MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

14-1

Human Activity Recognition Using Sequences of Postures

Vili Kellokumpu
Machine Vision Group
P.O. Box 4500
FIN-90014 University of Oulu
kello@ee.oulu.fi

Matti Pietikäinen
Machine Vision Group
P.O. Box 4500
FIN-90014 University of Oulu
mkp@ee.oulu.fi

Janne Heikkilä
Machine Vision Group
P.O. Box 4500
FIN-90014 University of Oulu
jth@ee.oulu.fi

Abstract

This paper presents a system, which is able to recognize 15 different continuous human activities in real-time using a single stationary camera as input. The system can recognize activities such as raising or waving hand(s), sitting down and bending down. The recognition is based on describing activities as a continuous sequence of discrete postures, which are derived from affine invariant descriptors. Using affine invariant descriptors makes our system robust against such differences in camera locations as distance from the object and change in viewing direction as these differences can be considered to have the affect of near affine transformations as human silhouettes are considered.

1 Introduction

The recognition of human gestures and activities has become a research area of great interest as it has many potential application domains including human-computer interfaces, sign language interpretation and automated surveillance of parking lots and ATMs. In recent years, many approaches to human activity recognition have been presented [1, 2]. Davis and Bobick [3] used temporal templates to represent and recognize aerobics actions. Ali and Aggarwal [4] used the angles of inclination of the torso, the lower and upper parts of the legs as features to recognize human activity. Eickeler et al. [5] used global motion features, whereas Iwai et al. [6] used image flow for recognizing various gestures. Elgammal et al. [7] used exemplar poses to recognize six simple arm gestures.

There are two major types of variation in the way people perform different activities: temporal and spatial variation. Temporal variation is caused by the difference in duration of performing activities and it is usually effectively compensated by using Hidden Markov Models (HMM). Spatial variation on the other hand is due to the fact that people have different ways of performing activities and different physics. Compensating for spatial variations is more difficult and usually requires lengthy and non-optimal training.

In general, motion-based activity recognition systems are vulnerable to spatial variation caused by unintended motion. We feel that this problem can be handled efficiently with an appearance-based approach. It is an interesting idea to discard motion information and use purely posture information without any explicit human model for activity recognition. However, the usage of exemplar poses becomes impractical as the number of activities increases. Our approach is that it is neither advantageous nor practical to try to recognize a posture in too detail, and that it is more important to classify the posture in to a superclass that is semantically correct. Steps to this direction have previously been taken by Liu and Lovell [8] and Leo et. al [9] who can recognize six and four activities, respectively, using only three basic postures. In this paper we describe a novel methodology capable of recognizing a wider range of postures and activities. Our method is to use two successive classifiers, a support vector machine (SVM) and a discrete HMM, to build a system, which is robust against spatial style variations and does not need lengthy and non-optimal training of continuous HMMs.

2 System Overview

The system implementation consists of the three parts shown in Fig 1. In this work, we make the assumption of a static background, and so we can do the background subtraction by thresholding the difference between the current frame and the static background image. After the background subtraction, we have a human silhouette, and as the last phase of pre-processing, we extract the contour from the silhouette.

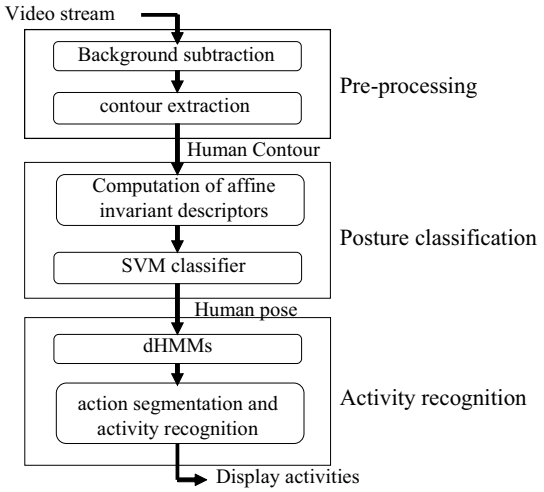


Figure 1. Illustration of the processing stages of the system.

In posture classification, we first calculate affine invariant Fourier descriptors from the contour. Then we use these descriptors as a feature vector to classify the posture with a radial basis SVM. The output of the posture classification module is thus a sequence of discrete postures.

The last part of the implementation is the activity recognition module. This module uses hidden Markov models to model different activities and calculates the probabilities of the activities based on the posture sequence from posture classification module. Action segmentation is done by using a windowing method.

3 Posture Classification

3.1 Affine invariant Fourier descriptors

Let $\mathbf{x}^0 = \{ \boldsymbol{x}(n) , n = 1,2,...,N\}$ be the reference image boundary, where $\boldsymbol{x}(n)=\{u(n),v(n)\}$ is a vector representation of the n^{th} pixel on the boundary, and similarly $\mathbf{x}$ be the contour in an observed image. Then if $\mathbf{x}$ represents the same boundary as $\mathbf{x}^0$, but has gone through an affine transformation, the relationship between $\mathbf{x}$ and $\mathbf{x}^0$ can be written as,

$$\mathbf{x} = \mathbf{A}\mathbf{x}^0 + \mathbf{b}\,. \tag{1}$$

where $\mathbf{A}$ is a 2x2 matrix that represents scaling, rotation and other linear transformations and has $|\mathbf{A}| \neq 0$, and $\mathbf{b}$ represents a 2x1-translation vector. In terms of efficient posture classification, we must design an algorithm that generates contour descriptors that are independent of all seven parameters (four elements of matrix $\mathbf{A}$, two elements of vector $\mathbf{b}$ and the starting point of the contour $\mathbf{x}$) of the transformation.

Affine invariant Fourier descriptors have been used for lip reading [10] and for recognition of aircrafts [11]. Usage of affine invariant Fourier descriptors in human posture estimation is a new approach especially to activity recognition. In our system, the extraction of affine invariant Fourier descriptors from a silhouette contour is similar to the work of Arbter et al. [11]. After Fourier transformation is applied to the contour vector $\mathbf{x}$, we get a matrix of coefficients

$$\mathbf{X} = \begin{bmatrix} \ldots & U_0 & U_1 & \ldots \\ \ldots & V_0 & V_1 & \ldots \end{bmatrix}. \tag{2}$$

Arbter et al. show that affine invariant Fourier descriptors Q_k can be calculated from these coefficients

$$Q_k = \frac{\left|\mathbf{X}_k, \mathbf{X}_p^*\right|}{\left|\mathbf{X}_p, \mathbf{X}_p^*\right|} = \left|\frac{U_k V_p^* - V_k U_p^*}{U_p V_p^* - V_p U_p^*}\right|, \tag{3}$$

where p, $k > 0$ and $\mathbf{X}_p \neq 0$. This is a complete set of invariants for $\mathbf{x}$ and $\mathbf{x}^0$ which have different starting points and satisfy Eq. 1 with arbitrary $\mathbf{A}$, $\mathbf{b}$. From this set of invariants, we use ten invariants to represent a human posture.

3.2 Support vector machine

Support vector machine is a two-class classifier that can be used to as a multiple class classifier by constructing a net consisting of two-class classifiers. Details on SVMs can be found from [12].

In our implementation we use a SVM with Gaussian radial basis kernel function to classify the invariant descriptors into a discrete posture. The idea behind this classification step is to extract the relevant information about peoples posture. This means that we are interested in describing postures with small variations into a super-posture which contains all the relevant information of the actual postures. This makes our system more robust against spatial variation in performing of activities and also makes the teaching phase of the HMMs easier.

4 Activity Recognition

Discrete Hidden Markov Models treat discrete time sequences as the output of a Markov process whose states cannot be directly observed. The basics of dHMMs are introduced next, but see [13] for more details, such as the training of the models and the forward algorithm.

A dHMM which has N states $Q = \{q_1, q_2,\ldots, q_N\}$ and M output symbols $V = \{v_1, v_2, ..., v_M\}$ is fully specified by the triplet $\lambda = \{A, B, \pi\}$. An example of a model where N = 3 and M = 2 is illustrated in Fig 2. Let the state at time step t be s_t, now the NxN state transition matrix $\mathbf{A}$ is

$$\mathbf{A} = \{a_{ij} \mid a_{ij} = P(s_{t+1} = q_j \mid s_t = q_i)\}, \tag{4}$$

the NxM state output probability matrix $\mathbf{B}$ is

$$\mathbf{B} = \{b_i(k) \mid b_i(k) = P(v_k \mid s_t = q_i)\} \tag{5}$$

and the initial state distribution vector π is

$$\pi = \{\pi_i \mid \pi_i = P(s_1 = q_i)\} \tag{6}$$

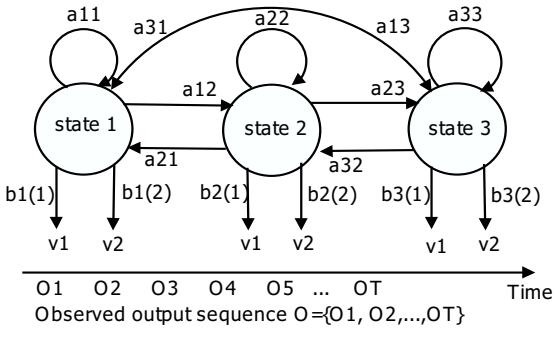


Figure 2. An example of a fully connected three-state, two-output dHMM.

The idea behind using the dHMMs is to construct a model for each of the activities that we want to track. DHMMs give a state based representation for each activity. For example, activity 'raise both hands' could be represented with a three-state model where the position of the hands (down, between shoulders, and above the head) determine the states and the classified posture is the one-dimensional output symbol.

After having the models for each activity, we take a posture sequence $O = \{O_1,O_2,\ldots,O_T\}$ from the SVM classifier and calculate the probability $P(O|\lambda)$, the probability of a model λ for the observation sequence, for every model. The probability $P(O|\lambda)$ can be solved by using the forward algorithm. We can then recognize the activity as being the one, which is represented by the most probable model.

In real situations, the transitions between activities make activity recognition much more complicated and determining the correct boundaries for the observation sequence in the calculation of model probabilities is difficult. Our approach to temporal segmentation is just simply to use multiple windows, with lengths ranging from τ_{min} to τ_{max}, which are sliding through the observation sequence. Activity segmentation can be done by thresholding the difference between two most probable models in every window. An activity is detected if any activity model rises above a threshold in any window.

5 Experiments

To test the performance of our approach we implemented a real time system capable of recognizing 15 basic activities and recorded a test set of five sequences containing continuous actions by five different persons. Each person made about 20 activities in varying orders and without any intentional pauses. Examples of the activities are presented in Fig 3. The viewing direction was the same in all test sequences as well as in training. Example views seen by the camera are shown in Fig 4.

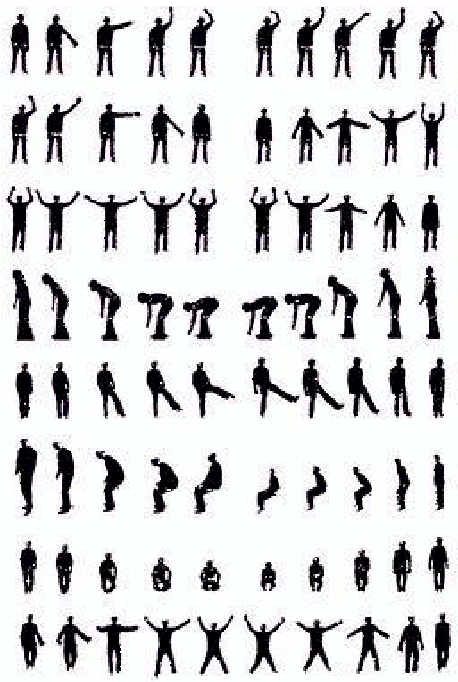
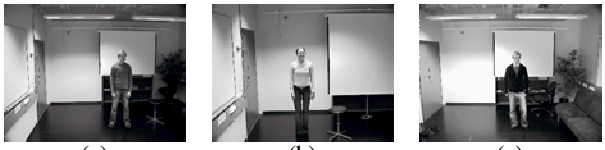


Figure 3. Examples of the 15 activities. Beginning from the top row: Raise one hand, wave one hand, lower one hand, raise both hands, wave both hands, lower both hands, bend down, get up, raise foot, lower foot, sit down, stand up, squat, rise from squat, and finally the last row contains x-hopping.



(a) (b) (c)

Figure 4. Examples of the view seen by the camera. (a) and (b) are views from the test set and (c) is one posture used in training.

The system was trained using only one person for both the posture classification module and the activity recognition module. The training sequences are not part of the test set though the person used on the training phase also appears in the test set. Table 1 gives the results on the test set.

Table 1. Confusion matrix for the experiment. Rows and columns represent the recognized activities and the ground truth, respectively.

	Raise one hand	Wave one hand	Lower one hand	Raise both hands	Wave both hands	Lower both hands	Bend down	Get up	Raise foot	Lower foot	Sit down	Stand up	Squat	Raise from squat	X-hopping	No activity
Raise one hand	6												1			
Wave one hand		6								1						1
Lower one hand		1	6													
Rise both hands				6											1	
Wave both hands					5											1
Lower both hands						6										1
Bend down							4							1		
Get up								5								
Raise foot									8							1
Lower foot										8						2
Sit down											5	1				
Stand up												5				
Squat													3			2
Raise from squat													1	2		2
X-hopping	1														16	
No detection							1		1	1			2	2		
Late detection														1		
Total	7	7	6	6	5	6	5	5	9	9	5	6	6	6	17	11

The total number of activities in the sequences was 101, the number of detections was 110, and the number of correct recognitions was 91, giving recognition rate of 90% and detection accuracy of 83%. The column sums in Table 1 exceed the total number of activities because at times, the temporal segmentation fails to work ideally and we get multiple detections (both correct and incorrect) from a single activity.

Even though the test set is rather small it proves that our methodology works robustly. The nature of the test is to show that even though the system is trained using only one person, it can still correctly classify activities by people with different body builds, as the tallest test subject was 185 cm tall the shortest was 160 cm. For the person that was used both in training and tests, the system gave only one false alarm and did not miss any activity, indicating that if more people were used in training, the recognition results would be even better.

The system is trained using only a single view and we have presented tests using that same view, but the system also tolerates some change in the viewing direction. Activities are still recognized even with the change of ±45

degrees in the viewing direction though the recognition rate slowly starts to fall as the angle increases.

Currently the implementation has a few restrictions. As mentioned earlier, the viewing direction is somewhat fixed and the background is assumed to be static making the segmentation of the silhouette easy. In addition, we assume that there is only one person in the field of view and that there is no occlusion.

6 Conclusions

We have presented a new robust method for real-time human activity recognition using only posture information, and we have implemented a system that can recognize gestures with 90% accuracy. The recognition accuracy could still be improved with intensive training. However, the presented tests show clearly that the system is effective against spatial variation although only one person was used for training. Using superpostures also helps to avoid ambiguities caused by spatial variation and therefore diminishes the variance in observation probabilities for the dHMMs making the action segmentation easier. When compared to continuous HMMs, the states and observation vectors of the dHMMs in our approach have a clear meaning. This is a very useful property when tracking down reasons behind misclassifications and it makes iterative upgrading of the system easier.

The system implementation is really fast as the computation time for processing one frame on an AMD Athlon 2200+ processor is less than 10 ms. This makes it possible to further develop the system in the future while still maintaining operation at video rate.

Acknowledgement

This research was sponsored by the Academy of Finland as a part of the PROACT program

References

[1] D.M. Gavrila: "The Visual Analysis of Human Movement: A survey". *Computer Vision and Image Understanding,* vol. 73, no.1, pp. 82–98, 1999.

[2] J.K. Aggarwal and Q. Cai: "Human Motion Analysis: a Review". *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428 – 440, 1999.

[3] J. W. Davis, A. F. Bobick: "The Recognition of Human Movement Using Temporal Templates". *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23 no.3, pp. 257 – 267, 2001.

[4] A. Ali, J. K. Aggarwal: "Segmentation and Recognition of Continuous Human Activity". In *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 28 – 35, 2001.

[5] S. Eickeler, A. Kosmala, G. Rigoll: "Hidden Markov Model Based Continuous Online Gesture Recognition". In: *International Conference on Pattern Recognition* (ICPR), Brisbane, vol. 2, pp.1206 – 1208, 1998.

[6] Y. Iwai, H Shimizu, M Yachida: "Real-Time Context-based Gesture Recognition using HMM and Automaton" . In *IEEE Proceedings for International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, Kerkyra, Greece, pp. 127 – 134, 1999.

[7] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis: "Learning Dynamics for Exemplar-based Gesture Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. 571 – 578, 2003.

[8] N. Liu, B.C. Lovell: "Gesture Classification Using Hidden Markov Models and Viterbi Path Counting". In *Proceedings Digital Image Computing: Techniques and Applications*, Sydney, pp 273 – 282, 2003.

[9] M. Leo, T. D'Oranzio, I. Gnoni, P. Spagnolo, A. Distante: "Complex Human Activity Recognition for Monitoring Wide Outdoor Environments". In: *International Conference on Pattern Recognition* (ICPR), vol. 4, pp. 913 – 916, 2004.

[10] S. Gurbuz, Z. Tufekci, E.K. Patterson, J.N. Gowdy: *"*Application of Affine-Invariant Fourier Descriptors to Lipreading for Audio-Visual Speech Recognition". In: *International Conference on Acoustics, Speech, and Signal Processing,* Salt Lake City, pp.177 – 180, 2001.

[11] K. Arbter, E. E. Snyder, H. Burkhardt, G. Hirzinger: "Application of Affine Invariant Fourier Descriptors to Recognition of 3-D Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 640 – 647, 1990.

[12] C. Burges: "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery Journal*, vol. 2, no. 2, pp. 121 – 167, 1998

[13] L.R. Rabiner: "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 – 285, 1989.