MULTI-VIEW HUMAN ACTION RECOGNITION UNDER OCCLUSION BASED ON FUZZY DISTANCES AND NEURAL NETWORKS

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Aristotle University of Thessaloniki, Depeartment of Informatics
Box 451, 54124 Thessaloniki, Greece
Email: {aiosif,tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

While action recognition methods exploiting information coming from multiple viewing angles have been proposed in order to overcome the known viewing angle assumption of single-view methods, they set the assumption that the person under consideration is visible from all the cameras forming the adopted camera setup. However, this assumption is not usually met in real applications and, thus, their applicability is limited. In this paper we propose a novel action recognition method that overcomes this assumption. The method exploits information coming from an arbitrary number of viewing angles. The classification procedure involves Fuzzy Vector Quantization and Artificial Neural Networks. Experiments on two publicly available action recognition databases evaluate the effectiveness of the proposed action recognition approach.

Index Terms— Action Recognition, Multi-camera setup, Fuzzy Vector Quantization, Artificial Neural Networks

1. INTRODUCTION

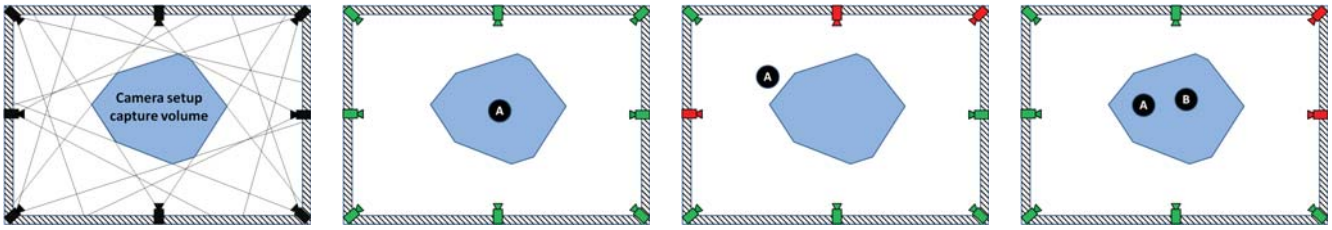
Human action recognition exploiting information coming from multiple viewing angles has been recently proposed in order to overcome the view-dependence restriction of single-view action recognition methods, i.e., of the methods utilizing one camera for action recognition. By capturing the human body, during action execution, from multiple viewing angles, a view-independent human body representation can be obtained, leading to view-invariant action representation and recognition. In order to capture the human body from different viewing angles, camera setups consisting of multiple cameras are employed. An example multi-camera setup consisting of $N_C = 8$ cameras is illustrated in Figure 1a. As can be seen in this Figure, the space that is captured by all the N_C cameras is referred to as camera setup capture volume.

The research leading to these results has received funding from the Collaborative European Project MOBISERV FP7- 248434 (http://www.mobiserv.eu), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly. We would like to thank Prof. K. Lyroudia, Dept. of Dentistry, AUTH, for fruitful discussions regarding eating/drinking activity recognition.

Multi-view action recognition methods can be categorized, depending on the adopted human body representation, in 2D and 3D methods. 2D methods, exploit the 2D image data corresponding to the projections of the human body on the planes of the cameras. In this way, multiple human body representations are obtained, each corresponding to one capturing viewing angle, which are, subsequently, combined in order to obtain a view-independent, multi-view, human body representation [1]. 3D methods, exploit the 2D projections of the human body on the planes of the cameras in order to calculate a 3D human body representation. Such 3D representations include visual hulls [2], 3D optical flow [3], and skeletal and super-quadratic body models [4]. Actions are, usually, described as sequences of successive human body poses. Action classification is, finally, performed by employing machine learning techniques, such as Artificial Neural Networks (ANNs) [5] and dimensionality reduction based classification schemes [6].

Most multi-view methods set the assumption that the person under consideration is visible from all the cameras forming the camera setup. However, this assumption is not met in several cases. Let us assume that the person under consideration, referred to as A, is free to move inside a room that is monitored by an eight-view camera setup, like the one shown in Figure 1a. As it is shown in Figure 1b, in the cases where the person is inside the camera setup capture volume, he/she is visible from all the eight cameras forming the camera setup. However, in the cases where the person moves outside the camera setup capture volume, as shown in Figure 1c, he/she is not visible from some of the cameras (three in this example). Furthermore, there are cases where the person under consideration may be inside the camera setup capture volume and not be visible from all the cameras. An example is illustrated in Figure 1d, where the person A is occluded by another person in two of the cameras. In these cases, most multi-view methods will probable fail to provide the correct action classification result, since the human body representation will be incorrect. This fact renders these methods to be applicable only in restricted action recognition settings.

Having these in mind, we propose an action recognition



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

Fig. 1. *a) An eight-view camera setup, b) a person inside the camera setup capture volume, c) a person outside the camera setup capture volume and d) a person inside the camera setup being occluded in two cameras by another person.*

method that can effectively overcome the above described restrictions. The proposed method utilizes a multi-camera setup in order to capture the human body from multiple viewing angles. Actions are described as sequences of successive human body poses, in terms of binary images denoting the human body Regions of Interest (ROIs). Such binary images can be efficiently obtained by applying video segmentation techniques [7] on the camera frames. In the training phase, labeled action instances are employed in order to train a single-view, view-invariant action recognition classifier. To this end, we employ an Artificial Neural Network. In the test phase, multiple body tracking techniques [8] can be used in order to determine the cameras in which the person under consideration is visible. Action classification is performed on the video streams coming from all these cameras independently, resulting to multiple action classification results. The classification results from different views are, subsequently, combined in order to provide the final action classification result.

The rest of the paper is organized as follows: Section 2, describes the proposed method. Section 3 presents experiments on two action recognition databases. Finally, Section 4 draws the conclusion of this work.

2. PROPOSED METHOD

The proposed method employs the dyneme based action representation which has been proposed in [9]. For the neural network training procedure, we employ the Extreme Learning Machine optimization method that has been recently proposed in [10], for single hidden layer feedforward neural networks. In the following we provide a comprehensive description of these methods.

2.1. Dyneme based Action Representation

As it was already mentioned, the proposed method operates on binary videos depicting a person performing an action. Such videos are centered to the person's ROIs center of mass, cropped to the ROIs size and rescaled in order to create binary images of fixed ($N_x \times N_y$ pixels) size. The resulted images are vectorized column-wise, in order to produce the so-called posture vectors $\mathbf{p}_{ij} \in \mathbb{R}^{N_x \times N_y}$, where i is the video index and j runs along the video frames of video i, i.e., $j = 1, ..., N_i$.

In the training phase, all training posture vectors $\mathbf{p}_{ij}$ corresponding to the N_T training videos, are employed in order to determine D posture vector prototypes $\mathbf{v}_d \in \mathbb{R}^{N_x N_y}$ $d = 1, ..., D$, the so-called dynemes. This is done by clustering the training posture vectors in D clusters, without exploiting the known action labels that are available for the training videos. In this work we employ the K-Means clustering algorithm [11]. Dynemes are determined to be the mean cluster vectors. After dyneme calculation, each posture vector $\mathbf{p}_{ij}$ is mapped to the so-called membership vector $\mathbf{u}_{ij} \in \mathbb{R}^D$, which denotes the fuzzy similarity of $\mathbf{p}_{ij}$ with all the dynemes $\mathbf{v}_d$ according to a fuzzification parameter $m > 1$:

$$u_{ijd} = (\parallel \mathbf{p}_{ij} - \mathbf{v}_d \parallel_2)^{-\frac{2}{m-1}},\ d = 1, ..., D. \qquad (1)$$

Membership vectors $\mathbf{u}_{ij}$ are normalized in order to have unit l_2 norm. The mean of the membership vectors $\mathbf{s}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{u}_{ij}$ corresponding to each video is calculated in order to represent the video. Vectors $\mathbf{s}_i \in \mathbb{R}^D$, which will be called action vectors hereafter, representing all the training videos are normalized to have zero mean and unit variance. Test action vectors are normalized accordingly.

2.2. ELM training algorithm

After obtaining the training action vectors $\mathbf{s}_i$, we exploit the available action labels in order to train a single hidden layer feedforward neural network. For a classification problem involving N_A action classes, the network consists of D input, L hidden and N_A output neurons. Let $\mathbf{T} \in \mathbb{R}^{N_A \times N_T}$ be a matrix containing the network's target values, i.e., $[T]_{ji} = 1$ in the case where $\mathbf{s}_i$ belongs to action class j and $[T]_{ji} = -1$ otherwise.

Let $\mathbf{W}_{in} \in \mathbb{R}^{D \times L}$ be a matrix containing network's input weights and $\mathbf{b} \in \mathbb{R}^L$ be a vector containing the hidden layer neurons bias values, which are randomly chosen. The hidden layer outputs for a given training action vector $\mathbf{s}_i$ are calculated by using the sigmoid function, i.e.:

$$G(\mathbf{w}_j, \mathbf{b}, \mathbf{s}_i) = \frac{1}{1 + \exp^{-(\mathbf{w}_j^T\ \mathbf{s}_i + b_j)}},\ j = 1, ..., L, \qquad (2)$$

where $\mathbf{w}_j$ denotes the j-th column of the input weights matrix $\mathbf{W}_{in}$. By storing the hidden layer outputs corresponding to all the training action vectors in a matrix $\mathbf{G} \in \mathbb{R}^{L \times N_T}$ and using linear activation function for the output neurons, the network's outputs corresponding to the training action vectors can be written in a matrix form as $\mathbf{O} = \mathbf{W}_{out}^T \mathbf{G}$, where $\mathbf{W}_{out} \in \mathbb{R}^{L \times N_A}$ is a matrix containing the network's output weights. By using linear activation function for the network's output neurons, $\mathbf{W}_{out}$ can be calculated by $\mathbf{W}_{out} = \mathbf{G}^{\dagger}\mathbf{T}^T$, where $\mathbf{G}^{\dagger}$ is the Moore-Penrose generalized pseudo-inverse of $\mathbf{G}^T$. By assuming zero training error, the generalization ability of standard ELM algorithm is sensitive to outliers that may appear in the training set. In order to enhance the generalization ability of the ELM network, an optimization based regularized ELM algorithm has been proposed in [10], where it has been shown that $\mathbf{W}_{out}$ can be calculated, according to a regularization parameter C, by:

$$\mathbf{W}_{out} = \left(\frac{1}{C}\mathbf{I} + \mathbf{G}\mathbf{G}^T \right)^{-1} \mathbf{G}\mathbf{T}^T, \tag{3}$$

After $\mathbf{W}_{out}$ calculation, a test action vector $\mathbf{s}_{t,i}$ can be introduced to the ELM network and be classified to the action class corresponding to the highest network's output, i.e.:

$$c_{t,i} = arg \max_j [\mathbf{o}_{t,i}]_j, \; j = 1, ..., N_A, \tag{4}$$

where i denotes the camera that has captured the action video corresponding to action vector $\mathbf{s}_{t,i}$ and $\mathbf{o}_{t,i}$ is the network's output for $\mathbf{s}_{t,i}$.

2.3. Test Phase

Let a person performing an action being captured by $N \leq N_C$ cameras. This results to the creation of N action videos. Binary action videos are created by applying video segmentation techniques on the action video frames. The resulted binary action videos are preprocessed by following the procedure described in subsection 2.1 and, thus, N test action vectors $\mathbf{s}_{t,i}, i = 1, ..., N$ are obtained. $\mathbf{s}_{t,i}$ are, subsequently, introduced to the neural network and N action classification results $c_{t,i}$ are obtained. $c_{t,i}$ are, finally, combined by using a majority voting algorithm in order to provide the final action classification result, i.e.:

$$c_t = arg \max_j \sum_{i=1}^{N} \alpha_{ij}, \; j = 1, ..., N_A. \tag{5}$$

where $\alpha_{ij} = 1$ if $c_{t,i} = j$ and $\alpha_{ij} = 0$ otherwise.

3. EXPERIMENTAL RESULTS

In this section we present experiments conducted in order to evaluate the performance of the proposed method. Since it can be directly applied on both multi-view and single-view action recognition problems, we conducted experiments on two publicly available action recognition databases. The first one is the i3DPost multi-view action recognition database [12] aiming at recognition of daily actions, while the second one is the AIIA-MOBISERV action recognition database [13, 14] aiming at recognition of actions appearing in meal intakes.

In all our experiments, we have performed the leave-one-person-out cross validation procedure. That is, we trained the algorithm by using the action videos depicting all but one persons in the database and tested it by using the action videos of the remaining one. This has been done multiple times, equal to the number of persons in the database, in order to complete an experiment. Regarding the method's parameters, we have used the following values: $N_x = N_y = 32$, $m = 1.1$ and $L = 1000$. The optimal number of dynemes D, as well as the optimal value of the regularization parameter C, were determined by performing the LOPO cross validation procedure. Specifically, we have tested the algorithm by using values of $D = 10k, \; k = 1, ..., 20$ and $C = 10^r, \; r = -6, ..., 6$.

3.1. Experiments on i3DPost database

The i3DPost multi-view action recognition database contains high resolution (1080×1920 pixels) image sequences depicting eight persons performing eight actions: 'walk' (wk), 'run' (rn), 'jump in place' (jp), 'jump forward' (jf), 'bend' (bd), 'sit' (st), 'fall' (fl) and 'wave one hand' (wo). The database camera setup consists of eight cameras, which were placed in a ring of 8m diameter at a height of 2m above the studio floor. The studio background was of blue color. Example action video frames are illustrated in Figure 2. Binary images sequences have been obtained by applying a color based image segmentation technique exploiting the properties of the HSV color space.

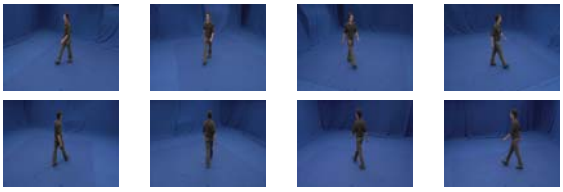
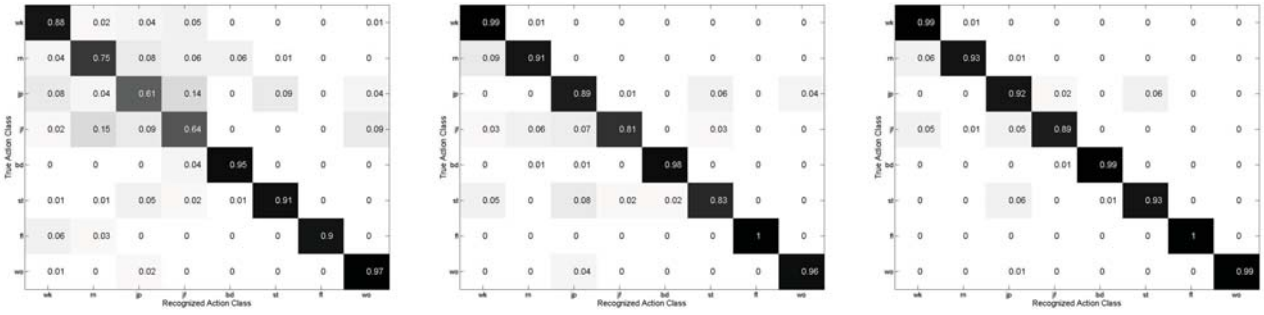


Fig. 2. *Example images of the i3DPost database depicting a person from different viewing angles.*

In order to simulate the case of performing action recognition in the appearance of total person occlusion, i.e., using an arbitrary number of cameras in the test phase, we have performed multiple experiments by training the algorithm using all the available cameras in the database and testing it using a subset of them. For example, we trained the algorithm by using all the eight cameras forming the database camera setup and tested it by using only two cameras, which were randomly chosen for each action sequence.



(a) (b) (c)

Fig. 4. *Confusion matrices on the i3DPost database obtained by using different number N of test cameras ($N_C = 8$): a) $N = 1$, b) $N = 4$ and c) $N = 8$.*

In Figure 3, we illustrate the action classification rates obtained for all these experiments. By using only one camera in the test phase, a classification rate equal to 82.56% has been obtained. By using two cameras, the action classification rate was increased to 84.35%. The use of four cameras resulted to an action classification rate equal to 92.18%. Finally, as can be seen, when using all the cameras of the database, an action classification rate equal to 95.5% is achieved.

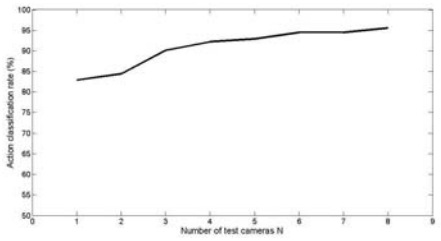


Fig. 3. *Action classification rates obtained by using different number N of randomly chosen test cameras ($N_C = 8$).*

As can be observed in Figure 3 the use of more than 2 cameras in the test phase leads to an action classification rate higher than 90%. In Figure 4 we also provide the confusion matrices corresponding to the cases where the method has been tested by using $N = 1$, $N = 4$ and $N = 8$ cameras. As can be seen in Figure 4a, by using only one camera, action classes 'walk', 'run', 'jump in place', 'jump forward' and 'sit' are confused to each other. This is reasonable, since, in this case, the human body is observed by only one, arbitrary viewing angle. As it was also shown in [15], the viewing angle that the human body is captured from plays a significant role on the classification performance. By using $N = 4$ test cameras, human body can be better represented, leading to higher action classification rates. As can be seen in Figure 4b, high classification rates have been obtained for most of the action classes. However, action classes 'jump in place', 'jump forward' and 'sit' are still confused to each other. This can be explained by the fact that these action classes contain a high number of common human body poses and, thus, it is difficult to be distinguished. Finally, by using all the available cameras, i.e., when $N = 8$, the action classes are better distinguished and, thus, higher classification rates are obtained for all of them.

Table 1. Comparison results in the i3DPost multi-view action recognition database ($N = N_C = 8$).

Method [6]	Method [16]	Method [5]	**Proposed method**
94.34%	95%	94.87%	**95.5%**

In Table 1, we compare the performance of the proposed method with that of other method, recently proposed in the literature, evaluating their performance on the i3DPost action recognition database, while in Table 2 we compare the performance of the proposed method with that of [5] for different numbers of test cameras N. As can be seen, the proposed method achieves state of the art performance in both experimental settings.

Table 2. Comparison results in the i3DPost multi-view action recognition database for different N ($N_C = 8$).

Number of cameras N	1	3	4
Method [5]	79%	84.9%	90%
Proposed method	**82.83%**	**90.8%**	**92.18%**

3.2. Experiments on the AIIA-MOBISERV database

The AIIA-MOBISERV single-view database [14, 13] contains low resolution (640×480 pixels) videos depicting twelve persons. A camera was placed at a distance of 2m in front of them during a meal. Four meals were recorded for all the persons in different days. The persons perform multiple instances of the following actions: 'eat', 'drink' and 'apraxia'. These actions contain several sub-actions. That is, the persons eat using a spoon, a fork, or cutlery. They drink

Fig. 5. *Example video frames of the AIIA-MOBISERV database depicting a person eating and drinking.*

Fig. 6. *Confusion matrix on the AIIA-MOBISERV database.*

from a cup or a glass. Finally, action class 'apraxia' contains actions 'slicing food' and 'rest'. Example video frames depicting persons of the database are illustrated in Figure 5.

Binary action videos denoting the persons' skin regions, i.e., their head and hands, have been obtained by applying a color-based image segmentation technique on the video frames exploiting the properties of the HSV color space. By applying the LOPO cross validation procedure using the proposed method, an action classification rate equal to 90% has been obtained. The confusion matrix of this experiment can be seen in Figure 6. As can be seen, high classification rates have been obtained. Despite the fact that all the three action classes contain high number of common human body poses, and, thus, it is difficult to distinguish to each other, high classification rates have been obtained.

4. CONCLUSION

In this paper we presented a novel multi-view action recognition method that can successfully operate in the cases where the person under consideration is not visible from all the cameras forming the recognition camera setup. Action representation involves fuzzy vector quantization and action recognition is performed by a neural network that is trained for view-invariant action classification. Action classification is performed to all the video streams depicting the person from different viewing angles independently. Action classification results from different views are combined in order to provide the final action classification. The proposed method has been evaluated in both single-view and multi-view action classification problems providing high action classification rates.

5. REFERENCES

[1] A. Iosifidis, A. Tefas, and I. Pitas, "Person specific activity recognition using fuzzy learning and discriminant analysis," *19th European Signal Processing Conference (EUSIPCO 2011)*, pp. 1974–1978, 2011.

[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] M.B. Holte, T.B. Moeslund, N. Nikolaidis, and I. Pitas, "3d human action recognition for multi-view camera systems," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*. IEEE, 2011, pp. 342–349.

[4] 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*. IEEE, 2008, pp. 1–9.

[5] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 412–425, 2012.

[6] 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.

[7] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," in *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[8] N. Papadakis and A. Bugeau, "Tracking with occlusions via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 144–157, 2011.

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

[10] G.B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, , no. 99, pp. 1–17, 2010.

[11] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification, 2nd ed*, Wiley-Interscience, 2000.

[12] 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.

[13] "http://poseidon.csd.auth.gr/mobiserv-aiia/index.html," .

[14] A. Iosifidis, A. Tefas, and I. Pitas, "Activity based person identification using fuzzy representation and discriminant learning," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, 2012.

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

[16] Michael B. Holte, Thomas B. Moeslund, N. Nikolaidis, and I. Pitas, "3D Human Action Recognition for Multi-View Camera Systems," *First Joint 3D Imaging Modeling Processing Visualization Transmission (3DIM/3DPVT) Conference*, 2011.