we use a mixture of 2 Gaussians to model their output (observation) probabilities, and estimate the parameters, namely means, covariances (we use a diagonal covariance matrix), and weighting coefficients for each mixture, based on 1.5 hour of hand labeled dining activities for 4 residents.

The prior probabilities for each state and the transition matrix between states are also estimated using the labeled data. In the training process, the initial parameter estimations are given by hand, and an EM algorithm is then used to give the final optimized estimation.

The advantage of using a HMM model for our purpose is as follows: First, mixture of Gaussians model gives an accurate probabilistic description of different stages in a dining process. For example, in eating instances, the distances between head and hands are typically shorter, and directions of motions are more along the head-hand axis, compared with motion incurred from moving between dishes and in gesture communications. These subtle differences can be accurately modeled using mixture of Gaussians.

Second, the temporal relation between each stages and probability of their occurrence also impose strong constraints in terms of temporal order of each state and the transitions between them. Fig. 4 shows topology of the HMM model described in this paper.

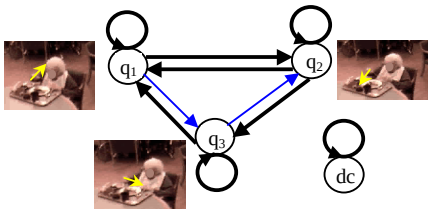


**Fig. 4.** Topology of the HMM model. The states: $q_1$-hand moving toward head, start of an eating event; $q_2$-hand moving away from head, end of an eating event; $q_3$-hand movements not directly related to eating, such as moving between dishes; dc-"don't care" state, no motion feature exists. Arrows denotes transitions between states, with blue arrows denoting largely unlikely transitions. State dc can transit to and from any other states.

## 5. Experimental results

We show results for 30 minutes of videos of 10 residents in a dining room. The segmentation and tracking results have been given in Fig.1.

Fig. 5 and Fig. 6 give results for one resident in a video of over 2 minutes long. Motion and distance features are shown as motion and distance curves along time axes. They are relative velocities and distances between head and hands.

Fig. 6 gives the result for identifying eating instances using the HMM. In both figures, the ground truth is labeled using magenta shadings on the time axes. Eating events are counted by hand motions toward the head, which is more clearly identifiable than the downward motion.

In Table 1 we give results of 10 residents, captured in different days. Total length of the videos is 30 minutes. The results by using HMM and the result

only based on relative motions between head and hands are compared. HMM gives much lower false positive rate while keeping a high detection rate at the same time.

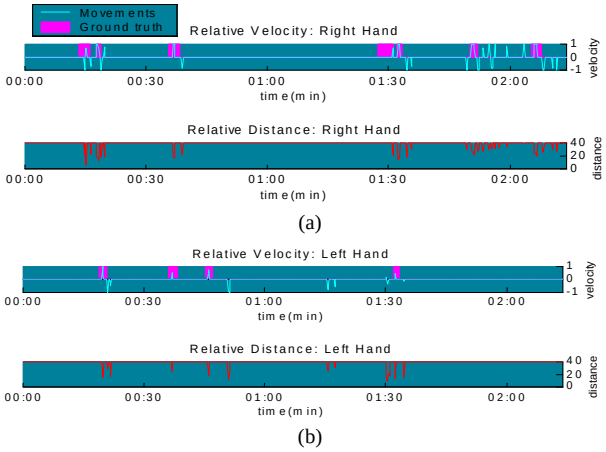


**Fig. 5.** Motion and distance curves for right hand (a) and left hand (b). The motion curves show hand motions mapped to the head-hand axes, with toward-head ground truth labeled. The distance curves are upper limited by 40.

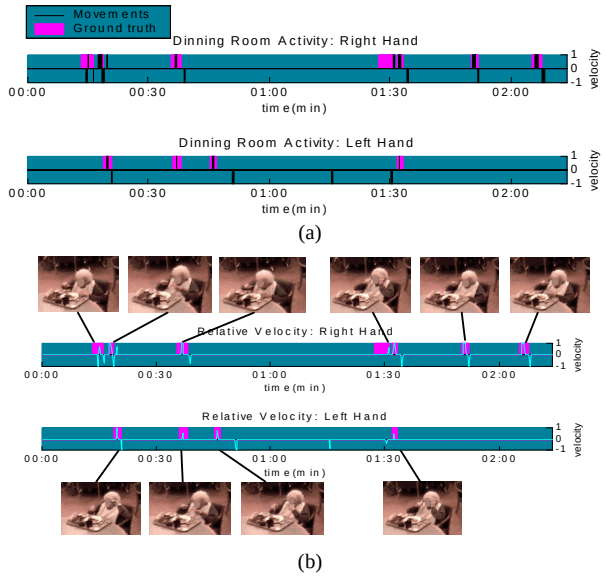


**Fig. 6.** Identification of eating events using a HMM. (a) Identified instances of hands moving toward and moving away from head, with toward head corresponding to positive velocity values. (b) Motion curves, with snap shots for each instances added.

**Table 1.** Dining room activities analysis results.

| Method | Correct Detection | Miss Detections | False Alarms |
|---|---|---|---|
| Relative Motions | 50 | 6 | 26 |
| HMM | 50 | 6 | 9 |

## 6. Conclusions

We proposed a hidden Markov model for dining room activity analysis. Our model combines both motion and distance features to identify individual dining events. HMM considerably reduces false positive rate while keeps high detection rate at the same time.

We are currently using the system to analyze dining activities in the nursing home across several days. The result can be used to assist caregivers by providing them with better measurements. We are also working on using the similar strategy to analyze other activities of interest, such as encountering and conversation in hallway at the same nursing home.

## Acknowledgements

This material is based on work supported by the National Science Foundation (NSF) under Grant No. IIS-0205219.

## References

[1] Special Section on Video Surveillance. *IEEE Trans. PAMI*, 22(8), 2000.

[2] L. Rabiner and B. Huang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] T. Starner, J. Weaver and A. Pentland, Real-time American sign language recognition using desk and wearable computer based video. *IEEE Trans. PAMI*, vol. 20, no.12, 1998.

[4] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov models. *Proc. ICCV*, 1992.

[5] H. Siebens, E. Trupe, A. Siebens, et al. Correlates and consequences of eating dependency in institutionalized elderly. *J Am Geriatr Soc* 34:192-198, 1986.

[6] J. Gao, R.T. Collins, A.G. Hauptmann and H.D. Wactlar, Articulated Motion Modeling for Activity Analysis. *IEEE Workshop on Articulated and Nonrigid Motion (ANM2004)*, 2004a.

[7] R. Watson, Measuring feeding difficulty in patients with dementia: perspectives and problems. *J. Adv. Nursing* 18:25-31, 1993.

[8] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, vol. 24: 381-395, 1981.

[9] J. Gao, A.G. Hauptmann and H.D. Wactlar, Combine Motion Segmentation with Tracking for Activity Analysis. *International Conference on Automatic Face and Gesture Recognition*, 2004b.

[10] H. Schneiderman, T. Kanade. A statistical method for 3D object detection applied to faces and cars. *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.