Person Identification from Actions based on Artificial Neural Networks

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 person identification method exploiting human motion information. A Self Organizing Neural Network is employed in order to determine a topographic map of representative human body poses. Fuzzy Vector Quantization is applied to the human body poses appearing in a video in order to obtain a compact video representation, that will be used for person identification and action recognition. Two feedforward Artificial Neural Networks are trained to recognize the person ID and action class labels of a given test action video. Network outputs combination, based on another feedforward network, is performed in the case of multiple cameras used in the training and identification phases. Experimental results on two publicly available databases evaluate the performance of the proposed person identification approach.**

I. Introduction

Person identification is one of the most heavily researched vision based pattern recognition tasks, due to its importance in many applications, such as human-computer interaction, video surveillance and security. The majority of methods proposed in the literature for person identification, employ face recognition techniques [1]. This approach leads to good person identification results, setting strong restrictions on the identification scenario. That is, the person under consideration should be in front of the camera and look at it, heaving a near frontal facial pose and neutral expression. However, there are several important applications, such as video surveillance and security, where these assumptions can not be met. In these cases, the cameras are, usually, placed in order to capture a wide area, capturing the person under consideration from a far distance, having various human body orientations.

In order to provide non-invasive person identification techniques, researchers have employed other biometrics for person identification. Gait recognition, i.e., the identification of individuals by the way they walk, has been extensively studied to this end [2]. The main idea behind gait recognition is the exploitation of the differences appearing in human body sizes and walking style variations among individuals. Persons are, usually, described by using binary images denoting the human body image locations. This human body representation has been widely adopted because it is computationally inexpensive and, thus, it can be used in the cases where the real-time operation is important. Gait recognition has been widely used in video surveillance applications, where it is assumed that the persons walk. However, this is a strong assumption and is not, usually, met in most applications. For example, in security applications, it is expected that the person under consideration will run, rather than walk, in order to escape. Even in the cases of video surveillance, it is not unusual that a person will bend in order to pick up something. Finally, consider a game where the user should perform a variety of actions, such as jump in place, wave his/her hands, etc. In these cases, gait recognition methods will, probably, fail to operate well.

In most applications, the camera viewing angle is not fixed and the persons are observed from arbitrary camera viewpoints. This generates several issues, which are related to the viewing angle effect [3], [4], that should be properly addressed in order to obtain a view-independent operation. Camera setups consisting of multiple cameras, which are referred as multi-camera setups, have been adopted to this end. By capturing the human body from multiple viewing angles and combining the obtained human body poses, a view-independent human body representation is obtained. Several combination schemes have been proposed for view-independent human body representation, like the multi-view postures proposed in [5]. However, such approaches set several restrictions. In these settings, the person under consideration should be visible from all the cameras forming the camera setup. Furthermore, the camera setups used in the training and identification phases should have the same properties. That is, both of them should be consisted of the same number of, synchronized, cameras, placed in the same positions.

In this paper, we propose a method aiming at person identification exploiting human motion information. Contrary to gait recognition methods, we do not set the assumption of one, known, human action in the recognition process. That is, the person under consideration is allowed to perform multiple actions, depending on the application scenario. For example, in surveillance and games, the person may walk, jump, bend, etc. In different applications, such as nutrition assistance, the person may eat, drink or slice his/her food. The persons are described by using binary images denoting their poses during action execution. In the training phase, multiple cameras ($N_C \geq 1$) are used in order to create videos depicting persons from different viewing angles performing actions. Such videos are referred as action videos hereafter. Human body poses coming from all the available viewing angles are employed in order to train a Self Organizing Map (SOM), resulting to the determination of human body pose prototypes. These pose prototypes are, subsequently, used in order to describe the training action videos. This is achieved by calculating the fuzzy similarities between all the human body pose prototypes and the human body poses appearing in each training action video. By using such an action video representation and the person ID and action class labels of all training action videos, we train two feedforward ANNs, one for person identification and one for action recognition. We employ single hidden layer feedforward networks trained by the, recently proposed, Extreme Learning Machine (ELM) algorithm to this end. In the case of $N_C > 1$, we exploit the available labeling information of the training action videos, in order to train another feedforward network that will be used in the test phase for networks' output combination. In the test phase, given $1 \leq N \leq N_C$ action videos depicting an unknown person performing an unknown action, we obtain N person ID and N action class classification results, which are introduced to the classification results combination network in order to provide the recognized person ID label. By capturing the human body from different viewing angles, the proposed method performs view-independent person identification. By performing classification of each test action video independently and, subsequently, combining the obtained classification results, the proposed method can operate by using different number of cameras in the training and test phases. Finally, by incorporating multiple action classes in the recognition process, the proposed method performs action-independent person identification.

Compared to our previous works [6], [7], the proposed method employs an ANN-based person ID and action class classification results fusion scheme, which can automatically learn nonlinear dependencies between person ID classes. Furthermore, by employing ANNs for person identification and action recognition, the method is able to better describe the nonlinear discrimination structure of the person ID and action classes, leading to increased person identification performance.

The remaining of the paper is structured as follows. We describe in detail the proposed method in Section II. Experimental results on two, publicly available, databases are illustrated in Section III. Finally, conclusions are drawn in Section IV.

II. Proposed Method

In this Section, each step of the proposed person identification method is described in detail. The preprocessing steps performed to both the training and test action videos are described in Subsection II-A. The action video representation scheme is presented in Subsection II-B. The ELM network training procedure for multi-class classification is, briefly, described in Subsection II-C. Finally, the procedures followed in both the training and identification phases are presented in Subsections II-D and II-E, respectively.

A. Action Video Preprocessing

The proposed method operates on binary action videos. In the case of color action videos, moving object segmentation [8], or color-based image segmentation techniques [9] are applied to the color action video frames in order to produce binary action videos denoting the person's regions of interest (ROIs) at each video frame. Depending on the application at hand, these ROIs may differ. For example, in video surveillance binary action video frames may depict the entire human body, while in nutrition support applications the ROIs may be the person's head and hands.

Binary action video frames are centered to the person's ROIs center of mass, cropped to the ROIs bounding box and resized to produce fixed size ($N_H \times N_W$ pixels) binary images depicting the human body poses. Example human pose images are illustrated in Figure 1. Human pose images are, finally, represented as matrices, which are vectorized in order to produce the so-called posture vectors. That is, each action video consisting of N_t video frames is represented by N_t posture vectors. In our experiments the human pose images have been vectorized column-wise.

B. SOM-based Action Video Representation

Let $\mathcal{V}$ be a video database containing N_I action instances performed by N_P persons belonging to one of the N_A action classes forming an action class set $\mathcal{A}$. Each action instance is depicted in N_C action videos captured by all the N_C cameras forming the training camera setup. These action videos are preprocessed following the procedure described in Subsection II-A in order to produce $\sum_{i=1}^{N_I} \sum_{j=1}^{N_C} N_{ij}$ posture vectors, $\mathbf{p}_{ijk} \in \mathbb{R}^{N_H \cdot N_W}$, $i = 1, ..., N_I$, $j = 1, ..., N_C$, $k = 1, ..., N_{ij}$. We use the notation N_{ij} to denote the number of posture vectors representing action video of the i-th action instance captured by camera j, since action videos may vary in duration. Even the action videos depicting the same action instance may differ in duration, due to differences in cameras properties, or synchronization errors.

Training posture vectors $\mathbf{p}_{ijk}$ are used to determine D action independent human body pose prototypes without exploiting the available person ID labels for the training action videos. We train a Self Organizing Map (SOM) [10] to this end consisting of $D = N_x \times N_y$ neurons. Let $\mathbf{w}_d \in \mathbb{R}^{N_H \cdot N_W}$, $d = 1, ..., D$ be the D SOM neurons. Its training procedure involves two phases:

- **Competition:** For each of the training posture vectors $\mathbf{p}_{ijk}$, its Euclidean distance from every SOM neuron $\mathbf{w}_d$ is calculated. Wining neuron is the one providing the smallest distance, i.e.:

$$d^* = arg \min_d \| \mathbf{p}_{ijk} - \mathbf{w}_d \|_2. \quad (1)$$

- **Co-operation:** Each SOM neuron is adapted with respect to its lateral distance from the winning neuron h_d, i.e.:

$$\mathbf{w}_d(n+1) = \mathbf{w}_d(n) + \eta(n) h_d(n)(\mathbf{p}_{ijk} - \mathbf{w}_d(n)), \quad (2)$$

where $h_d(n)$ is a function of the lateral distance between the winning neuron d^* and neuron d, $r_{d*,d}$, $\eta(n)$ is an adaptation rate parameter and n refers to the algorithms training iteration. Typical choice of $h_d(n)$ is the Gaussian function $h_d(n) = \exp\left(-\frac{r_{d^*,d}^2}{2\sigma^2(n)}\right)$.

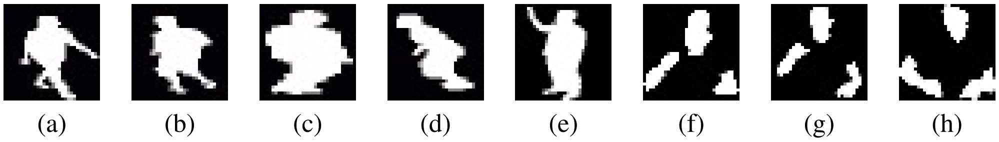


Fig. 1. *Binary human pose images of eight actions taken from various viewing angles: a) walk, b) run, c) jump in place , d) jump forward, e) wave one hand, f) eat, g) drink and h) slicing food.*

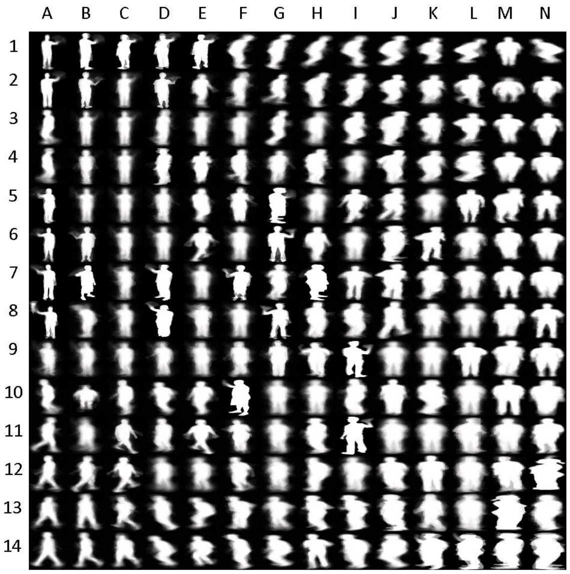


Fig. 2. *A* 14×14 *SOM obtained by using action videos depicting eight persons performing multiple instances of actions walk, run, jump in place, jump forward and wave one hand.*

The optimal SOM parameters are determined by performing the Leave-One-Instance-Out (LOIO) cross-validation procedure, which involves training the method by using all the action videos depicting all but one action instances in the database and testing it on the action videos depicting the remaining one. This procedure is performed multiple times (folds), equal to the number of action instances appearing in the database in order to complete one experiment. Multiple experiments are performed by using different SOM parameter values and the optimal values are determined to be those providing the highest person identification performance.

Figure 2 illustrates a SOM lattice obtained by using action videos depicting eight persons performing multiple instances of five action classes. As can be seen, the SOM neurons correspond to representative human body poses during action execution. Furthermore, it can be observed that each SOM neuron captures human body shape properties of different persons in the database. For example, it can be seen that neuron $\{14, L\}$ depicts a female waving her hand from a side view, while neuron $\{9, I\}$ depicts a male waving his hand from a back view.

After the SOM determination, each posture vector $\mathbf{p}_{ijk}$ is mapped to the so-called membership vector $\mathbf{u}_{ijk} = [u_{ijk,1} u_{ijk,2} ... u_{ijk,D}]^T$, which denotes the fuzzy similarity between $\mathbf{p}_{ijk}$ with all the $\mathbf{w}_d$, according to a fuzzification parameter $m > 1$. That is:

$$u_{ijk,d} = (\| \, \mathbf{p}_{ijk} - \mathbf{v}_d \, \|_2)^{-\frac{2}{m-1}}. \qquad (3)$$

The optimal value of m can be, also, obtained by applying the LOIO cross validation procedure [11]. Action vectors $\mathbf{s}_{ijk} \in \mathbb{R}^D$ are calculated as the mean normalized membership vectors of the corresponding action videos, i.e.:

$$\mathbf{s}_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} \frac{\mathbf{u}_{ijk}}{\|\mathbf{u}_{ijk}\|_1}. \qquad (4)$$

Finally, training action vectors $\mathbf{s}_{ij}$ are normalized in order to have unit norm, zero mean and unit variance. Test action vectors are normalized accordingly.

C. Extreme Learning Machine

Extreme Learning Machine (ELM) [12] is an efficient algorithm for single hidden layer feedforward network training. Let $\mathbf{x}_i$, $i = 1, ..., N_x$ be a set of training vectors, accompanied with the corresponding class labels $c_i \in \mathcal{C} = \{1, ..., C\}$. In our case, the training vectors set is the set of action vectors corresponding to the training action videos, i.e., $\mathbf{s}_{ij}$, $i = 1, \dots, N_I$, $j = 1, \dots, N_C$. In ELM, the network's input weights $\mathbf{W}_{in}$ are randomly chosen, while the output weights $\mathbf{W}_{out}$ are analytically calculated. The network's outputs corresponding to vector $\mathbf{x}_i$, $\mathbf{o}_i = [o_{i1}, ..., o_{iC}]^T$, are set equal to $o_{ik} = 1$ for vectors belonging to class k, i.e., when $k = c_i$, and $o_{ik} = -1$ otherwise.

Let us assume that the network's hidden layer consists of L neurons and that $\mathbf{b} \in \mathbb{R}^L$ is a vector containing the hidden layer neurons bias values, which are randomly chosen as well. For a given activation function $G()$, the output $\tilde{\mathbf{o}}_i$ of the ELM network corresponding to training vector $\mathbf{x}_i$ is given by:

$$\tilde{o}_{ik} = \sum_{j=1}^{L} \mathbf{W}_{out,k}^T \, G(\mathbf{W}_{in,j}, b_j, \mathbf{x}_i), \; k = 1, ..., C. \qquad (5)$$

Many activation functions $G()$ can be used for the hidden layer neurons' output calculation, such as sigmoid, sine, Gaussian and hard-limiting function. In our experiments we have used the sigmoid function. That is, in our case:

$$G(\mathbf{W}_{in,j}, \mathbf{b}, \mathbf{x}_i) = \frac{1}{1 + \exp^{-(\mathbf{W}_{in,j}^T \, \mathbf{x}_i + b_j)}}, \qquad (6)$$

where $\mathbf{W}_{in,j}$ is the j-th column of $\mathbf{W}_{in}$.

By storing the hidden layer neurons outputs in a matrix $\mathbf{G}$, the network's outputs can be written in a matrix form as $\tilde{\mathbf{O}} = \mathbf{G}^T \mathbf{W}_{out}$. Finally, by assuming that the network's predicted outputs $\tilde{\mathbf{O}}$ are equal to the network's desired outputs

$\mathbf{O}$, $\mathbf{W}_{out}$ can be analytically calculated, i.e., $\mathbf{W}_{out} = \mathbf{G}^{\dagger}\mathbf{O}$, where $\mathbf{G}^{\dagger}$ is the Moore-Penrose generalized pseudo-inverse of $\mathbf{G}^T$, i.e., $\mathbf{G}^{\dagger} = \left(\mathbf{G}\mathbf{G}^T\right)^{-1}\mathbf{G}$.

Huang et. al. [13] have, recently, proposed a constrained optimization ELM training procedure, aiming to increase the generalization properties of ELM network, where $\mathbf{W}_{out}$ can be calculated by:

$$\mathbf{W}_{out} = (\mathbf{G}\mathbf{G}^T + \frac{1}{\Lambda}\mathbf{I})^{-1}\mathbf{G}\mathbf{O}^T. \tag{7}$$

After $\mathbf{W}_{out}$ calculation, a test vector $\mathbf{x}_{test}$ can be introduced to the ELM network and be classified to the class corresponding to the highest network's output:

$$c_{test} = arg \max_{k} \tilde{o}_{test,k}. \tag{8}$$

D. Training Phase

Let us assume that the action videos appearing in the video database $\mathcal{V}$ have been preprocessed following the procedures described in Subsections II-A and II-B, resulting to $N_T = N_I \cdot N_C$ training action vectors $\mathbf{s}_{ij},\ i = 1, ..., N_I,\ j = 1, ..., N_C$, followed by the corresponding person ID and action class labels $h_{ij} \in \{1, ..., N_P\}$ and $a_{ij} \in \{1, ..., N_A\}$, respectively. By using $\mathbf{s}_{ij}$ and h_{ij}, we train one ELM network having L_P hidden neurons, that will perform person identification on action videos. Since action execution style variations may uniquely characterize persons, action class information will, probably, help the identification process. This coincides with the human perception, as humans firstly identify that there is a person walking, for example, and then they may identify the person's ID through his/her body shape and walking style properties. In order to incorporate the action class information in the identification process, we exploit the action class labels a_{ij} of the training action vectors $\mathbf{s}_{ij}$ in order to train a second ELM network having L_A hidden neurons, that will be used for human action recognition on action videos.

In the case of person identification using a multi-camera setup, it is expected that a test action instance will be captured by multiple cameras $N \leq N_C$ and, thus, it will be depicted in N action videos. Since all these N action videos correspond to the same action instance performed by the same person, we would like to combine the person identification and action recognition results corresponding to all these N action videos, in order to provide the final identification result. To this end, we exploit the available labeling information of the training action vectors $\mathbf{s}_{ij}$ in order to train a third ELM network, having L_C hidden neurons, that will be used for classification results combination.

After training the person identification and action recognition ELM networks, we introduce the training action vectors $\mathbf{s}_{ij}$ to both these networks and we obtain their outputs, $\tilde{\mathbf{o}}_{ij,h} \in \mathbb{R}^{N_P}$ and $\tilde{\mathbf{o}}_{ij,a} \in \mathbb{R}^{N_A},\ i = 1, ..., N_I,\ j = 1, ..., N_C$, respectively. ELM outputs $\tilde{\mathbf{o}}_{ij,h}$ and $\tilde{\mathbf{o}}_{ij,a}$ corresponding to the same action instance are used in order to produce feature vectors $\mathbf{q}_{i,h}$ and $\mathbf{q}_{i,a}$ corresponding to the mean action instance ELM outputs, that is:

$$\mathbf{q}_{i,h} = \frac{1}{N_C}\sum_{j=1}^{N_C} \tilde{\mathbf{o}}_{ij,h}, \tag{9}$$

$$\mathbf{q}_{i,a} = \frac{1}{N_C}\sum_{j=1}^{N_C} \tilde{\mathbf{o}}_{ij,a}. \tag{10}$$

$\mathbf{q}_{i,h}$ and $\mathbf{q}_{i,a}$ are concatenated in order to produce the third ELM network training vectors, i.e., $\mathbf{q}_i = [\mathbf{q}_{i,h}\ \mathbf{q}_{i,a}]^T \in \mathbb{R}^{N_P+N_A}$. By using $\mathbf{q}_i$ and the corresponding person ID labels h_i, we train the third ELM network for person ID and action class classification results fusion.

E. Person Identification (test phase)

Let a person appearing in the database $\mathcal{V}$ perform an action instance appearing in the action class set $\mathcal{A}$, which is captured by N cameras. Clearly, in the case of single-view person identification $N = 1$, while in the case of multi-view person identification $N \leq N_C$. The case $N < N_C$ may appear either when the test camera setup consists of fewer cameras, compared to the training camera setup, or when the person under consideration performs the action outside some cameras field of view, or when he/she is occluded. The resulted N action videos are preprocessed by following the procedures described in Subsections II-A and II-B and N action vectors $\mathbf{s}_{test\ i},\ i = 1, ..., N$ are obtained. These vectors are introduced to the person identification and action recognition ELM networks and the networks' outputs, $\tilde{\mathbf{o}}_{test\ i,h}$ and $\tilde{\mathbf{o}}_{test\ i,a}$, are obtained. In the case of single-view person identification, the recognized person ID is the one corresponding to the highest person identification ELM output. In the case of multi-view person identification, the mean ELM output vectors $\mathbf{q}_{test,h}$ and $\mathbf{q}_{test,a}$ are calculated and concatenated in order to produce the feature vector $\mathbf{q}_{test}$. Finally, $\mathbf{q}_{test}$ is introduced to the recognition results combination ELM network and the recognized person ID is the one that corresponds to its highest output.

III. Experimental Results

In this section we illustrate experiments conducted in order to assess the performance of the proposed person identification method. We have used two publicly available action databases aiming at different application scenarios. The first one is a multi-view database containing everyday actions, while the second one is a single-view database containing actions appearing in meal intakes. In all the experiments presented in this Section the following parameter values have been used: $N_H = N_W = 32$, $m = 1.1$, $\eta(0) = 0.1$, $\sigma(0) = \frac{N_x+N_y}{2}$. Regarding the optimal ELM regularization parameter value and number of hidden neurons, they have been determined by performing grid search using the values $\Lambda = 10^{\lambda}$, $\lambda = -6, \ldots, 6$, $L_P = L_A = [100, 250, 500, 1000]$ and $L_C = [50, 100, 200]$.

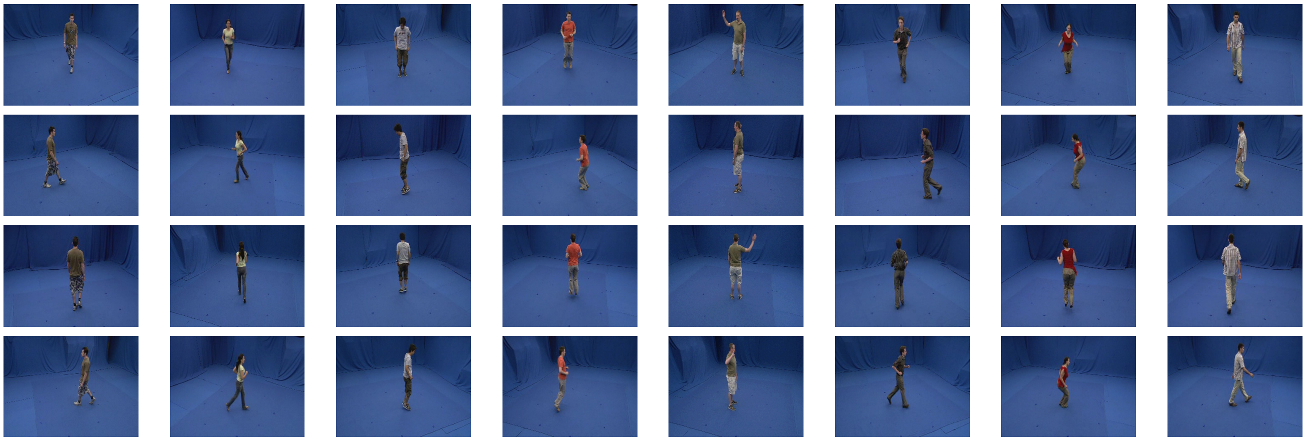


Fig. 3. *Video frames depicting the person of the i3DPost database captured by four viewing angles during action execution.*

A. The i3DPost Database

The i3DPost database [14] contains high resolution (1080×1920 pixels) videos depicting eight persons. 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. Each person performs one or more instances of eight action classes: 'walk', 'run', 'jump in place', 'jump forward', 'bend', 'sit', 'fall' and 'wave one hand'.

The LOIO cross-validation procedure has been performed to the i3DPost database by using the action videos belonging to action classes that contain more than one instances per person. That is, the action videos used in our experiments belong to one of the following action classes: 'walk', 'run', 'jump in place', 'jump forward' and 'wave one hand'. Actions 'bend', 'sit' and 'fall' were not used as each person performs these actions once. Binary action videos have been created by applying a color based image segmentation technique discarding the blue background. In these experiments, the entire human body is used to describe the human body poses. Example action video frames depicting all the persons of the database captured by four different viewing angles are illustrated in Figure 3.

In our first set of experiments, we have performed the LOIO cross-validation procedure by using different SOM topologies and all the eight cameras of the database in both training and test phases, i.e., $N = N_C = 8$. The optimal SOM topology was found to be equal to $N_x = N_y = 14$, which provided a person identification rate equal to 95.63%. The confusion matrix of this experiment is illustrated in Figure 4.

In a second set of experiments we tried to simulate the cases of different cameras setups between the training and test phases, as well as the cases of person occlusion in one or more cameras in the test phase. To this end, we performed the LOIO cross-validation procedure multiple times by using different numbers of cameras in the test phase. That is, during one experiment, the algorithm was trained by using all the eight cameras in the database. In the test phase, only $N < N_C$, randomly chosen, action videos were used to identify the depicted person. By using only one camera $N = 1$, an identification rate equal to 73.125% has been obtained. By using $N = 2$ and $N = 3$ cameras during the test phase, identification rates equal to $80.\%$ and 81.25% have been obtained. In Table I, we compare the performance of the proposed method with the methods in [7], [6] for different numbers of test cameras. As can be seen, by using only one camera in the test phase, i.e., when $N = 1$, the person identification rates are, relatively, low. This is closely related to the viewing angle effect, since it is possible that one person when captured from one viewing angle performing one action will be similar to another person captured from another camera performing the same or another action. By using a higher number of cameras during testing, the viewing angle effect is better addressed and, thus, higher person identification rates are obtained. Furthermore, in this Table, it can be seen that the use of nonlinear classification and classification results fusion schemes can better describe person ID, as well as, action classes, leading to increased, up to 9.46%, person identification performance.

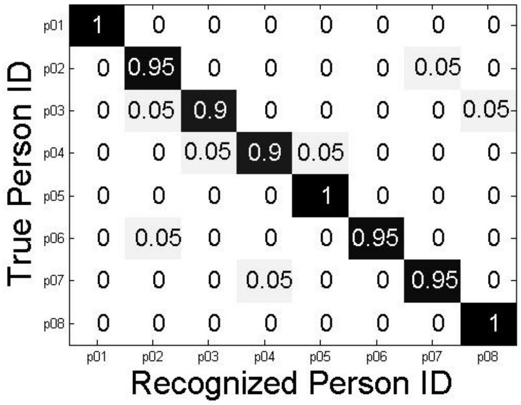


Fig. 4. *Confusion matrix on the i3DPost database.*

B. The AIIA-MOBISERV Database

The AIIA-MOBISERV database [15], [16] contains low resolution (640×480 pixels) videos depicting twelve persons.

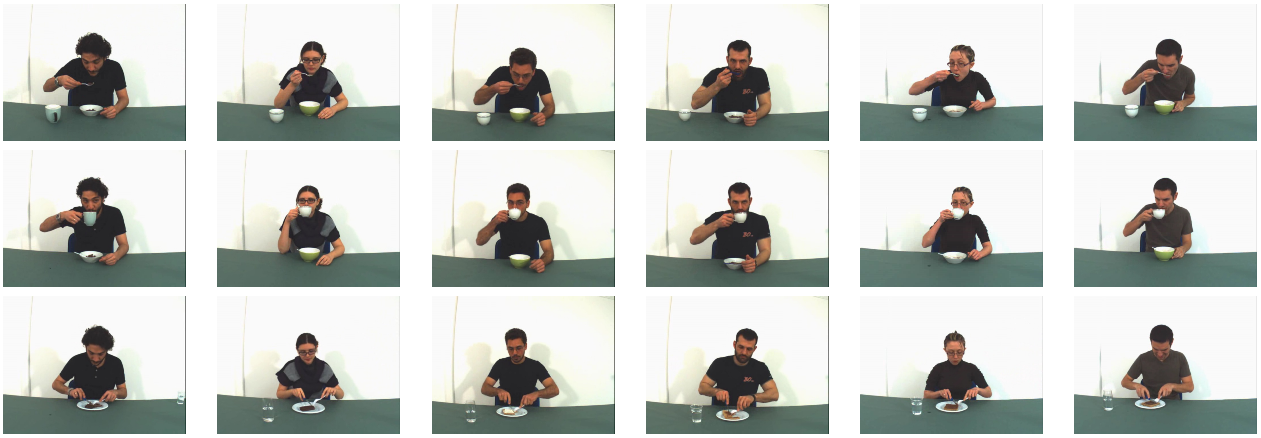
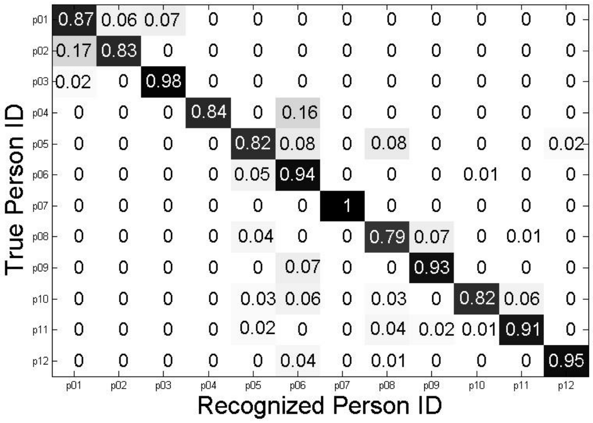


Fig. 5. *Video frames depicting six persons of the AIIA-MOBISERV database during a meal.*

TABLE I
Comparison results in the i3DPost database for different numbers of cameras N

N	1	2	3	8
Method [6]	–	–	–	94.37%
Method [7]	71.68%	70.54%	80.26%	94.34%
Proposed Method	**73.125%**	**80%**	**81.25%**	**95.63%**



	p01	p02	p03	p04	p05	p06	p07	p08	p09	p10	p11	p12
p01	0.87	0.06	0.07	0	0	0	0	0	0	0	0	0
p02	0.17	0.83	0	0	0	0	0	0	0	0	0	0
p03	0.02	0	0.98	0	0	0	0	0	0	0	0	0
p04	0	0	0	0.84	0	0.16	0	0	0	0	0	0
p05	0	0	0	0	0.82	0.08	0	0.08	0	0	0	0.02
p06	0	0	0	0	0.05	0.94	0	0	0	0.01	0	0
p07	0	0	0	0	0	0	1	0	0	0	0	0
p08	0	0	0	0	0.04	0	0	0.79	0.07	0	0.01	0
p09	0	0	0	0	0	0.07	0	0	0.93	0	0	0
p10	0	0	0	0	0.03	0.06	0	0.03	0	0.82	0.06	0
p11	0	0	0	0	0.02	0	0	0.04	0.02	0.01	0.91	0
p12	0	0	0	0	0	0.04	0	0.01	0	0	0	0.95

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

A camera was placed at a distance of 2m in front of them during a meal. Four meals (instances) were recorded for all the persons, each for a different day. The persons perform multiple iterations 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 their hands. The persons can drink from a cup or a glass. Finally, action class ’apraxia’ contains actions ’slicing food’ and ’rest’.

The LOIO cross-validation procedure has been performed to the AIIA-MOBISERV database by using the action videos depicting the persons eating using a fork and drinking from a cup. In this case, the human body ROIs were determined to be the human body skin regions, i.e., his/her head and hands. Binary action videos have been created by applying a color-based image segmentation technique on the action video frames [9]. Example action video frames depicting six of the persons of the database are illustrated in Figure 5.

The LOIO cross-validation procedure has been performed multiple times by using different SOM topologies. A person identification rate equal to 89.67% has been obtained by using a 15×15 SOM. The confusion matrix of this experiment is illustrated in Figure 6. Similar to the i3DPost case, it can be seen that nonlinear classification and fusion schemes can better describe the nonlinear structure of person ID and action classes, since the proposed method outperforms the method in [7], which achieved a person identification performance equal to 87.83%.

IV. Conclusion

In this paper, we presented a person identification method exploiting human motion information, based on fuzzy action video representation and ANN based action video classification. Action videos are represented by the fuzzy similarity between the human body poses appearing in them and representative human body poses determined by training a self organizing network. Feedforward networks are responsible for action video classification. The combination of multiple person identification and action recognition results corresponding to the same action instance captured by different viewing angles leads to view-independent person identification with high identification rates. The proposed method can be applied to different application scenarios without any modification.

Acknowledgements

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.

REFERENCES

[1] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.

[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] S. Yu, D. Tan, and T. Tan, "Modeling the effect of view angle variation on appearance-based gait recognition," *Phys. Rev. B.*, pp. 807–816, 2006.

[4] 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. 412–424, 2012.

[5] 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, pp. 347–360, 2012.

[6] 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.

[7] ——, "Activity based person identification using fuzzy representation and discriminant learning," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 530–542, 2012.

[8] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," *International Conference on Pattern Recognition*, pp. 1051–4651, 2009.

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

[10] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 2002.

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

[12] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," *IEEE International Joint Conference on Neural Networks*, pp. 985–990, 2004.

[13] G. 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*, vol. 42, no. 2, pp. 513–529, 2012.

[14] 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.

[15] A.-M. Eating and D. Database, "http://poseidon.csd.auth.gr/MOBISERV-AIIA/index.html."

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