IMPROVING ACTION CLASSIFICATION WITH VOLUMETRIC DATA USING 3D MORPHOLOGICAL OPERATORS

Eliana Frigerio, Marco Marcon, Stefano Tubaro

DEI - Politecnico di Milano
P.za L. Da Vinci, 32 - 20133 Milano - Italy
e-mail: efrigerio/marcon/tubaro@elet.polimi.it

ABSTRACT

This work deals with the definition of a framework for interpreting, modeling and classifying sequences of body movements into a pre-defined vocabulary of actions. Starting from sequences of volumetric reconstructions of the actor pose in each frame, we split action recognition into three separated tasks. The first task is the representation of the four-dimensional patterns reconstructed from each sequence, the second task is the extraction of motion descriptors, and the third task is the classification into action classes. In particular, we extract the *curve skeleton* from the reconstructed volumes in order to underly the actor movements and to reduce the system dependence from the actor gender and the body shape. The proposed method increases the action recognition rate.

Index Terms— Action recognition, Motion History Volume, Morphological Thinning, Hessian Invariant Descriptor

1. INTRODUCTION

Human action/gesture recognition is an important and challenging topic in Computer Vision, with many fundamental applications including video surveillance, video indexing and social sciences [1]. In this work, we investigate how to build models of human actions for categorization and recognition of simple action classes, independently from viewpoint, actor gender and body sizes. We use sequences of volumetric reconstructions of the actor poses performed in each frame. The choice of a 3D representation has several advantages over a single, or multiple, 2D view representation. A 3D representation: (i) is more informative than simple sets of 2D images, since additional calibration information is taken into account; (ii) is more robust to the object position relative to the devices, as it replaces a possibly complex matching between learned views and the actual observations by a 3D alignment; (iii) allows different device configurations.

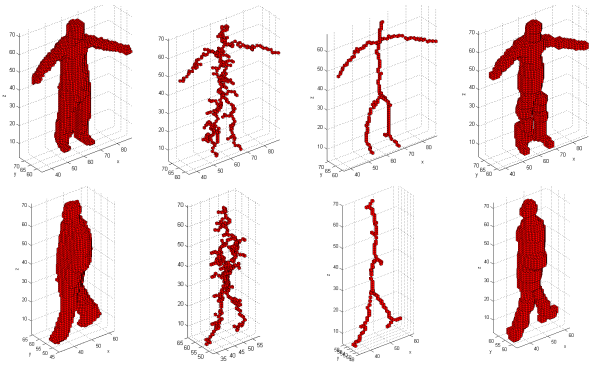
We develop and apply a novel 3D thinning algorithm that extracts the curve skeleton of the reconstructed volume. A curve skeleton is a one-voxel wide representation of a 3D object and it provides a compact and expressive characterization of the solid. We experimentally demonstrate that this tool underlies the actor movements and reduces the system dependence from the actor gender and body shape.

The paper is organized as follows. First, we recall some works related both to action recognition and to skeleton extraction (Section 2). In Section 3, we present an efficient representation of the action pattern based on the proposed thinning algorithm and the Motion History Volume [2]. In Section 4 the extraction of features invariant to position and orientation is described. Section 5 proposes the classification procedure. In Section 6, we discuss the obtained results. Conclusion ends the paper (Section 7).

2. RELATED WORKS

Inside the human action recognition systems that work from multi-viewpoint sequences, a first class of solutions extracts motion descriptors directly from videos. Such descriptors are not invariant to viewpoint, which can be partially resolved by multiplying the number of action classes by the number of possible viewpoints [3], relative motion directions [4], and point correspondences [5]. This leads to a poorer categorization and to an increased complexity. Problems such as viewpoint dependence and motion ambiguities are inherently solved by performing a volumetric reconstruction of the scene prior to the feature extraction and classification stages. A volumetric reconstruction of the human body is used by Trivedi *et al.* for the purpose of fitting a human body ellipsoid model [6]. Even if a gesture can be defined as a time series of body joint configurations (poses), the pose estimation is not strictly necessary in order to classify actions. Weinland *et al.* [2] propose a representation based on Fourier analysis of Motion History Volumes (MHV) in cylindrical coordinates, to build free-viewpoint invariant motion descriptors.

There are many algorithms in the literature describing curve skeleton extracting methodologies for different applications [7]: thinning [8], distance transform-based [9], geometric [10] and general field function-based [11] methods. The thinning process is usually very fast and produces a skeleton that is topologically equivalent to the object. Dur-



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

Fig. 1. Examples of (a) reconstructed volumes from the acquisition system [20]; (b) curve skeletons extracted with the Palágyi and Kuba algorithm [12]; (c) curve skeletons extracted with the proposed algorithm; (d) volumes recovered with the proposed algorithm.

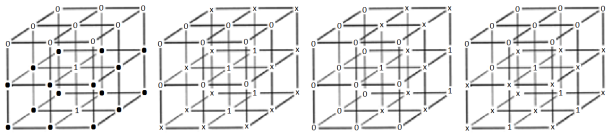
ing the thinning process, border points of a binary object, that satisfy certain topological and geometric constraints, are deleted in an iterative procedure. There are several subclasses of thinning methods based on how detachable points are detected and considered for removal. In particular, the n-Subiteration (or directional) algorithms divide each iteration into n subiterations. In each subiteration, only border points of certain kind can be deleted simultaneously. Since there are six kinds of major directions in 3D (see Section 3), 6-subiteration algorithms were generally proposed [12], [13].

3. ACTION REPRESENTATION

In the discrete space $\mathbb{Z}^3$, each point $\mathbf{p} = (x, y, z)$ is called *voxel*. It can be viewed as a cube, having 6 faces, 12 edges, and 8 corners. The original voxel set (A) is the volumetric reconstruction of the actor pose performed in a frame. This voxel set is approximated and substituted using the proposed *Thinning & Reconstruction procedure*: a ball growing representation computed on the extracted curve skeleton. Computing the MHV, not on the original voxel set (A) (as proposed in [2]) but on the reconstructed volume (R), underlies the body parts involved during the movement and reduces the body shape dependence of the representation.

3.1. Curve Skeleton

3D thinning is a morphological operator which aims at removing external foreground voxels in order to reduce the thickness of objects. The thinning of a set A by a structuring element SE, denoted by $A \otimes SE$, can be defined in terms of the Hit-or-Miss Transform [14]:



(SE_U^1) (SE_U^2) (SE_U^3) (SE_U^4)

Fig. 2. Base structuring elements $SE_U^1 - SE_U^4$ belonging to the deletion direction U.

$$A \otimes SE = A - (A \circledast SE) = A \cap (A \circledast SE)^c, \quad (1)$$

where $\circledast$ denotes the Hit-or-Miss operator. The structuring element origin scans each voxel of A and, if there is perfect overlap between the neighboring voxels with those of the structuring element, the voxel on which the structuring element lays is set to 0 (empty) otherwise it is left at 1 (full). The voxels of a structuring element are described by four kinds of values: "1" means full, "0" means empty, "x" means "do not care" and "." means that at least one point marked "." is full.

The usual process is to thin A using a sequence of structuring elements $\{SE\} = \{SE^1, SE^2, ..., SE^n\}$:

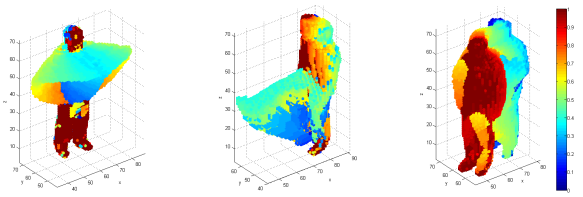
$$A \otimes \{SE\} = ((...((A \otimes SE^1) \otimes SE^2)...) \otimes SE^n). \quad (2)$$

Starting from the idea of Palágyi and Kuba [12], we implement a 6-subiteration 3D thinning algorithm. Our innovative contribution is the definition of new structuring elements in order to obtain a curve skeleton with few peripheral branches (Fig.1 (b)-(c)). We obtain a curve skeleton that is: homotopic with the original object, geometrically centered within the object boundary, close to the object shape, smooth, robust to noise on the surface, and fast to compute.

Considering the canonical space, right handed oriented, we call U and D the extremes, negative and positive respectively, of the y axis, W and E the extremes, negative and positive respectively, of the x axis, and S and N the extremes, negative and positive respectively, of the z axis. The 6 subiterations are sequentially applied following the directions: U, W, S, D, E, and N and iterated until no more points are deleted. The thinning in the U direction, is computed following Eq. 2 with structuring elements $\{SE_U\}$. The structuring elements $SE_U^1 - SE_U^4$ assigned to the direction U are given in Fig.2. Additionally, all rotations of 90°, 180° and 270° around the U-D axis of the base elements $SE_U^1 - SE_U^4$ are structuring elements too. Deletion conditions assigned to the directions W, S, D, E and N, can be derived from the appropriate rotations of the structuring elements in $\{SE_U\}$. Two examples of 3D curve skeletons extracted using the proposed algorithm are shown in Fig.1 (c).

3.2. Volume representation

Once the curve skeleton is extracted, the reconstructed voxel set (R) is computed using a ball growing approach. First the



(a) (b) (c)

Fig. 3. Examples of Motion History Volumes of the actions: (a) open, (b) kick, (c) march.

object surface (S) is extracted with a 3D morphological dilatation with a $3 \times 3 \times 3$ structuring element and then by subtracting the original 3D object (A) to the result. After that, the distance of each skeleton-voxel to the nearest point on the surface is computed and finally the skeleton-voxel is substituted with a discretized ball with center equal to the skeleton-voxel and with radius equal to the estimated distance to the surface. Two examples of reconstructed 3D frames are shown in Fig.1(d).

3.3. Motion History Volumes

Motion History Volume (MHV) represents the extension of Motion History Image, introduced by Bobick *et al.* [3] to capture motion information in images. MHV encodes the history of motion occurrences in the 3D space. Applying the aforementioned T&R procedure, the occupancy function $D(x, y, z, t)$ is obtained: $D = 1$ if the 3D point $\mathbf{p} = (x, y, z)$ is occupied at time t and $D = 0$ otherwise. Considering $D(x, y, z, t)$, the MHV is defined as [2]:

$$v_\tau(x, y, z, t) = \begin{cases} \tau & if D(x, y, z, t) = 1 \\ max\{0, v_\tau(x, y, z, t-1) - 1\} & otherwise \end{cases} . \quad (3)$$

In order to loose the dependency on the absolute execution speed, the templates are normalized with respect to the duration of an action:

$$v(x, y, z) = \frac{v_{\tau=t_{max}-t_{min}}(x, y, z, t_{max})}{t_{max} - t_{min}} \quad (4)$$

where t_{min} and t_{max} are start and end time of an action. Fig.3 shows three examples for Motion History Volumes computed on three sequences representing three different actions.

4. MOTION DESCRIPTORS

Our purpose is to compare body motions that are free in location, orientation and size. The location and scale dependencies can be removed by centering, with respect to the center of the mass, and by scale normalizing, with respect to a constant variance, motion templates, as it is usual in shape matching. The rotation dependence can be removed by using Fourier based features and by choosing coordinate systems that map rotations onto translations [2]. Using invariant motion descriptors is advantageous, because we do not need to align training examples for learning a class model, or to align test examples with all the class prototypes for recognition.

We express the motion templates in a cylindrical coordinate system (r, θ, z):

$$v\left(\sqrt{x^2 + y^2}, tan^{-1}\left(\frac{y}{x}\right), z\right) \rightarrow v(r, \theta, z), \quad (5)$$

where (r, θ, z) are the coordinates of the point $\mathbf{p} = (x, y, z)$ in the cylindrical coordinate system representing respectively radius, azimuth angle, and height. Thus, a rotation around the z-axis (of an angle equal to θ_0) results in a cyclical translation shifts along the azimuth angle axis.

Calling v the volumetric cylindrical representation of a motion template, as defined in Eq.4, we consider the point cloud composed by all the voxels that represent a time step, *i.e.*, for which $v(r, \theta, z) > 0$. We compute the mean μ and standard deviations σ_r and σ_z in r- and z-direction. The template is then shifted, so that $\mu = \mathbf{0}$, and is scale normalized so that $\sigma_z = \sigma_r = 10$. We choose to normalize in z and r directions focusing on the main directions human differ on, as suggested by Weinland *et al.* [2].

In order to construct a feature descriptor that is invariant to a rotation around the z-axis, instead of using only the modulus of the Fourier transform as in [2], we implement a Hessian Invariant Descriptor (HID) on the 1D Fourier transform:

$$V(r, k_\theta, z) = \int_{-\pi}^{\pi} v(r, \theta, z) e^{-j2\pi k_\theta \theta} d\theta, \quad (6)$$

for each value of r and z. The idea behind the HID [15] is to differentiate the phase spectrum twice to eliminate the linear phase terms that is the only difference between a motion template and its rotated counterpart. The invariant parts are then the modulus of the spectrum $|V|$ and the second order partial derivative, respect to k_θ, of the phase spectrum $\varphi_{k_\theta k_\theta}$:

$$F_H(r, k_\theta, z) = \left[|V(r, k_\theta, z)|, \varphi_{k_\theta k_\theta}(r, k_\theta, z)\right]. \quad (7)$$

The main advantage of the HID, with respect to the absolute value of the Fourier Transform alone, is that it does not avoid completely the phase information. The 1D-Fourier magnitude is ambiguous with respect to the reversal of the signal. Consequently, motions that are symmetric to the z-axis (*e.g.*, move left arm-move right arm) result in the same motion descriptors. This is a loss in information that can be avoided using the HID.

5. ACTION CLASSIFICATION

In this section we illustrate the classification process using Linear Discriminant Analysis (LDA) for dimensional reduction [16], combined with a minimum Mahalanobis distance classifier [17].

Table 1. Action classification results (percentages). Results based on PCA [2], LDA, and T&R + LDA methods are presented.

Action	**PCA** (%)	**LDA** (%)	**T&R + LDA** (%)
Open	100	100	100
Kick	100	100	100
Walk	76.19	90.48	100
Crouch	90.48	100	100
Grasp	100	100	100
March	100	90.48	100
PointAt	85.71	95.24	100
Push	95.24	100	100
MoveRt	71.43	80.95	100
Pull	100	100	100
Average Rate	91.9	95.71	100

In order to reduce the vectors dimensionality, we use LDA that projects data through a linear mapping that maximizes the between-class variance while minimizes the within-class variance. When the size of the original feature space D is much larger than the number of vectors available for training, the within-class scatter matrix is singular, so not invertible. To address this problem, we follow the approach proposed by Huang *et al.* [18] which belongs to the class of approaches in the null-space. The basic idea is to maximize the extended Fisher criterion [16]:

$$\hat{F}(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_B \mathbf{w} + \mathbf{w}^T S_W \mathbf{w}}, \tag{8}$$

where S_W is the within-class covariance matrix and S_B is the between-class covariance matrix. Each $\mathbf{w}$, such that $\mathbf{w}^T S_W \mathbf{w} = 0$ while $\mathbf{w}^T S_B \mathbf{w} \neq 0$, maximizes the function $\hat{F}(\mathbf{w})$ [19]. These vectors belong to the null-space of S_W. However not all the null-space of S_W is needed [16]: it is necessary to eliminate the null-space of S_t, which is the intersection between the null-space of S_W and the null-space of S_B. Then, in the complementary subspace, the null space of the new within-class scatter matrix is computed. In this subspace the vectors that maximize the distance between different classes are estimated.

Classification is realized using a Nearest Class Centroid Classifier: it assigns the test vector to the class c of the gallery vector from which the Mahalanobis distance is minimal.

6. RESULTS

We test the proposed action classification system with a database of actions developed in our laboratory, as described in [20]. It is composed by 10 different actions, each performed 5 times by 5 actors (different bodies and different orientations). The dataset is available at [20]. The main assumptions, taking the Efros *et al.* table as a reference [21], are: the subject remains inside the workspace during the whole capturing process; only one subject is present in the workspace at a time; gestures can possibly involve the whole body; no occlusion occurs in the acquired scene except self occlusions. A leave-one-out cross validation is implemented, where we successively use 4 actors (1 sequence for each performer and action) to learn the motions (40 sequences in total) and the other sequences (210 in total) for testing. One actor is not present in the sequences used for learning.

We compare our method with the method proposed by Weinland *et al.* [2] that use MHV and Principal Component Analysis (PCA) for classification. We obtain an average classification rate of 0.9190 with their method, with respect to a rate of 0.9571 using LDA applied on the original MHV and to a rate of 1 using the proposed method (T&R+LDA). Moreover, the original descriptor (Eq.7) is reduced from $D = 262144$ to $d = 9$ components using LDA.

The detailed results, given in Table 1, show, first of all, that the implemented classification method outperforms the classification method based on PCA, because PCA can cut some useful information. Moreover the T&R procedure reduces the body shape dependence, incrementing the similarity between movements even if performed by actors with different gender or different body structure.

7. CONCLUSION

In this paper we propose an Action Classification process, having the 3D human body reconstructions available for each frame. The developed T&R morphological procedure allows to obtain a better representation of the human body, highlighting the movement and reducing the body shape dependence. Its application increments the similarity between movements, even if performed by actors with different gender or different body structure. In order to obtain the invariance under rotation around the z axis, we employ the Hessian Invariant Descriptor: considering the voxel set in cylindrical coordinates, a rotation around the z axis can be tackled considering the module and the phase second order derivative of the 1D Fourier Transform respect to the k_θ axis. We demonstrate that, using the designed morphology operators, the performance greatly improves till 100% success in action classification. We would like to test our method on a bigger database, having a greater number of actions and a bigger number of performers.

8. REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no.

6, pp. 976–990, 2010.

[2] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[3] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[4] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 726–733.

[5] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 984–989.

[6] M.M. Trivedi, K.S. Huang, and I. Mikic, "Dynamic context capture and distributed video arrays for intelligent spaces," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 35, no. 1, pp. 145–163, 2005.

[7] K. Siddiqi and S.M. Pizer, *Medial representations: mathematics, algorithms and applications*, vol. 37, Springer Verlag, 2008.

[8] C. Lohou, "Detection of the non-topology preservation of ma's 3d surface-thinning algorithm, by the use of¡ i¿ p¡/i¿-simple points," *Pattern Recognition Letters*, vol. 29, no. 6, pp. 822–827, 2008.

[9] M.S. Hassouna and A.A. Farag, "Robust centerline extraction framework using level sets," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 458–465.

[10] N. Amenta, S. Choi, and R.K. Kolluri, "The power crust," in *Proceedings of the sixth ACM symposium on Solid modeling and applications*. ACM, 2001, pp. 249–266.

[11] N.D. Cornea, D. Silver, X. Yuan, and R. Balasubramanian, "Computing hierarchical curve-skeletons of 3d objects," *The Visual Computer*, vol. 21, no. 11, pp. 945–955, 2005.

[12] K. Palāgyi and A. Kuba, "A 3d 6-subiteration thinning algorithm for extracting medial lines," *Pattern Recognition Letters*, vol. 19, no. 7, pp. 613–627, 1998.

[13] W. Xie, R.P. Thompson, and R. Perucchio, "A topology-preserving parallel 3d thinning algorithm for extracting the curve skeleton," *Pattern Recognition*, vol. 36, no. 7, pp. 1529–1544, 2003.

[14] R.C. Gonzalez, R.E. Woods, et al., "Digital image processing. 2002," *ISBN: 0-201-18075-8*, 2002.

[15] R.D. Brandt and F. Lin, "Representations that uniquely characterize images modulo translation, rotation, and scaling," *Pattern Recognition Letters*, vol. 17, no. 9, pp. 1001–1015, 1996.

[16] K. Liu, Y.Q. Cheng, J.Y. Yang, and X. Liu, "An efficient algorithm for foley-sammon optimal set of discriminant vectors by algebraic method," *IJPRAI*, vol. 6, no. 5, pp. 817–829, 1992.

[17] R. De Maesschalck, D. Jouan-Rimbaud, and DL Massart, "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[18] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of lda," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. IEEE, 2002, vol. 3, pp. 29–32.

[19] L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[20] M. Marcon, A. Sarti, and S. Tubaro, "http://www-dsp.elet.polimi.it/ispg/index.php/description.html," 2010.

[21] T.B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.