DISCRIMINANT ACTION REPRESENTATION FOR VIEW-INVARIANT PERSON IDENTIFICATION

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

In this paper we propose a novel person identification method exploiting human motion information. Persons are described by using their poses during action execution. Identification process involves Fuzzy Vector Quantization and Discriminant Learning. In the case of multiple cameras used in the identification phase, single-view identification results combination is achieved by employing a Bayesian combination strategy. The proposed identification approach does not set the assumptions of known action class and number of capturing cameras in the identification phase. Experimental results on two publicly available video databases denote the effectiveness of the proposed approach.

Index Terms— Person identification, Discriminant Learning, Bayesian Learning

1. INTRODUCTION

Person identification from video streams is an important task finding applications in many fields, such as human-computer interaction, visual surveillance and security. This task has been approached, mainly, by applying face recognition techniques on the video frames depicting the persons under consideration [1]. This is a reasonable approach, since it is expected that the human facial features do not change in short time periods. However, there are applications where the face recognition approach can not be applied due to the limitations that it sets. For example, in visual surveillance, the person under consideration may be at a distance from the camera, captured by a side or a back view. In these cases, face recognition based person identification will fail, since face recognition methods assume that the person is in front of the capturing camera, having a near frontal facial pose.

In order to address such issues, researchers have focused their attention on identifying individuals by exploiting information provided by different human biometrics. Gait recognition, i.e., the identification of individuals by the way they walk, has been extensively studied in the last decade [2]. This approach exploits the information appearing in the person's body shape and his/her walking style. Persons, usually, are described by using binary images denoting their body poses and identification is performed in video snapshots consisting of consecutive binary body images. Gait recognition based person identification, however, sets the assumption that the person under consideration walks. This is a reasonable assumption in visual surveillance applications, which is the main application field of such methods. However, there are several applications where this assumption does not met. For example in human-computer interaction and games the person may perform different actions, such as jump, or bend.

It has been shown that the viewing angle that the human body is captured from plays a significant role on the performance of person identification methods [3]. By using one camera, single-view recognition methods usually set the assumption of known human body orientation. In order to provide view-independent person identification, the use of multi-camera setups has been recently proposed [4]. By observing the human body from different viewing angles, multi-view methods create a view-independent human body representation which is, subsequently, used for person description and classification. However, most multi-view methods assume that the human body is visible from all the synchronized cameras forming the identification camera setup. Furthermore, the training and identification camera setups should be consisted of the same number of cameras, having the same properties. These assumptions are quite restrictive, since there are applications, such as visual surveillance, where the persons are visible from some of the cameras forming the identification camera setup.

In this paper, we propose a novel person identification method exploiting human motion information. We do not set the assumption of known action class in the identification process. That is, the persons are allowed to perform several actions, which are defined by an action class set. Persons are described by binary images denoting their body poses. Depending on the application scenario, one or multiple cameras can be used in order to capture the information needed for human body representation. In order to address the above mentioned issues related to the multi-view methods, we choose to perform single-view person identification on the video streams depicting the person under consideration independently and, subsequently, combine the person identification results. The single-view identification process involves Fuzzy Vector Quantization and Discriminant Analysis based classification. Single-view identification results combination, based on Bayesian learning, leads to high person identification rates.

2. PROPOSED METHOD

The proposed method operates upon binary videos denoting the human body poses. In the cases of color videos, the corresponding binary videos are created by applying image segmentation techniques. In our experiments we have used a color-based image segmentation method exploiting the properties of the HSV color-space [5].

2.1. Single-view Person Identification

Let $\mathcal{V}$ be a video database containing N_T binary videos depicting N_P persons performing multiple instances of N_A action classes captured by $N_C \geq 1$ cameras. The binary video frames are centered to the human body center of mass, cropped to the ROI size and rescaled to fixed size ($N_H \times N_W$ pixels) binary images, which are called posture images. Posture images are vectorized column-wise in order to produce the so-called posture vectors $\mathbf{p}_{ij} \in \mathbb{R}^D,\ i = 1, ..., N_T,\ j = 1, ..., N_i,\ D = N_H \times N_W$, where N_i is the number of video frames forming video i.

By using all the $\sum_{i=1}^{N_T} N_i$ training posture vectors $\mathbf{p}_{ij}$ we determine K posture vector prototypes $\mathbf{v}_k \in \mathbb{R}^D,\ k = 1, ..., K$ without exploiting the known person ID labels of the training posture vectors. We employ K-Means clustering algorithm [6] to this end, minimizing the between cluster scatter, i.e.:

$$\sum_{k=1}^{K}\sum_{i=1}^{N_T}\sum_{j=1}^{N_i} \alpha_{ijk}\|\mathbf{p}_{ij} - \mathbf{v}_k\|^2, \quad (1)$$

where $\alpha_{ijk} = 1$, if $\mathbf{p}_{ij}$ is assigned to cluster k and $\alpha_{ijk} = 0$, otherwise. $\mathbf{v}_k,\ k = 1, ..., K$ are defined as the mean cluster vectors, $\mathbf{v}_k = \frac{1}{n_k}\sum_{i=1}^{N_T}\sum_{j=1}^{N_i} \alpha_{ijk}\mathbf{p}_{ij}$, where $n_k = \sum_{i=1}^{N_T}\sum_{j=1}^{N_i} \alpha_{ijk}$ is the number of posture vectors $\mathbf{p}_{ij}$ assigned to cluster k. The optimal number of posture vector prototypes K is determined by performing the cross-validation procedure.

After the posture vector prototypes $\mathbf{v}_k$ calculation, we map the training posture vectors to the so-called membership vectors $\mathbf{u}_{ij} \in \mathbb{R}^K, \mathbf{u}_{ij} = [u_{ij1}, ..., u_{ijK}]^T$, which denote the fuzzy similarity between each posture vector $\mathbf{p}_{ij}$ with all the posture vector prototypes $\mathbf{v}_k$. $u_{ijk},\ k = 1, ..., K$ are calculated according to a fuzzification parameter $m > 1$ by:

$$u_{ijk} = \frac{(\parallel \mathbf{p}_{ij} - \mathbf{v}_k \parallel_2)^{-\frac{2}{m-1}}}{\sum_{l=1}^{K}(\parallel \mathbf{p}_{ij} - \mathbf{v}_l \parallel_2)^{-\frac{2}{m-1}}}. \quad (2)$$

The mean of the membership vectors corresponding to each video is calculated in order to represent the video, i.e. $\mathbf{s}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{u}_{ij},\ i = 1, ..., N_T$. Vectors $\mathbf{s}_i$, 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.

By exploiting the person ID labels available for the training videos, we perform Linear Discriminant Analysis (LDA) [6] in order to map the action vectors $\mathbf{s}_i$ in a lower-dimensional feature space, where action vectors belonging to different person ID classes are better discriminated. LDA is a dimensionality reduction technique which is used to determine the optimal projection matrix $\mathbf{W}_{opt}$ by minimizing Fither's criterion:

$$\mathbf{W}_{opt} = \arg\min_{\mathbf{W}} \frac{trace\{\mathbf{W}^T\mathbf{S}_w\mathbf{W}\}}{trace\{\mathbf{W}^T\mathbf{S}_b\mathbf{W}\}}. \quad (3)$$

$\mathbf{S}_w$ and $\mathbf{S}_b$ are the within class and between class scatter matrices, respectively:

$$\mathbf{S}_w = \sum_{n=1}^{N_P}\sum_{i=1}^{N_T} \frac{\beta_i^n(\mathbf{s}_i - \boldsymbol{\mu}_n)(\mathbf{s}_i - \boldsymbol{\mu}_n)^T}{N_n} \quad (4)$$

$$\mathbf{S}_b = \sum_{n=1}^{N_P} \frac{(\boldsymbol{\mu}_n - \boldsymbol{\mu})(\boldsymbol{\mu}_n - \boldsymbol{\mu})^T}{N_P} \quad (5)$$

where $\beta_i^n = 1$ for action vectors $\mathbf{s}_i$ belonging to class n and $\beta_i^n = 0$ otherwise, $\boldsymbol{\mu}_n$ is the mean vector of class n having $N_n = \sum_{i=1}^{N_T} \beta_i^n$ action vectors and $\boldsymbol{\mu}$ is the total mean vector of the training action vector set. After calculating $\mathbf{W}_{opt}$, the discriminant action vectors $\mathbf{z}_i \in \mathbb{R}^{N_P - 1}$, are obtained by $\mathbf{z}_i = \mathbf{W}_{opt}^T\mathbf{s}_i$.

Finally, each person ID class is represented in the discriminant feature space by the corresponding mean discriminant action vector $\mathbf{q}_n \in \mathbb{R}^{N_P-1},\ n = 1, ..., N_P$, i.e., $\mathbf{q}_n = \sum_{i=1}^{N_T} \beta_i^n\mathbf{z}_i$.

2.2. Person Identification Results Combination

Let us assume that a person appearing in the video database $\mathcal{V}$ performs an action and that he/she is captured by $N \leq N_C$ cameras. This results to the creation of N action videos depicting him/her from different viewing angles. All these N videos can be processed by following the above described procedure and, thus, they can be represented by the corresponding discriminant action vectors $\mathbf{z}_{test,i},\ i = 1, ..., N$. Each test discriminant action vector $\mathbf{z}_{test,i}$ can be classified to one person ID class by using a nearest class centroid algorithm, i.e.: $l_{test,i} = \underset{n}{argmin}\|\mathbf{z}_{test,i} - \mathbf{q}_n\|^2$, where $l_{test,i}$ denotes the recognized person ID class label corresponding to test video i.

Since all these N test videos refer to the same action instance performed by the same person, we would like to combine all these classification results in order to recognize the depicted person. In order to do so, we exploit the labeling information that is available for the training videos. After determining the discriminant vectors $\mathbf{q}_n$ representing the person ID classes, we classify all the training discriminant vectors $\mathbf{z}_i$ to the nearest class centroid, and, thus, the corresponding person ID labels $l_i,\ i = 1, ..., N_T$ are obtained. Using these labels, the a posteriori probabilities $P(l_i|n)$, i.e., the probabilities that video i belonging to the person ID class n was classified to person ID class l_i, can be calculated. Assuming that both the training and test discriminant action vectors follow the same distributions, a test action instance depicted in N action videos can be classified to the person ID class that provides the maximum a posteriori probability sum [7], i.e. $l_{test} = \underset{n}{argmax} \sum_{i=1}^{N} P(n|l_{test,i})$.

Probabilities $P(n|l_{test,i})$ are calculated by using the Bayes formula:

$$P(n|l_{test,i}) = \frac{P(l_{test,i}|n) \cdot P(n)}{\sum_{l=1}^{N_P} P(l_{test,i}|l) \cdot P(l)}. \qquad (6)$$

The a priori probabilities $P(n),\ n = 1, ..., N_P$ can be calculated by using the ratio of the training videos belonging to person ID class n and the total number of training videos forming the training set, i.e., $P(n) = \frac{N_n}{N_T}$.

3. EXPERIMENTAL RESULTS

In this Section we present experiments conducted in order to assess the performance of the proposed person identification method. Since it can be applied in application scenarios adopting one, or multiple cameras, we present experiments on a single-view and a multi-view publicly available video databases containing daily actions.

3.1. The i3DPost multi-view database

The i3DPost eight-view database [8] contains videos depicting eight persons (six males and two females) performing eight daily actions: ’walk’, ’run’, ’jump in place’, ’jump forward’, ’bend’, ’fall’, ’sit’ and ’wave one hand’. Example video frames depicting a person of the database are illustrated in Figure 1. In our experiments we have used the videos belonging to action classes ’walk’, ’run’, ’jump in place’, ’jump forward’ and ’wave one hand’, since each person performed the remaining actions only once. Binary action videos have been produced by applying an image segmentation technique to the action video frames exploiting the properties of the HSV color-space.

The Leave-One-Instance-Out (LOIO) cross validation procedure has been performed. It involves training the algorithm by using the videos depicting all but one action instances of all the persons in the database and testing it on the videos depicting the remaining action instance. This procedure is performed multiple times, equal to the number of action instances appearing in the database. The LOIO cross-validation has been performed for different number of posture prototypes K. In order to assess the contribution of the proposed person ID classification results combination strategy, a second set of experiments has been conducted following a majority voting person ID classification results combination strategy. The identification rates obtained for these experiments are illustrated in Figure 2. A person identification rate equal to 94.38% has been obtained by using $K = 110$ posture vector prototypes. The confusion matrix of this experiment is illustrated in Figure 3. In Figure 2 it can be seen that, the proposed combination strategy outperforms the majority voting one in all the presented experiments. The best person identification rate, equal to 93.38%, for the majority voting combination approach has been obtained by using $K = 110$ posture vector prototypes as well.

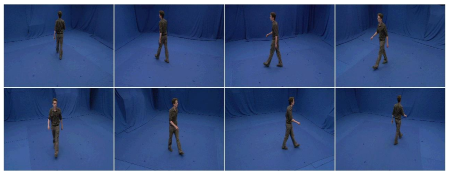

Fig. 1. *Action video frames of the i3DPost database depicting a person walking from multiple viewing angles.*

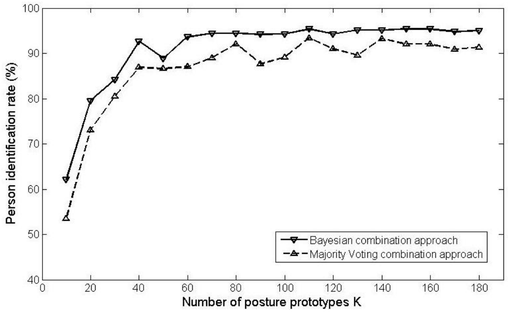

Fig. 2. *Person identification rates on the i3DPost database.*

3.2. The AIIA-MOBISERV single-view database

The AIIA-MOBISERV single-view database [9, 10] contains videos depicting twelve persons (six males and six females) having a meal. Four meal intakes have been recorded for each the person in different days. Thus, four sessions for each person are available. The persons eat using fork, cutlery and spoon and drink from a cup and a glass. Furthermore, the persons perform other actions appearing in meal intakes, such as slicing their food or chewing it. Each person performed multiple instances of these actions in each meal. We have applied a color-based image segmentation technique to the video frames in order to create binary videos depicting the person’s head and hands [5]. Example binary video frames

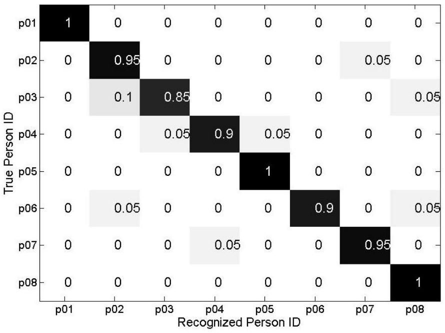

Fig. 3. *Confusion matrix on the i3DPost database.*

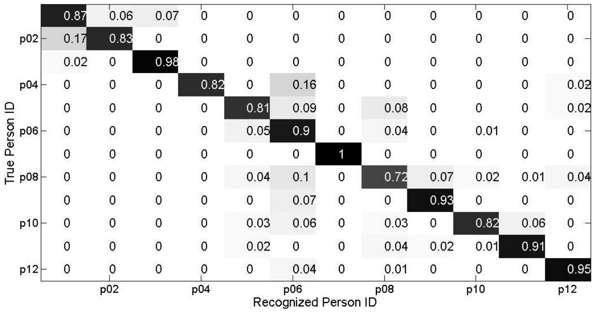

Fig. 4. *Confusion matrix on the AIIA-MOBISERV database.*

are illustrated in Figure 5.

In all the experiments we have used Leave-One-Session-Out (LOSO) cross validation. That is, the proposed algorithm is trained using data from the three meals in the training set and the fourth meal is used for testing. This procedure is applied four times in order to complete an experiment. The LOSO cross validation has been performed for different number of posture prototypes K. A person identification rate equal to 87.83% has been obtained by using $K = 230$ posture vector prototypes. The confusion matrix of this experiment is illustrated in Figure 4. It can be seen by observing the confusion matrices in Figures 3 and 4, that person identification in the case of actions appearing in meal intakes is more difficult, compared to person identification from daily actions. This is reasonable, since variations in execution style between individuals on such actions are fewer. However, as can be seen, even in this case, high identification rates have been obtained.

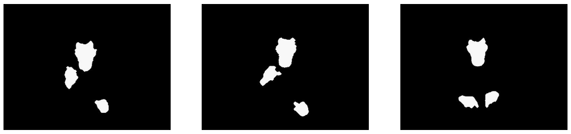

Fig. 5. *Binary video frames depicting a person of the AIIA-MOBISERV database a) eating, b) drinking and c) slicing his food.*

4. CONCLUSION

In this paper we proposed a person identification method exploiting human motion information. The method performs single-view person identification on multiple videos depicting a person from different viewing angles during action execution. Single-view identification results combination based on Bayesian learning in the test phase, leads to view-independent person identification with high 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. We would like to thank Prof. K. Lyroudia, Dept. of Dentistry, AUTH, for fruitful discussions regarding eating/drinking activity recognition.

5. REFERENCES

[1] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” *Acm Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] J. Wang, M. She, S. Nahavandi, and A. Kouzani, “A Review of Vision-Based Gait Recognition Methods for Human Identification,” in *International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2010, pp. 320–327.

[3] S. Yu, D. Tan, and T. Tan, “Modeling the effect of view angle variation on appearance-based gait recognition,” in *Proceedings Asian Conf. Computer Vision*, vol. 1, Jan. 2006, pp. 807–816.

[4] 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*, 2011.

[5] E. Marami, A. Tefas, and I. Pitas, “Nutrition assistance based on skin color segmentation and support vector machines,” *Man-Machine Interactions 2*, pp. 179–187, 2011.

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

[7] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On combining classifiers,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[8] 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, pp. 159–168.

[9] “http://poseidon.csd.auth.gr/mobiserv-aiia/index.html.”

[10] A. Iosifidis, E. Marami, A. Tefas, and I. Pitas, “Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2012, pp. 2201–2204.