HMM-Based Action Recognition Using Contour Histograms

M. Ángeles Mendoza and Nicolás Pérez de la Blanca

University of Granada
Department of Computer Science and Artificial Intelligence
{nines,nicolas}@decsai.ugr.es

Abstract. This paper describes an experimental study about a robust contour feature (*shape-context*) for using in action recognition based on continuous hidden Markov models (HMM). We ran different experimental setting using the KTH's database of actions. The image contours are extracted using a standard algorithm. The *shape-context* feature vector is build from of histogram of a set of non-overlapping regions in the image. We show that the combined use of HMM and this feature gives equivalent o better results, in term of action detection, that current approaches in the literature.

1 Introduction

Automatic action recognition is a constantly expanding research area due to the number of applications for surveillance (behaviour analysis), security (pedestrian detection), control (human-computer interfaces), content-based video retrieval, etc. It is, however, a complex and difficult-to-resolve problem because of the enormous differences that exist between individuals, both in the way they move and their physical appearance and the environment where the action is carried out [5]. An increasingly popular approach to this problem is HMM. One action is sequences of events ordered in space and time, and HMM exploits this ordering to capture structural and transitional features and therefore the dynamic of the system. The contour of the subject, and more explicitly how the contour shape succeeds another in the action cycle is a powerful discriminating signal for HMM-based action recognition (available even though there is no texture or colour information) and may be rapidly extracted with simple techniques. This information, however, can be very noisy in cluttered environments and when there are shadows. Some authors choose to use the full silhouette to increase robustness, Kale et al. [4] identify people by their gait based on the width of binary silhouettes; Sundaresa et al [8] using the sum of silhouette pixels; Yamato [9] also uses binary silhouettes to recognize 6 different tennis strokes with discrete HMM using the ratio of black pixels; Ahmad and Lee [1] combine shape information (applying PCA to the silhouettes to reduce dimensionality) and the optical flow in a discrete HMM in multiple views. But by doing so, they are introducing redundancy (the inside of the pedestrian does not give

J. Martí et al. (Eds.): IbPRIA 2007, Part I, LNCS 4477, pp. 394–401, 2007.

additional information), extraction is more expensive, morphological pre-processing is necessary, and in addition, silhouettes are too affected by background and occlusions.

Belongie and Malik [2] introduced *shape contexts* as shape descriptors, and these are basically log-polar histograms of distances and angles of a point to the neighbouring points of the uniformly sampled contour. Since these are based on histograms, they are reasonably tolerant to variations in pose and occlusion. In this paper, we will focus on *shape contexts* to define a feature that extracts contour information for HMM-based action recognition. This feature is robust to occlusions or bad background segmentation, cluttered environments and shadows, and exploits the contours' good qualities (i.e. lower sensibility to lighting changes and to small viewpoint variations than methods based on optical flow or pixel brightness, similarity at different scales, ease and speed of extraction with simple techniques).

The rest of this paper is organized as follows: Section 2 describes the coding of the contour information in our approach. Section gives an overview of our experimentation and results. Finally, conclusions are drawn in Section 3.

2 Feature Vector

Our features use *shape contexts* [2] to describe the shape of the contour of the subjects performing the actions. We don't calculate the *shape contexts* relative to subject's whole shape but we divide the region where the person is into uniform-shaped tiles, and the *shape contexts* will describe the contour that appears in each tile independently. In other words, for each point sampled in a tile, the log-polar histogram of this point's relative positions in relation to the other points in the same tile is calculated. This is because we want to add position information, although contour histograms encode local shape robustly, they do not preserve spatial information about the image, which would increases the discriminant power of feature vector, for example between walking and waving. When an action is performed, movement is mainly carried out by some of the human body's limbs. Movement is going to be characterized by movement of theses specific limbs. In order to introduce this information, the person's area is divided into four vertical sections corresponding approximately to the area of the head-shoulders, trunk-arms, hips-thighs and knees-feet, which enable a wide range of movements to be covered, and two horizontal sections to exploit human symmetry. If smaller tiles are used, it is not possible to capture the characteristic contours of the human figure. This coding is illustrated in Figure 1.

Once the subject's rectangle is extracted, we divided this rectangle in meaningful tiles (see before paragraph). In each tile, we sampled uniformly a set of points on the contour, for every point a log-polar histogram of relative positions and angles of that point to the remaining point is calculated, *shape context*'s mean histogram in each tile generates a feature vector which will be the HMM input The use of log gives to *shape contexts* greater sensibility to nearby points than to far points, more probably affected by noise. Invariance to uniform scale is achieved by normalizing all the radial distances by the mean distance between all the contour-point pairs. We referred the angles to an absolute axis (positive x-axis) because invariance to rotation is not

desirable for our problem, but to exploit this information. In order to minimize redundancy and to compress the data, we compare two widely extended techniques; principal component analysis (PCA) and the discrete cosine transform (DCT). We apply PCA on the eight tile's median histograms and obtain eight eigenvectors of dimension 8, in total 64 coefficients (8x8). With DCT, we calculate the 2D discrete cosine transform of median histogram in each tile and took the first eight coefficients, for obtaining a feature vector of 64 coefficients too (8 coefficients DCT x 8 tiles), and to compare PCA and DCT performance, the number of required training samples is increased with feature vector's size.

We employ continuous HMM (CHMM) with mixed Gaussian output probability, simple left to right topology, where a state can only be reached from itself or from the previous state. Since certain actions have a common pose, since allowing higher jumps between states could result in recognition errors.

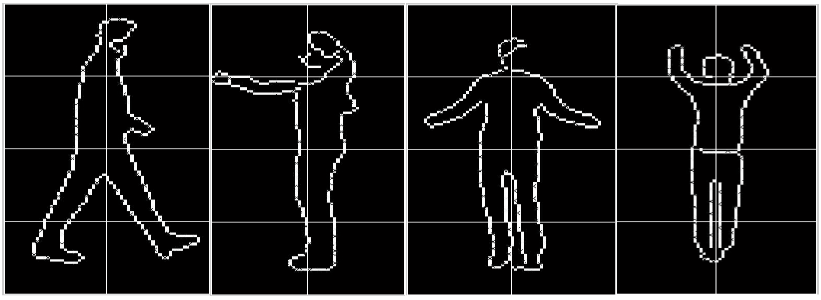

Fig. 1. The eight independent regions in which we split each image. From left to right, walking to the right, boxing to the left, clapping, waving.

3 Experimentation and Results

3.1 Data

We have evaluated our proposal using KTH's database [7] on 10 different types of actions: walking to the right, walking to the left, jogging to the right, jogging to the left, running to the right, running to the left, boxing to the right, boxing to the left, clapping, and waving. These actions were performed outdoors by 25 different people of both sexes with different lighting conditions and shadows. The sequences are made up of grey images with a resolution of 160x120 pixels to 25fps. Figure 2 shows the original images from the database and their corresponding outline images. The outline images are images, which would be obtained, in realistic conditions affected by shadows, people's own clothes, background elements, and bad segmentation. In some sequences of boxing to right there is a continuous zoom during the recording, therefore the subject appears with different size.

Each video was segmented into activity cycles, whereby we defined the boxing cycle as punches with both fists, clapping as the movements corresponding to a single clap, waving to raising the arms above the head and returning to the initial position, the walking and jogging cycle as a step with both legs (as most pedestrian recognition

approaches do and so as to increase robustness), but the running cycle only considers one step with either leg since this recorded the most sequences because of the pedestrian's speed. We used a total of 44 samples for walking to the right, 46 for walking to the left, 34 for jogging to the right, 38 for jogging to the left, 51 for running to the right, 54 running to the left, 104 for boxing to the right, 113 for boxing to the left, 164 for clapping, and 119 for waving.

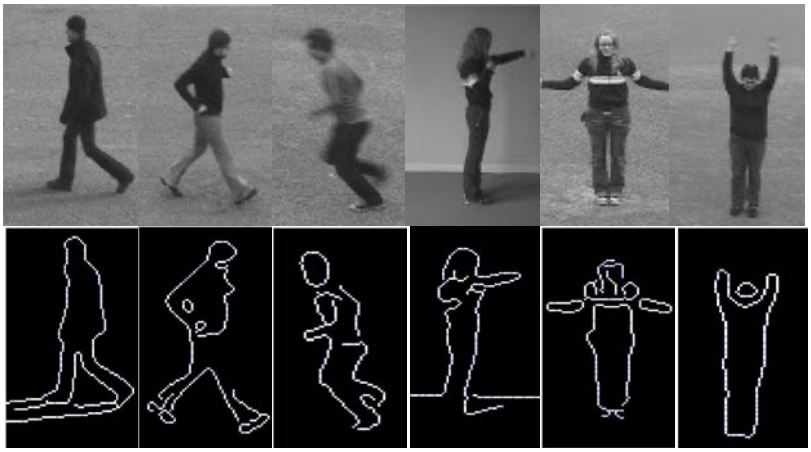

Fig. 2. Top row, original images from KTH's database; bottom row, outline images for the activities (from left to right): walking to the left, jogging to the right, running to the left, boxing to the right, clapping and waving.

3.2 Experimentation

Once the frames representing an activity cycle have been obtained, we extracted a fixed size feature vector for each frame. This vector represents the HMM set of observable. In order to calculate the feature vector proposed in this paper, we first divided the area containing the person performing the action into eight identical tiles, 2 horizontal ones and 4 vertical ones, and we consider 8 main areas of movement: knees-feet, hips-thighs, trunk-arms, shoulders-head (see Section 2.2). In each tile, the shape contexts were calculated. In order to choose the number of sampled contour points and the number of log-polar histogram bins, it is necessary to balance the information we wish to capture with the size of the feature vector, bigger size more training samples are necessary. We have taken 10 contour points in each tile, 4 distance bins (log bin size (d/4), where the diagonal pixels of the tile are represented by d) and 8 angle bins (45 degrees). Finally, we assessed the median histogram of shape contexts in each tile to obtain a 256 D vector (8 tiles x 4 distance bins x 8 angle bins). We referred the angles to an absolute axis (positive x-axis) because invariance to rotation is not desirable but exploitation of this information. In order to reduce the size of the vector to 64 coefficients, we employed PCA or DCT.

The vectors extracted in this way constitute the input to the hidden Markov model toolkit (HTK), which is an integrated environment of software tools for building

HMM, and this facilitates HMM learning for each activity and subsequent recognition. HTK consists of a set of libraries written in ANSI C and a detailed description of its operation can be found in [10].

3.3 Results and Discussion

The proposed coding is validated on two multi-subject open tests, one with PCA and the other with DCT. The samples are divided into three groups, two of which are used for training and the other for testing by cross validation (about 250 samples for testing and 500 for training). For each action, a simple left-right CHMM with mixed Gaussian output probability is learned. New components are incrementally added to the output probability at each stage, one at a time, until 20 Gaussians. HMM were tested with 8, 10 and 12 states since the number of states cannot be higher than the number of frames for the action cycle and for fewer than 8 states, the recognition rate was lower than 60%. Table 1 shows the average accuracy of HMM with 8, 10 and 12 states for PCA and DCT.

Table 1. Accuracy with PCA and DCT: the number of output probability Gaussians with the greatest average accuracy is shown in brackets

	No. states		
	8	**10**	**12**
PCA	73.19 (11 Gauss)	81.04 (9 Gauss)	84.06 (4 Gauss)
DCT	88.47 (13 Gauss)	92.35 (8 Gauss)	93.11 (4 Gauss)

The best average accuracy was obtained for an HMM with 12 emitting states and a mixed output probability of 4 Gaussians, 84.06% with PCA and 93.11% with DCT (about 10% higher than PCA). In the literature, the recognition rate of HMM-based actions which use human shape-based features, for example, is 81.3% when only shape features are used and 87.5% when shape and optical flow are combined in [1] for four actions (walking, running, bowing, raising the hand). In [9] the recognition rate is 96% for 6 tennis strokes where training and testing is performed for a single person. Table 2 shows the confusion matrix for the best recognition rate which is obtained with a CHMM of 12 states and an output probability of 4 Gaussians for DCT, where H is the number of correct actions, D the ignored actions (deletion errors), I the erroneously entered actions between two activity cycles (insertion errors), S the wrongly recognized actions (substitution errors), and N the total number of samples, Accuracy=(H-I)/N.

As the actions of walking, jogging and running have similar characteristic and jogging is between the other two, they are often confused. Consequently, walking and running are sometimes confused with jogging, and jogging can be confused with either of the others (as the confusion matrix shows). It is worth mentioning that the two walking actions (to the left and to the right), which are erroneously classified as jogging, correspond to the same person (the 6th person in KTH's database). At times, as the person is centred in the rectangular area, the predominant movement in either jogging or running is the arm movement, and this can be misinterpreted by the system which wrongly classifies it as boxing. Nevertheless, it can be seen that the confusion

matrix obtained using our coding is almost diagonal and gives a greater error rate for those actions with a smaller number of samples. We would therefore expect that an increase in the number of samples would be accompanied by an increase in the recognition rate for these actions. Finally, we should mention that

Table 2. Confusion matrix for the best case: for a 12-state HMM and an output probability of 4 Gaussians for DCT

WORD: %Corr=95.59, Acc=94.71 [H=217, D=0, S=10, I=2, N=227]											
	WALKING_L	WALKING_R	JOGGING_L	JOGGING_R	RUNNING_L	RUNNING_R	BOXING_L	BOXING_R	CLAPPING	WAVING	Del [%c / %e]
WALKING_L	13	0	1	0	0	0	0	0	0	0	0[92.9/0.4]
WALKING_R	0	13	0	1	0	0	0	0	0	0	0[92.9/0.4]
JOGGING_L	1	0	9	0	1	0	0	0	0	0	0[81.8/0.9]
JOGGING_R	0	0	0	10	0	0	0	1	0	0	0[90.9/0.4]
RUNNING_L	0	0	0	0	9	0	0	0	0	0	0
RUNNING_R	0	0	0	2	0	6	1	0	0	0	0[66.7/1.3]
BOXING_L	0	0	0	0	0	0	33	0	0	0	0
BOXING_R	0	0	0	0	0	0	0	34	0	0	0
CLAPPING	0	0	0	0	0	0	0	0	54	0	0
WAVING	0	0	0	0	0	0	0	0	2	36	0 [94.7/0.9]
Insertions	0	0	0	0	1	0	0	1	0	0	

misclassification could result in two insertions appearing in the matrix (e.g. boxing to the right and running to the left) and in a file with only one activity cycle, the system considers there to be two cycles. In the case of running, for example, this is due to a jogging cycle having been wrongly classified as running, and since a jogging cycle consists of two steps and a running cycle of one, the system considers the second jogging step to be another running cycle, and it therefore appears to have two cycles.

If we examine the outline images, we can see that our recognition system is capable of handling noisy images which are affected by shadows, occlusion and bad background segmentation. In addition, thanks to the locality introduced in our proposal, when we divide the region where the person performs the action into meaningful movement areas, it is possible to distinguish between one direction and another.

In order to study the independence to people, we ran a subject independent test for a system with 12-state HMM for DCT, where the testing's subjects are all different to the training's ones. This was, 11 subjects for training and 3 for testing in walking, 15 for training and 4 for testing in jogging, 18 for training and 4 for testing in running, 9 for training and 2 for testing in clapping, 7 for training and 2 for testing in waving, 6 for training and 2 for testing in boxing to the left, 4 for training and 1 for testing in boxing to the right; with from 2 to 8 sequences by each subject for walking, jogging and running, and from 8 to 15 sequences by subject for clapping, waving and boxing. The accuracy obtained was 81.29% in a 12-state HMM for DCT. The confusion matrix is shown in the table 3.

Table 3. Confusion matrix for the subject independent test with a 12-state HMM and an output probability of 1 Gaussians for DCT

WORD: %Corr=81.29, Acc=81.29 [H=113, D=0, S=26, I=0, N=139]											
	WALK IL	WALK IR	JOG IL	JOG IR	RUN IL	RUN IR	BOX IL	BOX IR	CLAP	WAV IN	Del [%c / %e]
WALKING_L	8	0	0	0	0	0	0	0	0	0	0
WALKING_R	0	6	0	2	0	0	0	0	0	0	0[75.0/1.4]
JOGGING_L	0	0	8	0	0	0	0	0	0	0	0
JOGGING_R	0	0	0	7	0	0	0	0	0	0	0
RUNNING_L	0	0	0	0	9	0	0	0	0	0	0
RUNNING_R	0	0	0	0	0	9	0	0	0	0	0
BOXING_L	0	0	0	0	0	0	26	0	1	0	0[96.3/0.7]
BOXING_R	0	0	0	0	0	0	0	8	0	0	0
CLAPPING	0	0	0	0	0	1	0	2	20	7	0[66.7/7.2]
WAVING	0	0	0	0	0	0	2	0	11	12	0[48.0/9.4]
Insertions	0	0	0	0	0	0	0	0	0	0	

How can see the matrix is almost diagonal. The confusion is mainly due to waving that is erroneously classified as clapping. But it's worth noting that we only used 7 different subjects for training, too few subjects for building subject independent systems. Even so we obtained promising results, for example, it's obtained 100% of accuracy for boxing to the right, with only 4 training's subjects.

4 Conclusions

Our coding is applied on a rectangular box tracking the subject. How can see in Figure 2, there are body's parts don't detected, shadows and background's noise, by its definition our coding is more robust than other ones based on the subject's exact silhouette. The results show that the use of contour features for HMM-based action recognition can achieve a recognition system capable of operating with low-level features which are easily and rapidly extractable directly from the image, and which can be obtained in realistic training and testing conditions without many *a priori* assumptions.

The proposal coding for incorporating contour information enables us to border the drawbacks suffered by other contour-based coding. (occlusions, shadows, noises resulting from cluttered environments and the subject's own clothes). Allowing us to exploit its advantages: greater robustness to lighting changes and to slight variations in viewpoint than other approaches (such as those based on optical flow or pixel brightness), similarity at different scales (the same approach can therefore be used for different depths in the field of vision), and its speed of extraction.

The speed of feature extraction is also an important issue, since the final aim for most action recognition applications is real-time operation. The recognition speed in HMM-based systems will depend on the number of HMM states, currently there are studies which accelerate the recognition time of the HMM-based system by hardware

[3]. Using features with a computationally efficient calculation and which do not require much pre-processing (as in the case of the contours) will help us achieve our objective.

One possible improvement to our coding would be to use overlapped tiles of sizes (rather than uniformly sized tiles) to divide the area where the subject performs the action using our knowledge of the human body.

Acknowledgments. This research was supported by the Spanish Ministry of Education and Science (TIN2005-01665).

References

1. Ahmad, M., Lee, S.: Human Action Recognition Using Multi-View Image Sequence Features. FGR, pp. 10–12 (2006)
2. Belongie, S., Malik, J.: Matching with Shape Contexts. CBAIVL (2000)
3. Fahmy, A., Cheung, P., Luk, W.: Hardware Acceleration of Hidden Markov Model Decoding for Person Detection. In: Proc. of the Design, Automation and Test in Europe Conference and Exhibition (2005)
4. Kale, A., Rajagopalan, A.N., Cuntoor, N., Krueger, V.: Gait-Based Recognition of Humans Using Continuous HMMS. FGR, pp. 336–341 (2002)
5. Moeslund, T.B., Hilton, A., Krüger, V.: A Survey of Advances in Vision-Based Human Motion Capture and Analysis. CVIU 104, 90–126 (2006)
6. Rabiner, L.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE, vol. 2, pp. 257–286 (1989)
7. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: a Local SVM Approach. ICPR III, 32–36 (2004)
8. Sundaresan, A., RoyChowdhury, A., Chellappa, R.: A Hidden Markov Model based Framework for Recognition of Humans from Gait Sequences. ICIP, pp. 93–96 (2003)
9. Yamato, J., Ohya, J., Ishii, K.: Recognizing Human Action in Time-Sequential Images Using a Hidden Markov Model. CVPR, pp. 379—385 (1992)
10. Young, S.J.: The HTK Hidden Markov Model Toolkit: Design and Philosophy. In: Technical Report CUED/F-INFENG/TR.152, Department of Engineering, Cambridge University, Cambridge (1994)