PANORAMIC CAPTURING AND RECOGNITION OF HUMAN ACTIVITY

Xinding Sun, B. S. Manjunath
Department of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106
{xdsun, manj}@ ece.ucsb.edu

ABSTRACT

This paper presents a unified approach to human activity capturing and recognition. It targets applications such as a speaker walking, turning around, sitting and getting up from a chair in a classroom setting. A panoramic camera capturing system is designed for video capture. Virtual camera control outputs the region of interest (ROI) video that covers the speaker. Given a ROI sequence, the virtual camera control parameters are used for the recognition of activities like walking, and the motion parameters of each frame are used for the recognition of other activities like turning around, sitting down and getting up etc. For motion parameter based recognition, the likelihood of the motion parameters is represented using a multivariate Gaussian model. The temporal change of the likelihood is characterized using a continuous density hidden Markov model (HMM). Experimental results show that the method works well in recognizing the above mentioned human body activities.

1. INTRODUCTION

Analysis of typical activities such as a speaker walking, turning around, sitting down on a chair, and getting up from a chair in a classroom settings is the main concern in this paper. It has many potential applications in the areas of indexing classroom and seminar presentations, and in human computer interfaces. While recognition of these activities is important, capturing of the regions of interest corresponding to these activities is essential to the success of the overall application. While most previous work discuss the two problems separately, we present a framework that integrates the capturing and recognition processes.

There have been quite a few systems that use active camera like Sony's EVI camera, or combine it with other wide-angle cameras for seminar capturing [7]. In general these systems involve camera motion, which is not helpful in the recognition process. Therefore, we choose the Flycam [3] panoramic system to capture the activities. For the system, we have designed virtual camera control methods [9][10] to output ROI video that covers the speaker. The advantage of this system is that it is automatic, and the speaker is always in the scene. There is no physical camera motion in the system and the virtual camera parameters are readily available for recognition purpose. Figure 1 gives one example of such a capturing result.

The human activities usually involve changes in the environment, object occlusion, etc. Therefore, feature point based or region-based techniques that work well on facial expression, gesture recognition, etc. [8] cannot be directly applied here. Given the complexity of human body motion, techniques that do not require explicit image feature detection or segmentation are of much interest. Among the early work is Davis and Bobick [2], wherein they use temporal templates for human movement recognition. Their method requires less computation, but is sensitive to variances in the movement. Little and Boyd [4] use the moments of moving points to represent the optic flow for the purpose of periodic human gait recognition. Yacoob and Black [11] propose recognition of activities based on matching of principal components under global temporal change.

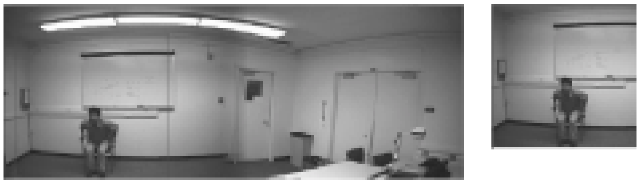

(a) Panoramic view (b)ROI output

Figure 1. Panoramic capturing and ROI output.

Our proposed method integrates the capturing and recognition processes. The virtual camera control parameters are used for the recognition of some activities like walking, and the motion parameters of each frame are used for the recognition of other activities. Similar to those using global motion fields that do not require image feature tracking or segmentation, we introduce a multivariate Gaussian model to represent the likelihood of the motion parameters. The temporal change of the likelihood is characterized using a HMM for activity recognition. Motion parameter based recognition of activity is then posed as a maximum likelihood parameter estimation problem. The virtual camera control parameters and HMM are designed to work on different types of activities. Experimental results show that the method works well in recognizing such complex human body activities.

2. PANORMAIC CAPTURING OF HUMAN ACTIVITIES

The Flycam [3] panoramic system is used to capture the speaker. The camera system is fixed and covers all the area where the speaker activities take place. The panoramic system produces real time panoramic video output. While we can compress the panoramic video first and extract ROI video from compressed domain for later activity recognition [10], here we choose to extract ROI video in real time [9] from the panoramic video. By doing this, we can avoid storing extra large amount of redundant

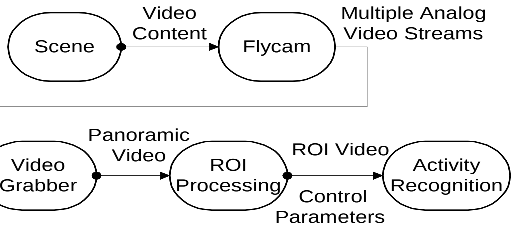


Figure 2. General system architecture for activity capturing and recognition.

data outside the ROI area in the panoramic video and still do not lose any information about the activities. Figure 2 shows the general system architecture for activity capturing and recognition. The ROI video output and its associated virtual camera control parameters are used for activity recognition.

Our initial experiments consist of ten activities. We separate these activities into three groups. In the first group, we have: walking toward left and walking toward right. In the second group, we have: turning of the body from left to front ("l2f"), front to left ("f2l"), front to right ("f2r") and right to front ("r2f"). In the third group we have: standing up ("su"), sitting down ("sd"), starting to sit down but returning to the standing position without sitting down ("bu"), and starting to get up (from a sitting position) but returning to the sitting position without getting up ("bd"). The third group is designed in such a way that the sequences have similar sub-processes. Figure 3(a-c) shows three representative frames (r-frames) from a "su" ROI video sequence.

The speaker is simply modeled as a point object corresponding to the centroid of the body. The ROI output is a predetermined rectangular region that surrounds this point. Thus, the ROI basically tracks the body centroid. We have a simple centroid model $\mathbf{F}(k)=\left[x(k),y(k),v_x(k),v_y(k)\right]^T$, where $x(k),y(k)$, are the positions of the centroid, and $v_x(k),v_y(k)$ are the velocities of the centroid in x and y direction respectively. The ROI detection results are processed through a Kalman filter. The Kalman filter output $\overset{\wedge}{\mathbf{F}}(k)=[\hat{x}(k),\hat{y}(k),\hat{v}_x(k),\hat{v}_y(k)]^T$ is then used to steer a virtual camera to create smooth ROI video output. According to [9], the virtual camera control has three regimes. When the speaker is motionless or moving only in a small region, ROI is kept at the same position (stabilization control). When the speaker changes his position by a large distance, an IIR filter is used to steer ROI to catch up with the speaker (transition control). After the speaker has been centered, ROI is changed according to the estimate $[\hat{x}(k),\hat{y}(k)]$ (following control).

3. HUMAN ACTIVITY RECOGNITION BASED ON VIRTUAL CAMERA CONTROL PARAMETERS

Further observation of the activities described in the last section show that the first group of activities basically correspond to the virtual camera control in the "transition control" and "following control" regimes in x direction, while the second and the third group correspond to the "stabilization control" in x direction.

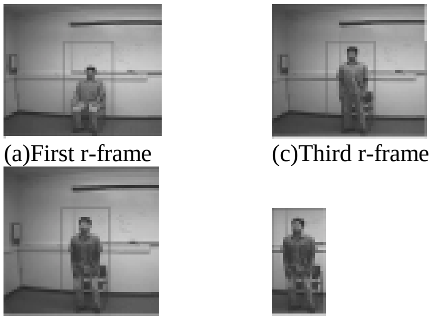

(a)First r-frame (c)Third r-frame
(b) Second r-frame (d) Object window in (b)

Figure 3.Representative frames from a 'su' ROI sequence.

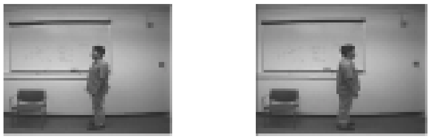

(a)Walking toward left (b)Walking toward right

Figure 4. Representative frames from walking sequences.

Figure 4 shows representative frames of the walking sequences. Since walking is a periodic process, modeling it is hard [4] . However, from above observation, we conclude that a decision on pattern of walking activity can be made if virtual camera control process falls into the categories of "transition control" and "following control" in x direction.

4. HUMAN ACTIVITY RECOGNITION BASED ON MOTION PARAMETERS

The second and the third group of activities correspond to virtual camera motions that are not consistent in one direction. It is not straightforward to do activity recognition directly based on the virtual camera parameters as discussed in the previous section. Therefore, we propose to use a probabilistic model to characterize these types of activities.

4.1. A Gaussian motion parameter model

Here, we use a model-based approach proposed in [1] to compute the motion parameters. Here we use simplest model, i. e. the 2-D optic flow or the motion vector for recognition. The motion vector $\mathbf{V}$ at a given position $\mathbf{x}=(x,y)$ can be represented as: $\mathbf{V}(\mathbf{x})=(v_x,v_y)$. These parameter values are then organized into a vector by row scanning the image. Let L be the number of pixels in a video frame or a region of interest in a frame (ordered according to a row scan). Let

$$\mathbf{Z}=[v_x^1,v_x^2,......v_x^L,v_y^1......v_y^L]^T \tag{1}$$

Note that $\mathbf{Z}$ is a $N=2\times L$ dimensional vector. We model $\mathbf{Z}$ as a multivariate Gaussian. Let the mean of this Gaussian be $\mathbf{m}$ and the covariance be $\mathbf{Q}$. Then, given $\mathbf{Z}$ from an observation class Ω, we can write the conditional probability $P(\mathbf{Z}\mid\Omega)$ as:

$$P(\mathbf{Z}\mid\Omega)=\frac{\exp(-\frac{1}{2}(\mathbf{Z}-\mathbf{m})^{\mathbf{T}}\mathbf{Q}^{-1}(\mathbf{Z}-\mathbf{m}))}{(2\pi)^N\left|\mathbf{Q}\right|^{1/2}} \tag{2}$$

If we have activity in class Ω, then (2) gives the likelihood of the motion parameters for a given frame. This is the basis for later statistical modeling of the activities using HMM. This approach to modeling the observation is inspired by the work in [5], where the observation vector is the image intensity, and the application is object recognition. In the following discussion, we will refer to $\mathbf{Z}$ as the *motion object (MO)*.

Let $\tilde{\mathbf{Z}} = \mathbf{Z} - \mathbf{m}$,the covariance matrix can then be decomposed as: $\mathbf{Q} = \mathbf{\Phi \Lambda \Phi^T}$, where the columns of $\mathbf{\Phi}$ are the orthonormal eigenvectors of $\mathbf{Q}$, and $\mathbf{\Lambda}$ corresponds to the diagonal eigenvalue matrix of $\mathbf{Q}$. Let $\mathbf{Y} = \mathbf{\Phi^T} \tilde{\mathbf{Z}}$, (2) can be computed as:

$$P(\mathbf{Z}|\Omega) = P_P(\mathbf{Z}|\Omega)P_{\underset{p}{-}}(\mathbf{Z}|\Omega)$$

$$= \left[\frac{\exp(-\frac{1}{2}\sum_{1}^{M} y_i^2/\alpha_i)}{(2\pi)^{M/2}\prod_{1}^{M}\alpha_i^{1/2}} \right] \left[\frac{\exp(-\frac{1}{2}\sum_{M+1}^{N} y_i^2/\alpha_i)}{(2\pi)^{(N-M)/2}\prod_{M+1}^{N}\alpha_i^{1/2}} \right] \qquad (3)$$

Where M is the dimension of the principal subspace, y_i is the ith component of $\mathbf{Y}$, and α_i is the ith eigenvalue of $\mathbf{Q}$.

In (3), we divide the likelihood for a MO into two parts; the first part, $P_P(\mathbf{Z}|\Omega)$, corresponds to the likelihood of the MO in the principal subspace as used in principal component analysis (PCA); the second part, $P_c(\mathbf{Z}|\Omega)$ corresponds to the likelihood of the MO in the complementary orthogonal subspace of the principal subspace. PCA has been successfully used activity analysis [11]. The principal space is enough for general representation and approximation purposes. However, note that the likelihood in the principal space $P_P(\mathbf{Z}|\Omega)$ does not provide an optimal approximation of the likelihood $P(\mathbf{Z}|\Omega)$ in the whole space. The second part $P_c(\mathbf{Z}|\Omega)$ plays an important role in the recognition process. This is also observed in our experiments (discussed in section 5). To reduce the expense of computation of $P_c(\mathbf{Z}|\Omega)$, following [5], we use an optimal approximation of it:

$$P_c(\mathbf{Z}|\Omega) \approx \left[\frac{\exp(-\frac{1}{2}\sum_{M+1}^{N} y_i^2/\rho)}{(2\pi\rho)^{(N-M)/2}} \right] \qquad (4)$$

where $\rho = \frac{1}{N-M}\sum_{M+1}^{N}\alpha_i$.

It is generally not necessary to use the motion parameters of the whole ROI video frame for activity recognition. Instead, an even smaller region that covers the object, called object window, is chosen in our experiment. Figure 3(d) shows such an object window in the frame of Figure 3(b).

4.2. HMM for activity modeling

In the context of human motion recognition, promising results have been obtained using the HMM [12]. A generic HMM [6] can be represented as $\lambda = \{\Xi, A, B, \pi\}$, where $\Xi = \{q_1, q_1, ...q_{N'}\}$ denotes the N' possible states, $A = \{a_{ij}\}$ denotes the transition probabilities between the hidden states, $B = \{b_j(.)\}$ denotes the observation symbol probability corresponding to the state j, and π denotes the initial state distribution.

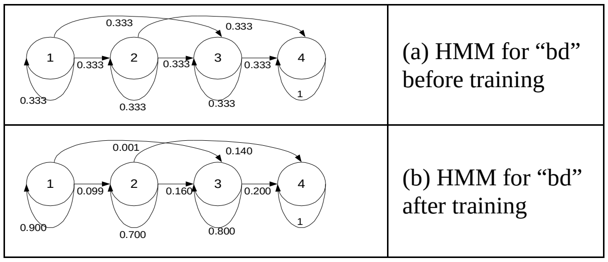
We choose a four-state continuous density HMM for activity recognition here. The number of states is empirically determined and we observed that an increase to a larger number of states did not result in any performance gains on our initial data sets. The first step of our HMM training is to obtain the observation model B . Since the motion pattern at any given short interval can be regarded as unchanged, we can divide the sequence into temporal segments where each segment corresponds to a state. We uniformly segment each training sequence into four segments before clustering. Each segment is assigned a state number that is the same as its segment order in the sequence. As in speech recognition, this method provides a good initial clustering of states. The position of the MO in each frame is manually selected around the moving subject. Then, we compute $\mathbf{m}$ and $\mathbf{Q}$, and consequently $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ for each state. After this step, we follow the conventional *K-means* clustering method to iteratively classify the frames based on their likelihood computed using (3). Any misclassified initial segmentation can be corrected in the clustering process. Note that we have one set of bases for each hidden state.

The next step is to obtain the state transition matrix A . This is done using the EM algorithm. A is initialized as shown in Figure 5(a). Note that we do not need to compute π , as in our model we always start in state 1. The trained HMM structure for the "bd" activity is shown in Figure 5(b). Given a video sequence $\{O_1, O_2, ...O_T\}$, where T is the length of the sequence, we then want to find one model λ_{i*} from a given dictionary $\{\lambda_1, \lambda_2, ...\lambda_E\}$; the recognition of the activity λ_{i*} follows from the maximum likelihood estimate:

$$i* = \arg\max_{1 \le i \le E}[P(O/\lambda_i)] \qquad (5)$$

5. EXPERIMENTAL RESULTS

The Flycam system is used to capture the speaker activities and produces panoramic video of size 800x300 in pixel resolution. The output ROI window size is 200x200. The size of object widow for motion parameter based recognition is 64x160. We collect 20 sequences for each activity. Each sequence contains 20 to 30 frames. Half of the video sequences are used for training, while the other half are used for evaluation. For simplicity, the subjects are asked not to wave hands or make other gestures while recording the video. They also pause for a while between two consecutive activities. This creates artificial zero motion frames in the video, and thus simplifies the segmentation of activities. Therefore, it makes the recognition similar to isolated-word instead of connect-word recognition in speech processing [6].



(a) HMM for “bd” before training

(b) HMM for “bd” after training

Figure 5. An example HMM for the “bd” sequence.

In recognizing the activities, the ROI sequences are first segmented into smaller sequences containing one single activity each based on the temporal position of zero motion frames. The recognition of walking activity is processed using the virtual camera control parameters first. For the rest of the video sequences, motion parameter based recognition method is used. The optic flow vectors of a ROI frame are computed first based on [1] to obtain the MO. The MOs are normalized to a zero-mean unit-norm. Since the subject is kept at a constant distance from the camera, no normalization is needed here. However, if the video is captured at different scales, we can use the bilinear transformation to normalize the parameters.

Table 1 summarizes recognition results. Results on group 1 activities are stable as expected. Results for group 2 are better than those for group 3 activities. This is partly due to the fact that group 3 activities share similar sub-processes, making their estimation more difficult. Also, group 3 activities are more complex. For example, the first state of “su” is the same as the first state of “bu”. In addition, the transitions in “bu” and “bd” are also more complicated than those in group 1.

Two different numbers of principal subspace dimensions are also tested. In general, larger dimensions of principal subspaces perform better than smaller ones, but we did not observe significant differences here between six and ten dimensions. PCA based method is also tested here. It is done by taking $P_{\bar{P}}(\mathbf{Z} \mid \Omega)$ out of computation in (3). It is essentially the same feature used in [11]. It can be seen from experiment that in general MO method outperforms the PCA method.

6. CONCLUSIONS

While most previous solve the problem of capturing and recognition of human activities separately, in this paper we present a unified approach that integrates the capturing and recognition processes. A panoramic camera capturing system is designed for capturing purpose. Virtual camera control outputs the ROI video that covers the person. Given a ROI sequence, the virtual camera control parameters are used for the recognition of some activities, and the motion parameters of each frame are used for the recognition of other activities. The likelihood of the motion parameters is optimally approximated based on a multivariate Gaussian model. The temporal change of the likelihood is characterized using a continuous HMM for activity recognition. Experimental results show that the method works well in recognizing such complex human body activities.

For simplicity, we have worked on activity sequences that have explicit shot boundaries. Our on-going work is the recognition of activity sequences without explicit boundaries. We are also investigating the research on a more general Bayesian Network for even more complex human activities.

Activity		Group1	Group2	Group3
Virtual Camera Control Parameters		100%	----	----
PCA	6 bases	----	75%	60%
	10 bases	----	80%	75%
MO	6 bases	----	90%	80%
	10 bases	----	90%	85%

Table 1. Experimental results on the test sequences.

7. ACKNOWLEDGEMENT

This research is in part supported by the following grants/awards: Office of Naval Research grant #N00014-01-1-0391, an NSF instrumentation award #EIA-9986057, an NSF infrastructure award #EIA-0080134. FX Palo Alto Laboratory provides the panoramic camera for the experiment.

8. REFERENCES

[1]. J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical Model-Based Motion Estimation,” in *ECCV’92*, pp.237-252, 1992.

[2]. J. W. Davis and A. F. Bobick, " The representation and recognition of human movement using temporal templates,” in *CVPR’97*, pp.928-34, 1997.

[3]. J. Foote, and D. Kimber, “FlyCam: practical panoramic video and automatic camera control,” *Proc. ICME’2000*, pp. 1419-1422, 2000.

[4]. J. J. Little and J. Boyd, “Recognizing People by Their Gait: the Shape of Motion, ” Videre, 1(2), the MIT press, 1998.

[5]. B. Moghaddam and A. Pentland, “Probabilistic Visual Learning for Object Representation,” *IEEE PAMI*, 19(7), pp.696-710, 1997.

[6]. L. R. Rabiner,"A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77(2), pp. 257-286, 1989.

[7]. Y. Rui, L. He, A. Gupta, and Q. Liu, “Building an intelligent camera management system,” *Proc. ACM Multimedia*, pp.2-11, 2001.

[8]. M. Shah, and R. Jain, “Motion-based Recognition,” Kluwer-Academic Publishers, Computational Imaging and Vision Series, 1997.

[9]. Xinding Sun, Jonathan Foote, Don Kimber, B. S. Manjunath, "Recording the Region of Interest from FlyCam Panoramic Video," *Proc. ICIP*, 2001.

[10]. X. Sun, J. Foote, D. Kimber, B. S. Manjunath, "Panoramic Video Capturing and Compressed Domain Virtual Camera Control", *Proc. ACM Multimedia*, pp. 329-338, 2001.

[11]. Y. Yacoob, M.J. Black, **“**Parameterized modeling and recognition of activities,**”** in *ICCV’98*, pp.120-127, 1999.

[12]. I. Yamato, I.Ohya, and K. Ishii, “Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model,” in *CVPR’92*, pp. 379-385, 1992.