HMM Based Action Recognition with Projection Histogram Features

Roberto Vezzani, Davide Baltieri, and Rita Cucchiara

Dipartimento di Ingegneria dell'Informazione - University of Modena and Reggio Emilia, Via Vignolese, 905 - 41100 Modena - Italy
{roberto.vezzani,davide.baltieri,rita.cucchiara}@unimore.it

Abstract. Hidden Markov Models (HMM) have been widely used for action recognition, since they allow to easily model the temporal evolution of a single or a set of numeric features extracted from the data. The selection of the feature set and the related emission probability function are the key issues to be defined. In particular, if the training set is not sufficiently large, a manual or automatic feature selection and reduction is mandatory. In this paper we propose to model the emission probability function as a Mixture of Gaussian and the feature set is obtained from the projection histograms of the foreground mask. The projection histograms contain the number of moving pixel for each row and for each column of the frame and they provide sufficient information to infer the instantaneous posture of the person. Then, the HMM framework recovers the temporal evolution of the postures recognizing in such a manner the global action. The proposed method have been successfully tested on the UT-Tower and on the Weizmann Datasets.

Key words: HMM, Projection Histograms, Action Classification

1 Introduction

Action classification is a very important task for a lot of automatic video surveillance applications. The main challenge relies on developing a method that is able to cope with different types of action, even if they are very similar to each other and also in the case of cluttered and complex scenarios. Occlusions, shadows and noise are the main problems to be faced.

In video surveillance applications the actions should usually be recognized by means of an image stream coming from a single camera. Common 2D approaches analyze the action in the image plane relaxing all the environmental constraints of 3D approaches but lowering the discriminative power of the action-classification task. The action classification can be performed in the image plane by explicitly identifying feature points [1], or considering the whole silhouette [2, 3]. Other approaches directly map low-level image features to actions, preserving spatial and temporal relations. To this aim, feature choice is a crucial aspect to obtain a discriminative representation. An interesting approach that detects human action in videos without performing motion segmentation

was proposed by Irani et al. in [4]. They analyzed spatio- temporal video patches to detect discontinuities in the motion-field directions. Despite the general applicability of this method, the high computational cost makes it unusable for real-time surveillance applications.

After their first application in speech recognition [5], HMMs have been widely used for action recognition tasks. In a recent and comprehensive survey on action recognition [6] several HMM based methods are presented. Yamato *et al* in [7] used HMMs in their most simpler shape: a set of HMM, one for each action, is trained. The observation probability function is modeled as a discrete distribution adopting a mesh feature computed frame by frame on the data [8], and finally, the learning was based on the well known Baum-Welch approach. Similarly, Li [9] proposed a simple and effective motion descriptor based on oriented histograms of optical flow field sequence. After dimensional reduction by principal component analysis, it was applied to human action recognition using the hidden Markov model schema. Recently, Martinez *et al* [10] proposed a framework for action recognition based on HMM and a silhouette based feature set. Differently from the other proposals, their solution lies on an 2D modeling of human actions based on motion templates, by means of motion history images (MHI). These templates are projected into a new subspace using the Kohonen self organizing feature map (SOM), which groups viewpoint (spatial) and movement (temporal) in a principal manifold, and models the high dimensional space of static templates. The higher level is based on a Baum-Welch learned HMM.

In this work we adopt the common HMM framework with a feature set particularly suitable for low quality images. We firstly segment and track the foreground images by means of the Ad-Hoc system [11]. Thus, the projection histograms of the foreground blobs are computed and adopted as feature set [2]. To avoid the course of dimensionality we sub-sampled the histograms, in order to obtain a feature set with a reasonably limited number of values. Ad-Hoc includes a shadow removal algorithm [12]; nevertheless shadows can contain information about the current posture and can be adopted as additional data to recover missing one.

In Section 2 the traditional HMM action classification framework is reported. Section 3 describes the Projection Histogram feature set as well as a shape based feature set used as reference. Finally, comparative tests and the results of the proposed schema over the UT-Tower dataset are reported in Section 4.

2 HMM Action Classification

Given a set of C action classes $\Lambda = \lambda^1 \dots \lambda^C$, our aim is to find the class λ^* which maximise the probability $P(\lambda|O)$, where $O = \{o_1 \dots o_T\}$ is the entire sequence of frame-wise observations (features). In his famous tutorial [5], Rabiner proposed to use hidden Markov models to solve this kind of classification problems. An HMM should be learned for each action; the classification of an observation sequence O is then carried out selecting the model whose likelihood is highest, $\lambda^* = \arg\max_{1 \leq c \leq C} [P(O|\lambda^c)]$. If the classes are equally likely, this solution is

optimal also in a Bayesian sense.

$$\lambda^* = \arg\max_{1 \leq c \leq C} \left[P\left(O|\lambda^c\right) \right] \tag{1}$$

Since the decoding of internal state sequence is not required, the recursive forward algorithm with the three well known initialization, induction and termination equations have been applied.

$$\begin{aligned} \alpha_1(j) &= \pi_i b_j(o_1), 1 \leq i \leq N \\ \alpha_{t+1}(j) &= \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j\left(o_{t+1}\right) \\ P\left(O|\lambda\right) &= \sum_{j=1}^{N} \alpha_T\left(j\right) \end{aligned} \tag{2}$$

The term $b_j(o_t)$ depends on the type of the observations. We adopted the K-dimensional feature set described in the following, which requires to model the observation probabilities by means of density functions. As usual, we adopt a Gaussian Mixture Model, which simplifies the learning phase allowing a simultaneous estimation of both the HMM and the Mixtures parameters using the Baum-Welch algorithm, given the numbers N and M of hidden states and Gaussians per state respectively. In this case, the term $b_j(o_t)$ of Eq. 2 can be approximated as:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}^K\left(o_t|\mu_{jm}, \Sigma_{jm}\right) \tag{3}$$

where $\mathcal{N}^K(\mu, \Sigma)$ is a K-dimensional Gaussian distribution having mean vector μ and covariance matrix Σ; μ_{jm},Σ_{jm} and c_{jm} are the mean, the covariance matrix and the mixture weight of the m-th component for the action j.

3 Feature sets

The selection of the feature set to use is very important for the final classification rate. In particular, the adopted features should capture and follow the action peculiarities, but, at the same time, they should allow the action generalization.

In this paper we propose and compare two different feature sets. The first is based on the so called Projection Histograms and it is based on the shape of the foreground mask only; position and global motion of the person are not considered. The projection histograms have been used in the past for frame by frame posture classification [2]. The second feature set, instead, is composed by a mix of different measures, some of them based on the appearance and some on the person position and speed [13]. Independently from the semantics and the computation schema, the input for the HMM framework is a K-dimensional vector $o_t^1 \ldots o_t^K \in \mathbb{R}^K$.

3.1 Projection Histograms Feature Set

Since the videos were acquired by a fixed camera, each frame $I_t(x, y)$ is processed to extract the foreground mask (F) by means of a background subtraction step

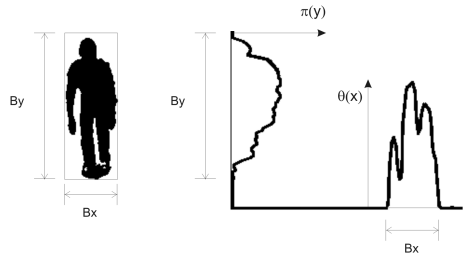

Fig. 1. Vertical and Horizontal Projection histograms of a sample blob

[12]. For this contest, we directly used the foreground images furnished within the dataset [14]. The feature vectors o_t are then obtained from the projection histograms of the foreground mask [2], i.e. projections of the person's silhouette onto the principal axes x and y.

Examples of projection histograms are depicted in Fig. 1.

Given the boolean foreground mask $F(x, y)$, the projection histograms θ and π can be mathematically defined as:

$$\theta(x) = \sum_{y=0}^{F_y} \phi(F(x,y)) \; ; \; \pi(y) = \sum_{x=0}^{F_x} \phi(F(x,y)) \tag{4}$$

where the function ϕ is equal to 1 if $F(x, y)$ is true, 0 otherwise, while F_x and F_y are the width and the height of the foreground mask F respectively.

In practice, θ and π can be considered as two feature vectors and the final feature vector $O_t \in \mathbb{R}^K$ used to describe the current frame is obtained from θ and π normalizing each value such as they sum up to 1, resampling the two projection histograms to a fixed number $S = K/2$ of bins, and concatenating them into a unique vector.

3.2 Model based Feature Set

Projection histograms do not depend on any assumption on the people shape and they can be used to describe a generic object. We propose another simple feature set, which is based on a simplified body model, discriminative enough to obtain reasonable classification rates, but not too complex to permit fast processing. The foreground silhouettes are divided into five slices $S^1 \dots S^5$ using a radial partitioning centered in the gravity center $\{x_c(t), y_c(t)\}$. These slices should ideally correspond to the head, the arms and the legs. Calling A_t and $\{A_t^i\}_{i=1\dots5}$ the areas of the whole silhouette and of each slice $\{S^i\}$ respectively, the 17-dimensional feature set is obtained as reported in Fig. 3. The features contain both motion (o^1 and o^2) and shape information ($o^3 \dots o^{17}$).

$$o_t = \left\{o_t^1 \dots o_t^{17}\right\}, = \left\{ \begin{array}{l} o_t^1 = x_c(t) - x_c(t-1); \\ o_t^2 = y_c(t) - y_c(t-1); \\ o_t^{3\dots7} = \frac{A_t^i}{A_t}, i = 1 \dots 5; \\ o_t^{8\dots12} = \frac{\max_{(x,y)\in S_i} x}{\sqrt{A_t^i}}, i = 1 \dots 5; \\ o_t^{13\dots17} = \frac{\max_{(x,y)\in S_i} y}{\sqrt{A_t^i}}, i = 1 \dots 5; \end{array} \right\} \quad (5)$$

4 Experimental Results

The proposed method have been tested on the UT-Tower Dataset [14] and on the Weizmann dataset [15].

The **UT-Tower Dataset** [14] contains 112 videos of 9 actions performed 12 times each, Some actions are performed in different ways, thus in the on-line recognition we used all the 16 specific classes (Some frames of the dataset are reported in Fig. 4).

We tested the system using the projection histogram feature set. The number of bins have been sub-sampled to 10 for each direction, obtaining a 20-dimensional feature set. The classification precision achieved using a leave-one-out test schema is around 96%. The confusion matrix is reported in table 5.

The low quality of the segmentation masks and the too limited size of the blobs make the alternative feature set ineffective. Moreover, shadows play an important role in the classification results. In Fig. 2(d) and 2(e) the projection histograms obtained by including shadows or removing them are shown: shadows strongly affect projection histograms based on blob's silhouette, and thus they usually must be removed. Anyway, if the shadow characteristics (i.e., size, position, direction) are not changing among sequences, they can be leaved; on the contrary, information about the performed action are also embedded in the shadow. Thus, we can avoid any shadow removal step if the shadows are always in the same direction and if the adopted feature set is not model based (such as the projection histograms). The model-based feature set described in section

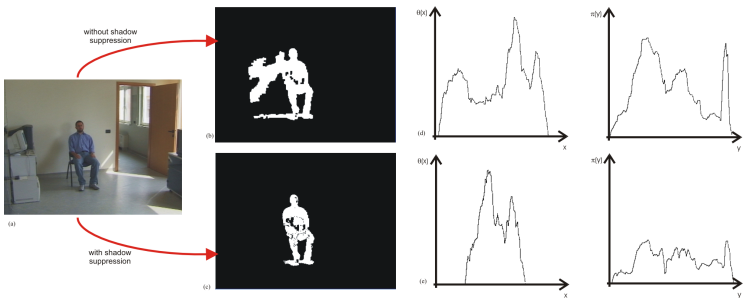

Fig. 2. Comparison of the projection histograms achieved by preserving (top) or removing (bottom) shadows.

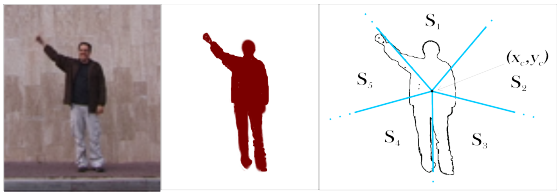

Fig. 3. Model-based 17-dimensional Feature set

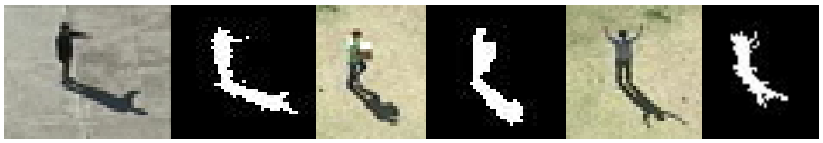

Fig. 4. Sample input frame of the UT-Tower dataset

3.2, instead, starts with the estimation of the body center. Shadows strongly compromise this estimation and the overall action classification rate, achieving performance around the 60% on the same dataset.

Ground thruth - Action ID

Recognized Action ID	1	2	3	4	5	6	7	8	9
1	10	0	0	0	0	0	2	0	0
2	0	10	0	1	0	0	1	0	0
3	0	0	12	0	0	0	0	0	0
4	0	0	0	12	0	0	0	0	0
5	0	0	0	0	12	0	0	0	0
6	0	0	0	0	0	12	0	0	0
7	0	0	0	0	0	0	12	0	0
8	0	0	0	0	0	0	0	12	0
9	0	0	0	0	0	0	0	0	12

Fig. 5. Confusion matrix of the Projection Histograms Feature set on the UT-Tower dataset

The **Weizmann dataset** [15] contains 90 videos of 10 main actions performed by 9 different people. Some actions are performed in different ways, thus in the on-line recognition we used all the 16 specific classes. Example frames of this well known dataset are shown in Figure 7.

With this dataset the model based feature set performs better than the projection histograms one. The confusion matrix obtained using the model based feature set is shown in Figure 6.

We empirically tuned the HMM parameters. In particular the number N of hidden states and the number M of Gaussians of the mixture model of Eq. 3

Recognized Action ID \ Ground thruth - Action ID	1	2	3	4	5	6	7	8	9	10
1	**100**	0	0	0	0	0	0	0	0	0
2	0	**99**	0	0	0	0	0	0	1	0
3	0	0	**68**	0	4	0	27	1	0	0
4	0	12	0	**87**	0	0	0	0	1	0
5	0	0	0	0	**81**	0	19	0	0	0
6	0	0	0	0	5	**95**	0	0	0	0
7	0	0	12	0	31	0	**57**	0	0	0
8	0	0	0	0	0	0	0	**100**	0	0
9	0	0	0	0	0	0	0	0	**86**	14
10	0	0	0	0	0	0	0	0	6	**94**

Fig. 6. Confusion matrix of the Model Based Feature set on the Weizmann dataset

have been set to 5 and 3 respectively to maximize the recognition rates based on some experiments we carried out on the Weizmann dataset.

The complete system, including the background subtraction and updating step, the object tracking, feature extraction and action classification is working in real time, processing about 15 frames per second.

5 Conclusions

In this paper, a traditional HMM framework for action recognition is presented. We proposed and compared two different feature sets, based on projection histograms and shape descriptors respectively. The framework was initially developed for the participation to the ICPR 2010 Contest on Semantic Description of Human Activities - "Aerial View Activity Classification Challenge" [14]. Using the projection histogram feature set the classification precision is around 96%. The system was also tested on the **Weizmann dataset** [15], on which the shape descriptors performs better than projection histograms. Given the temporal segmentation of the actions and a well representative training set, the Hidden Markov Model approach still guarantees good performances both in terms of precision and computational load.

Acknowledgments

This work has been done within the THIS project with the support of the Prevention, Preparedness and Consequence Management of Terrorism and other

Fig. 7. Sample input frame of the Weizmann dataset

Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security. The authors also thank Massimo Piccardi for several discussions on the work and for his valuable advice.

References

1. Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE Int. Conf. on Computer Vision (ICCV'03), Nice, France (2003)
2. Cucchiara, R., Grana, C., Prati, A., Vezzani, R.: Probabilistic posture classification for human behaviour analysis. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans **35** (2005) 42–54
3. Ke, Y., Sukthankar, R., Hebert, M.: Spatiotemporal shape and flow correlation for action recognition. In: Proc. of IEEE Intl Conference on Computer Vision and Pattern Recognition. (2007) 1–8
4. Shechtman, E., Irani, M.: Space-time behavior-based correlation -or- how to tell if two underlying motion fields are similar without computing them? IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 2045 –2056
5. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. In: Proc. of the IEEE. Volume 77. (1989) 257–286
6. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine recognition of human activities: A survey. IEEE Trans. on Circuits and Systems for Video Technology **18** (2008) 1473–1488
7. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden markov model. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. (1992) 379–385
8. Umeda, M.: Recognition of multi-font printed chinese. characters. In: Proc. of 6th International Conference on Pattern Recognition. (1982) 793–796
9. Li, X.: Hmm based action recognition using oriented histograms of optical flow field. Electronics Letters **43** (2007) 560 –561
10. Martinez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S.: Recognizing human actions using silhouette-based hmm. In: Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on. (2009) 43 –48
11. Vezzani, R., Cucchiara, R.: Ad-hoc: Appearance driven human tracking with occlusion handling. In: First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS'2008), Leeds, UK (2008)
12. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003) 1337–1342
13. Vezzani, R., Piccardi, M., Cucchiara, R.: An efficient bayesian framework for online action recognition. In: Proceedings of the IEEE International Conference on Image Processing, Cairo, Egypt (2009)
14. Chen, C.C., Ryoo, M.S., Aggarwal, J.K.: UT-Tower Dataset: Aerial View Activity Classification Challenge. http://cvrc.ece.utexas.edu/SDHA2010 (2010)
15. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence **29** (2007) 2247–2253