VIEW-INDEPENDENT HUMAN ACTION RECOGNITION BASED ON MULTI-VIEW ACTION IMAGES AND DISCRIMINANT LEARNING

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Greece
{aiosif,tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper a novel view-independent human action recognition method is proposed. A multi-camera setup is used to capture the human body from different viewing angles. Actions are described by a novel action representation, the so-called multi-view action image (MVAI), which effectively addresses the camera viewpoint identification problem, i.e., the identification of the position of each camera with respect to the person's body. Linear Discriminant Analysis is applied on the MVAIs in order to to map actions to a discriminant feature space where actions are classified by using a simple nearest class centroid classification scheme. Experimental results denote the effectiveness of the proposed action recognition approach.

Index Terms— Human Action Recognition, Multi-camera Setup, Multi-view Action Images, Discriminant Learning

1. INTRODUCTION

Human action recognition is an active research field with a wide range of applications. Some of the most important tasks involving human action recognition include intelligent visual surveillance, content-based video retrieval, human-computer interaction, augmented reality and games. The term action can be described in various and different ways. Its most widely adopted interpretation describes actions as single periods of human motion patterns, e.g. a walking step. Action recognition is not an easy task. Action recognition methods should be able to address issues relating to action execution style and human body size variations appearing between different persons. Furthermore, speed variations between different action realizations and the relative position of the person under consideration and the camera(s) used to capture the needed visual information should not affect the action recognition performance. All the previously mentioned issues result to high intra-class, and possibly low inter-class, variations for human action classes.

Action recognition methods can be categorized depending on the number of cameras used to obtain the available visual information in single-view and multi-view methods. Single-view methods, utilize one camera and, usually, require the same viewing angle during both the training and recognition phases. This angle should, ideally, be the one that captures the most discriminant motion information and is, usually, the side view of the human body. In different cases their performance will, probably, decrease. In order to overcome the known viewing angle assumption of single-view methods, multi-view methods have been proposed. By capturing the human body from multiple viewing angles, a view-independent human body representation can be obtained, leading to view-independent action recognition.

Most multi-view methods proposed in the literature have focused their attention on finding a convenient multi-view human body representation. In [1] human body poses are described by 3D skeletal and super-quadratic body models. In [2], human body poses are described by 2D multi-view posture images, resulted by combining binary body images corresponding to different viewing angles. After obtaining a convenient human body representation, actions are described as series of successive human body poses. Another multi-view approach can be found in [3], where the transition between successive human body poses is directly described by calculating the corresponding 3D optical flow.

An issue that multi-view methods should be able to address is the camera viewpoint identification problem, which is important since actions are quite different when observed by different viewing angles [4] and, thus, it is possible for the same action instance when observed from different viewing angles to be quite different. Taking into consideration the high intra-class variations of action classes, as was previously discussed, the camera viewpoint identification problem may decrease the action recognition performance. Another important issue that multi-view methods should be able to address is related to the action representation dimensionality. Clearly, the use of multi-view human body representations leads to high dimensional action representations. Since action classification usually involves the application of statistical learning

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007 − 2013) under grant agreement number 316564 (IMPART). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

techniques, like Linear Discriminant Analysis (LDA), the dimensionality of action representation may decrease the action recognition performance. This is the well-known Small Sample Size (SSS) problem [5].

In this paper we propose a method aiming at view-independent human action recognition utilizing a multi-camera setup, like the one shown in Figure 2b. Binary body images denoting the video frame locations of the person under consideration during action execution are employed in order to create a novel multi-view action representation, the so-called multi-view action image (MVAI). The proposed multi-view action representation implicitly addresses the camera identification problem by exploiting the circular shift invariance property of the 2D Discrete Fourier Transform (2D DFT). LDA is, subsequently, applied to the MVAI-based action representations of the training action videos in order to determine a low-dimensional discriminant feature space, where action classification can be performed by using a simple nearest action class centroid classification scheme. Since, as it will be discussed in the following, the high MVAI dimensionality leads to low action recognition performance in this setting, we evaluate three LDA-based dimensionality reduction schemes aiming at overcoming the SSS problem. The use of such dimensionality reduction schemes leads to high action classification performance.

The paper is structured as follows. Section 2 describes the proposed action recognition method. Section 3 illustrates experimental results conducted in order to evaluate its performance. Finally, conclusions are drawn in Section 4.

2. PROPOSED METHOD

As it was already mentioned, the proposed method operates on binary videos depicting a person performing an action. Such binary videos can be efficiently obtained by applying image segmentation techniques on the camera frames. In the preprocessing phase, videos depicting multiple action instances are manually segmented to smaller ones depicting one action instance only, e.g., a walking step. In the case of continuous action recognition, i.e., recognition in videos containing multiple action instances, smaller videos are automatically produced by using a sliding window consisting of N_{tw} video frames, in both the training and recognition phases.

2.1. Multi-view Action Images creation

The video frames of each binary video are centered to the human body center of mass, cropped and resized in order to produce binary images of fixed ($H \times W$ pixels) size, the so-called posture images. Posture images are concatenated with respect to time and resized to fixed ($H \times W \times N_t$ pixels) size volumes using linear interpolation. In the experiments presented in this paper the values $H = W = 16$ and $N_t = 13$ have been used. The resulting volumes are, subsequently, split in order to produce single-view action images. This procedure is illustrated in Figure 1. Single-view action images corresponding to all the cameras forming the camera setup are, finally, concatenated using the cameras ID information in order to produce the MVAIs, as illustrated in Figure 2a.

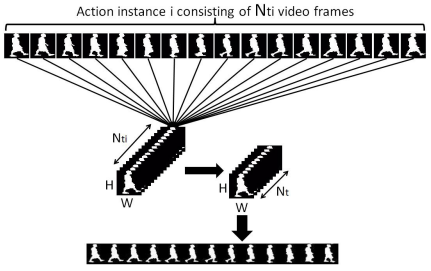

Fig. 1. *Single-view action image creation.*

Since the person under consideration is free to move, the cameras observation angle with respect to the human body orientation is not a priori known. Furthermore, since actions are usually periodic, they can be described by using different starting human body poses and they may have different durations. For example, in a 20fps video, a walking step may be depicted in $10-20$ video frames, while a bend sequence may be depicted in $30-70$ video frames. Two walking steps starting from different human body poses and consisting of different numbers of human body poses are illustrated in Figure 3. As can be observed in Figures 1,2a, the obtained MVAIs are time invariant in the sense that all the action instances are represented as fixed size images. In order to ignore any temporal information relating to the starting human body pose of actions and in order to obtain a view-independent action representation, we exploit the circular shift invariance property of the 2D DFT. Let $\mathbf{Q}_j \in \mathbb{R}^{N_C H \times N_t W}$ denote the i-th MVAI and $[\mathbf{Q}_i]_{n_1,n_2}$ denote the grayscale value of pixel (n_1, n_2). By applying 2D DFT on $\mathbf{Q}_i$, we obtain:

$$[\tilde{\mathbf{P}}_i]_{k_1,k_2} = \sum_{n_1=0}^{N_C H-1} \sum_{n_2=0}^{N_t W-1} [\mathbf{Q}_i]_{n_1,n_2} e^{\left(\frac{-2\pi n_1 k_1}{N_C H} \frac{-2\pi n_2 k_2}{N_t W}\right)}. \quad (1)$$

where $\tilde{\mathbf{P}}_i \in \mathbb{C}^{N_C H \times N_t W}$ is a matrix containing the (com-

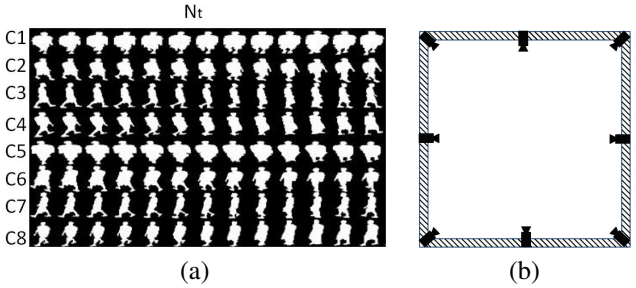

Fig. 2. *a) A multi-view action image resulted by concatenating single-view images corresponding to eight cameras and b) An eight-view camera setup ($N_C = 8$).*

plex) DFT coefficients corresponding to MVAI i. After $\tilde{\mathbf{P}}_i$ calculation, MVAI i is represented by a matrix $\mathbf{P}_i \in \mathbb{R}^{N_C H \times N_t W}$ containing the magnitudes of $\tilde{\mathbf{P}}_i$.

Fig. 3. *Two walking steps starting from different human body poses.*

2.2. Multi-view Action Images classification

Let $\mathcal{U}$ be a video database containing N_T videos depicting persons performing actions belonging to an action class set $\mathcal{A}$ captured by N_C viewing angles. These videos are preprocessed following the above described procedure and $N_I = \frac{N_T}{N_C}$ training MVAIs are obtained, which are represented by the corresponding matrices $\mathbf{P}_i$, $i = 1, ..., N_I$. $\mathbf{P}_i$ are vectorized column-wise in order to produce the so-called action vectors $\mathbf{p}_i \in \mathbb{R}^D$, $D = H \cdot W \cdot N_C \cdot N_t$.

In order to discriminate action classes, the labeling information available in the training phase can be exploited. We employ LDA [6] in order to map action vectors $\mathbf{p}_i$ to a low-dimensional discriminant space. LDA specifies an optimal discriminant subspace by minimizing:

$$\mathbf{W}_{opt} = \arg\min_{\mathbf{W}} \frac{trace\{\mathbf{W}^T \mathbf{S}_w \mathbf{W}\}}{trace\{\mathbf{W}^T \mathbf{S}_b \mathbf{W}\}}. \quad (2)$$

The matrix $\mathbf{W}_{opt} \in \mathbb{R}^{D \times d}$, $d = N_A - 1$, where N_A is the number of action classes forming $\mathcal{A}$, represents a linear transformation and $\mathbf{S}_b$, $\mathbf{S}_w$ are the between-class and within-class scatter matrices of the training action vectors $\mathbf{p}_i$. $\mathbf{W}_{opt}$ is formed by the eigenvectors corresponding to the $N_A - 1$ highest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$ in the case where $\mathbf{S}_w$ is non-singular, or the $N_A - 1$ lowest eigenvalues of $\mathbf{S}_b^{-1}\mathbf{S}_w$ in the case where $\mathbf{S}_b$ is non-singular. After obtaining $\mathbf{W}_{opt}$, discriminant action vectors $\mathbf{z}_i \in \mathbb{R}^d$ are obtained by $\mathbf{z}_i = \mathbf{W}_{opt}^T \mathbf{p}_i$.

However, since $\mathbf{p}_i$ dimensionality is high ($D = 16 \cdot 16 \cdot 13 \cdot 8 = 26624$ in our experiments), an enormous training set should be used in order to avoid singularity of $\mathbf{S}_b$, $\mathbf{S}_w$. In order to address this issue three approaches have been proposed. The first one, is regularized LDA (RLDA) [6] which employs a regularized version of the within scatter matrix, i.e., $\tilde{\mathbf{S}}_w = \mathbf{S}_w + r\mathbf{I}$, where r is a small positive value and $\mathbf{I}$ is the identity matrix. The second one involves a Principal Component Analysis (PCA) based dimensionality reduction step in order to determine a low-dimensional feature space where the corresponding within-class scatter matrix is non-singular. Finally, the third one is Weighted Piecewise LDA (WPLDA) [7], which involves the creation of multiple LDA based projection problems formed by subspaces of the action vector feature space.

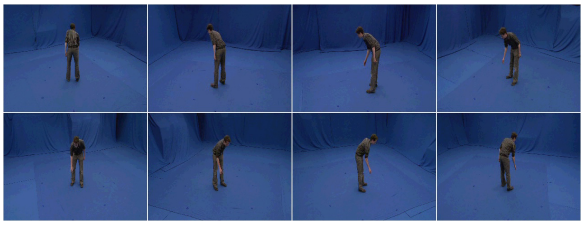

Fig. 4. *Action classification rates on the i3DPost database.*

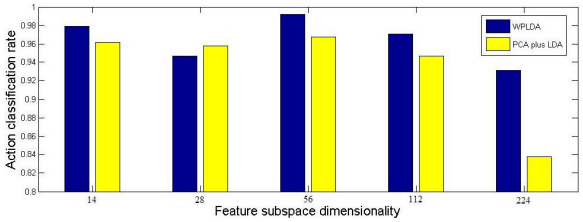

Fig. 5. *Action classification rates on the i3DPost database.*

After obtaining the discriminant feature space resulted by the LDA projection, a test action vector $\mathbf{p}_{test}$ is mapped to the corresponding discriminant action vector $\mathbf{z}_{test}$ by applying $\mathbf{z}_{test} = \mathbf{W}_{opt}^T \mathbf{p}_{test}$ and is assigned the label of the closest action class centroid using the Euclidean distance.

3. EXPERIMENTAL RESULTS

In this section we present experiments conducted on the i3DPost [8] multi-view action recognition database in order to evaluate the performance of the proposed action recognition method. The database camera setup consists of eight cameras (Figure 2b). Eight persons have been captured during the execution of eight actions: walk, run, jump in place, jump forward, bend, sit, fall and wave one hand. Example video frames depicting a person from all the eight viewing angles are illustrated in Figure 4. The Leave-One-Person-Out cross-validation procedure has been performed in order to evaluate the performance of the proposed method to generalize on data that it was not trained on. That is, the algorithms have been trained multiple times (folds) using the MVAIs depicting seven of the persons in the database and tested on the MVAIs depicting the remaining one. Multiple folds, one for each test person, have been performed in order to complete an experiment.

Multiple experiments have been performed by employing all the three LDA variants discussed in Section 2.2. In the cases of WPLDA, multiple experiments have been conducted, using 2, 4, 8, 16, and 32 pieces. In the case of PCA dimensionality reduction, the dimensionality of the resulted feature space was equal to 14, 28, 56, 112 and 224, in order to directly compare its performance with that of WPLDA. By using the regularized LDA algorithm an action classification rate equal to 74.39% has been obtained. PCA followed by LDA based

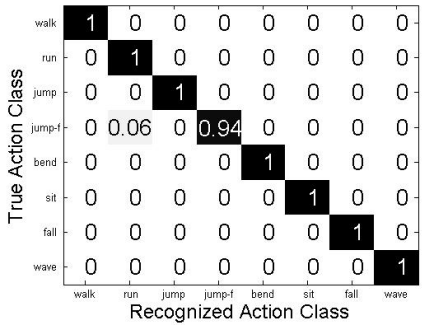

Fig. 6. *Confusion matrix on the i3DPost database.*

Table 1. Comparison results.

Method [2]	Method [9]	Method [3]
94.34%	94.87%	95%
Method [10]	Method [11]	**proposed method**
95.5%	98.44%	**99.22%**

action classification resulted to a best action classification rate equal to 97.56%. As can be seen in Figure 5, the use of a high number of PCA dimensions results in a drop of the obtained action classification rate. This is due to the fact that, in this case, both the within- and between-class scatter matrices in (2) are singular. WPLDA is the overall winner providing a best action classification rate equal to 99.22%. The action classification rate of all the conducted experiments are illustrated in Figure 5. The confusion matrix corresponding to the best obtained action classification rate is illustrated in Figure 6. Finally, in Table 1 we compare the performance of the proposed method with that of other multi-view methods recently proposed in the literature evaluating their performance on the i3DPost database. As it can be seen, the proposed method provides state of the art performance, outperforming all the other methods.

4. CONCLUSION

In this paper we presented a multi-view action recognition method employing a novel multi-view action representation. The proposed action representation implicitly addresses the camera viewpoint identification problem by exploiting the circular shift invariance property of the 2D Discrete Fourier Transform. Action classification is performed by mapping the proposed action representation to a discriminant subspace by applying Weighted Piecewise Linear Discriminant Analysis and following the nearest class centroid classification rule. Experimental results indicate the effectiveness of the proposed action recognition approach.

5. REFERENCES

[1] C. Tran and M.M. Trivedi, "Human body modelling and tracking using volumetric representation: Selected recent studies and possibilities for extensions," in *Second ACM/IEEE International Conference on Distributed Smart Cameras*, 2008.

[2] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347–360, 2012.

[3] M.B. Holte, T.B. Moeslund, N. Nikolaidis, and I. Pitas, "3d human action recognition for multi-view camera systems," in *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, 2011.

[4] S. Yu, D. Tan, and T. Tan, "Modelling the effect of view angle variation on appearance-based gait recognition," in *Asian Conference on Computer Vision*, 2006.

[5] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Boosting linear discriminant analysis for face recognition," in *International Conference on Image Processing*, 2003.

[6] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.

[7] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise lda for solving the small sample size problem in face verification," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 506–519, 2007.

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

[9] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–425, 2012.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view human action recognition under occlusion based on fuzzy distances and neural networks," *European Signal Processing Conference*, 2012.

[11] M.B. Holte, B. Chakraborty, J. Gonzalez, and T.B. Moeslund, "A local 3d motion descriptor for multi-view human action recognition from 4d spatio-temporal interest points," .