Human Activity Recognition in Archaeological Sites by Hidden Markov Models

Marco Leo, Paolo Spagnolo, Tiziana D'Orazio, and Arcangelo Distante

Institute of Intelligent Systems for Automation - C.N.R.
Via Amendola 122/D, 70126 Bari, ITALY
{leo,spagnolo,dorazio,distante}@ba.issia.cnr.it

Abstract. This work deals with the automatic recognition of human activities embedded in video sequences acquired in an archeological site. The recognition process is performed in two steps: first of all the body posture of segmented human blobs is estimated frame by frame and then, for each activity to be recognized, a temporal model of the detected postures is generated by Discrete Hidden Markov Models. The system has been tested on image sequences acquired in a real archaeological site meanwhile actors perform both legal and illegal actions. Four kinds of activities have been automatically classified with high percentage of correct decisions. Time performance tests are very encouraging for using the proposed method in real time applications.

1 Introduction

Automatic recognition of human activities is one of the most important and interesting open area in computer vision. Automatic visual surveillance, multimodal interfaces and automatic indexing of multimedia data are some of the most common and relevant applications of this research field.

In this paper we focus on the automatic surveillance of archeological sites. Monitoring archaeological sites is becoming a crucial problem in order to preserve buried and unburied property from thefts and vandalic actions. Nowadays archeological sites are monitored by using passive systems based on a set of large view cameras sending the acquired streams to an headquarter where one or more people, looking at the monitors, have to detect suspicious behaviours.

A large portion of open literature is devoted to human activity recognition in limited know spaces where the subjects dominate the image frame so that the individual body components (head, hands, etc.) can be reliably detected. Detailed reviews of these works can be found in [6,7]. Few works dealt, instead, with the problem of human activity recognition in large areas. CMU's Video Surveillance and Monitoring (VSAM) project [1], MIT AI Lab's Forest of Sensors Project [2] and VIGILANT project [5] are three of the most appreciate examples of recent research efforts in this field. In [2], the patterns (cars and humans) and their activities are learned by motion analysis. In [1], measurements based on a simple skeleton of the target are used to distinguish running people from walking

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3332, pp. 1019–1026, 2004.

ones. In [5] velocity and width-to-height ratio of the patterns (car and human) are supplied as input to an HMM procedure.

Other considerable works in this area, like Pfinder [3] and W4 [4], try to classify humans and their activities by detecting features such as hands, feet and head, tracking and fitting them to an a prior human model.

The analysis of the related works reveals that these algorithms for large area monitoring can recognize very simple activities like vehicle and person entering and exiting form a parking area, people running or walking and so on. The automatic recognition of these simple actions could not be adequate to meet the requirements of the automatic surveillance of an archaeological site.

In this paper we propose a new approach for human activity recognition that works on binary patches extracted from the images containing human blobs. At first the horizontal and vertical histograms of human blobs are computed and supplied as input to an unsupervised clustering algorithm in order to detect the human posture in each frame. Then a statistical approach based on Discrete Hidden Markov Models is applied to temporal modelling the sequence of detected postures and to discriminate between legal and illegal activities. The last point that has been addressed in this work concerns the ability of the method to recognize in a long test sequence the beginning of the known activities. We have used a sliding window that has been overlapped to the test sequence to extract a fixed length observation sequence provided to the behavior classification step. The proposed approach has been validated using 165 long test sequences acquired in a real archeological site.

In the rest of the paper, first a description of the proposed activity recognition approach is explained and then the experimental results obtained on image sequences acquired in a real archaeological site meanwhile actors perform both legal and illegal actions are reported.

2 Human Activity Recognition

The human activity recognition system proposed in this paper works on the binary patches containing the human blob. For this reason, a preliminary people segmentation algorithm is required. Since the description of this algorithm is beyond the scope of this paper, we refer to the significant work proposed in the last years [9,12,13].

The behavior classification algorithm executes two steps: first of all the human body postures have to be estimated in each frame and then the temporal sequence of detected postures has to be modelled by discrete HMMs. In the pose estimation step horizontal and vertical histograms of the binary shapes are evaluated and supplied as input to an unsupervised clustering algorithm named BCLS (Basic Competitive Learning Scheme) [11]. In this work the proximity measure among two postures Im_1 and Im_2 is calculated as follows:

$$D(Im_1, Im_2) = d_1(X_1, X_2) + d_2(Y_1, Y_2) \tag{1}$$

where d_1 and d_2 are the Manhattan distances between the horizontal and vertical projections respectively. In particular a modified version of the Manhattan distance has been implemented; it was defined as:

$$d_2(Y1, Y2) = min \begin{pmatrix} \sum_{j=0}^{DimY-1} |Y1(j) - Y2(j+1)| \\ \sum_{j=0}^{DimY-1} |Y2(j) - Y1(j+1)| \end{pmatrix} \tag{2}$$

$$d_1(X1, X2) = min \begin{pmatrix} \sum_{j=0}^{DimX-1} |X1(j) - X2(DimX1 - j - i)| \\ \sum_{j=0}^{DimX-1} |X2(j) - X1(DimX1 - j - i)| \\ \sum_{j=0}^{DimX-1} |X1(j) - X2(j+i)| \\ \sum_{j=0}^{DimX-1} |X2(j) - X1(j+i)| \end{pmatrix} \tag{3}$$

where the minimum is evaluated when i changes respectively in the interval [0, DimY-1] and [0, DimX-1].

In this new definition the vertical and horizontal histograms of an image are compared, by the proximity measure, with all the translated (and mirrored for the horizontal) versions (on the left and on the right) of the same histograms of another one. The minimum values are taken as the proximity measure.

In this way the proximity measure becomes invariant to the translation and mirroring of the binary target in the scene. Using the proposed proximity measure, the BCLS algorithm groups the available training images and then it classifies unknown new images on the base of their relative distances with respect to the built prototypes.

The recognition of human behavior is then performed by fully connected HMM in order to statistically analyze the temporal sequence of detected postures. In this step the number of different postures determines the number of the HMM codebook symbols (i.e the possible state values M) and each activity is associated to an HMM: this means that the number of HMM is always equal to the number of different activities of interest. Otherwise, the number of states N is fixed experimentally.

In the training phase the parameters of each HMM are updated in order to maximize the output probability of the training sequences. The training procedure based on the multiple observation sequence proposed in [10] has been used. This training solution has been adopted considering that different people perform the same activity in different ways. The algorithm proposed in [10] expresses the multiple observation probability as a combination of individual observation probabilities. In particular we have implemented a generalizing Baum's auxiliary function and we have built an associated objective function using Lagrange multiplier method. For each different activity an HMM model λ_i has been generated. In the test phase unknown sequences are provided as input to the HMMs. The probability to have the activity A given the observation sequence X of postures is computed by evaluating the forward backward probability. A decision criterion based both on maximum likelihood measure:

$$A^* = argmax P(X|\lambda_i) \tag{4}$$

and a set of proper thresholds to manage unknown behavior has been introduced. Indeed each HMM has associated a threshold equal to the minimum probability value obtained during the training phase. The sequence X of posture observations is labeled as activity A if both its corresponding HMM gives the maximum likelihood measure among the whole set of HMMs and at the same time this probability value is greater than the relative HMM threshold. If this second condition is not satisfied the observation X cannot be associated to any of the known activities and is labeled as unknown.

The length of each observation sequence supplied to the HMMs is fixed in both training and testing phases and it has to be experimentally evaluated. In the training phase the observation sequences are segmented by hand whereas in the testing phase a sliding window (of the same length of training sequences) is used to cover the whole acquisition sequence.

3 Experimental Results

The proposed human activity recognition approach has been tested on real sequences acquired in an archaeological site. The images were acquired with a static TV camera Dalsa CA-D6. In order to consider only significant frames for the activity recognition process we have sampled the acquisition sequence tacking two frames per second. The software was implemented by using Visual C++ on a Pentium III 1 Ghz and 128 Mb of RAM. The archaeological site considered is a wide country area where some legal or illegal activities need to be discerned. In particular illegal activities are executed by people that first probe the subsoil using simple tools (such as sticks, tanks) and then they excavate to dig up some attracting objects. The people segmentation algorithms produces for each person in the scene a binary patch of 175x75 pixels. Starting from these patches, the BCLS algorithm detects three kinds of different postures: “standing”, “squatted” and “bent”. One example of each detected posture can be found in figure 1. Sequences composed by a temporal succession of these three postures are supplied as input to the HMMs in order to identify 4 kinds of activities:

1. Walking
2. Probing the subsoil by a stick
3. Damping the ground with a tank
4. Picking-up some objects from the ground.

The first activity, the simpler one, is legal; while the remaining ones are more complicated and illegal. The figure 2 shows some frames for each of the possible sequences of the different activities. In particular it can be note that the second and the third activities are very similar: they are composed by sequences of the same two postures, but with different temporal variations. The statistical modelling step is then composed by 4 HMMs. Each HMM is associated with a different kind of activity and it is trained with three different examples (performed by different people) of the associated activity. The training set, composed by 4x3=12 sequences is not changed during all the experiments described

Fig. 1. Three fundamental postures classified in the archaeological site.

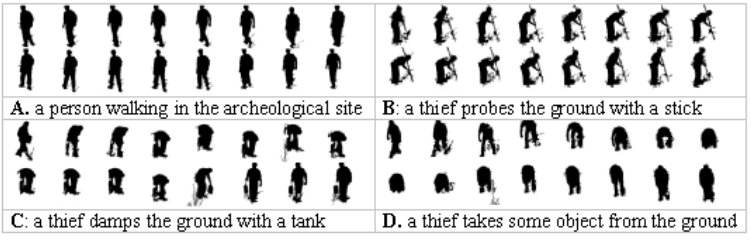

Fig. 2. Some frames extracted from 4 of the 12 sequences used to train the HMMs.

below. Each training sequence consists of 50 frames (so 50 is also the length of the sliding window used in the test phase). The experimental tests have shown that a greater number of training sequences decreases the generalization ability of the HMM, as asserted in [10].

In the first experiment, the system has been tested using 160 sequences. Each sequence contains one of the 4 activities to be recognized (just 40 for each kind of activity), but the beginning and the ending frames are not known. The length of the input sequence ranges from 400 to 1500 frames. If N_{TOT} is the total number of frame in each test sequence and n is the length of the sliding window then $N_W = N_{TOT} - (n - 1)$ is the number of windowed observation sequences O_w supplied as input to the HMMs for each test sequence.

An activity is recognized in a test sequence when at least one of its observation sequence O_w extracted by the sliding window, satisfies the recognition procedure described in the previous section (bayesan criterion + adaptive threshold).

In table 1 HMMs with 2-3-4-5-6-7-10 and 12 states have been tested in order to determine the optimal number N of HMM states in our application domain.

Each test sequence is given as input to the four HMMs with the same number of hidden states. For this reason, the results of every row of the table have to be considered altogether. The last column shows the mean percentage of correct classification; it sums up the best classification results obtained with 4 and 6 hidden states. In this case the percentage of right classification is 84.37%. HMM with a larger number of states have not been considered because the HMM's theory [8] suggests the use of a number of states much smaller than the number of symbols in each observation sequence (50 in our case). In the second experiment, in order to further improve the classification results, the four HMMs with the best classification performances have been selected and tested on the same 160

Table 1. The activity recognition results when the number of HMM states changes

HMM States	Activity								
	Walking People		Probing People		Damping People		Picking-up People		% of correct classification
2	**40/40**	**100%**	**40/40**	**100%**	0/40	0.0%	**40/40**	**100%**	75
3	40/40	100%	25/40	62.5%	29/40	72.5%	40/40	100%	83.75
4	40/40	100%	25/40	62.5%	30/40	75%	40/40	100%	84.375
5	40/40	100%	28/40	70%	26/40	65%	25/40	62.5%	74.375
6	40/40	100%	27/40	67.5%	28/40	70%	40/40	100%	84.375
7	40/40	100%	27/40	67.5%	26/40	65%	40/40	100%	83.125
10	40/40	100%	20/40	50%	**31/40**	**77.5 %**	40/40	100%	81.875

Table 2. The activity recognition results when the best hmm architecture of the exp.1 was used

Walking person HMM with 2 states		Probing people HMM with 2 states		Damping people HMM with 10 states		Pickig up people HMM with 2 states		Mean Percentage of corr. classification
40/40	100%	29/40	72.5%	30/40	75%	40/40	100%	139/160 **86.87%**

sequences used in the experiment 1. Notice that in this case the performances of the proposed approach can change with respect to the ones reported in table 1, since both the relative maximum and corresponding threshold are used to classify each sequence. For the sequences "walking people" and "Picking up People" two hidden states have been selected because, under the same conditions, a smaller number of states makes simpler the training and test algorithms. For the sequence "Probing people" two states have also been selected because this case is the only one that ensures a classification performance of 100%. For the sequence "Damping People" the HMM with ten states has been selected since it ensures the best classification performance (77.50%).

The classification values relative to the selected HMMs are reported in cursive and bold type in table 1. The mean percentages of correct recognitions of the experiment 2 are reported in table 2 whereas table 3 shows the relative scatter matrix. The results demonstrate the effectiveness of the proposed approach based on a combination of HMM with different state numbers. The scatter matrix shows that the system mistakes the activities "probing people" and "Damping People". Actually these two activities are very similar and hard to distinguish also for a human beings.

A further experiment was performed: we have supplied to the HMM architecture used in the experiment 2 a set of 5 sequences containing none of the 4 activities used in the training phase. In this case no false positives have been found (meaning that the threshold constraint relative to the winner HMM is never satisfied).

Finally, in order to evaluate the possibility of using the proposed approach for real time applications, some considerations about the computational load have

Table 3. Details of the activity recognition results when the best HMM architecture of the exp. 1 was used

Scatter Matrix	HMM Classification			
	Walking People	Probing People	Damping People	Picking up People
Walking Person	**40**	0	0	0
Probing Person	0	**29**	11	0
Damping Person	0	10	**30**	0
Picking up People	0	0	0	**40**

Table 4. Distribution of the computational load

Segmentation	Pose Estimation	Activity Recognition	**Estimated Total Time per frame**
$\sim 4 \times 10^{-2}$	$\sim 4 \times 10^{-2}$	$\sim 5 \times 10^{-5}$	$\sim 14 \times 10^{-2}$

been done. Each frame can be processed in about 14x10^{-2}s and the distribution of the computational load in the four subsystems is reported in table 4. The total amount allows the processing of 6 frames/sec. This can be a satisfying result taking in account that normally the human movements are slow.

4 Conclusions and Future Works

In this paper we have presented a reliable approach to recognize complex human activities performed by human beings in an archeological site. In particular we have addressed some of the problems concerning this kind of application domains.

Starting from the detection of moving people, the proposed approach addresses the problem of recognizing four different activities from temporal variations of postures. The postures have been detected using an unsupervised clustering algorithm that is able to separate the binary shapes in the required number of classes. Fixed length sequences (50 frames) of postures have been used both in training and test phase to model the four different activities and to classify new examples of the same behavior.

The experiments have demonstrated the effectiveness of using HMM to recognize activities based on sequence of temporal postures. Besides, the computational times have been evaluated for each step of the whole system: they are very encouraging for using the system in real time applications. Future work will be addressed to evaluate how a larger number of postures can improve the results of the activity classification, also considering that the same position of a person can be perceived in a different way from the camera according to the relative orientations. Besides, we will face the problem of selecting variable length observation sequences from the test sequences, in order to overcome the constraint imposed in this work of having the same behavior in quite the same number of frames.

References

1. R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, "A System for Video Surveillance and Monitoring", Technical Report CMU-RI-TR-00-12, Carnagie Mellon University, 2000.
2. C. Stauffer and W.E.L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking" IEEE transactions on PAMI, vol. 22, n.8, pp. 747-757, August 2000.
3. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. IEEE Transactions on PAMI, 19(7):780–785, 1997.
4. I. Haritaoglu, D. Harwood, and L. S. Davis, "W-4: Real-time surveillance of people and their activities," IEEE Transactions PAMI, vol. 22, no. 8, pp. 809-830, 2000.
5. P. Remagnino and G.A. Jones, "Classifying Surveillance Events from Attributes and Behaviors" in the Proceeding of the BMVC, Sept. 10-13, Manchester, pp. 685-694,2001.
6. D. Ayers and M. Shah, "Monitoring human behavior from video taken in an office environment", Image and Vision Computing, Vol. 19 (12) (2001) pp. 833-846.
7. M. Petkovic, W. Jonker and Z. Zivkovic, "Recognizing Strokes in Tennis Videos Using Hidden Markov Models", In proceedings of VIIP, Marbella, Spain, 2001.
8. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Processing", Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.
9. T. Kanade, T. Collins and A. Lipton, "Advances in cooperative multi sensor video surveillance, DARPA Image Understanding Workshop, Morgan Kaufmann, Nov. 1998, pp.3-24.
10. X. Li, M. Parizeau, R. Plamondon, "Training Hidden Markov Models with Multiple Observations - A Combinatory Method", IEEE Trans. on PAMI, vol. 22,4, pp.371-7, Apr.2000.
11. S. Theodoridis, K. Koutroumbas, "Pattern Recognition", Academic Press, San Diego, 1999, ISBN 0-12-686140-4.
12. A. Branca, G. Attolico, A. Distante "Cast Shadow Removing in Foreground Segmentation". In Proc. Int. Conf. on Pattern Recognition, 2002.
13. M. Leo, G. Attolico, A. Branca, A. Distante "People detection in dynamic images" In the proceedings of the IEEE WCCI, Honolulu, Hawaii, May 12-17, 2002.