Behavior Recognition with HMM and Feature Analysis using a Wide-view Camera and a PTZ Camera

Norimichi Ukita, Akihito Kotera, and Masatsugu Kidode
Nara Institute of Science and Technology

Abstract

This paper proposes a method for recognizing human behaviors based on a combination of "overall motion analysis using Hidden Markov Model (HMM)" and "detailed feature analysis in a specific area by using an active (pan-tilt-zoom, PTZ) camera". The PTZ parameters are controlled depending on the intermediate result of analyzing the target's motions based on HMM. As a result, our system can distinguish between similar behaviors by observing high-resolution characteristic features (e.g., motion/shape). Experimental results demonstrated the effectiveness of the proposed system.

1 Introduction

Recently, a number of home appliances are endowed with intelligent technologies such as sensing and communication functions. Similarly, it is expected that personal-use robots at home will become popular. If these systems can understand people's situations (e.g., what people are doing and want to do), the systems can support their daily activities. For these systems, human-behavior recognition is essential.

Human-behavior recognition is an important technique for various kinds of human-computer interaction systems in the real world. Most of the algorithms for behavior recognition first extract motion features (e.g., a motion history image[1] and optical flow[2]) from observed images. These motion features are then classified into predefined categories each of which defines one behavior. However, the difference between the motion features of similar behaviors is imperceptible in the images. This results in difficulty in classifying the similar behaviors.

We tackle this problem with a vision system consisting of a fixed wide-view camera and an active pan-tilt-zoom (PTZ) camera as follows:

1. Common behavior analysis with Hidden Markov Model[3] (HMM) is applied to temporal images observed by the wide-view camera in order to (1) classify the history of observed large motions (see [4, 5], for example) and (2) control the active camera for capturing high-resolution images of small but characteristic appearances/motions of each behavior.

2. The high-resolution appearances/motions are analyzed in order to identify the observed behavior.

The idea of active vision is classical and popular in target tracking[6], object recognition[7], and many other vision-based algorithms. In our approach, the active camera is controlled for capturing images suitable for discriminating between similar behaviors.

2 Basic Schemes

In this work, two conditions are assumed. All image sequences are observed in the same configuration between a camera system and a person (e.g., web cam on a monitor). A long sequence is segmented so that each segmented sequence shows one of behaviors[1].

Under these assumptions, the following three steps for learning behavior samples are executed:

1. Acquiring motion data (Sec. 3.1):

   Optical flow is computed in wide-view images.

2. Generating Behavior Models (Sec. 3.2):

   With a number of acquired motion data of each behavior, its behavior model is generated using the Baum-Welch algorithm[3].

3. Detecting features for discriminating between similar behaviors (Sec. 3.3):

   Each sample sequence is evaluated by all the behavior models to find similar behaviors. Then, features that can discriminate between these behaviors are extracted with the PTZ camera.

With the learning data, the system first tries to recognize an observed behavior by HMM analysis.

1. Acquiring motion data (Sec. 3.1):

   Motion data is computed online from an observed image sequence as well as in the learning step.

2. Recognition using HMM (Sec. 4.1):

   The motion data is classified into one of the predefined behaviors using the behavior models.

If the system cannot identify the observed behavior with HMM, the following feature analysis is executed.

1. Target selection for camera control (Sec. 4.2):

   A target region is detected in a wide-view image. The PTZ camera is then controlled towards it.

2. Integrating wide-view HMM and high-resolution feature analyses (Sec. 4.3):

   The recognition result is estimated by integrating two results, HMM and feature analysis with high-resolution images captured by the active camera.

In our method, selective regions are analyzed depending on the result of HMM. Our goal is (1) to select target features suitable for identifying a current behavior depending on the result of HMM analysis and (2) to control an active camera in order to capture high-resolution images of the target features.

[1]For this assumption (i.e., segmented sequences), spotting from a long image sequence[8, 9] is proposed as is well known.

Figure 1: Computed optical flows: each arrow is extended for visualization.

3 Learning Models and Features

3.1 Acquiring Motion Data for HMM

As with many behavior-recognition systems, our system first tries to identify an observed behavior from motion data by employing HMM. In our system, the motion data is represented by optical flows computed by the Lucas-Kanade algorithm[10] (Fig. 1). The image is divided into blocks at regular intervals, each of which has a unique ID ($\in \{1, \cdots, N_D\}$, where N_D indicates the number of the blocks). Then, the average of the optical flows in each block is obtained. The average that has the maximum norm is employed for learning and recognition. In practice, the 3D vector consisting of "the ID of the block that has the maximum average" and "the 2D vector representing the maximum average" is employed.Many behavior-recognition algorithms need the temporal history of positions of a body part (e.g., head/hand) as motion data. Body-part tracking is, therefore, required. In our approach, on the other hand, tracking is not required.

3.2 Generating Behavior Models

Motion data mentioned in Sec. 3.1 is acquired from all sample image sequences observed by a wide-view camera. Each behavior model is then generated by employing the Baum-Welch algorithm[3]. The left-to-right model is used for simplification. In practice, we used the Hidden Markov Model Toolkit (HTK)[11] for modelling. For generating each behavior model with the HTK, the number of states must be given. We test several patterns of the number of states and select the one that has the most optimized parameters.

3.3 Detecting Detailed Features

If it is possible to find a minute difference between similar behaviors from zoom-in images captured by the PTZ camera, these behaviors can be discriminated. Two examples are as follows:

Motion difference: In similar behaviors, the motions of a whole body are almost the same. The motions of a certain body part (e.g., hand), however, may be minutely different from each other.

Appearance difference: While the trajectories of a hand are almost the same, there may be a different item in the hand depending on the behavior.

These differences must be able to be detected by image-analysis algorithms implemented in the system. If complex image recognition algorithms are employed, wide variety of differences can be detected, but, recognition may sometimes fail. Therefore, the prototype system proposed in this paper has the following simple and robust functions in order to confirm the effectiveness of the basic idea of our system:

Optical flow estimation: As in the case of wide-view analysis, the difference of motion data represented by computed optical flows[10] is useful also in the case of analyzing zoom-in images. For stable computation, the temporal average and deviation of optical flows are obtained.

Skin detection: Since the item in a hand is changed depending on each behavior, the behavior can be identified by recognizing the item. It is, however, difficult to recognize all kinds of items; for example, a number of kinds of cups must be able to be recognized in order to identify *drinking*. Instead of detecting items themselves, we focus on the changes in skin areas due to the difference of the shapes of the items. Practically, skin detection is implemented using [12] and the number of skin pixels is obtained and considered to be the characteristic feature.

To find the difference between similar behaviors, we design the following learning scheme:

Step 1: Behavior recognition using HMM with sample images observed by a wide-view camera is executed as well as the recognition step. If one or more sample sequence of a certain behavior get high likelihoods in recognition using the model of another behavior by mistake, these two behaviors are considered to be *similar behaviors*.

Step 2: In the prototype system presented in this paper, therefore, we scope out the differences between the similar behaviors manually. What we scope out for each combination of the similar behaviors are (1) the PTZ parameters for observing the image region where the difference is observed and (2) the timing (i.e., the frame number) when the difference is observed.

Step 3: Although the timing (i.e., the frame number) for camera control is found in the sample sequence in Step 2, the frame number might be changed among image sequences depending on the motion speed of a person in each sequence. To cope with this problem also, HMM is useful because it can find a typical temporal feature robust to time warping. To choose the timing of camera control online, we prepare HMMs trained by image sequences, each of which shows one of the behaviors from its beginning to the timing when or right before the difference is observed. With these short HMMs, the system can choose the timing of camera control and then the corresponding target region for capturing characteristic features of the behavior can be also determined. We call these short HMMs *Camera-Control HMMs* (*CCHMMs*). The number of the CCHMMs is $\sum_{i=1}^{N_s} N_b^i$, where N_s and N_b^i denote the numbers of "the combinations of the similar behaviors" and "the similar behaviors in each combination", respectively. Hereafter, on the other hand, we call the HMMs representing the whole image sequences *Whole-Sequence HMMs* (*WSHMMs*). The number of the WSHMMs is equal to the total number of the behaviors.

Step 4: The sample features indicating the difference of the similar behaviors are observed for learning by employing a PTZ camera as follows. For learning each feature, a wide-view camera observes a person who repeats one of the similar behaviors and the CCHMM corresponding to this behavior is evaluated. The PTZ camera is controlled when this CCHMM gets a high likelihood at the end. The PTZ parameters are associated with this CCHMM in advance (i.e., Step 2). The captured zoom-in images are evaluated by the algorithm for analyzing the target feature in order to prepare the samples of this feature indicating the difference of the similar behaviors.

In the recognition step, the likelihoods in the wide-view sequence and the zoom-in sequence are integrated if the zoom-in sequence is captured (described in Sec. 4.3). Note that, in zoom-in image analysis for an online input sequence, the same image-analysis algorithm must be applied to the same region in order to compute the likelihoods of all behaviors for equitable evaluation.

4 Recognition with the Combined HMM and Feature Analysis

4.1 Recognition using HMM

Behavior recognition using each HMM is implemented with the Viterbi algorithm[3]. Note that the probabilities of both WSHMMs and CCHMMs are evaluated simultaneously in our system.

If the highest probability of one WSHMM is above a predefined threshold, the behavior represented by this model is regarded as a candidate of the observed behavior. Depending on the number of the candidates and whether or not active camera control is occurred, the recognition result is determined as follows:

- If the candidate is only one, this candidate is regarded as the observed behavior.
- If the active camera is controlled (described in Sec. 4.2), the recognition result is determined taking into account the result of analyzing images observed by the PTZ camera (described in Sec. 4.3).

4.2 Target Selection for Camera Control

When the highest probability of one of the CCHMMs is higher than a predefined threshold, camera control is started. By the learning step, the corresponding target region, in which the characteristic feature of the behavior represented by this CCHMM is observed, and image-analysis algorithm (i.e., optical flow estimation or skin detection in our prototype system) are already determined. After obtaining the feature in this region, the PTZ camera is ready to be controlled again. This camera control is repeated until the behavior is finished. That is, the camera is controlled whenever the highest probability of one of the CCHMMs is above the predefined threshold.

The likelihoods of the obtained feature with respect to the samples of all behaviors are computed by the corresponding image-analysis algorithm.

4.3 Integrating Wide-view HMM and High-resolution Feature Analyses

For equitable evaluation, the likelihoods of all the behaviors must be computed from the same combination of feature analyses. For this purpose, in the learning step, the samples of all the behaviors for all feature analyses are already acquired. With these samples, the definite recognition result is determined from the integrated likelihoods of all behaviors as follows:

Step 1: Assume that the total number of behaviors is N_B. When the active camera is controlled N_F times and N_F characteristic features are captured, $N_B \times N_F$ likelihoods $L_b(f)$ are computed, where $L_b(f)$ denotes the likelihood of the f-th feature ($f \in \{1, \cdots, N_F\}$) with respect to the sample of the b-th behavior ($b \in \{1, \cdots, N_B\}$).

Step 2: The definite likelihood of the b-th behavior is represented by the equation below:

$$l_b = l_b^W \prod_{f=1}^{N_f} L_b(f),$$

where l_b^W denotes the likelihood of the b-th behavior estimated by the WSHMM of the b-th behavior.

Step 3: The largest of $l_1, \cdots, l_{N_B}$ is selected and the corresponding behavior is considered to be the observed behavior.

5 Experimental Results

We conducted experiments using a Pentium4 2.8GHz computer and SONY EVI-D30 (PTZ camera; 640×480 pixel, 30 fps) $\times$ 2. The distance between two cameras, which were located horizontally, was 20cm. This camera system was placed at a distance of 200cm from a target person.

For evaluating the proposed system, the following four kinds of behaviors in daily life were observed: (1) *drinking*, (2) *eating*, (3) *reading*, and (4) *writing*, all of which are observed when a subject is in a sitting position. Figure 2 shows examples of these behaviors.

Although the number of target behaviors is only four, *drinking* and *eating* are similar because the right hand moves up and down between the lips and a table, and *reading* and *writing* are similar because the whole body remains almost stationary.

In the learning step, the following features were selected manually and obtained as the differences between the above similar behaviors:

Temporal change in the number of skin pixels around the block with the maximum average of optical flows: Since the cup overlap the face while the cup is raised to person's lips, some skin pixels is occluded around the block with the maximum average of optical flows, namely around the moving hand region, in *drinking*. Accordingly, we extracted the image sequence for the CCHMM from each sequence for the WSHMM so that the extracted sequence is finished when the hand begins moving upwards to control the PTZ camera towards the following motion.

Temporal average and deviation of optical flows in bottom skin regions: In contrast to *reading*, *writing* shows small hand motions at any time. Therefore, the temporal average and deviation of optical flows in the hand region, which corresponds to the bottom of all skin regions as shown in Fig. 2, can be characteristic features. Accordingly, we extracted the short image sequence for the CCHMM from the beginning of each sequence for the WSHMM to control the PTZ camera towards the bottom skin region after the system starts.

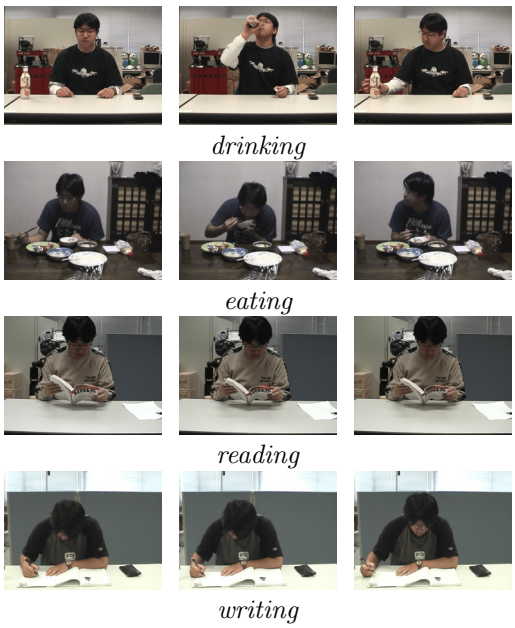

Figure 2: Example sequences of four behaviors.

Table 1: Recognition results of the WSHMM analysis and the proposed system.

behaviors	WSHMM analysis		proposed system	
	success	miss	success	miss
drinking	13	7	18	2
eating	12	8	17	3
reading	6	14	14	6
writing	10	10	15	5

First, image sequences with a single subject were used for both learning and recognition. In these experiments, ten sequences of each behavior were prepared for learning and other twenty sequences were for recognition. Table 1 shows the recognition results of the WSHMM analysis and the proposed system. It can be confirmed that the proposed system improves the recognition performance well especially in the case of *reading* and *writing*.

In our method, the results of optical flow estimation and skin detection are employed for recognition. Using [12], skin regions could be stably detected even under small changes in illumination. On the other hand, acquired optical flows were unstable. Error recognition results in *reading* and *writing* were caused by failure in estimating optical flows[10]. One of the reasons of why these detection failures were occurred was that the time-interval for observing the target region (i.e., hand region) by the active camera was too short to observe hand motions reliably; for example, the hand motion for *writing* sometimes stops. However, if the observation interval is too long, the next camera control cannot be executed when it is required. To solve this problem, (1) a suitable observation interval should be determined automatically depending on each behavior and (2) multiple active cameras are useful for reliably observing multiple behaviors that are happened in series or simultaneously.

Next, image sequences of other two subjects were recorded and recognized using the training data employed in the above described experiments. Ten sequences of each behavior for each subject were prepared. Table 2 shows the recognition results. The recognition results were inferior those of the experiments with a single subject. However, the superiority of the proposed system can be confirmed.

Table 2: Recognition results of other subjects.

behaviors	WSHMM analysis		proposed system	
	success	miss	success	miss
drinking	10	10	16	4
eating	11	9	17	3
reading	6	14	13	7
writing	11	9	16	4

6 Concluding Remarks

This paper proposes a method for recognizing human behaviors based on a combination of "overall motion analysis using Hidden Markov Model with optical flows" and "detailed feature analysis in a specific area by using a PTZ camera". HMM analysis is employed to determine how to control the PTZ camera. As a result, our system can distinguish between similar behaviors by observing characteristic features of each behavior by using the active camera.

References

[1] A. Bobick and J. Davis, "Real-time Recognition of activity using temporal templates," the Workshop on Application of Computer Vision, pp.39–42, 1996.

[2] M. Black and A. Jepson, "A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions," ECCV, pp.909–924, 1998.

[3] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in Proc. of IEEE, pp.257–286, 1989.

[4] J. Yamato, J. Ohya, and K. Ishii, "Recognizing Human Action in Time-Sequential Images using Hidden Markov Models," CVPR, pp.379–385, 1992.

[5] T. Mori, Y. Nejigane, M. Shimosaka, Y. Segawa, T. Harada, and T. Sato "Online Recognition and Segmentation for Time-Series Motion with HMM and Conceptual Relation of Actions," IROS, pp.2568–2574, 2005.

[6] G. Sandini and M. Tistarelli, "Active tracking strategy for monocular depth inference over multiple frames," PAMI, Vol.12, pp.13–27, 1990.

[7] S. J. Dickinson, H. I. Christensen, J. K. Tsotsos, and G. Olofsson, "Active Object Recognition Integrating Attention and Viewpoint Control," CVIU, Vol.67, No.3, pp.239–260, 1997.

[8] H.-D., Yang, A.-Y., Park, and S.-W. Lee, "Human-Robot Interaction by Whole Body Gesture Spotting and Recognition," ICPR, Vol.4, pp.774–777, 2006.

[9] J. Song, D. Kim, "Simultaneous Gesture Segmentation and Recognition based on Forward Spotting Accumulative HMMs," ICPR, Vol.1, pp.1231–1235, 2006.

[10] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," IJCAI, pp.674–679, 1981.

[11] http://htk.eng.cam.ac.uk/

[12] T. Wada "Color Target Detection based on Nearest Neighbor Classifier," IPSJ Transactions on CVIM, Vol.44, No.SIG17, pp.126–135, 2003.