Learning Human Identity using View-Invariant Multi-View Movement Representation

Alexandros Iosifidis, Anastasios Tefas, Nikolaos Nikolaidis and Ioannis Pitas

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

Abstract. In this paper a novel view-invariant human identification method is presented. A multi-camera setup is used to capture the human body from different observation angles. Binary body masks from all the cameras are concatenated to produce the so-called multi-view binary masks. These masks are rescaled and vectorized to create feature vectors in the input space. A view-invariant human body representation is obtained by exploiting the circular shift invariance property of the Discrete Fourier Transform (DFT). Fuzzy vector quantization (FVQ) is performed to associate human body representation with movement representations and linear discriminant analysis (LDA) is used to map movements in a low dimensionality discriminant feature space. Two human identification schemes, a movement-specific and a movement-independent one, are evaluated. Experimental results show that the method can achieve very satisfactory identification rates. Furthermore, the use of more than one movement types increases the identification rates.

Keywords: View-invariant Human Identification, Fuzzy Vector Quantization, Linear Discriminant Analysis.

1 Introduction

Human identification from video streams is an important task in a wide range of applications. The majority of methods proposed in the literature approach this issue using face recognition techniques [6], [11], [1], [12], [8]. This is a reasonalbe approach, as it is assumed that human facial features do not change in significantly small time periods. One disadvantage of this approach is the sensitivity to the deliberate distortion of facial features, for example by using a mask. Another approach, for the human identification task is the use of human motion information [9], [7], [5]. That is, the identity (ID) of a human can be discovered by learning his/her style in performing specific movements. Most of the methods that identify human's ID using motion characteristics exploit the information captured by a single-static camera. Most of these methods assume the same viewing angle in training and recognition phases which is obviously a significant constraint.

In this paper we exploit the information provided by a multi-camera setup in order to perform view-invariant human identification exploiting movement style information. We use different movement types in order to exploit the discrimination capability of different movement patterns. A movement-independent and a movement-specific human identification scheme are assessed, while a simple procedure that combines identification results provided by these schemes for different movement types is used in order to increase the identification rates.

The remainder of this paper is organized as follow. In Section 2, we present the two human identification schemes proposed in this work. In Section 3, we present experiments conducted in order to evaluate the proposed method. Finally, conclusions are drawn in Section 4.

2 Proposed Method

The proposed method is based on a movement recognition method that we presented in [4]. This method has been extended in order to perform human identification. Movements are described by a number of consecutive human body postures, i.e., binary masks that depict the body in white and the background in black. A converging multi-camera setup is exploited in order to capture the human body from various viewing angles. By combining the single-view postures properly a view-invariant human posture representation is achieved. This leads to a view invariant movement recognition and human identification method. By taking into account more than one movement types we can increase the identification rates, as is shown in Subsection 3.3. In the remaining of this paper the term movement will denote an elementary movement. That is, a movement will correspond to one period of a simple action, e.g., a step within a walking sequence. The term movement video will correspond to a video segment that depicts a movement, while the term multi-view movement video will correspond to a movement video captured by multiple cameras.

2.1 Preprocessing

Movements are described by consecutive human body postures captured by various viewing angles. Each of the N_{t_m}, $m = 1, ..., M$ (M being the number of movement classes) single-view binary masks comprising a movement video, is centered to the human body center of mass. Image regions, of size equal to the maximum bounding box that encloses the human body in the movement video, are extracted and rescaled to a fixed size ($N_x \times N_y$) images, which are subsequently vectorized column-wise in order to produce single-view posture vectors $\mathbf{p}_{jc} \in \mathcal{R}^{N_p}$, $N_p = N_x \times N_y$, where j is the posture vector's index, $j = 1, ..., N_{t_m}$ and c is the index of the camera it is captured from $c = 1, ..., C$. Five single-view posture frames are illustrated in Figure 1.

Fig. 1. *Five single-view posture frames.*

2.2 Training Phase

Let $\mathcal{U}$ be an annotated movement video database, containing N_T C-view training movement videos of M movement classes performed by H humans. Each multi-view movement video is described by its $C \times N_{t_m}$ single-view posture vectors $\mathbf{p}_{ijc}, i = 1, ..., N_T, j = 1, ..., N_{t_m}, c = 1, ..., C$. Single-view posture vectors that depict the same movement instance from different viewing angles are manually concatenated in order to produce multi-view posture vectors, $\mathbf{p}_{ij} \in \mathcal{R}^{N_)P}$, $N_P = N_x \times N_y \times C, i = 1, ..., N_T, j = 1, ..., N_{t_m}$. A multi-view posture frame is shown in Figure 2.

Fig. 2. *One eight-view posture frame from a walking sequence.*

To obtain a view-invariant posture representation the following observation is used: all the C possible camera configurations can be obtained by applying a block circular shifting procedure on the multi-view posture vectors. This is because each such vector consists of blocks, each block corresponding to a single-view posture vector. A convenient, view-invariant, posture representation is the multi-view DFT posture representation. This is because the magnitudes of the DFT coefficients are invariant to block circular shifting. To obtain such a representation, each multi-view posture vector $\mathbf{p}_{ij}$ is mapped to a vector $\mathbf{P}_{ij}$ that contains the magnitudes of its DFT coefficients.

$$P_{ij}(k) = |\sum_{n=0}^{N_P-1} p(n) e^{-i\frac{2\pi k}{N_P}n}|, \quad k = 1, ..., N_P - 1. \tag{1}$$

Multi-view posture prototypes $\mathbf{v}_d \in \mathcal{R}^{N_P}, d = 1, ..., N_D$, called dynemes, are calculated using a K-Means clustering algorithm [10] without using the labeling information available in the training phase. Fuzzy distances from all the multi-view posture vectors $\mathbf{P}_{ij}$ to all the dynemes $\mathbf{v}_d$ are calculated and the membership vectors $\mathbf{u}_{ij} \in \mathcal{R}^{N_D}$, $i = 1, ..., N_T$, $j = 1, ..., N_{t_m}$, $d = 1, ..., N_D$, are obtained:

$$\mathbf{u}_{ij} = \frac{(\parallel \mathbf{P}_{ij} - \mathbf{v}_d \parallel_2)^{-\frac{2}{m-1}}}{\sum_{d=1}^{N_D} (\parallel \mathbf{P}_{ij} - \mathbf{v}_d \parallel_2)^{-\frac{2}{m-1}}}. \tag{2}$$

where $m > 1$ is the fuzzification parameter and is set equal to 1.1 in all the experiments presented in this paper.

The mean membership vector $\mathbf{s}_i = \frac{1}{N_{t_m}} \sum_{i=1}^{N_{t_m}} \mathbf{u}_{ij}$, $\mathbf{s}_i \in \mathcal{R}^{N_D}, i = 1, ..., N_T$, is used to represent the movement video in the dyneme space and is noted as movement vector. Using the known labeling information of the training movement vectors LDA [2] is used to map the movement vectors in an optimal discriminant subspace by calculating an appropriate projection matrix $\mathbf{W}$. Discriminant movement vectors, $\mathbf{z}_i \in \mathcal{R}^{M-1}$, $i = 1, ..., N_T$, are obtained by:

$$\mathbf{z}_i = \mathbf{W}^T \mathbf{s}_i. \tag{3}$$

2.3 Classification Phase

In the classification phase single-view posture vectors consisting single-view movement videos are arranged using the camera labeling information and the multi-view posture $\mathbf{p}_j$ vector is mapped to its DFT equivalent $\mathbf{P}_j$, as in the training phase. Membership vectors $\mathbf{u}_j \in \mathcal{R}^{N_D}, j = 1, ..., N_{t_m}$ are calculated and the mean vector $\mathbf{s} \in \mathcal{R}^{N_D}$ represents this multi-view movement video in the dyneme space. The discriminant movement vector $\mathbf{z} \in \mathcal{R}^{M-1}$ is obtained by mapping $\mathbf{s}$ in the LDA space. In that space, the multi-view movement video is classified to the nearest class centroid.

2.4 Human Identification

As previously mentioned, the movement videos of the database $\mathcal{U}$ are labeled with movement class and human identity information. Thus, a classification scheme can be trained and subsequently used in order to provide the ID of a human depicted in an unlabeled movement video that depicts one of the H known humans in the database performing one of the M known movements.

In this paper we examine two classification procedures in order to achieve this. In the first one, we apply the procedure described above using one classification step. That is, the labeling information exploited by the classification procedure is that of the humans' IDs. Each multi-view movement video in the training database is annotated by the ID of the depicted human. Using this approach a movement-independent human identification scheme is devised. A block diagram of the classification procedure applied in this case is shown in Figure 3.

The second procedure consists of two classification phases. In the first phase, the multi-view movement video is classified to one of the M known movement classes. The movement classifier utilized in this phase is trained using the movement class labels that accompany the videos. Subsequently, the use of a movement-specific human identification classifier provides the ID of the depicted human. More specifically M human identification classifiers are used in this phase. Each of them is trained to identify humans using videos of a specific movement class. Human ID labels are used for the training of these classifiers. A block diagram of the classification procedure applied in this case is shown in Figure 4.

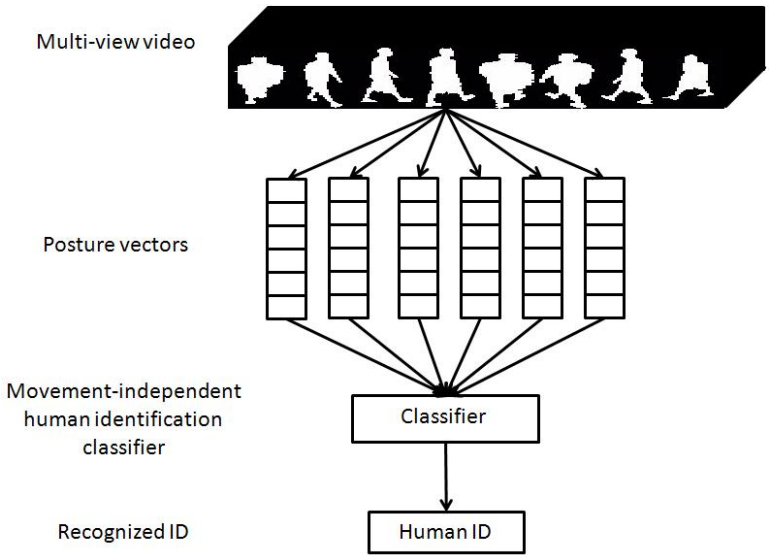

Fig. 3. *Movement-independent human identification procedure.*

2.5 Fusion

Video segments that depict single movement periods are rare. In most real-world videos a human performs more than one movement periods of the same or different movement types. In the case where a movement video depicts N_s movement periods, of probably different movement classes, the procedures described above will provide N_s identification results. By combining these results, the human ID correct identification rates increase. A simple majority voting procedure can be used for this procedure. That is the ID of the human depicted in a video segment is set to that of the mostly recognized human.

3 Experimental Results

In this section we present experimental results in the i3DPost multi-view video database described in [3]. This database contains high definition image sequences depicting eight humans, six males and two females, performing eight movements, walk, run, jump in place, jump forward, bend, sit, fall and wave one hand. Eight cameras were equally spaced in a ring of 8m diameter at a height of 2m above the studio floor. The studio background was uniform. Single-view binary masks were obtained by discarding the background color in the HSV color space. Movements that contain more than one periods were used in the following experiments. That is movements walk (wk), run (rn), jump in place (jp), jump forward (jf) and wave one hand (wo) were used, while movements bend (bd), sit (st) and fall (fl) were not used as each human performs the movement once For each movement class

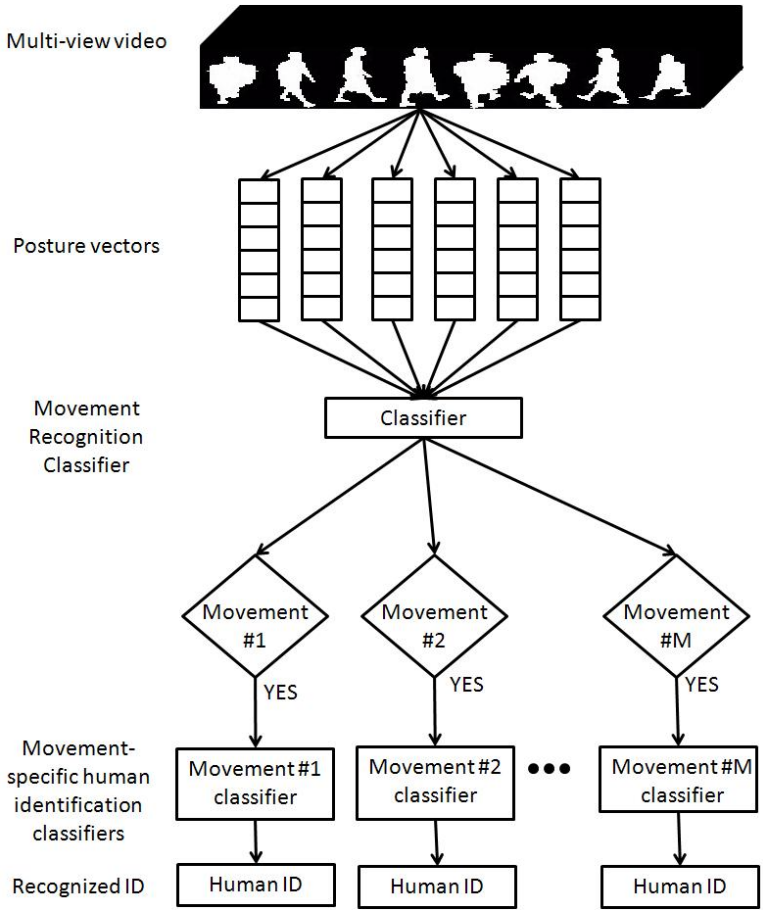

Fig. 4. *Movement-specific human identification procedure.*

four movement videos were used in order to perform a four-fold cross-validation procedure in all the experiments presented.

3.1 Movement-independent human identification

In this experiment we applied the procedure illustrated in Figure 3. In this case the multi-view training movement videos were labeled with human ID information. At every step 40 multi-view movement videos, one of each movement class (5 classes) depicting each human (8 humans), were use for testing and the remaining 120 multi-view movement videos were used for training. This procedure was applied four times, one for each movement video set. A 82.5% identification rate was obtained using 70 dynemes. The corresponding confusion matrix is presented in Table 1. As it can be seen, some of the humans are confused with others.

Table 1. Confusion matrix containing identification rates in the movement-independent case on the I3DPost database.

	chr	hai	han	jea	joe	joh	nat	nik
chr	0.95			0.05				
hai		0.85		0.05	0.05		0.05	
han	0.1		0.5		0.15	0.05	0.15	0.05
jea	0.05			0.9	0.05			
joe					1			
joh					0.05	0.85	0.05	0.05
nat	0.1	0.05			0.1	0.05	0.65	0.05
nik		0.1						0.9

3.2 Movement-specific human identification

In order to assess the discrimination ability of each movement type in the human identification task we applied five human identification procedures, each corresponding to one of the movement type. For example, in the case of movement walk, three multi-view movement videos depicting each of the eight humans walking were used for training and the fourth multi-view movement video depicting him/her walking was used for testing. This procedure was applied four times, one for each movement video. Identification rates provided for each of the movement types are illustrated in Table 2. As it can be seen, all the movement types provide high identification rates. Thus, such an approach can be used in order to obtain the identity of different humans in an efficient way.

Table 2. Identification rates of different movement classes.

Movement	Dynemes	Identification Rate
wk	14	0.90
rn	29	0.90
jp	18	1
jf	21	0.93
wo	17	0.96

In a second experiment, we applied the procedure illustrated in Figure 4. That is, the multi-view movement videos were firstly classified to one of the M movement classes and were subsequently fed to the corresponding movement-specific classifier provided that the human's ID. An identification rate equal to 94.37% was achieved. The optimal number of dynemes, for the movement recognition classifier was equal to 25. The optimal number of dynemes for the movement-specific classifiers were 14, 29, 18, 21 and 17 for movements wk, rn, jp, jf and wo, respectively. Table 3 illustrates the confusion matrix of the optimal case. As it can be seen, most of the multi-view videos were assigned correctly

to the person they depicted. Thus, the movement-specific human identification approach is more effective than the movement-independent approach.

Table 3. Confusion matrix containing identification rates in the movement-specific case on the I3DPost database.

	chr	hai	han	jea	joe	joh	nat	nik
chr	1							
hai		0.9			0.05		0.05	
han			0.85				0.05	0.1
jea				1				
joe					1			
joh						0.95	0.05	
nat	0.05						0.95	
nik	0.05					0.05		0.9

3.3 Combining IDs of different movement types

In this experiment we combined the identification results provided the movement-independent and the movement-specific classification schemes (Figures 3 and 4). At every step 40 multi-view movement videos, each depicting one human performing one movement, were used for testing and the remaining 120 multi-view movement videos were used for training. In the movement-independent identification procedure, training multi-view movement videos were labeled with the human ID information, while in the movement-specific identification procedure the training multi-view movement videos were labeled with the movement and the human ID information. At every fold of the cross-validation procedure, the test multi-view movement videos of each humans of the database were fed to the classifier and a majority voting procedure was applied to the identification results in order to provide the final ID. Using this procedure identification rates equal to 90.62% and 96.87% were achieved for the movement-independent and movement-specific classification procedures, respectively. Tables 4 and 5 illustrate the confusion matrices of these experiments. As can be seen, a simple majority voting procedure increases the identification rates. This approach can be applied to real videos, where more than one action periods are performed.

4 Conclusion

In this paper we presented a view-invariant human identification method that exploits information captured by a multi-camera setup. A view-invariant human body representation is achieved by concatenating the single-view postures and computing the DFT equivalent posture representation. FVQ and LDA provides a generic classifier which is subsequently used in a movement-independent and

Table 4. Confusion matrix containing identification rates in the movement-independent case on the I3DPost database using a majority voting procedure.

	chr	hai	han	jea	joe	joh	nat	nik
chr	1							
hai		0.75		0.25				
han			0.75		0.25			
jea				1				
joe					1			
joh						1		
nat	0.25						0.75	
nik								1

Table 5. Confusion matrix containing identification rates in the movement-specific case on the I3DPost database using a majority voting procedure.

	chr	hai	han	jea	joe	joh	nat	nik
chr	1							
hai		1						
han			0.75					0.25
jea				1				
joe					1			
joh						1		
nat							1	
nik								1

a movement-specific human identification scheme. The movement-specific case seem to outperform the movement-independent one. The combination of identification results provided for different movement types increases the identification rates in both cases.

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) and COST Action 2101 on Biometrics for Identity Documents and Smart Cards.

References

1. Ahonen, T., Hadid, A., et al.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 2037–2041 (2006)
2. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd ed. Wiley-Interscience (2000)

3. Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I.: The i3dpost multi-view and 3d human action/interaction database. In: 6th Conference on Visual Media Production. pp. 159–168 (Nov 2009)
4. Gkalelis, N., Nikolaidis, N., Pitas, I.: View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation. In: Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. pp. 394–397. IEEE (2009)
5. Gkalelis, N., Tefas, A., Pitas, I.: Human identification from human movements. In: Image Processing (ICIP), 2009 16th IEEE International Conference on. pp. 2585–2588. IEEE (2010)
6. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 328–340 (2005)
7. Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., Bowyer, K.: The humanid gait challenge problem: Data sets, performance, and analysis. IEEE transactions on pattern analysis and machine intelligence pp. 162–177 (2005)
8. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. pp. 586–591. IEEE (2002)
9. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. Pattern Analysis and Machine Intelligence, IEEE Transactions on 25(12), 1505–1518 (2003)
10. Webb, A.: Statistical pattern recognition. A Hodder Arnold Publication (1999)
11. Wiskott, L., Fellous, J., Kuiger, N., Von der Malsburg, C.: Face recognition by elastic bunch graph matching. Pattern Analysis and Machine Intelligence, IEEE Transactions on 19(7), 775–779 (2002)
12. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. Acm Computing Surveys (CSUR) 35(4), 399–458 (2003)