Movement recognition exploiting multi-view information

Alexandros Iosifidis [1], Nikos Nikolaidis [2], Ioannis Pitas [3]

Informatics and Telemtatics Institute, Center for Research and Technology Hellas, Greece
Department of Informatics, Aristotle University of Thessaloniki, Greece
[1] aiosif@aiia.csd.auth.gr
[2] nikolaid@aiia.csd.auth.gr
[3] pitas@aiia.csd.auth.gr

***Abstract*—In this paper a novel view-invariant movement recognition method is presented. A multi-camera setup is used to capture the movement from different observation angles. Identification of the position of each camera with respect to the subject's body is achieved by a procedure based on morphological operations and the proportions of the human body. Binary body masks from frames of all cameras, consistently arranged through the previous procedure, are concatenated to produce the so-called multi-view binary mask. These masks are rescaled and vectorized to create feature vectors in the input space. Fuzzy vector quantization is performed to associate input feature vectors with movement representations and linear discriminant analysis is used to map movements in a low dimensionality discriminant feature space. Experimental results show that the method can achieve very satisfactory recognition rates.**

I. Introduction

Human behavior analysis is an active research field with a wide range of applications. In visual surveillance it can be used to recognize human action patterns and increase robustness and area coverage in security systems. In entertainment industry it can provide high quality multiperspective viewing experiences and 3D scene/actor reconstructions for digital cinema movies and interactive games. Important part in this analysis is the movement recognition procedure. The term movement can be described in various and different ways. In this paper we use the description proposed in [1]. According to this, every movement is composed by a number of dynemes. A dyneme is defined as the elementary constructive unit of movement. In this way every movement can be described in a unique chain of dynemes with some contextual meaning, e.g., a walking step.

In order to be applicable to real world problems a movement recognition system should be able to deal with motion speed changes and style variations both between different subjects performing the same movement and between different realizations of a movement by the same subject [2]. Furthermore the camera position relatively to the observed human body is a determinant factor for such algorithms [3]. The majority of algorithms that have been proposed use one camera and require the same view angle during both training and recognition phases. This angle must ideally be the one that captures the most discriminant motion information and is usually the side one. Such algorithms will fail if the subject under study is captured from a different view and angle or changes motion direction over time. In order to overcome this limitation, researchers have come up view invariant movement representations and recognition approaches. Algorithms that use more than one views have been also proposed, since such information can improve the recognition ability, as indicated in [4].

Recent research results in the area of single-view view-independent human movement recognition are reviewed in [5]. In this review, methods are divided into two categories: State-space and template-based methods. In the first category, each posture specifies one possible state. Hidden Markov Models (HMM) [6][7][8] are often used to describe the transition between successive states, assuming independence between observations. A variant of HMM, Conditional Random Fields (CRF) overcome this independence assumption and experimental results show that they can model dependencies between features and observations better than HMMs [9]. In template-based methods, view invariant features are used to uniquely describe the movements. A computational representation of human movement that captures significant changes in the speed and direction of motion represented by the spatio-temporal curvature of a 2D trajectory is proposed in [10]. This representation is compact and view-invariant. Movements with the same number of instants are compared using a matching criterion and similar movements are grouped together. In [11] contour point correspondences are used to obtain a compact movement representation which is invariant to the camera view angle. Differential geometric surface properties, such as peaks, pits, valleys and ridges, are utilized to produce movement descriptors capturing both spatial and temporal properties. In [12] human movements are represented by three-dimensional shapes formed by the body silhouettes in the space-time volume. The method exploits the solution to the Poisson equation to extract various shape properties, such as local space-time saliency, movement dynamics, shape structure and orientation and produce features to represent and classify movements.

View invariant movement representation and recognition can be also achieved through the usage of a multi-camera set-up. In [13] visual hulls are computed from multiple view video and accumulated within a time period to form Motion History Volumes (MHVs). MHVs are transformed into cylindrical coordinates around their vertical axes and view-invariant features in the Fourier domain are extracted.

In this paper we present a novel view invariant method that utilizes a synchronized multi-camera set-up. The view/camera correspondence problem, namely the identification of the view angle relative to the human body for each camera, is solved using an efficient method based on morphological operations and anthropometric ratios. Using this information, single- or multi-view movement recognition is achieved through fuzzy vector quantization (FVQ) and linear discriminant analysis (LDA), similar to the approach proposed in [14].

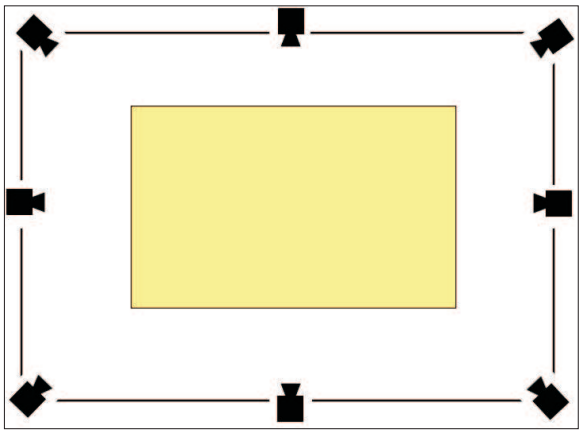

Fig. 1. *A convergent eight-view camera setup and its capture volume.*

II. Proposed Method

The proposed method operates upon data captured by a convergent multi-camera setup such as the eight-camera setup depicted in Figure 1. A collection of frames from all synchronized cameras acquired at the same time instance are referred as a multi-view frame (Figure 2), while a number of consecutive multi-view frames form a multi-view video. The common camera coordinate system is placed at the center of the capture volume and its axes are fixed and known in advance. A subject performing a movement inside the capture volume is observed from all eight cameras and every single-view frame depicts the same movement instance captured from a different viewpoint. The coordinate system rigidly attached to the subject's body is assumed to have axes that are parallel to the cameras coordinate system, while its orientation can vary, depending on subject's movement direction as he may move freely within the view volume.

Clearly, the eight views captured at a certain time instance can depict a different view of the subject as he moves freely in space. For example, whereas camera #1 can at a certain time instance depict a frontal view of the subject, a change in movement direction might result in this camera depicting a side view. Thus, the camera correspondence problem i.e., the identification of the cameras position with respect to the subject's coordinate system, must be solved before the recognition process. Since the positions of the cameras are fixed and known in advance, identifying the location (view angle) with respect to the body of a single camera is enough to derive locations (relative to the body) of all cameras and rearrange them in a consistent manner.

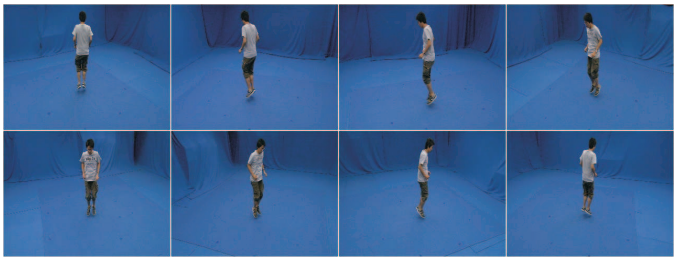

Fig. 2. *A multi-view frame depicting an instance of a jumping movement.*

A. Camera correspondence problem

In order to solve the camera correspondence problem binary body masks are extracted from every single-view frame using a background subtraction [15][16], or chroma keying technique. The bounding boxes (BB) of these masks are evaluated and downscaled. Subsequently, the lower and the upper parts of the body mask (roughly corresponding to head and legs) are rejected. This is done by rejecting the upper 20% and the lower 30% of each BB. On the resulting masks, a number of erosions and dilations are performed in order to reject the limbs and keep only the torso. The number of erosions and dilations is adaptively found. In more detail, for every binary mask, its distance map is computed and the number of erosions and dilations is set equal to half the maximum distance found in the distance map. The result of this procedure can be seen in Figure 3, where bounding boxes of the torsos are depicted in red color.

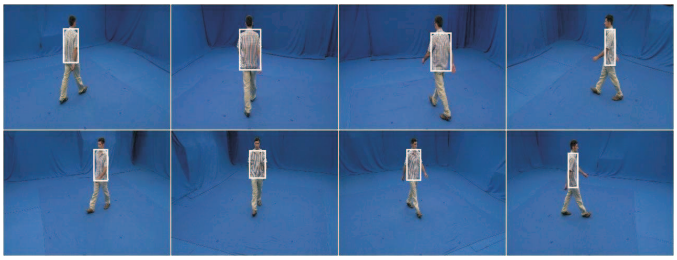

Fig. 3. *Body bounding boxes.*

For every multi-view frame, ratios R between width and height of the resulting torso BBs are computed. Due to the torso proportions the view in which the smallest ratio is observed is labeled as the side view. Since the cameras setup is symmetrical, it is expected that for each pair of opposite facing cameras (e.g. $0^o - 180^o, 45^o - 225^o$) this ratio will be the same. Thus, we cannot distinguish between the right and left

side view, or the frontal and rear view. However, the translation of the center of mass of the torso's BBs signifies direction of motion and can be used to dissolve this ambiguity. In more detail for those views that the smallest R ratio is observed the torso translation is calculated. The right side camera is the one in which a left to right translation is observed. Since the cameras spatial arrangement is a priori known, the function that provides a new view-consistent index f() for each camera whose original index is i, is $f(i) = (d-2+i)_Q$, where d is the original index of the camera identified as the right side view at 90 degrees, Q the total number of cameras and $()_Q$ denotes the modulo Q operator. Through this circular rearrangement, the camera with new index 1 always corresponds to the frontal view.

B. Preprocessing

After the solution of the camera correspondence problem each camera is consistently labeled with the view, in relation to the body, that it depicts. We can perform movement recognition utilizing one or more views, based on a procedure similar to the one described in [14]. The number of used views can vary. All eight available views were included in our implementation. By using more than one views, the information captured from different perspectives is exploited and leads to a better movement representation.

Every binary mask that resulted from background subtraction is centered at the body's center of mass and rescaled in order to produce binary posture frames with the same size. Single-view posture frames corresponding to five movements are shown in Figure 4.

Fig. 4. *Five single-view posture frames of actions walk, run, jump in place, jump forward and bend.*

Multi-view binary posture frames are then created by taking into account the new camera indices, i.e. by placing the frontal camera first, followed by all other cameras in a clockwise manner. An eight-view posture frame from a walking sequence can be seen in Figure 5. The resulting n-view posture frames are scanned column-wise to produce the n-view posture vector, where n denotes the number of views used in training and recognition procedures. In our case, $n = Q = 8$.

Fig. 5. *An eight-view posture frame of action walk. The first posture frame corresponds to the frontal view and the order is clock-wise.*

C. Action Representation

In the training phase the n-view posture vectors of the training sequences are clustered to a fixed number of classes using a fuzzy c-means (FCM) algorithm [17]. This procedure considers unlabeled data and is based on the minimization of the following objective function:

$$\mathbf{J}_{FCM}(\mathbf{\Phi}, \mathbf{V}) = \sum_{c=1}^{C}\sum_{i=1}^{N}(\phi_{c,i})^m(\parallel \mathbf{x}_i - \mathbf{v}_c \parallel_2)^2 , \quad (1)$$

where, N, C are the number of samples and class centers respectively, $\mathbf{x}_i$ is the i-th sample in the training set, $\mathbf{V} = [v_{j,c}] = [\mathbf{v}_1, \ldots, \mathbf{v}_C]$ is the matrix of class centers, $\mathbf{\Phi} = [\phi_{c,i}]$ is the membership matrix with $\phi_{c,i} \in [0,1]$ being the degree by which the i-th sample belongs to the c-th class, $m > 1$ is the fuzzification parameter and $\parallel \cdot \parallel_2$ denotes the Euclidean vector norm.

The resulting cluster centers correspond to dyneme vectors used in subsequent stages. Each dyneme can be thought of as the average of similarly looking body postures. Since no labeling information is used, resulting dynemes can represent movement postures appearing in more than one movements.

After the computation of the dyneme vectors, every posture vector is expressed through its membership vector. This vector denotes the relationship between a posture and the various dynemes.

$$\phi_{c,i} = \frac{(\parallel \mathbf{x}_i - \mathbf{v}_c \parallel_2)^{-2(m-1)^{-1}}}{\sum_{j=1}^{N}(\parallel \mathbf{x}_j - \mathbf{v}_c \parallel_2))^{-2(m-1)^{-1}}}, \quad (2)$$

Finally, every image sequence depicting a single movement is represented with its mean membership vector evaluated over all frames:

$$\mathbf{s}_r = \frac{1}{L_r}\sum_{j=1}^{L_r}\phi_j^{(r)} . \quad (3)$$

where $\phi_j^{(r)}$ is the j-th membership vector of sequence r and L_r is the number of posture frames that correspond to the r-th image sequence. Mean membership vectors are called movement vectors.

In order to discriminate movement classes, labeling information available in the training phase is exploited. A linear discriminant analysis (LDA) algorithm [18] is used to project movement vectors in a discriminant subspace. This projection, represented through the projection matrix $\mathbf{J}_{LDA}$, must maximize the criterion:

$$\mathbf{J}_{LDA} = \underset{\mathbf{G}}{\text{argmax}}\frac{\mid \mathbf{G}^T\mathbf{S}_b\mathbf{G} \mid}{\mid \mathbf{G}^T\mathbf{S}_w\mathbf{G} \mid} \quad (4)$$

$$\mathbf{S}_b = \sum_{c=1}^{C} N_c(\mu^{(c)} - \mu)(\mu^{(c)} - \mu)^T \quad (5)$$

$$\mathbf{S}_w = \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{x}_n^{(c)} - \mu^{(c)})(\mathbf{x}_n^{(c)} - \mu^{(c)})^T \quad (6)$$

where the matrix $\mathbf{G}$ represents a linear transformation, $\mathbf{S}_w$ and $\mathbf{S}_b$ are the within and between class scatter matrices respectively, μ is the mean movement vector of the entire

training set, $\mu^{(c)}$ is the mean movement vector of class c and N_c the number of samples in class c. Movement vectors of the various image sequences are projected with LDA and the average of projected movement vectors of all sequences depicting the same movement (e.g. all walking sequences) is computed to represent this movement class.

$$\mathbf{Z}_r = \frac{1}{L_r} \sum_{i=1}^{L_r} \mathbf{p}_i \ . \tag{7}$$

where $\mathbf{Z}_r$ is the vector representing class r, L_r is the number of samples (image sequences) of class r and $\mathbf{p}_i$ is the movement vectors projected in LDA subspace.

D. Movement Recognition

In the recognition phase, binary masks of n single-view image sequences depicting a subject performing one movement are rearranged/ordered in each time instance according to the procedure described in Section II-A. The frames are also centered with respect to the center of mass of the depicted binary body masks and rescaled to produce single-view posture frames. Ordered single-view posture frames create n-view posture frames which are then vectorized to produce posture vectors. These are expressed in the dyneme space constructed in the training phase by the computation of membership vectors and subsequently projected to the discriminant subspace that resulted by the application of LDA in the training phase. Finally, Euclidean or Mahalanobis distances between the unknown (to be classified) projected movement vector $\mathbf{x}$ and all class vectors $\mathbf{Z}_r$ are calculated:

$$\mathbf{d}(\mathbf{x}, \mathbf{Z}_r) = \| \ \mathbf{x} - \mathbf{Z}_r \ \|_2 \tag{8}$$

These distances are ordered and the unknown movement is classified to the nearest class.

III. Experimental Results

To evaluate the performance of proposed method we tested it to an eight-view video database consisting of eight subjects performing eight everyday movements (walk (wk), run (rn), jump in place (jp), jump forward (jf), bend (bd), sit (st), fall down (fl) and wave one hand (wo)). The video capture took place at the Visual Media Laboratory of the University of Surrey. The capture volume dimensions were approximately $4 \times 3 \times 2$ cubic meters. Additional details can be found in [19].

Camera correspondence problem was solved through the procedure described in Section II.A. Single-view binary masks were extracted by thresholding on the blue color using the HSV color space. Those masks were further processed to produce binary posture frames with size of 64×64 pixels and then were combined to create multi-view posture frames.

The leave-one-out cross-validation (LOOVC) procedure was used to test the performance of the proposed method. In every run seven subjects were used for training and one for testing. This procedure was repeated eight times, one for every subject. Videos were manually segmented in smaller temporal segments, each one consisting of a single movement period, e.g., one walk cycle. Thirty five dyneme vectors, and a fuzzification parameter equal to 1.1 were utilized. The confusion matrices of this experiment are shown in Tables I and II for Euclidean and Mahalanobis distances respectively. Overall correct classification rates of 89.88% for Mahalanobis and 90.88% for Euclidean distances evaluated by averaging all the per class correct classification rate were attained. It can be seen that the only movements that resulted in quite significant recognition errors were jump in place and wave one hand, whereas other movements were perfectly or almost perfectly recognized.

	wk	rn	jp	jf	bd	st	fl	wo
wk	18							
rn	1	18						
jp			43	2		3		
jf				20		1		
bd					8			
st			2			6		
fl							8	
wo		2	4	5				29

TABLE I
Confusion matrix for eight movements using Euclidean distance. A row represents the actual movement and a column the movement recognized by the algorithm.

	wk	rn	jp	jf	bd	st	fl	wo
wk	18							
rn	1	17		1				
jp			38	1		8		1
jf		1		19		1		
bd					8			
st			2			6		
fl							8	
wo			6					34

TABLE II
Confusion matrix for eight movements using Mahalanobis distance. A row represents the actual movement and a column the movement recognized by the algorithm.

IV. Conclusion

In this paper we presented a view-invariant human movement recognition method that exploits information captured by a multi-view camera setup. The correspondence problem is solved using morphological operations and anthropometric ratios. Multi-view representation of movement is achieved by concatenating ordered binary masks extracted from every available view, whereas recognition process involves FVQ and LDA. The usage of a low dimensional feature representation reduces the computational cost. Experiments show that this technique performs action recognition with sufficiently good correct classification rate.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211471 (i3DPost).

REFERENCES

[1] R. D. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images - part i: A new framework for modeling human motion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 179–190, February 2004.

[2] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image and Vision Computing*, vol. 21, no. 8, pp. 729–743, 2003.

[3] S. Yu, D. Tan, and T. Tan, "Modelling the effect of view angle variation on appearance-based gait recognition," in *Proc. Asian Conf. Computer Vision*, 2006, pp. I:807–816.

[4] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on r transform," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[5] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *SMC-C*, vol. 40, no. 1, pp. 13–24, January 2010.

[6] P. Patrick, W. Svetha, and V. Geoff, "Tracking-as-recognition for articulated full-body human motion analysis," pp. 1–8, 2007.

[7] M. S. T. Mori, Y. Segawa and T. Sato, "Hierarchical recognition of daily human actions based on continuous hidden markov models," in *Proc. IEEE Conf. Automatic Face and Gesture Recognition*, 2004, pp. 779–784.

[8] Y. Shi, A. Bobick, and I. Essa, "Learning temporal sequence model from partially labeled data," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. II: 1631–1638.

[9] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[10] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *International Jurnal of Computer Vision and Pattern Recognition*, vol. 50, no. 2, pp. 203–226, 2002.

[11] A. Yilmaz and M. Shah, "Actions as objects: A novel action representation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 984–989.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE Conf. Computer Vision*, vol. 2, 2005, pp. 1395–1402.

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

[14] N. Gkalelis, A. Tefas, and I. Pitas, "Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 18, no. 11, pp. 1511–1521, 2008.

[15] M. Seki, T. Wada, H. Fujiwara, and K. Sumi, "Background detection based on the cooccurrence of image variations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 65–72.

[16] W. G. C. Stauffer, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[17] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[18] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[19] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," in *6th Conference on Visual Media Production*, Nov 2009.