View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation

Nikolaos Gkalelis, Nikos Nikolaidis, Ioannis Pitas

Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece
Department of Informatics, Aristotle University of Thessaloniki, Greece
{galelis,nikolaid,pitas}@aiia.csd.auth.gr

***Abstract*—In this paper a novel method for view independent human movement representation and recognition, exploiting the rich information contained in multi-view videos, is proposed. The binary masks of a multi-view posture image are first vectorized, concatenated and the view correspondence problem between train and test samples is solved using the circular shift invariance property of the discrete Fourier transform (DFT) magnitudes. Then, using fuzzy vector quantization (FVQ) and linear discriminant analysis (LDA), different movements are represented and classified. This method allows view independent movement recognition, without the use of calibrated cameras, a-priori view correspondence information or 3D model reconstruction. A multi-view video database has been constructed for the assessment of the proposed algorithm. Evaluation of this algorithm on the new database, shows that it is particularly efficient and robust, and can achieve good recognition performance.**

I. Introduction

Human behavior understanding using multiple view video streams is a relatively unexplored topic compared to its single-view counterpart. However, multiple view video technology is currently attracting growing attention [1], for instance, in the entertainment industry [2], where it can be used to provide high quality multiperspective viewing experiences and 3D scene/actor reconstructions for digital cinema movies and interactive games, in security applications [3], for view independent non-invasive event detection, and in other areas. Human movement recognition can play an important role in such applications, e.g., by providing "anthropocentric" semantic information for the characterization, summarization, indexing and retrieval of multi-view and 3D data, or for the detection of unusual activities in video surveillance systems.

In such scenarios, a convergent multi-view camera setup (Figure 1) is often used, to simultaneously acquire various views of the same scene and produce a number of single-view video sequences. The set of the synchronized single-view video streams is the so-called multi-view video, and the set of the images (frames) at each time instance is called multi-view image.

The task of motion classification encompasses the recognition of several types of human motion of different complexity, e.g., running or playing soccer. In this work we adopt the taxonomy proposed in [4]. In the lower level of this taxonomy, a dyneme is described as the most elementary constructive unit of motion, while one level above, a movement is conceived as a sequence of dynemes. In this paper we will deal with the task of movement recognition.

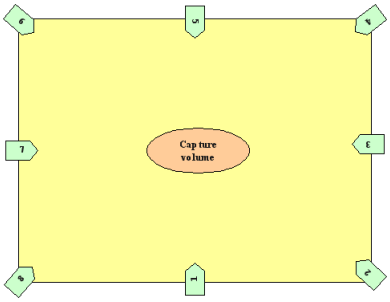

Fig. 1. *A convergent multi-view camera setup.*

Motion classification algorithms are differentiated by the kind of information exploited to describe a human posture in the input space of the recognition task. Several researchers, e.g. [5], use local motion features extracted by applying a tracking algorithm on major body parts. However, these techniques are still not very robust and a certain amount of manual intervention is still needed. An alternative approach is the use of global motion information, e.g. through the extraction of binary body posture masks, as done for example in [6], [7]. This, especially in cases of an almost static background, is a relatively easier task. Here, we concentrate on the recognition of simple human movements using binary body masks, thus exploiting global motion information.

The most recent literature review regarding human action recognition is given in [8]. Since we are interested in view independent movement recognition, a short review of related approaches will be provided. We categorize such approaches with respect to the type of videos (i.e., single or multiple view) used in the training and testing phase of the recognition algorithm.

Methods using single-view videos in both the training and the testing phase employ view invariant feature descriptors, e.g., image moments, to provide a view invariant representation for each movement type [5], [9], [10]. However, these

descriptors are usually invariant only across a relatively small range of viewing angles and thus are not appropriate in cases that full invariance is necessary.

Many researchers, e.g., in [6], [11], based on the fact that the same movement seen from a different viewing direction is considerably different, exploit multi-view videos in the training stage in order to build a model for each movement type and view. In the testing phase, a single-view video stream is used and the computed features are compared with all movement type and view models so as to recognize the unknown movement. Thus, the view correspondence problem, i.e., the estimation of the view angle with respect to the moving person (i.e. frontal, 45 degrees, side view, etc.), for the camera, is implicitly solved. A similar strategy is to use the multi-view videos in the training phase for the computation of a 3D representation of each movement. In the testing phase the view correspondence problem is solved, e.g., by using a probabilistic model and information derived from the camera calibration parameters. Then, the respective single-view representation of the movement that corresponds to the found view is generated and compared to the test video, e.g., as done in [7]. The drawback of these methods, if they are to be used on multi-view videos during testing, is that the discriminant information contained in the non-selected views of the test multi-view video is not exploited. Moreover, the view correspondence problem should be solved implicitly by exhaustive searching or explicitly, e.g., using camera calibration data and a probabilistic model as done in [7].

Finally, methods that use multi-view videos in both the testing and training stage have been proposed. These methods usually compute a 3D model of the moving human and use it to compute free viewpoint features, e.g., by exploiting the circular invariance property of discrete Fourier transform (DFT) in log-polar coordinates as done in [12]. During the testing phase, the 3D model of the moving human in the test video is evaluated and consequently its view invariant feature vector is computed and compared with the respective feature vectors of movement prototypes. However, computation of the 3D moving model is quite expensive and requires calibrated cameras in both the training and testing phase.

Most of the above methods do not fully exploit the rich information contained in the available test multi-view sequences, which may enhance the recognition rate, e.g., as shown in [7]. To exploit the full discriminant information within the multi-view streams, the view correspondence problem between different multi-view images should be efficiently solved, and efficient recognition algorithms should be utilized to allow real-time processing of multi-view data. In this paper we address the above issues and provide a real-time view independent human movement recognition algorithm applicable in cases, where multi-view videos are available. In particular, we implicitly overcome the problem of view correspondence using the circular invariance of the discrete Fourier transform (DFT) and efficiently recognize a number of human movements using a version of the single-view algorithm we proposed in [13].

II. Movement representation and recognition using multi-view videos

A convergent multi-view camera setup consisting of Q cameras is shown in Figure 1. The Q simultaneous video streams produced from the Q cameras, $\{\mathbf{I}_{p,q}\},\ q = 1,\ldots,Q,\ p = 1,\ldots,P$, are referred to as multi-view video where $\mathbf{I}_{p,q}$ is the frame produced from the q-th camera at the p-th time instance. The set of images taken from the cameras at a specific time instance t, $\mathbf{I}_{t,q},\ q = 1,\ldots,Q$, is referred to as multi-view image. The term multi-view movement video is used to denote a video that depicts a person performing a single instance of a particular movement whereas multi-view posture image denotes the set of images that depict the same posture captured from the Q viewpoints.

From each single-view posture image the binary body mask is extracted, and all body masks are preprocessed to create body posture regions of interest (ROIs), which have the same dimensions and are centered with respect to the centroid of the body posture. The posture ROIs are scanned column-wise to produce the so called single-view posture vector $\mathbf{x}_{p,q}^{d} \in \Re^{F}$, where F equals the number of pixels in the ROI, and the index d denotes the 3D posture captured from the d-th viewing angle, and should not be confused with the camera index q. More specifically, the index $d = 1,\ldots,Q$ is used to denote the view angle index with respect to a coordinate system attached to the person. For example, when 8 cameras are used, this index is used to denote the views of the person starting from the front view ($d = 1$), and continuing with the other views in 45 degrees increments and in a clockwise manner, i.e., right side frontal view at 45 degrees ($d = 2$), right side view at 90 degrees ($d = 3$) and so on. Obviously, since the subject may move freely within the view volume, the view index that corresponds to the q-th camera view is unknown and the problem of view correspondence, i.e., which view of the person is captured by each camera, can be defined as finding the mapping $d = f(q)$. It should be noted however that once this correspondence is evaluated for one camera (e.g. the q-th camera) then the correspondences for the rest of the cameras can be easily calculated as $f(q+i) = (d+i)_{Q}, -q+1 \leq i \leq Q-q$, where $()_{Q}$ denotes the modulo Q operator.

All the single-view posture vectors belonging to the same multi-view posture image are concatenated to produce the so called multi-view posture vector $\mathbf{x}_{p}^{d} \in \Re^{N}$:

$$\mathbf{x}_{p}^{d} = \left[\mathbf{x}_{p,1}^{d\ T}, \ldots, \mathbf{x}_{p,Q}^{(d+Q)_{Q}\ T} \right]^{T}, \tag{1}$$

where $N = QF$ and d is the (unknown) view index for the first camera.

A. View invariant multi-view posture vector computation

One can observe that all Q possible configurations, $\mathbf{x}_{p}^{d}, d = 1,\ldots,Q$, of a multi-view posture vector can be obtained by block circularly shifting its elements, with a block corresponding to a single-view posture vector. By denoting the elements of a multi-view posture vector as $x_{p}^{j}(n), n = 1,\ldots,N$,

this circualr shifting can be formally written as $x_p^{\imath}(n) = x_p^{\jmath}((n-|\imath-\jmath|F)_N)$, where $|\cdot|$ denotes absolute value. Based on this observation, the view correspondence problem can be implicitly solved using the magnitude of the discrete Fourier transform (DFT) coefficients of the multi-view posture vector:

$$x_p(k) = |\sum_{n=0}^{N-1} x_p^d(n)\exp^{-j2\pi kn}|,\ k = 1,\ldots,N\,, \quad (2)$$

where $x_p^d(n)$ is the n-th element of the posture vector $\mathbf{x}_p^d$. By concatenating all the N samples of $x_p(k)$ in a vector we get the multi-view posture vector in the DFT domain $\mathbf{x}_p \in \Re^N$. Since the magnitude of the DFT coefficients is invariant to circular shifts, $\mathbf{x}_p$ is view invariant, i.e. it is the same regardless of the ordering of the views in it. The superscript index d has been dropped from $\mathbf{x}_p$ due to its view invariance. This representation enables view-independent human movement recognition, which in practice means that we do not require from the person executing the movement to move in a specific, predefined direction within the camera space, e.g., so that the first camera always captures his/her frontal view.

B. Movement representation and recognition

Let $\mathcal{U}$ be an annotated database of multi-view movement videos, where each video belongs to one of R different movement classes. We represent the $\imath$-th video sequence of the r-th class with length $L_\imath$ as a set of view independent, multi-view posture vectors $\{\mathbf{x}_{\imath,1}^{(r)},\ldots,\mathbf{x}_{\imath,L_\imath}^{(r)}\}$. Thus, the whole set of posture vectors is $\{\mathbf{x}_{1,1}^{(1)},\ldots,\mathbf{x}_{1,L_1}^{(1)},\ldots\ldots,\mathbf{x}_{O_R,1}^{(R)},\ldots,\mathbf{x}_{O_R,L_{O_R}}^{(R)}\}$, where O_r is the number of sequences of the r-th class, in $\mathcal{U}$, therefore, $O = \sum_{r=1}^{R} O_r$, is the total number of movement sequences in the database. View independent, multi-view posture vectors are computed using the magnitude of their DFT coefficients as described in section II-A.

jump1 jump2

Fig. 2. *Overlapping of human actions: jump in place (jump1) and jump forward (jump2).*

Postures in human movements clearly overlap as shown in Figure 2, where jump in place and jump forward can be confused even from a human observer. Therefore, we model each movement as a mixture of densities, where the mixture components are represented by their centers, the so-called dyneme vectors, and use a variant of the single-view recognition methods published in [13] in order to learn and recognize the R different movements. This method initially considers unlabeled data and utilizes the fuzzy c-means (FCM) clustering algorithm with a certain fuzzification parameter m and a number of clusters C to identify the dyneme vectors, $\{\mathbf{v}_1,\ldots,\mathbf{v}_C\}$ and then expresses the multi-view posture vectors with the respective membership vectors, i.e., $\mathbf{x}_\imath \rightarrow \boldsymbol{\phi}_\imath \in \Re^C$. Then, the arithmetic mean of all membership vectors of a multi-view movement video (which will be called movement vector) is used to represent the movement depicted in the video

$$\mathbf{s}_i^{(r)} = \frac{1}{L_\imath}\sum_{\jmath=1}^{L_\imath} \boldsymbol{\phi}_{\imath,\jmath}^{(r)}\,. \quad (3)$$

where $\boldsymbol{\phi}_{\imath,\jmath}^{(r)}$ is the membership vector that resulted from the posture vector $\mathbf{x}_{\imath,\jmath}^{(r)}$, and $\mathbf{s}_i^{(r)}$ is the movement vector representing the $\imath$-th movement video of the r-th movement class in the database.

Therefore, the multi-view movement video database is described by the respective set of movement vectors, $\{\mathbf{s}_1^{(1)},\ldots,\mathbf{s}_{O_1}^{(1)},\ldots\ldots,\mathbf{s}_1^{(R)}\ldots\mathbf{s}_{O_R}^{(R)}\}$. The labelling information available in the training phase can be exploited using a subspace method, e.g. linear discriminant analysis (LDA) or one of its variants, to project the movement vectors in a discriminant subspace and, thus, further reduce the dimensionality of the multi-view movement video feature vectors. Assuming that $\boldsymbol{\Psi} \in \Re^{C\times R-1}$ is the projection matrix computed using LDA in the annotated movement vector database, the LDA-projected feature vector of the $\imath$-th movement video belonging to the r-th class will be given by $\mathbf{y}_i^{(r)} = \boldsymbol{\Psi}^T\mathbf{s}_i^{(r)}$. The r-th movement type can then be represented by the mean of all movement vectors belonging to this movement type, i.e.,

$$\boldsymbol{\zeta}^{(r)} = \frac{1}{O_r}\sum_{\imath=1}^{O_r}\mathbf{y}_i^{(r)}, r = 1,\ldots,R\,. \quad (4)$$

During the testing phase, in order to classify a test movement video we first compute the view invariant posture vectors of the video, as explained in section II-A, and subsequently the movement vector of the video and its projection in the LDA space $\boldsymbol{\tau}$. Then, the video is assigned to the movement type whose mean movement vector $\boldsymbol{\zeta}^{(r)}$ exhibits the smallest Mahalanobis distance or the maximum cosine distance from $\boldsymbol{\tau}$. The number of dynemes C (number of FCM centers) and the fuzzification parameter m are initially not known. The leave one out cross validation (LOOCV) procedure is combined with a global-to-local search strategy, similar to [13], in order to identify C, m, as it will be detailed in the next section.

III. Experimental results

To evaluate the proposed method and due to the limited availability of suitable databases we created a multi-view video database consisting of various everyday human movements, using the high definition (HD) convergent eight-camera setup ($Q = 8$) shown in Figure 1. The video capture was done in a studio with blue backdrop background at the Visual Media Laboratory of the University of Surrey. The capture volume dimensions were about $4 \times 3 \times 2$ cubic metres.

From this database we selected 40 high resolution multi-view videos ($1920 \times 1080, 25$fps), where each video depicts one of eight persons performing one of five movements,

namely, walk (wk), run (rn), jump in place (jp or jump1), jump forward (jf or jump2) and bend (bd). The body binary masks were extracted by thresholding on the blue channel. One frame (single view) and the respective binary mask from each of the five movements are shown in Figure 3.

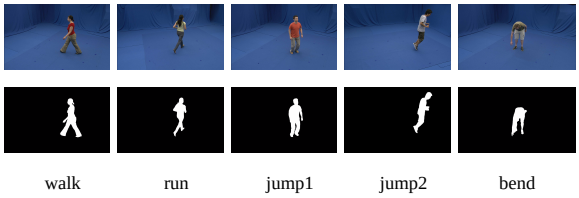

Fig. 3. *One frame and the respective binary mask for the five movements used in the experiments.*

Then, as explained in section II, each single-view binary mask was further processed and scaled with bicubic interpolation to produce ROIs of size 32×64 pixels, which were then scanned column-wise to form 2048-dimensional single-view posture vectors.

We used the LOOCV procedure and a global to local search strategy similar to [13] to identify the optimal parameters C, m, regarding the five movements described above. At each cycle, the video of a specific person executing a specific movement is removed from the training set in order to form the test set. From the remaining videos, those that depict a person performing more that one instances of one particular movement are temporally segmented manually to their constituting single period movements, to create a training set of movement videos, and the five movement prototypes $\boldsymbol{\zeta}^{(r)}$ are computed. Similarly, if necessary, the test video is manually segmented to its constituting single period movement videos. Each of the single period test movement videos are classified, and the final decision for the whole test movement video is taken using majority voting. Using this procedure, the optimal values were found to be $C = 20$, $m = 1.1$. For these values a very good correct classification rate (CCR) of 90% was attained, since only 2 videos were misclassified (i.e., 5% misclassification rate), while 2 videos remained unclassified. The respective confusion matrix is shown in Table I.

	wk	rn	jf	jp	bd	–
wk	7					1
rn		8				
jf			7			1
jp			1	6	1	
bd					8	

TABLE I
Confusion matrix for five movements ($m = 1.1, C = 20$). The last column corresponds to videos that remained unclassified.

IV. Conclusions

A novel multi-view human movement recognition method has been proposed. The method exploits the circularly shift invariance property of the DFT to compute view independent feature vectors of multi-view posture images. This representation is used as input to a movement recognition algorithm that involves fuzzy vector quantization and LDA. The proposed method is in overall very efficient due to the use of FFT and simple nearest centroid classification in a low dimensional feature subspace and achieves good recognition rates in a new database of multi-view videos.

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). The authors would like to thank Prof. Adrian Hilton and Dr. Hansung Kim of the Centre for Vision, Speech and Signal Processing, University of Surrey for providing the capturing studio and equipment and for their valuable assistance in the creation of the multi-view video, within the framework of i3DPost project.

References

[1] J. G. Lou, H. Cai, and J. Li, "A real-time interactive multi-view video system," in *Proc. 13th ACM Int. Conf. Multimedia*, Singapore, Nov. 2005, pp. 161–170.

[2] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 66–76, Nov. 2007.

[3] C. Shen, C. Zhang, and S. Fels, "A multi-camera surveillance system that estimates quality-of-view measurement," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, San Antonio, Texas, USA, Sep. 2007, pp. III–193 – III–196.

[4] R. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images-part I: A new framework for modeling human motion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 179–190, Feb. 2004.

[5] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *Int. Journal of Computer Vision*, vol. 66, no. 1, pp. 83–101, Jan. 2006.

[6] M. Ahmad and S. W. Lee, "HMM-based human action recognition using multiview image sequences," in *Proc. 18th Int, Conf. Pattern Recognition*, vol. 1, Hong Kong, China, Aug. 2006, pp. 263–266.

[7] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. 11th IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–7.

[8] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[9] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[10] I. Junejo, E. Dexter, I. Laptev, and P. Perez, "Cross-view action recognition from temporal self-similarities," in *Proc. 10th European Conf. Computer Vision*, Marseille, France, Oct. 2008.

[11] A. Ogale, A. Karapurkar, and Y. Aloimonos, "View-invariant modeling and recognition of human actions using grammars," in *Workshop on Dynamical Vision at 10th IEEE Int. Conf. Computer Vision*, Beijing, China, Oct. 2005.

[12] D. Weinland, R. Ronfard, and E. Boye, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, no. 2-3, pp. 249–257, Nov-Dec 2006.

[13] 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 Technol.*, vol. 18, no. 11, pp. 1511–1521, Nov. 2008.