CLASSIFICATION OF HUMAN BODY ACTIONS BY INVARIANT BODY SHAPE DESCRIPTOR

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

We propose a human body action classifier based on a 3D representation of the body in terms of volumetric coordinates. Features representing body postures are extracted directly from *3D data*, making the system inherently insensitive to viewpoint dependence, motion ambiguities and self-occlusions. An *Invariant Shape Descriptor* of human body is obtained in order to capture only posture-dependent characteristics, despite possible differences in translation, orientation, scale and body size. Frame-by-frame descriptions, generated from a gesture sequence, are collected together in matrices. Clustering of action matrices is eventually performed, and through DTW (*Dynamic Time Warping*) (while computing the distance metric), we gain independence from possible temporal nonlinear distortions among different instances of the same gesture.

1. INTRODUCTION

Systems that are able to recognize human gestures and actions, without any invasive device, have recently raised a great deal of interest not only in the research community, but also for industrial applications. All these techniques could have direct applications to video surveillance problems [1], human-computer gestural interaction projects, robot skill learning and to all fields in which activity recognition is needed. Multi-camera systems are considered nowadays among the most promising techniques used in computer vision. 3D reconstructions derived from different views are inherently able to solve ambiguities and viewpoint dependencies, which are unavoidable in systems based on monocular views (see [2]. In this paper we consider volumetric 3D reconstruction of a moving human body, in terms of voxel occupancy in an assigned voxelset. This is the starting point for developing a reliable set of features representing an actor performing a natural gesture.

A good selection of salient features from voxels coordinates of the “actor’s body” has a great importance for the overall performance of the recognition system. In order to succeed in obtaining a robust representation of an action, we developed a feature extraction method similar to [3], based on a spherical *Shape Descriptor* obtained from a sampled shape function, a cylinder, adapted on the fly to the size of the body. Features are invariant to scale, translation and rotation and constitute a meaningful representations of body postures. These features vary continuously with body motions. The recognition stage is then performed through a clustering of different instances of gestures formed by a collection of shape distributions, one for each considered time instance. Distance metric between sequences of features is computed through the use of *Dynamic Time Warping*, a method that accounts for possible nonlinear distortions in action delivery speed.

1.1. Previous Work

In the past few years a great deal of research has been done in the field of activity recognition with 3D data, for examples see [4] and [5]. Major effort has been put into the research of invariant features with respect to viewpoint and trajectory variations (see [6]). The classifier design is an important part in recognition system projects, but it cannot be considered separately from the evaluation of features. Many recognition methods have been proposed, most of them based on HMMs (Hidden Markov Models) theory, such as in [7], [8] and in [9]. In our approach we decided to adopt a simpler recognition algorithm, which is more computationally efficient and is able to exploit the discrimination properties of our features.

2. DATA ACQUISITION AND FEATURE EXTRACTION

2.1. 3D Volumetric Reconstruction

In order to have a 3D reconstruction of the moving body into the scene, we apply the so called *Volumetric Intersec-*

tion method (see [10], [11]). Starting from eight different viewpoints, represented by eight synchronized cameras, we compute, frame by frame, the segmentation of body silhouettes using a *Chroma Keying* algorithm [12]. Then, in a virtual 3D environment, we build the generalized cones starting from the optical center of each camera and intercepting each respective silhouette. The volumetric intersection of these cones, called *Visual Hull*, approximates the 3D reconstruction of the actor and, sampling its convolution with a smoothing filter, can be transformed in a 3D representation compound of voxels coordinates (Fig. 1).

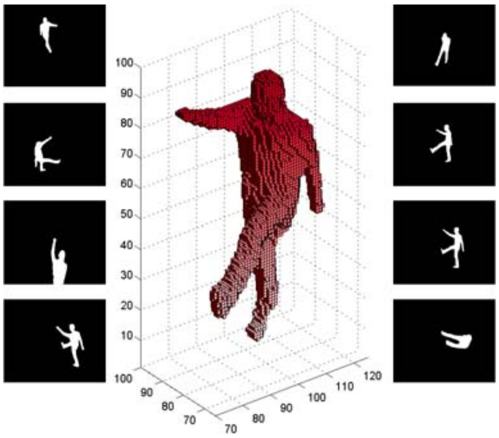

Fig. 1. *Volumetric intersection. Example of voxelsets creation by 3D intersection of Visual Hulls projected from segmented edges. This method is performed for each frame of a gesture action sequence.*

2.2. Invariant Shape Descriptor Method

The set of features that we extract is an extension of the *Shape Descriptor* explained in [3], already used to infer a body posture in a static environment. In this work we propose an adaptation of the method to a dynamic context: a collection of postures across time.

Let us describe the general *Shape Descriptor* applied to a volumetric voxelset:

- *Shape Descriptor* describes a 3D volumetric object with regard to a *reference shape*, Θ: normally a surface like a cylinder or a sphere is used.
- The surface of the reference shape is sampled regularly in a sufficient number N of points, called *control points*, according to some empiric criteria.
- For each control point, P_n:
  - Each voxel is encoded in a spherical frame of reference centered in P_n with dimensions ρ (from 0 to a suitable value), θ (from 0 to π rad) and φ (from 0 to 2π rad).
  - Each polar coordinate is uniformly sampled into ten parts, obtaining a set of 1000 elements $\{(\rho_i, \theta_j, \varphi_k) \ : \ 0 \leq i \ , \ j \ , \ k \leq 9\}$.
  - For each volume in spherical coordinates, defined by a particular $(\rho_i, \theta_j, \varphi_k)$, we count the voxels contained and build a *spherical histogram* $f_n(i,j,k)$ containing these values (for more details see [3]).
- A spherical *Shape Descriptor* $F(i,j,k)$ is computed summing up all the corresponding values in the histograms of the control points and normalizing all to the maximum value:

$$F(i,j,k) = \sum_{n=1}^{N} \frac{f_n(i,j,k)}{\max_{\bar{i},\bar{j},\bar{k}} \left(\sum_{l=1}^{N} f_l\left(\bar{i},\bar{j},\bar{k}\right) \right)}$$

Using a sphere centered in the body centroid with a radius that is proportional to body's main direction, we obtain a description of the body shape with complete loss of information about position in space, actor's height and 3D orientation. In our project we use, as suggested in [3], a cylinder with the axis crossing the centroid, vertically oriented and fitting the body's height. In our approach, instead of inscribing the body inside the reference shape, we optimized the cylinder radius using a suitable value. The used value is the radius of the major circle inscribed inside the projection of the entire voxelset on the floor (Fig. 2 right). This way we obtain a representation that is independent from position, size, scale, body proportions and, possibly, invariant to rotations on its own axis. We call it *Invariant Body Shape Descriptor*.

We would like to point out an important aspect that confirms the rotational invariance of the shape descriptor: for each polar reference frame, centered in its respective control point, we assume as zero-elevation and zero-azimuth the direction of the segment lying on the horizontal plane (zero-elevation) projecting the control point on the cylinder axis.

Following the described method, we compute an *Invariant Body Shape Descriptor* for each frame and the collection of these 1000×1 vectors throughout a sequence is the data set that we use to represent a gesture (six examples are shown in Fig. 4).

3. ACTION CLUSTERING STAGE

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

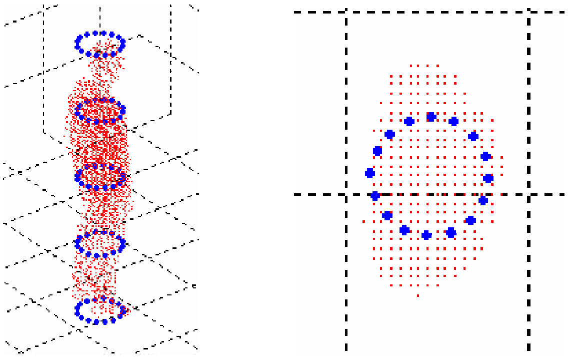

Fig. 2. *Cylindrical Reference Shape.* Left*: Example of a cylindrical reference shape adapted to body proportions. The voxelset is here sub-sampled by a rate of 4 and each voxel is represented only with its center in order to make internal points visible.* Right*: Cylinder radius is adapted to the major circle inscribed inside planar projection.*

and a test pattern. This method accounts for the non-linear distortions that could affect two sequences of features. If we take two gestures, a direct comparison between two feature vectors at a given time is clearly 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. 4). 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 index.

DTW is based on the *Dynamic Programming* theory. 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:

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

Using this formula we can compute the so-called Minimum Distance Grid (Fig. 3-left), in which every node is now associated to the minimum cost from the initial node. This matrix is computed incrementally in such a way that its node (i_F, j_F) contains the minimum cost $C_{min}(i_F, j_F)$ to reach the final node starting from the initial one, (i_0, j_0). Besides, we can take into account each optimal predecessor for each node of the grid in order to be able to construct the optimal path backtracking from (i_F, j_F) (Fig. 3-right).

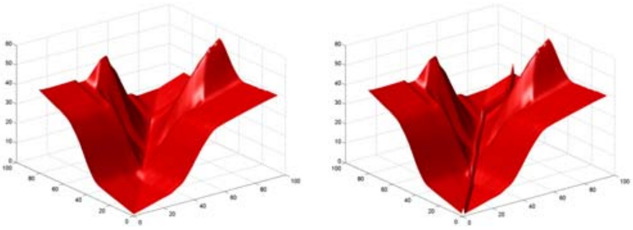

Fig. 3. *Dynamic Time Warping. Left: Minimum Distance grid between the two "KICK" sequences of Fig. 4. Right: the same grid with the overall optimal path (the one across the valley from* $(0, 0)$ *to* (I, J)*)*

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. Note that each sequence is composed of an isolated instance of a single action.

4. EXPERIMENTAL RESULTS

We tested the system with different instances, performed differently by the same person or by another one, of three simple actions: "POINTING AT", "CROUCHING 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. For each gesture we collected at least two different realizations. This constraint avoids problems due to the low-level comparison made by the DTW. Only by computing a statistical model of a gesture we can get rid of this limitation.

The first recognition can be made as shown in Fig. 4, where similarities between instances of the same action are quite apparent.

Using the DTW algorithm we built a matrix in which each element (n, m) has the distance value from the sequence n to the sequence m (Fig. 5). In Fig. 5(left) elements 1, 2, 3 correspond to "POINTING 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 "POINTING AT" cluster (3×3 dark upper-left sub-matrix). The farthest ones from these sequences are the "CROUCHING 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 4, 5 for "CROUCHING DOWN" action

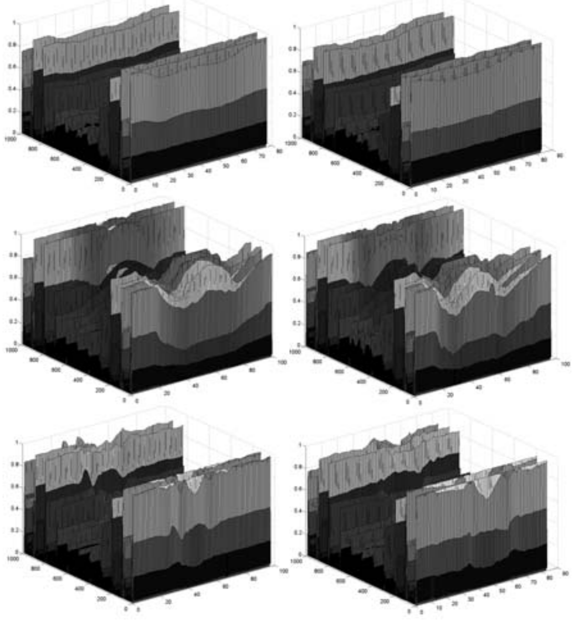

Fig. 4. *Examples of feature matrices. The upper two matrices are instances of “POINTING AT” gestures, the middle ones are two “CROUCHING DOWN” actions while the lower graphs correspond to two “KICK” sequences.*

(note the central dark square) and the elements 6, 7 for “KICK”(lower-right corner square). The only difference among distances from “CROUCHING DOWN” is that “KICK” gestures are a bit closer (grey columns or rows) than “POINTING AT” ones. In conclusion “KICK” has an intermediate position between “POINTING AT” and “CROUCHING DOWN” according to DTW-computed distance.

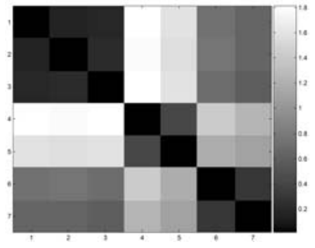

Fig. 5. *Distances between sequences: Comparison among 7 sequences: {1, 2, 3} = “POINTING AT”; {4, 5} = “CROUCHING DOWN”; {6, 7} = “KICK”.*

5. SUMMARY AND CONCLUSIONS

In this paper we proposed an action-clustering system based on volumetric 3D data. Features have been represented by a *Shape Descriptor* computed frame-by-frame and adapted in order to be independent from position, size, scale, body proportions and, possibly, be invariant to rotations. We used a rather simple, but robust, pattern recognition algorithm, *Dynamic Time Warping*, to compute distances among gestural actions.

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 [3], but also to be tuned up in a dynamic context. The simulations that have been carried out have demonstrated the ability of the proposed method in classifying the different considered actions.

6. REFERENCES

[1] Y. Ivanov, C. Stauffer, A. Bobick and W. E. L. Grimson, “Video surveillance of interactions,” *In IEEE Proc. of the CVPR’99*

[2] D. DiFranco, T. Cham and J. Rehg, “Reconstruction of 3D figure motion from 2D correspondences,” *In IEEE Proc. of CVPR’01*

[3] I. Cohen and H. Li, “Inference of human postures by classification of 3D human body shape,” *In IEEE Proc. of AMFG’03*

[4] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick and A. Pentland, “Invariant features for 3-D gesture recognition,” *In IEEE Proc. of FG’96*

[5] S. Seitz and C. Dyer, “View-invariant analysis of cyclic motion,” *International Journal of Computer Vision*, vol. 25, num. 3, pp. 231-251, 1997.

[6] C. Rao, A. Yilmaz and M. Shah, “View-invariant representation and recognition of actions,” *International Journal of Computer Vision*, vol. 50, num. 2, pp. 203-226, 2002.

[7] F. Cuzzolin, A. Sarti, S. Tubaro, “Invariant action classification with volumetric data”, *In IEEE proc. of MMSP’04*

[8] A. D. Wilson and A. F. Bobick, “Parametric Hidden Markov Models for gesture recognition,” *In IEEE Trans. on PAMI’99*

[9] M. Brand, N. Oliver, and A. Pentland, “Coupled HMM for complex action recognition,” *In IEEE Proceedings of CVPR’97*

[10] Z. Yue, L. Zhao, R. Chellappa, “View synthesis of articulating humans using visual hull,” *In IEEE Proc. of ICME’03*

[11] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2000-2003

[12] O. Grau, T. Pullen, G.A. Thomas, “A combined studio production system for 3-D capturing of live action and immersive actor feedback,” *In IEEE Trans. on CSVT’04*