3-D BODY POSTURE TRACKING FOR HUMAN ACTION TEMPLATE MATCHING

Massimiliano Pierobon, Marco Marcon, Augusto Sarti and Stefano Tubaro

Image and Sound Processing Group
Dipartimento di Elettronica e Informazione - Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
Email: massimiliano.pierobon@poste.it, marcon/sarti/tubaro@elet.polimi.it

ABSTRACT

In this paper we present a novel approach to 3-D human action classification based on the analysis of volumetric data obtained form the joint processing of video sequences acquired by a multiple-camera system. The use of volumetric data makes the system very robust and avoids problems related the typical human body self-occlusions and motion ambiguities, very common in an independent camera-by-camera analysis. A Shape Descriptor of human body is obtained in order to capture only posture-dependent characteristics and its outputs at each time instant are collected together in action feature matrices. The use of Dynamic Time Warping approach for action template matching accounts for possible temporal nonlinear distortions among different instances of the same gesture and allows gesture classification.

1. INTRODUCTION

Gestures and actions are among the principal ways through which a human being interacts with reality. A gesture is usually performed to communicate something while an action is carried out in order to achieve a material purpose. This distinction is merely semantical as gestures and actions are both carried out performing sequences of body postures. The possibility of building an automatic machine capable of receiving and classifying this type of information has been one of the most fascinating spur for the research community in recent years.

Potential applications of this type of research projects can be easily found in the fields of automatic video surveillance systems, human-computer gestural interaction researches, robot skill learning and in many others. Automatic recognition and classification of suspicious movements and gaits [1] in sensitive areas is perhaps one of the most important recent needs demanding for applications at the cutting edge of human action recognition technology.

Despite the capability of the human brain to recognize postures only on the basis of image data, information on body joints configuration is 3-D in nature. The natural way of dealing with posture representation is, thus, in the 3-D environment [2]. In the work presented in this paper we used a *multi-camera input device and a 3-D Visual-Hull reconstruction technique* [3] to provide volumetric information to the system. In this way, problems such as viewpoint dependence, motion ambiguities and self-occlusions are inherently solved before the body posture tracking stage.

Frame-by-frame 3-D representations of the scene in terms of voxels (volumetric pixels) have been the input data from which extracting posture-dependent features [4]. In the Sec. 2 we introduce a new method for performing the *tracking of body postures throughout an action sequence*, mainly based on the dynamic adaptation of the technique used by Cohen and Li [5] for static posture estimation. Through experimental sessions, we developed a technique able to extract a posture-dependent signal, independent from actor's position, orientation, size and voxel-set resolution.

Once a suitable feature set representing body postures during an action execution is computed, we apply a well-known Template Matching technique in order to perform action sequence classification (Sec. 3). It is possible to consider postures as the atoms of gestures in the same way as phonemes are often considered the bricks that form words. Exploiting the similarity with the speech recognition problem, we used the *Dynamic Time Warping* (DTW) *procedure* [6] *to compute a distance metric suitable for performing action comparisons*. The novelty of our DTW implementation consists in the use of the Kullback-Leibler distance applied to posture descriptions as cost function. Experimental results, shown through the use of "confusion matrices", confirm the reliability of our approach.

2. BODY POSTURE TRACKING FROM VOLUMETRIC DATA

2.1. The shape of volumetric data

In order to perform a 3-D reconstruction procedure using sequence frames from multiple views of the scene, the system has to distinguish the actor silhouette from the rest of the image. We used a well-known method to perform this kind of segmentation: the **Chroma Keying**. This procedure accounts for differences in colour between the actor and the scene background. Once the object silhouette is extracted for each view, the so called **Visual-Hull** *volumetric reconstruction* of the scene shot by cameras is computed frame-by-frame before any tracking procedure. In this method, 3-D reconstruction is performed using the *volume intersection* approach, which recovers the volumetric description of the object from multiple silhouettes by back projecting from each viewpoint the corresponding silhouette for perspective projections [3] (Fig. 1 left). The intersection volume is then sampled regularly across the three dimensions in order to generate a volume made of binary voxels (ON/OFF). Body posture tracking is then computed directly on volumetric action sequence frames (Fig. 1 right).

### 2.2. Tracking postures using Shape Descriptor technique

The core of our body posture tracking procedure is based on the method proposed by Cohen and Li in [5]. They used the Shape
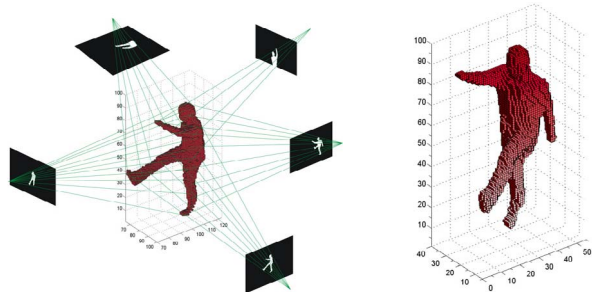
Fig. 1. *Volumetric intersection. Example of voxel-set creation by 3D intersection of Visual Hulls projected from segmented edges.*

Descriptor to compute features suitable for static posture recognition. Our purpose is slightly different because we need features to perform classification of human actions. Thus, *our shape description has to represent meaningfully not only body postures, but also their frame-by-frame dynamic changes.*

The procedure starts from a **3-D voxel-based representation of the scene** containing the human body volume. This is the data on which we compute the shape description (Fig. 1). The second step is the definition of a **reference shape** Θ consisting of a *vertically oriented cylinder*. It is adapted to the *actor's height* and its *axis passes through the body 3-D centroid* (Fig. 2 (b)). The use of
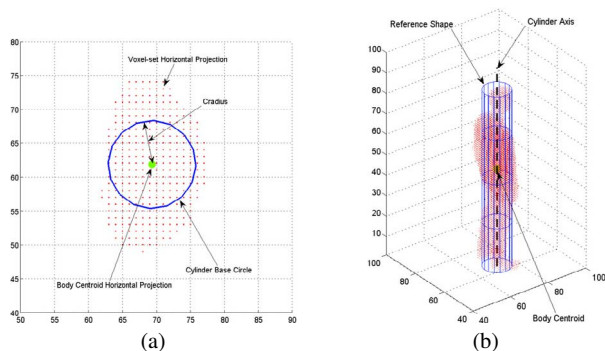
Fig. 2. *Example of Shape Descriptor reference shape. In (a) the body horizontal projection silhouette is used to adapt the base circle. In (b) it is shown the reference cylinder surface. Each voxel is represented only by its center point.*

the cylinder allows discriminating between different orientations of the object (body) with respect to the horizontal plane. The base of the adopted cylinder is the *major circle inscribed inside the projection of the body ON-voxels on the horizontal plane* (Fig. 2 (a)). The main advantages of this choice will be explained later.

Once the reference shape surface is gauged on the current voxel-set, we are able to apply the 3-D Shape Descriptor algorithm: we sample the reference cylinder surface into a number S of **control points** (p_s, $s \in \{1, \ldots, S\}$). S is a user-defined parameter chosen according to computational cost and representation accuracy criteria.

For each control point p_s:

- Define a spherical coordinates system (ρ, θ, φ) with ori-

gin fixed in the p_s location where: $0 \leq \rho \leq \rho_{max}$, $0 \leq \theta \leq \pi$ rad and $0 \leq \varphi \leq 2\pi$. $\theta = 0$ corresponds to the vertical direction, $\varphi = 0$ is the direction of the segment orthogonal to the cylinder axis passing through p_s and ρ_{max} is a value higher than the maximum distance of voxels from the control points.

- Sample uniformly the polar coordinates into parts, respectively S_ρ, S_θ and S_φ. This way we obtain a set of coordinate values $\{(\rho_i, \theta_j, \varphi_k)\}$.

- Assign to p_s a 3-D histogram f_s initially represented by a zero-valued matrix with $S_\rho \times S_\theta \times S_\varphi$ dimensions.

- For each elementary volume in spherical coordinates, defined by a particular $(\rho_i, \theta_j, \varphi_k)$, count how many ON-voxels are contained and store this number in the corresponding histogram location $f_s(i, j, k)$.

3-D *Shape Descriptor* $F(i, j, k)$ is obtained summing up the corresponding values taken from all the histograms of the control points and normalizing these quantities to the maximum value obtained:

$$F(i, j, k) = \sum_{s=1}^{S} \frac{f_s(i, j, k)}{\max_{\bar{i}, \bar{j}, \bar{k}} \left(\sum_{l=1}^{S} f_l \left(\bar{i}, \bar{j}, \bar{k} \right) \right)}$$

The Shape Descriptor $F(i, j, k)$ is invariant with respect to **body translations** in the voxel-set cartesian frame of reference. The reference cylinder, in fact, *follows the body centroid movements.* Furthermore, the *use of control points lying on the cylindrical surface* allows invariance with respect to **body rotations** on the cylinder axis. The particular procedure we used to *adapt the reference cylinder to the human body* ensures an invariance with respect to the **body proportions** of the actor who is performing the posture. The *final normalization of the Shape Descriptor values* removes the proportional relation to how many pixels the body volume is made up and possible effects due to **different sizes of volumes** in spherical coordinates, derived from the use of different reference cylinders.

After having computed the cylindrical surface, the cylinder follows the motion of the body's centroid but its size remains unchanged for the rest of the sequence. This way we obtain an harmonious variation of features throughout the motion. Following the described method, we compute a Body Shape Descriptor $F(i, j, k)$ for each frame and the collection of all its values ($S_\rho \times S_\theta \times S_\varphi$ values) in vectors (*feature vectors*) throughout a sequence (*feature matrix*) is the data set that we use to represent a gesture (six examples are shown in Fig. 3, where each action begins and ends with the same standing up position, with arms hanging on the hips). In our experiments, as suggested in [5], we sampled each spherical coordinate ten times, obtaining feature vectors with 1000 values.

3. ACTION TEMPLATE MATCHING

In order to evaluate the discriminatory abilities of the extracted features we use one of the simplest template matching algorithms. The *DTW* is a definition of a distance metric for measuring similarity between a known reference pattern and a test pattern. This method accounts for the non-linear distortions that could affect two sequences of features. If we took two gestures, a direct comparison between two feature vectors at a given time would be clearly
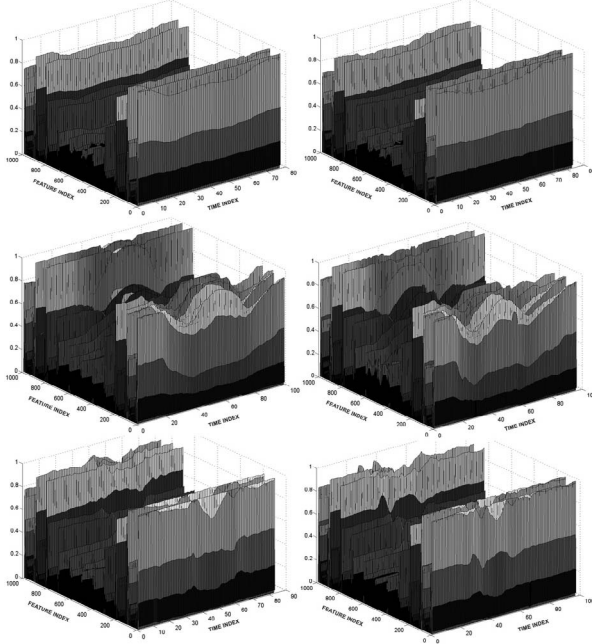
Fig. 3. *Examples of feature matrices. The lower right axis corresponds to the frame index, while the left one is associated to the feature index (a $F(i,j,k)$ value). The upper two matrices are instances of "POINTING AT" gestures, the middle ones are two "CROUCHING DOWN" actions while the lower graphs correspond to two "KICK" sequences.*

impossible: this is mainly due to the different duration of the gesture's steps. It follows that the whole action length has to be considered (Fig. 3). Through DTW we are able to find optimal correspondences between feature vectors of different matrices according to an agreed cost function. In other words, we can compare sequences of similar body postures in two actions independently from their time extension.

3.1. Dynamic Time Warping

If we have a reference pattern, say $r_i, i = 0, \cdots, I$, and a test pattern $t_j, j = 0, \cdots, J$, where, in the general situation, $I \neq J$, we can find a distance measure between the two sequences building a 2D grid with points on respective axis assigned to their feature vectors. Each node (i, j) is associated with a specific value of a cost function $c(i, j)$ measuring the "distance" between the respective elements of the strings, r_i and t_j. We are now looking for a path through the grid from an initial node (i_0, j_0) to a final one (i_F, j_F) that minimize the overall cost C defined as:

$$C = \sum_{k=0}^{F} c(i_k, j_k)$$

In order to obtain the optimal path with the overall minimum cost, we apply the *Bellman's Optimality Principle* [6]: for each node of the grid (i_k, j_k) we only have to find a node (i_{k-1}, j_{k-1}), from a list of possible predecessors, that leads to minimum cost:

$$C_{min}(i_k, j_k) = \min_{i_{k-1}, j_{k-1}} [C_{min}(i_{k-1}, j_{k-1}) + \\ + c(i_k, j_k | i_{k-1}, j_{k-1})]$$

Using this formula we can compute the so-called Minimum Distance Grid (Fig. 4), in which every node is now associated to the minimum cost from the initial node.

In this work we consider $(i_0, j_0) = (0, 0)$ and $(i_F, j_F) = (I, J)$, which means that we are searching for the optimal path from the initial node to the node corresponding to final feature vectors of both sequences. Thus, $C_{min}(i_F, j_F)$ represents the distance between the two sequences. We assume that $c(i_k, j_k | i_{k-1}, j_{k-1}) = c(i_k, j_k)$. $\{(i_k - 1, j_k), (i_k - 1, j_k - 1), (i_k, j_k - 1)\}$ is the set of possible predecessors (i_{k-1}, j_{k-1}).

4. EXPERIMENTAL RESULTS

We tested the DTW-based distance metric with different instances, performed differently by the same person or by another one, of the three simple actions: "POINT AT", "CROUCH DOWN" and "KICK". With the word "simple" we refer to actions that are not repeated for a random number of times, therefore different instances must contain corresponding feature vectors. This is the real limit of this DTW-based template matching approach: we are able to make a comparison between two feature matrices provided that the pattern of feature vector values in a matrix, although stretched or compressed in time, has the same order of subsequence with respect to the one in the other matrix.

We have carried out the DTW procedure using two different cost functions between feature vectors:

The Euclidian distance : $c(i_k, j_k) = \|\mathbf{x}_{i_k} - \mathbf{y}_{j_k}\|$

The Kullback-Leibler distance : $c(i_k, j_k) = KL(\mathbf{x}_{i_k}, \mathbf{y}_{j_k})$

We applied K-L distance to compute the cost function between two feature vectors, $\mathbf{x}$ and $\mathbf{y}$, formed by posture-dependent Shape Descriptor values, using this formula:

$$\mathbf{KL}(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{l=1}^{\#} (x_l - y_l) \ln \left(\frac{x_l}{y_l} \right)$$

where $\# = S_\rho \times S_\theta \times S_\varphi = 1000$ in our case.

(a) (b)
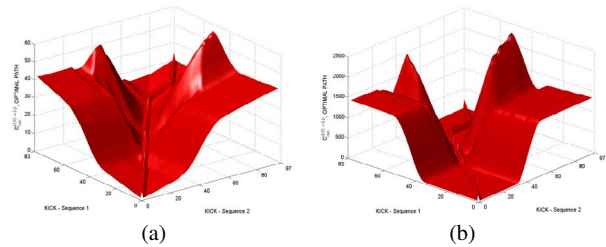
Fig. 4. *Example of DTW minimum distance grid and optimal path computed using Euclidian distance (a) and K-L distance metric (b) between feature vectors. The minimum distance grid and the overall optimal path (the one across the valley) are computed between the two "KICK" feature matrices of Fig. 3.*

An example of minimum distance grid, with the corresponding path leading to minimum overall cost, between the two "KICK" matrices of Fig. 3 computed using Euclidian norm can be found in Fig. 4 (a), while in Fig. 4 (b) the same grid is obtained using the K-L distance. The minimum distance grid computed using the Euclidian norm has less steep slopes that the one computed with K-L distances. This behaviour is due to the K-L computation: it has more discriminating power between Shape Descriptors corresponding to different postures, while retaining the same closeness for Shape Descriptors representing the same posture. Thus, when we compute the minimum cost paths from $(0,0)$ to any (i,j), and the Shape Descriptors compared are different, the minimum cost raises in a few steps: this creates the steep slopes. Using Euclidian norm all the costs associated to the nodes are lower and the distance computed between Shape Descriptors raises more slowly as their difference increases.

The first example of DTW-computed distances between sequences of actions can be seen through a plot of the so called *confusion matrix*. This is a matrix in which each element (n,m) has the DTW-computed distance value $(C_{min}(i_F, j_F))$ from the sequence n to the sequence m.

(a)   (b)
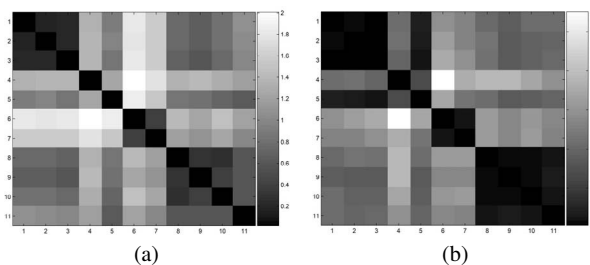
Fig. 5. *Confusion matrix computed using DTW procedure with Euclidian distance (a) and with K-L distance metric (b) between feature vectors. These are distances among 11 sequences:* $\{1, 2, 3, 4, 5\}$ = *"POINT AT";* $\{6, 7\}$ = *"CROUCH DOWN";* $\{8, 9, 10, 11\}$ = *"KICK". The distance values correspond to the black-white scale on the side.*

In Fig. 5 (a) there is a confusion matrix computed using **Euclidian norm** as feature vectors cost function. Elements from 1 to 3 correspond to "POINT AT" actions: we can see that the minimum distances between each one of these sequences and another one (notice that the distance of a sequence from itself is zero, hence the black main diagonal) are concentrated inside the "POINT AT" cluster (3×3 dark upper-left sub-matrix). Sequences 4 and 5 are two more "POINT AT" sequences, this time performed by another person. It is noticeable that these "POINT AT" sequences tend to fall outside clusters boundaries. More precisely, the fourth action has the same distance (fourth white-light grey column or row) from each of the others and the fifth seems to be closer to "KICK" instances. The farthest ones from these sequences are the "CROUCH DOWN" actions (white and light grey columns or rows) while the "KICK" gestures are a bit closer (grey sub-matrices). The same behavior is underlined by the other two clusters represented by the elements 6, 7 for "CROUCH DOWN" action (note the central dark square) and the elements from 8 to 11 for "KICK"(lower-right corner square). Sequences 10 and 11 are related to another actor: the eleventh is closer to the other "KICK" sequences but it is even very close to "POINT AT" actions. An explanation of

the above described phenomena could be twofold: the adaptive technique performed on the reference shape (in order to make the features person-independent) could still be non optimal, and, at the same time, each actor could probably perform the same action in a dramatically different fashion.

All the problems highlighted by the confusion matrix of Fig. 5 (a) are solved using **Kullback-Leibler distance** when computing the minimum distance grids. The results are shown in Fig. 5 (b). Using K-L, the distances between realizations of different gestures increase with respect to the one between sequences with the same action. The dark upper 5×5 matrix, corresponding to the five "POINT AT" sequences, forms now a well defined cluster, with evident boundaries. The first three sequences form a black sub-matrix showing their closeness. The fourth "POINT AT" realization has always farther distances with respect to the other sequences in the "POINT AT" cluster, but now the lowest values are inside the cluster boundaries. The fourth "KICK" sequence has now an highest difference between distances from the "POINT AT" realizations and distances from the other "KICK" actions.

5. SUMMARY AND CONCLUSIONS

In this paper we proposed an action-clustering system based on volumetric 3D data. The performance shown by the experiments have highlighted the abilities of this system based on *Shape Descriptor* not only to recognize postures, as shown in [5], but also to be tuned up in a dynamic context. The simulations that have been carried out demonstrated the ability of the proposed method in classifying the different considered actions. The K-L distance in the DTW context has proved its capability of being a suitable distance metric between posture-dependent Shape Descriptor outputs. Moreover, the Shape Descriptor algorithm can be parallelized and, in our opinion, after an optimization procedure, it will reach real-time performance.

6. REFERENCES

[1] J.J. Little and J.E. Boyd, "Recognizing people by their gait: the shape of motion," *Journal of Computer Vision Research*, vol. 1, no. 2, 1998.

[2] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman, "Articulated body posture estimation from multi-camera voxel data," *IEEE Proc. of the Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 455–460, 2001.

[3] A. Laurentini, "The Visual Hull concept for silhouette-based image understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence PAMI'94*, vol. 16, no. 2, pp. 150–162, February 1994.

[4] F. Cuzzolin, A. Sarti, and S. Tubaro, "Invariant action classification with volumetric data," *IEEE Workshop on Multimedia Signal Processing MMSP'04*, pp. 395–398, September 2004.

[5] I. Cohen and H. Li, "Inference of human postures by classification of 3d human body shape," *IEEE Proc. of Intl. Workshop on Analysis and Modeling of Faces and Gestures (AMFG'03)*, pp. 74–81, October 2003.

[6] S. Theodoridis and K. Koutroumbas, *Pattern Recognition - second edition*, Elsevier Academic Press, 2003.