PERSON IDENTIFICATION FROM ACTIONS BASED ON DYNEMES AND DISCRIMINANT LEARNING

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

In this paper we present a view-independent person identification method exploiting motion information. A multi-camera setup is used in order to capture the human body during action execution from different viewing angles. The method is able to incorporate several everyday actions in person identification. A comparative study of the discriminative ability of different actions for person identification is provided, denoting that several actions, except walk, can be exploited for person identification.

***Index Terms*—** Action-based person identification, Dyneme video representation, Discriminant learning, Classification results fusion

1. INTRODUCTION

The identification of persons based on visual information is an active research field due to its importance in a wide range of applications, including intelligent visual surveillance, human-computer interaction and content based video retrieval. Most methods proposed in the literature employ face recognition techniques [1] requiring a restricted identification scenario, in which the person under consideration should stand in front of a camera, having a (near-) frontal facial pose and neutral expression. Gait recognition [2], i.e., the identification of persons by the way they walk, has gained researchers' attention in the last two decades, since it leads to non-invasive person identification.

The major disadvantage of gait recognition is the, underlying, assumption that the person under consideration walks. Based on the fact that gait recognition, mainly, focuses on visual surveillance, this assumption is reasonable. However, there are several application scenarios where this assumption is not met. For example, consider a game where the person is free to perform several actions, like jump, bend and/or wave his/her hands, and where the game is automatically adapted based on the person playing the game. In such cases, most gait recognition methods would, probably, fail. In order to overcome this disadvantage of gait recognition techniques, action-based person identification techniques have been, recently, proposed [3, 4]. These techniques regard 'walk' as a special case of actions appearing in an action class set. By adopting such an identification approach, a less restricted identification scenario is required, since the person under consideration is free to perform several other actions, except 'walk'.

An important issue that action based person identification methods should be able to address is the fact that the human body is quite different when it is observed by arbitrary viewing angles. This is the so-called viewing angle effect [5] and is closely related with the person identification performance. In order to overcome this issue, the use of multi-camera setups, i.e., camera setups formed by multiple cameras, has been proposed. By capturing the human body from different viewing angles, the enriched visual information can be exploited in order to obtain a view-independent human body representation, leading to view-independent person identification.

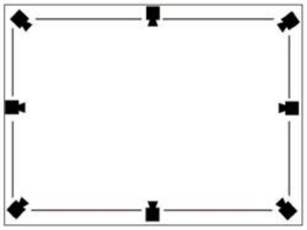


Fig. 1. *An eight-view ($N = 8$) camera setup.*

In this paper we present a person identification method exploiting motion information. The person under consideration is free to perform several actions, like 'walk', 'run' and 'jump', appearing in an action class set. In order to achieve view-independent operation, the method employs a multi-camera setup, like the one shown in Figure 1, in order to capture the human body from different viewing angles. In the training phase all the training videos depicting the persons performing actions from different viewing angles are employed in order to determine a discriminant feature space for view-independent person representation. In the test phase, multiple videos depicting the person under consideration performing the same action instance from different viewing angles are mapped to the discriminant space determined in the training phase and classified independently. The obtained classification results are, subsequently, combined in order to

978-1-4673-4989-5/13/$31.00 © 2013 IEEE

provide the final identification result. By adopting this multi-view person identification approach, a comparative study for different actions is provided, denoting that several actions, except from 'walk', contain enough discriminative ability for person identification.

The remaining of the paper is structured as follows. We describe the method in Section 2. Experiments conducted in order to evaluate its performance are provided in Section 3. Finally, conclusions are drawn in Section 4.

2. PROPOSED METHOD

2.1. Training Phase

Let $\mathcal{U}$ be a video database, created by using a camera setup formed by N cameras, containing videos depicting P persons, each performing several instances of actions appearing in an action class set $\mathcal{A}$ formed by A action classes. Image segmentation techniques, like color-based image segmentation or background subtraction [6], are applied to the video frames of these videos in order to produce binary videos, called action videos hereafter, depicting the video frame locations corresponding to the human body in white and the background in black. The obtained binary video frames are centered to the human body center of mass, cropped to the person's ROI and rescaled in order to produce fixed size ($H \times W$ pixels) images, the so-called posture images. Example posture images for five actions observed by different viewing angles are illustrated in Figure 2.

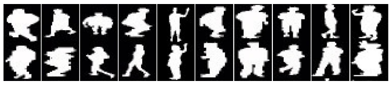


Fig. 2. *Posture images obtained by processing videos depicting actions 'walk', 'run', 'jump in place', 'jump forward' and 'wave one hand' observed by different viewing angles*

Posture images corresponding to each action video i are represented as matrices, which are vectorized in order to produce the so-called posture vectors $\mathbf{p}_{ij}$, $i = 1, \ldots, N_T$, $j = 1, \ldots, N_i$, where N_T is the number of the training action videos and j runs along the video frames of action video i. The posture vectors corresponding to all the N_T training action videos are clustered in order to determine D representative human body pose prototypes $\mathbf{v}_d$, $d = 1, , D$, the so-called dynemes. We have employed the K-Means [7] algorithm for dynemes calculation, minimizing the intra-cluster scatter:

$$\sum_{d=1}^{D}\sum_{i=1}^{N_T}\sum_{j=1}^{N_i} a_{ijd}\|\mathbf{p}_{ij} - \mathbf{v}_d\|_2^2, \tag{1}$$

where $a_{ijd} = 1$ if posture vector belongs to cluster d and $a_{ijd} = 0$ otherwise. Dynemes $\mathbf{v}_d$ are determined to be the mean cluster vectors:

$$\mathbf{v}_d = \sum_{i=1}^{N_T}\sum_{j=1}^{N_i} a_{ijd}\mathbf{p}_{ij}. \tag{2}$$

After dynemes calculation, each posture vector $\mathbf{p}_{ij}$ is mapped to the so-called membership vector $\mathbf{u}_{ij}$, by employing the fuzzy similarities between $\mathbf{p}_{ij}$ and all the dynemes $\mathbf{v}_d$:

$$\mathbf{d}_{ij} = \left[d_{ij1} \ \ldots \ d_{ijD}\right]^T, \tag{3}$$

$$d_{ijd} = \|\mathbf{p}_{ij} - \mathbf{v}_d\|_2^{-\frac{2}{m-1}}. \tag{4}$$

m is the fuzzification parameter ($m > 1.0$), which is set to $m = 1.1$ in all the presented experiments. Membership vectors $\mathbf{u}_{ij}$ are obtained by normalizing the similarity vectors $\mathbf{d}_{ij}$:

$$\mathbf{u}_{ij} = \frac{\mathbf{d}_{ij}}{\|\mathbf{d}_{ij}\|_1}. \tag{5}$$

Membership vectors calculated for action video i are used in order to calculate the so-called action vector:

$$\mathbf{s}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{u}_{ij}. \tag{6}$$

Action vectors $\mathbf{s}_i$, $i = 1, \ldots, N_T$ are normalized in order to have unit l_2 norm, zero mean and unit standard deviation.

After the calculation of the normalized action vectors, a discriminant feature space for view-independent person representation is determined by applying Linear Discriminant Analysis (LDA) [8] exploiting the person ID labels l_i, $i = 1, , N_T$ available for the training action videos. LDA determines an optimal projection matrix $\mathbf{W}^*$ by using the following optimization problem:

$$\mathbf{W}^* = \underset{\mathbf{W}}{argmin}\frac{tr\{\mathbf{W}^T\mathbf{S}_w\mathbf{W}\}}{tr\{\mathbf{W}^T\mathbf{S}_b\mathbf{W}\}}, \tag{7}$$

In the above optimization problem, $tr\{\mathbf{A}\}$ is the trace of $\mathbf{A}$ and $\mathbf{S}_w$, $\mathbf{S}_b$ are the within-class and between-class scatter matrices:

$$\mathbf{S}_w = \sum_{k=1}^{P}\sum_{i=1}^{N_T} r_i^k\left(\mathbf{s}_i - \bar{\mathbf{s}}_k\right)\left(\mathbf{s}_i - \bar{\mathbf{s}}_k\right)^T, \tag{8}$$

$$\mathbf{S}_b = \sum_{k=1}^{P} N_k\left(\bar{\mathbf{s}}_k - \bar{\mathbf{s}}\right)\left(\bar{\mathbf{s}}_k - \bar{\mathbf{s}}\right)^T, \tag{9}$$

where r_i^k is an index denoting if action vector $\mathbf{s}_i$ belongs to action class k, $\bar{\mathbf{s}}_k$ is the mean vector of class k, having cardinality equal to N_k, and $\bar{\mathbf{s}}$ is the mean vector of the entire action vector set.

The abovementioned optimization problem is approximated by solving the optimization problem $\mathbf{S}_w\mathbf{w} = \mathbf{S}_b\mathbf{w}, \ \neq 0$ [9], which can be solved by performing eigenanalysis to the matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$, in the case where $\mathbf{S}_w$ is invertible, or $\mathbf{S}_b^{-1}\mathbf{S}_w$, in the case where $\mathbf{S}_b$ is invertible. The optimal projection matrix $\mathbf{W}^*$ is formed by the eigenvectors corresponding to the $d = P - 1$ nonzero eigenvalues.

After obtaining $\mathbf{W}^*$, the training action vectors $\mathbf{s}_i$ are mapped to the so-called discriminant action vectors $\mathbf{z}_i$ by applying $\mathbf{z}_i = \mathbf{W}^{*\,T}\mathbf{s}_i$. Each person ID class is, finally, represented by the corresponding mean discriminant action vector:

$$\bar{\mathbf{z}}_k = \frac{1}{N_k}\sum_{i=1}^{N_T} r_i^k \mathbf{z}_i \tag{10}$$

and classification is performed by employing the nearest person ID class centroid classifier.

2.2. Identification (Test) Phase

Let us assume that a person appearing in the video database $\mathcal{U}$ performs an instance of an action appearing in the action class set $\mathcal{A}$. Let us, also, assume that the person is captured by all the N cameras forming the adopted N-camera setup. This results to the creation of N test videos depicting the same action instance from different viewing angles. Image segmentation techniques are applied to the video frames of these videos in order to produce N binary test action videos. The video frames of the binary action videos are centered to the person's center of mass, cropped to the person's ROI and rescaled in order to produce fixed size ($H \times W$ pixels) posture images. These posture images are vectorized in order to produce the corresponding posture vectors $\mathbf{p}_{ij,t}$, $i = 1,\ldots,N$, $j = 1,\ldots,N_i, t$. $\mathbf{p}_{ij,t}$ are, subsequently, used in order to produce the corresponding action vectors $\mathbf{s}_{i,t}$, representing the N test action videos. $\mathbf{s}_{i,t}$ are mapped to the test discriminant action vectors $\mathbf{z}_{i,t}$ by applying $\mathbf{z}_{i,t} = \mathbf{W}^{*\,T}\mathbf{s}_{i,t}$. Each test discriminant action vector $\mathbf{z}_{i,t}$ is assigned the person ID class label of the closest mean discriminant action vector, using the Euclidean distance:

$$l_{i,t} = \underset{k}{argmin}\|\mathbf{z}_{i,t} - \bar{\mathbf{z}}_k\|_2. \tag{11}$$

In order to obtain the final person identification result, the obtained person ID labels are combined by following the maximum probability Sum combination rule [9].

3. EXPERIMENTS

We have performed experiments on the i3DPost action database [10] containing everyday actions. Eight persons (six males and two females) have been asked to perform several instances of eight actions: 'walk', 'run', 'jump in place', 'jump forward', 'bend', 'sit on a chair', 'fall down' and 'wave one hand'. The database camera setup, Figure 1, is formed by eight cameras having a wide 45o baseline in order to provide 360o coverage of the capture volume. The studio was covered by a uniform blue background. Example video frames depicting a person jumping from all the eight cameras are illustrated in Figure 4. Example video frames depicting all the persons performing an action instance from arbitrary viewing angles are illustrated in Figure 4.

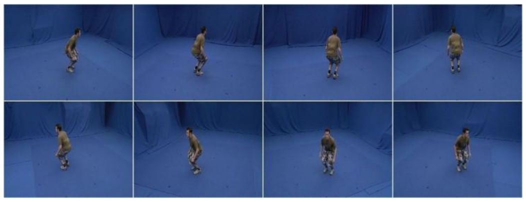


Fig. 3. *Example video frames of the i3DPost database depicting a person jumping from all the eight available viewing angles.*

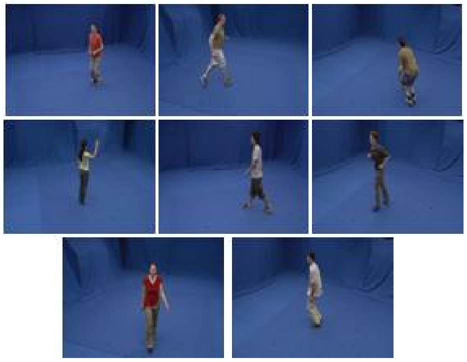
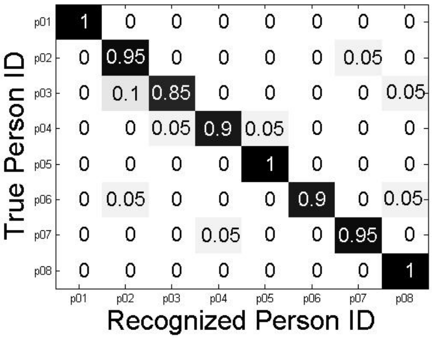


Fig. 4. *Example video frames depicting all the persons in the i3DPost database performing an action.*

In our experiments we have used the action videos depicting the persons performing five actions, i.e., 'walk', 'run', 'jump in place', 'jump forward' and 'wave one hand', since the persons performed the rest actions only once. Binary action videos have been created by applying a color-based image segmentation technique on the color video frames in order to discard the blue background. Four instances of each action class have been used for each person. The algorithm has been trained by using three instances of each action class and tested by using the remaining action instances. This procedure has been repeated multiple times (folds), one for each set of test action instances, in order to complete an experiment.

Regarding the optimal number of dynemes D, it has been determined by performing multiple experiments and using different values, i.e., $D = 10d$, $d = 1,\ldots,20$. A person identification rate equal to 94.37% has been obtained by using $D = 110$ dynemes. The confusion matrix of this experiment is illustrated in Figure 5. As can be seen in this Figure, high person identification rates have been obtained for all the persons in the database

In a second set of experiments we have investigated the



	p01	p02	p03	p04	p05	p06	p07	p08
p01	1	0	0	0	0	0	0	0
p02	0	0.95	0	0	0	0	0.05	0
p03	0	0.1	0.85	0	0	0	0	0.05
p04	0	0	0.05	0.9	0.05	0	0	0
p05	0	0	0	0	1	0	0	0
p06	0	0.05	0	0	0	0.9	0	0.05
p07	0	0	0	0.05	0	0	0.95	0
p08	0	0	0	0	0	0	0	1

Fig. 5. *Confusion matrix on the i3DPost action database.*

walk	run	jump in place
84.37%	90.62%	100%

jump forward	wave one hand
96.87%	100%

Fig. 6. *identification rates for different actions.*

discriminative ability of each action. Multiple experiments have been performed by using the previously defined optimal number of dynemes $D = 110$ to this end. The algorithm has been trained by using the action videos of all the persons depicting three instances of an action from all the available viewing angles and tested by using the action videos depicting them performing the fourth instance of the same action. The person identification rates obtained in all these experiments are illustrated in Figure 6. As can be seen in this Figure, several actions, other than walk, contain significant discriminative information for person identification. Specifically, it can be seen that the actions 'jump in place' and 'wave one hand' provided the best identification performance, equal to 100%. Action 'jump forward' resulted to a high person identification rate, equal to 96.87%, while action 'run' resulted to a person identification rate equal to 90.62%. Finally, it can be seen action 'walk' has been rated fifth, providing a person identification rate equal to 84.37%.

4. CONCLUSION

In this paper we presented a person identification method exploiting motion information. It employs a multi-camera setup in order to capture the human body from multiple viewing angles and to achieve view-independent operation. The method can, naturally, incorporate several actions in the identification process. A comparative study of the discriminative ability of five actions denotes that several actions, except walk, can be exploited for person identification. The method has been evaluated on a publicly available database containing everyday actions providing satisfactory identification rates

Acknowledgment

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.

5. REFERENCES

[1] A. Serrano, I.M. Diego, C. Conde, and E. Cabello, "Recent advances in face biometrics with gabor wavelets: A review," *Pattern Recognition Letters*, vol. 31, pp. 327–381, 2010.

[2] J. Wang, M. She, S. Nahavandi, and A. Kouzani, "A review of vision-based gait recognition methods for human identification," *International Conference on Digital Image Computing: Techniques and Applications*, pp. 320–327, 2010.

[3] A. Iosifidis, A. Tefas, and I. Pitas, "Learning human identity using view-invariant multi-view movement representation," *Biometrics and ID Management Workshop*, pp. 217–226, 2011.

[4] A. Iosifidis, A. Tefas, and I. Pitas, "Person specific activity recognition using fuzzy learning and discriminant analysis," *European Signal Processing Conference*, pp. 1–5, 2011.

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

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

[7] S. Theodoridis and K. Koutroumbas, "Pattern recognition," 2006.

[8] R. Duda, P. Hart, and D. Stork, "Pattern classification," 2001.

[9] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligenece*, vol. 20, pp. 226–239, 1998.

[10] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," *Conference on Visual Media Production*, pp. 159–168, 2009.