Markerless Body Part Tracking for Action Recognition

Simone Calderara[1], Andrea Prati[2], Rita Cucchiara[1]

[1]Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia, Italy
[2]Department of Engineering Science and Methods, University of Modena and Reggio Emilia, Italy

Abstract: This paper presents a method for recognizing human actions by tracking body parts without using artificial markers. A sophisticated appearance-based tracking able to cope with occlusions is exploited to extract a probability map for each moving object. A segmentation technique based on mixture of Gaussians (MoG) is then employed to extract and track significant points on this map, corresponding to significant regions on the human silhouette. The evolution of the mixture in time is analyzed by transforming it in a sequence of symbols (corresponding to a MoG). The similarity between actions is computed by applying global alignment and dynamic programming techniques to the corresponding sequences and using a variational approximation of the Kullback-Leibler divergence to measure the dissimilirity between two MoGs. Experiments on publicly available datasets and comparison with existing methods are provided.

Keywords: Action recognition, mean tracking, mixture of Gaussians, dynamic programming.

Reference to this paper should be made as follows: Calderara, S. and Prati, A. and Cucchiara, R. (2010) 'Markerless Body Part Tracking for Action Recognition', *Int. J. of Multimedia Intelligence and Security (IJMIS)*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Simone Calderara received the M.S.I. in computer science in 2004 from the University of Modena and Reggio Emilia, Italy and the Ph.D. degree in Computer Engineering and Science in 2009. He is currently Research Assistant in the Imagelab Computer Vision and Pattern Recognition Lab in University of Modena and Reggio Emilia. His current research interests are Computer Vision and Pattern Recognition applied to human behavior analysis, video surveillance and tracking and time series analysis for forensic applications.

Andrea Prati (Laurea '98, PhD '01 in Computer Engineering at University of Modena, Italy) is Assistant Professor in the Department of Engineering Science and Methods of University of Modena and Reggio Emilia, Italy. His research interests include video surveillance and machine vision. He is author of more than 100 papers in international conferences and journals.

Rita Cucchiara (Laurea ’89, Ph.D. ’92 in Computer Engineering at University of Bologna, Italy) is Full Professor in Computer Engineering at University of Modena and Reggio Emilia, Italy. She is currently vice-dean of the Faculty of Engineering of Modena, coordinator of the Phd Curricula in Computer Engineering of the Phd School of ICT of Modena and heads the ImageLab (http://www.imagelab.unimore.it) of Modena. Her current interests include pattern recognition and computer vision for video surveillance, machine vision and multimedia for image and video retrieval. She is responsible for many Italian and International projects (she coordinates This EU Project, and the NATO project Besafe in people surveillance). She is author of more than 150 papers in international journals and conference proceedings, and she is part of PC of many international conferences (CVPR, ICME, ACM MM, ICPR, ...). She is member of GIRPR (Italy-associated with IAPR), AixIA (Ital. Assoc. Of Artificial Intelligence), ACM and IEEE Computer Society.

1 Introduction

Labeling actions taking place in a given scene is a task of paramount importance for behavior analysis. The main challenge relies on developing a method able to cope with almost every type of action, even if they are very similar to other one and also in the case of cluttered and complex scenarios. In the recent past, many researchers have addressed action recognition in video sequences in different contexts and with different purposes, ranging from sports video analysis to video surveillance to human-centred computing. For several years, researchers have concentrated on ad-hoc solutions to identify, often with heuristic rules, specific actions, such as fighting, talking, etc. [Cupillard et al., 2002]. However, recent advances in computer vision and statistical pattern recognition offer an effective and often efficient help for the recognition of higher-level actions, such as abandoned luggage detection, repetitive and abnormal path detection, or people-to-people interactions.

Basic approaches for recognizing human actions are based on either the analysis of body shape (in 2D or 3D) or the analysis of the dynamics of prominent points or parts of the human body. More specifically, action recognition approaches can be divided into two main groups [Gavrila, 1999] depending on whether the analysis is performed directly in the image plane (*2D approaches*) or using a three dimensional reconstruction of the action itself (*3D approaches*). The latter ones have been widely adopted where building and fitting a 3D model of the body parts performing the action is relatively simple due to controlled environmental conditions and high-resolution view of the object. For instance, Regh and Kanade in [Regh and Kanade, 1995] used a 27 degree-of-freedom (DOF) hand model to recognize poses and gestures, while Goncalves *et al.* in [Goncalves et al., 1995] addressed the problem of analyzing human arm positions against a simple uncluttered background.

These methods are sometimes unfeasible in many real-time surveillance applications. Gavrila and Davis in [Gavrila and Davis, 1996] adopted a 22-DOF human-body model to detect actions against complex background but their approach constrains the user to wear a tight-fitting body suit with contrasting limb colors to simplify the edge detection problem in case of self-occlusions. Despite the complexity of the approach used, these methods can be applied only if a more or less sophisticated model of the target exists.

On the contrary, 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. People action classification can be performed in the image plane by either observing and tracking explicitly feature points (*local feature approaches* [Laptev and Lindeberg, 2003]), or considering the whole shape-motion as a feature itself (*holistic approaches* [Cucchiara et al., 2005; Ke et al., 2007]).

Yilmaz and Shah in [Yilmaz and Shah, 2005] exploited people contour-points tracking to build a 3D volume describing the action and their work represents an example of local feature approaches. A compact representation of this action-specific volume was presented and proved to be effective in distinguishing among several predefined actions. Although this proposal results effective in most situations, contour-points tracking is a difficult task to achieve in real-time systems leading to a NP-hard optimization problem when points are occluding each other and one-to-one matching is impossible.

Fei Fei *et al.* in [Niebles et al., 2006] proposed a feature-based approach that searches for "spatio-temporal words" as a time-collection of points of interest and classify them into actions using a pLSA (probabilistic latent semantic) graphical model. Training a complex graphical model requires many examples of the desired action and the "bag of words" feature extractor can be imprecise when a close view of the subject is not available.

Holistic approaches, instead, directly map low-level image features to actions, preserving spatial and temporal relations. Feature choice is a crucial aspect to obtain a discriminative representation. An interesting holistic approach that detects human action in videos without performing motion segmentation was proposed by Irani *et al.* in [Shechtman and Irani, 2007]. 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.

Some approaches tackled the action recognition problem by using multiple cameras. In fact, several actions cannot be easily caught by a single view. The use of multiple cameras looking simultaneously at the scene from different viewpoints allows to consistently recognize the action. Most of the previous works using multiple cameras tackle the problem by using a view-independent representation of the action [Li and Fukui, 2007; Gritai et al., 2004]. Li and Fukui in [Li and Fukui, 2007], for instance, considered the case of an action seen by a moving camera: by continuously changing the point of view, the action can be confused. The authors used non-rigid factorization [Torresani et al., 2001] of non-rigid shapes (like the human body is) and HMMs to model the action as a dynamic linear combination of basis shapes, whose weights contain the crucial information to recognize (in a view-invariant way) the action. Gritai *et al.* in [Gritai et al., 2004], instead, use anthropometric spatial constraints among body parts and proportions of them as cues for matching actions performed from different viewpoints and in different environments. Non-linear time warping is used to perform temporally invariant matching. Despite the generality of these two approaches, both of them exploit artificial markers or manual intervention to extract human body points to be analyzed.

Some other works explicitly use multiple sources of information and thus tend more to a "consistent" recognition of the actions. For instance, in [Syeda-Mahmood et al., 2001] the action was modeled by the so-called "action cylinder", i.e. a 3D object that models the evolution of the 2D object shape in time. View-independent action recognition is achieved by recovering the geometrical transformation between the action model and a given action cylinder. Cupillard *et al.* [Cupillard et al., 2002], instead, developed a combination mechanism to fuse recognition achieved in different views by combining the moving region graphs obtained by each view.

HMM-based techniques are commonly adopted to classify actions and learn motion semantics directly from data. Lee *et al.* in [Ahmad and Lee, 2008] proposed a method for recognizing actions from multiple views exploiting a combined local and global optical flow computed on segmented blobs. The view invariance was achieved using the normalize Zernike moments and recognition performed by a set of HMM-based classifiers, one for each known action, in a ML fashion. Although the method seems to perform correctly in many cases, optical flow approaches are unsuitable to recognize actions in low contrast environments and especially when single body parts are moving over the person silhouette, e.g. "drinking from a glass" action. In addiction, HMM-based classifiers need an exhaustive training set to avoid model overfitting and poor classification rates. In [Cuntoor et al., 2008] an HMM-based classifier that exploits state-level transition probabilities to model an action was proposed. One HMM was build for each action to classify and state sequence used to discriminate among them using a space time trajectory of marked point in the image plane as observed data. The authors described the classification stage avoiding the data extraction discussion. The proposed classifier could perform well if sufficient training data are available but the extraction of precise trajectories is the major weakness of trajectory-based classifiers because markers or body-part tracking is often unfeasible in real surveillance scenarios.

The system proposed in this paper is meant to solve the problem of using artificial markers by automatically segmenting the human silhouette into a certain number of relevant areas found in the image describing the motion evolution. The tracking of the areas' centroids produces 3D trajectories describing on a fine grain the action of the person. The evolution of the mixture in time is analyzed by transforming it in a sequence of symbols. Therefore, in order to compare two actions, we define a novel approach for comparing two sets of trajectories based on *sequence global alignment* and *dynamic programming.* Performed experiments on the publicly available dataset used in [Gorelick et al., 2007] showed an excellent discriminative power of this approach.

The rest of the paper is structured as follows: Section 2 presents a system overview and introduces the main three steps of the proposed approach; a brief description of the object detection and tracking step is also reported; Section 3 explains in details how the action space-time trajectories are extracted, whereas Section 4 explains how the STTs are compared to compute a similarity measure; eventually, Section 5 reports the experimental results.

2 System Overview

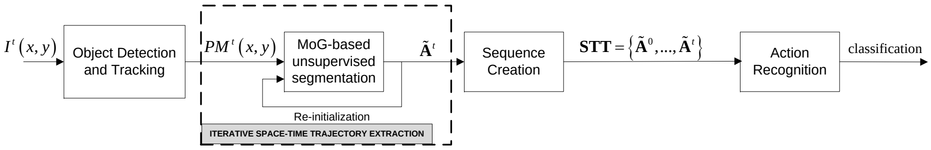

$I^t(x,y)$
Object Detection and Tracking
$PM^t(x,y)$
MoG-based unsupervised segmentation
$\tilde{\mathbf{A}}^t$
Re-initialization
ITERATIVE SPACE-TIME TRAJECTORY EXTRACTION
Sequence Creation
$\mathbf{STT}=\{\tilde{\mathbf{A}}^0,...,\tilde{\mathbf{A}}^t\}$
Action Recognition
classification

Figure 1 Scheme of the proposed system.

The proposed system is based on three main steps (Fig. 1):

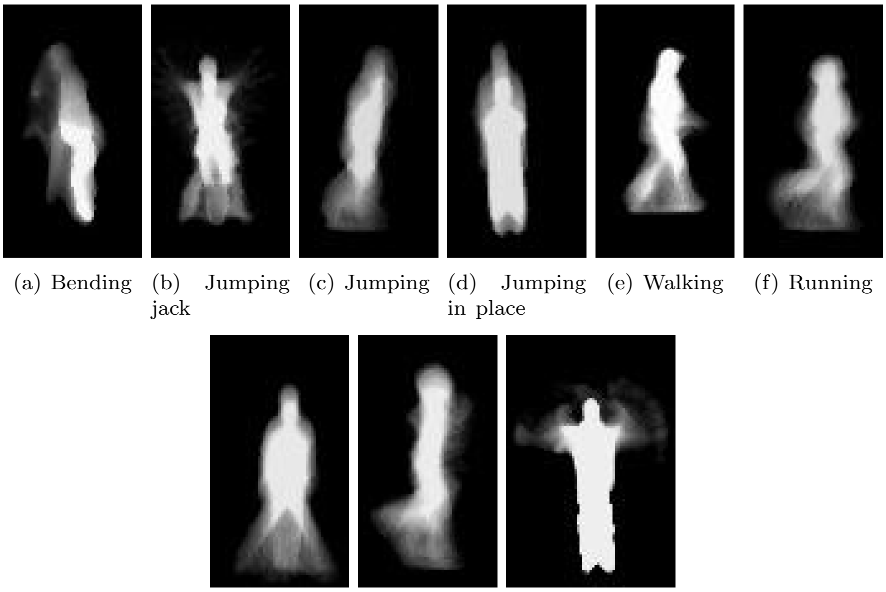
- *object detection and tracking*: the moving people are segmented and tracked from the image $I^t(x,y)$ at instant time t; by this step, a probability map $PM^t(x,y)$ is obtained for each moving person (section 2.1);
- *iterative space-time trajectory extraction*: K main components of PM^t corresponding to K main body parts are automatically extracted and used to model the action; EM algorithm is used to infer the parameter set $\widetilde{\mathbf{A}}^t$ of a 3-variate mixture of K Gaussians (MoG) on PM^t; finally, the sequences of the MoGs along time represent space-time trajectories $\mathbf{STT}=\left\{\widetilde{\mathbf{A}}^0,\cdots,\widetilde{\mathbf{A}}^t\right\}$ (section 3);
- *action recognition*: a new action modeled as a **STT** is compared using *global alignment* [Gusfield, 1997] to compute a measure of distance/similarity from all the existing actions; the cost for aligning two MoGs corresponding to two different actions is computed using a variational approximation of the Kullback-Leibler divergence; the classification is performed with a minimum distance classifier (section 4).

2.1 Object Detection and Tracking

For the sake of brevity, we report here only the key concepts of the object detection and tracking step in order to provide sufficient information for further steps. Complete description of the object detection step and tracking algorithm can be found in [Cucchiara et al., 2003] and [Cucchiara et al., 2004], respectively. These steps have a twofold scope: the first is to separate foreground/moving objects from the background; the second is to obtain a rich feature set characterizing the action.

The first scope is achieved with the approach called SAKBOT (Statistical And Knowledge-Based Object Tracker) [Cucchiara et al., 2003] which is based on background suppression where the background model updating is performed with an adaptive model and temporal median filtering. The updating is empowered by selectivity, i.e. the background model should not include the interesting moving objects if their motion is low or zero for a short period.

Once that moving objects/people have been segmented, they need to be tracked along time. For our purposes, tracking also represents a way for incorporating action evolution in a single observation by integrating in time single pixel membership. In other words, we create a *probability map* $PM(p)$ where each value defines the probability that the point p belongs to the object. The value of $PM(p)$ is updated with the segmentation results of the last n frames: further details can be found in [Cucchiara et al., 2004]. As shown in Fig. 2, PM represents a fine-grain description of the action, removing useless information such as the person's appearance.



(a) Bending (b) Jumping jack (c) Jumping (d) Jumping in place (e) Walking (f) Running

(g) Gallop sideways (h) Skip (i) Hands wave

Figure 2 Examples of PM computed for different actions. Brighter pixels correspond to higher probability.

3 Action Space-Time Trajectories

To model the complete evolution of the action we use a 3D observation $\mathbf{O}^t = (x, y, z)$, where $z = PM^t(x, y)$. In other words, the PM image is treated as 3D data to be clustered jointly in space and probability domains. Data clustering has the objective to identify main areas in the person's silhouette characterized by a

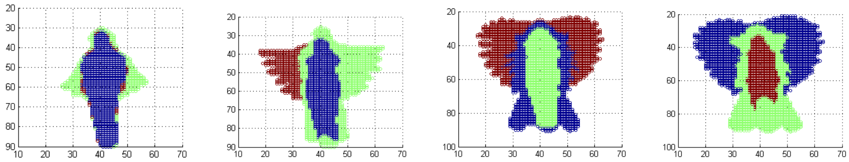
3D distribution similar to a given model. The 3D data are modeled with a 3-variate mixture of K Gaussians:

$$p\left(\mathbf{O}^t|\mathbf{A}^t\right) = \sum_{k=1}^{K} \pi_k^t \mathcal{N}\left(\mathbf{O}^t|\boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t\right) \tag{1}$$

where $\mathbf{A}^t = \{\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\pi}^t\}$ is the set of parameters of the MoG, including the mean vector $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$ for each of the K components, the covariance matrix $\boldsymbol{\Sigma}$ and the weight vector $\boldsymbol{\pi}$. The single 3-variate Gaussian can be written as:

$$\begin{aligned}\mathcal{N}\left(\mathbf{O}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = & \frac{1}{\sqrt[3]{2\pi}} \frac{1}{\left(\det\left(\boldsymbol{\Sigma}\right)\right)^{1/2}} \\ & \exp\left\{-\frac{1}{2}\left(\mathbf{O}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{O}-\boldsymbol{\mu}\right)\right\}\end{aligned} \tag{2}$$

By using the EM algorithm the set of estimated parameters $\widetilde{\mathbf{A}}^t = \left\{\widetilde{\boldsymbol{\mu}}^t, \widetilde{\boldsymbol{\Sigma}}^t, \widetilde{\boldsymbol{\pi}}^t\right\}$ can be easily inferred. To initialize the EM on the first frame ($t = 0$) we use the k-means clustering on PM^0. Conversely, the initialization for the subsequent frames ($t > 0$) is based on the estimate on the previous step, i.e. $\mathbf{A}^t = \widetilde{\mathbf{A}}^{t-1}$ (see also the feedback arrow in Fig. 1). This re-initialization process is based on the assumption that body parts do not move significantly between two consecutive frames (supposing a reasonably-high frame rate). Some examples of the segmentation achieved by this process with $K = 3$ components are reported in Fig. 3, where a person performing the “jumping jack” action.



(a) (b) (c) (d)

Figure 3 Example of the segmentation with MoG.

The set of parameters $\widetilde{\mathbf{A}}^t$ represents the locations (in the 3D space) of the body parts (through the means $\widetilde{\boldsymbol{\mu}}^t$), their reliability (through the means $\widetilde{\boldsymbol{\Sigma}}^t$) and their importance (through the weights $\widetilde{\boldsymbol{\pi}}^t$). How these values change in time is representative of the evolution of the action and constitutes a good descriptor of it. Think for instance to the action “walking”: some body parts (e.g., those corresponding to the torso) will follow a straight path with almost-constant covariance and weight; others (such as those associated to the legs) will vary significantly in all the parameters. Therefore, in order to model properly the action's dynamics the sequence of $\widetilde{\mathbf{A}}$ is collected to form the *space-time trajectory* (STT): $\mathbf{STT} = \left\{\widetilde{\mathbf{A}}^0, \cdots, \widetilde{\mathbf{A}}^t\right\}$

4 STT Comparison for Action Recognition

The previous sections describe our methodology to model an action as a sequence (or trajectory) of MoG (modelled by the set of the parameters $\widetilde{\mathbf{A}}$). In order to cluster or classify similar actions, single trajectories $\mathbf{STT}$ from different actions should be compared. To this aim, defined a and b the actions to be compared, a similarity measure $\Omega\left(\mathbf{STT}_i, \mathbf{STT}_j\right)$ between every trajectory i of action a and every trajectory j of action b is needed. Due to segmentation inaccuracies and different velocities in performing actions, the MoG distributions composing $\mathbf{STT}_i$ and $\mathbf{STT}_j$ cannot be directly compared point by point. Since STTs are indeed time series, we can borrow from bioinformatics a method for comparing sequences of data in order to find the best inexact matching between them, also accounting for time shifts, difference sequence lengths and statistical uncertainty.

Time shifts and different lengths can be accounted by using data alignment techniques, which find the best (with respect to a cost/similarity measure) alignment of single symbols composing the sequence. Among the many existing techniques, we used the Needleman-Wunsch *global alignment* algorithm [Needleman and Wunsch, 1970; Gusfield, 1997]. On the other hand, the statistical uncertainty is afforded by not matching directly the data, but using a statistical pdf (i.e., a 3-variate MoG) and a measure of similarity between pdfs. In this way, noisy data and inaccuracies can be handle easily.

Global alignment of two sequences S and T is obtained by first inserting spaces (*gaps*), either into or at the ends of S and T so that the length of the sequences will be the same; then, every symbol/space in one of the sequences is matched to a unique symbol/space in the other. The algorithm is based on the concept of "modification" to the sequence (analogous to the mutation in a DNA sequence), which can be due to *indel* operations (insertion or deletion of a symbol) or to *substitutions.* By assigning different weights/penalties to these operations it is possible to measure the degree of similarity of the two sequences.

Unfortunately, this algorithm is very onerous in terms of computational time if the sequences are long. For this reason, *dynamic programming* is used to reduce computational complexity to $O\left(n_T \cdot n_S\right)$, where n_T and n_S are the lengths of the two sequences. Dynamic programming overcomes the problem of the recursive solution to global alignment by not comparing the same subsequences for more than one time, and by exploiting tabular representation to efficiently compute the final similarity score.

Each element (p, q) of the table contains the alignment score of the distribution $p\left(\mathbf{O}|\widetilde{\mathbf{A}}^p\right)$ of sequence T with the symbol $p\left(\mathbf{O}|\widetilde{\mathbf{A}}^q\right)$ of sequence S, where $p\left(\mathbf{O}|\widetilde{\mathbf{A}}\right)$ is the same of equation 1 using the parameter $\widetilde{\mathbf{A}}$ estimated through the EM described in Section 3. The score can be measured statistically as a function of the distance between the corresponding distributions. If the two distributions result sufficiently similar the score should be high and positive, while if they differ much the score (penalty) should be negative. The best alignment can be found by searching for the alignment that maximizes the global score.

Specifically, in our specific case the symbols of the sequences correspond to a mixture of Gaussian distribution, represented by its parameters. Thus, we need a measure of distance between mixtures. The commonly-used Bhattacharyya distance for comparing pdfs cannot be used in this case. In fact, due to the presence of summation in the mixture expression, a closed-form solution of the Bhattacharyya distance for mixtures is not available. Iterative solutions exist which approximate the integral with a summation, but they are either too approximated or too slow to be computed. An alternative technique for comparing distributions is provided by the Kullback-Leibler divergence (see equation 3).

$$KL\left(p\left(\mathbf{O}|\widetilde{\mathbf{A}}^p\right)\|p\left(\mathbf{O}|\widetilde{\mathbf{A}}^q\right)\right) = -\int\int\int \sum_{k=1}^{K} \widetilde{\pi}_k^p \mathcal{N}\left(x, y, z|\widetilde{\boldsymbol{\mu}}_k^p, \widetilde{\boldsymbol{\Sigma}}_k^p\right) \ln\left(\frac{\sum\limits_{k=1}^{K} \widetilde{\pi}_k^q \mathcal{N}\left(x, y, z|\widetilde{\boldsymbol{\mu}}_k^q, \widetilde{\boldsymbol{\Sigma}}_k^q\right)}{\sum\limits_{k=1}^{K} \widetilde{\pi}_k^p \mathcal{N}\left(x, y, z|\widetilde{\boldsymbol{\mu}}_k^p, \widetilde{\boldsymbol{\Sigma}}_k^p\right)}\right) dxdyd$$

Unfortunately, in our case $p\left(\mathbf{O}|\widetilde{\mathbf{A}}\right)$ is a mixture, thus the KL divergence is not analytically tractable, nor does any efficient computational algorithm exist. However, many approximated approaches have been proposed [Hershey and Olsen, 2007]. Among the approaches proposed, the most suitable (since it preserves the majority of the KL properties) is based on the variational computation of the lower bound of the log likelihood. In fact, given two generic mixtures $f(x) = \sum_a \pi_a f_a(x|P_a)$ and $g(x) = \sum_b \omega_b g_b(x|P_b)$, with f_a and g_b the unimodal distributions and P_a and P_b their corresponding parameters, the KL divergence can be written as:

$$\begin{aligned} KL(f|g) &= \int f(x)\ln f(x)dx - \int f(x)\ln g(x)dx \\ &= \mathbb{E}_{f(x)}\left[\ln f(x)\right] - \mathbb{E}_{f(x)}\left[\ln g(x)\right] \end{aligned} \tag{4}$$

Being $\mathcal{L}_f\left(g\right) \equiv \mathbb{E}_{f(x)}\left[\ln g(x)\right] = \mathbb{E}_{f(x)}\left[\ln \sum_b \omega_b g_b(x)\right]$ the log likelihood, we can use a variational approach to find a (more tractable) lower bound. Introducing the variational parameters $\phi_{b|a} > 0$ such that $\sum_b \phi_{b|a} = 1$, we obtain the following equations:

$$\begin{aligned} \mathcal{L}_f\left(g\right) &\equiv \mathbb{E}_{f(x)}\left[\ln \sum_b \omega_b g_b(x)\right] \\ &\geq \mathbb{E}_{f(x)}\left[\sum_b \phi_{b|a} \ln \frac{\omega_b g_b(x)}{\phi_{b|a}}\right] \equiv \mathcal{L}_f\left(g, \phi\right) \end{aligned} \tag{5}$$

Since the variational log likelihood $\mathcal{L}_f\left(g, \phi\right)$ is a lower bound of the log likelihood $\mathcal{L}_f\left(g\right)$, the best bound can be found by maximizing it with respect to ϕ. Thus, the following equation [Hershey and Olsen, 2007] can be derived as:

$$\hat{\phi}_{b|a} = \frac{\omega_b e^{-KL(f_a|g_b)}}{\sum\limits_{b'} \pi_{b'} e^{-KL(f_a|g_{b'})}} \tag{6}$$

Applying the same procedure for $\mathcal{L}_f(f)$, the KL divergence of eq. (4) can be approximated by the following variational KL divergence:

$$KL_{var}(f|g) = \sum_a \pi_a \ln \frac{\sum\limits_{a'} \pi_{a'} e^{-KL(f_a|f_{a'})}}{\sum\limits_b \omega_b e^{-KL(f_a|g_b)}} \tag{7}$$

Please note that the KL values reported in eq. (7) are now KL divergences between single unimodal distributions (either f_a or g_b), which can be treated analytically. In our specific case, the KL divergences of eq. (7) can be brought back to the KL between multivariate normal distributions. An analytical formulation of the KL divergence between normal distributions has been reported in [Hershey and Olsen, 2007]. Given the two distributions $\mathcal{N}_1(\mathbf{O}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_1(\mathbf{O}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, as defined in equation 2, we can write:

$$\begin{aligned} KL(\mathcal{N}_1|\mathcal{N}_2) &= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - \frac{N_1}{2} + \\ &+ \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \frac{1}{2}\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) \end{aligned} \tag{8}$$

where N_1 is the dimension of $\boldsymbol{\Sigma}_1$.

Considering that KL divergence vary from 0 to $+\inf$, while global alignment requires a score which must be positive and less than 1 (with 1 corresponding to identical sequences), we transform the distance KL in the coefficient $c_{KL} = \exp\{-KL\}$. Assuming that two distributions are sufficiently similar if the coefficient is above 0.5 and that the score for perfect match is +2, whereas the score (penalty) for the perfect mismatch is -1 (that are the typical values used in DNA sequence alignments), we can write the general similarity score as follows:

$$\sigma(p,q) = \begin{cases} 2 \cdot (c_{KL}) & \text{if}\, c_{KL} \geq 0.5 \\ 2 \cdot (c_{KL} - 0.5) & \text{if}\, c_{KL} < 0.5 \\ 0 & \text{if}\, p \text{ or } q \text{ are gaps} \end{cases} \tag{9}$$

After computing the similarity score, it must be normalized to obtain the distance $\Omega(T,S)$.

5 Experimental results and discussion

We tested the proposed similarity measure for action analysis on the publicly available dataset used in [Gorelick et al., 2007]. The dataset* is composed by 9 actions, described in Fig. 2, performed by several actors in an unconstrained way for a total number of 93 repetitions. Since the main goal of these experiments is to demonstrate the robustness of the proposed measure for action classification, we used directly the segmentation masks provided on the website (for every action sequence), without using the SAKBOT system for object detection and tracking.

*Downloadable at the web address http://www.wisdom.weizmann.ac.il/˜vision/SpaceTimeActions.html

	Bend	Jack	Jump	Jmp in Place	Walk	Run	Side	Skip	Wave.
Bend	**9**	0	0	0	0	0	0	0	0
Jack	0	**9**	0	0	0	0	0	0	0
Jump	0	0	**9**	0	0	0	0	0	0
Jmp in Place	0	0	0	**9**	0	0	0	0	0
Walk	0	0	0	0	**9**	0	0	0	0
Run	0	0	0	0	0	**9**	0	0	0
Side	0	0	0	0	0	1	**8**	0	0
Skip	0	0	0	0	2	1	0	**6**	0
Wave.	0	0	0	0	0	0	0	0	**18**

Table 1 Confusion matrix.

The probability masks are then clustered in the spatio-probability space using the multivariate Gaussian approach of Section 3 and obtaining the sequences of pdfs that describe the action. The chosen dataset contains several challenges: first, the poor resolution with which the actions are performed (video frames are $130x90$ pixels); additionally, several actions are similar and may cause difficult or ambiguous classification for a completely automatic technique. In particular, looking at the probability mask of the "skip" action (Fig. 2(h)), it looks very similar to the one of the "running" (Fig. 2(f)) or "walking" (Fig. 2(e)) actions. This visual similarity can affect the precision of the system that may confuse these actions, leading to inaccuracies in the final classification results.

The test campaign has been performed by processing all the videos and computing the probability masks for every sequence. The similarity measure was then tested adopting a nearest-neighbor classifier in a leave-one-out scheme. More precisely, every sequence was modeled singularly, compared against the remaining sequences of the whole dataset and assigned to the closest one using the similarity measure presented in Section 4.

The system exhibits a overall accuracy of 96% using the one-nearest neighbour classification. Moreover, the accuracy raises to the 98% using a voting scheme where the five nearest neighbours are taken into account, each of them voting for an action and the most voted action is selected. The voting procedure is feasible only when more than a single repetition for every action is present in the dataset, at least equal to the half of the number of the considered nearest neighbours plus one. When this condition does not hold, the voting procedure tends to produce errors even if the first nearest neighbor, that is the closest one to the considered action, is correct. The errors are motivated by the lack of possible correct votes due to the poor presence of the action in the training data. For this motivation, the voting procedure is applicable only when a prior knowledge about the training data is available and a sufficiently high number of repetitions for every action is present. Adversely, it should be avoided when the classification is performed on a dataset not rich or incomplete.

Looking at the confusion matrix shown in Tab. 1 we can observe that most of the errors are related to the "skip" action that, as previously stated, exhibits visual similarities in terms of probability masks with both the "walk" and "run" actions. We can still raise the number of components of the mixture used to cluster the probability masks in order to obtain a more fine grain action description but when the video resolution is poor, as may often occur in video surveillance data, the experimentally-chosen value of 3 components appears to be a good trade-off between performances and accuracy.

Regarding the efficiency of the algorithm the computational time for modeling and comparing two actions is around 10 seconds[+], where most of the time is spent performing the EM algorithm on each probability mask of the sequence. A possible solution to reach quasi-realtime performances is to subsample the action probabilities sequence, keeping in mind that the probability masks is a time integral of the person motion itself with the effect of being a redundant descriptor if considered in a frame-by-frame fashion. For example, one could divide the action sequence in three equal parts (one at the beginning, one in the middle and one concluding part) and compute the similarity for these three parts separately. This leads to a computational time of less then 1 second for modeling and comparing two actions, and to a slight degradation of the accuracy (from 96% to 91%).

6 Conclusions

We presented a statistical model for action recognition that uses a holistic approach based on probability masks and Gaussian Mixture clustering as action descriptor as well as an inexact alignment-based measure for effectively comparing actions by similarity. The proposed measure demonstrates good results both in terms of accuracy and performances, and exhibits the desirable property of working in real time when the action sequence is subsampled and few descriptors are considered. The experimental results on a publicly available action dataset demonstrate that the algorithm accuracy is satisfying even when the actions are visually similar and that the proposed measure is robust to changes in the actors, resulting similar also for the same action performed by different people in different ways.

References

M. Ahmad and S.-W. Lee. Human action recognition using shape and clg-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7):2237–2252, July 2008.

R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams. *IEEE Trans. on PAMI*, 25(10):1337–1342, Oct. 2003.

R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Probabilistic posture classification for human-behavior analysis. *IEEE Trans. on Systems, Man, and Cybernetics - Part A*, 35(1):42–54, Jan. 2005.

R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Proceedings of IAPR International Conference on Pattern Recognition (ICPR 2004)*, volume 1, pages 132–135, Aug. 2004.

N. P. Cuntoor, B. Yegnanarayana, and R. Chellappa. Activity modeling using event probability sequences. *IEEE Trans. on Image Processing*, 17(4):594–607, Apr. 2008.

F. Cupillard, F. Brémond, and M. Thonnat. Group behavior recognition with multiple cameras. In *WACV*, pages 177–183, 2002.

[+] Times computed for an unoptimized Matlab implementation on a Core 2 Duo PC.

D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, Jan 1999.

D. M. Gavrila and L. S. Davis. 3d model-based tracking of humans in action: A multiview approach. In *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, pages 73–80, 1996.

L. Goncalves, E. DiBernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3d. In *Proc. of IEEE Intl Conference on Computer Vision*, pages 764–770, 1995.

L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

A. Gritai, Y. Sheikh, and M. Shah. On the use of anthropometry in the invariant analysis of human actions. In *Proc. of Int'l Conference on Pattern Recognition*, volume 2, pages 923 – 926, Aug. 2004.

D. Gusfield. *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, 1997.

J. R. Hershey and P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–317–IV–320, 2007.

Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. of IEEE Intl Conference on Computer Vision*, volume 1, pages 432 – 439, 2003.

X. Li and K. Fukui. View-invariant human action recognition based on factorization and hmms. In *Proc. of MVA*, pages 227–230, 2007.

S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3): 443–453, 1970.

J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *Proc of British Machine Vision Conference*, volume 3, pages 1249–1259, 2006.

J. Regh and T. Kanade. Model based tracking of self occluding articulated objects. In *Proc. of IEEE Intl Conference on Computer Vision*, pages 612–617, 1995.

E. Shechtman and M. Irani. Space-time behavior correlation -or- how to tell if two underlying motion fields are similar without computing them? *IEEE Trans. on PAMI*, 2007.

T. Syeda-Mahmood, A. Vasilescu, and S. Sethi. Recognizing action events from multiple viewpoints. In *Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, pages 64 – 72, 2001.

L. Torresani, D. B. Yang, E. J. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *CVPR (1)*, pages 493–500, 2001.

A. Yilmaz and M. Shah. Action sketch: A novel action representation. In *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, volume 1, pages 984–989, 2005.