Activity based Person Identification using Fuzzy Representation and Discriminant Learning

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

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

Abstract

In this paper, a novel view invariant person identification method based on human activity information is proposed. Unlike most methods proposed in the literature, in which 'walk' (i.e., gait) is assumed to be the only activity exploited for person identification, we incorporate several activities in order to identify a person. A multi-camera setup is used to capture the human body from different viewing angles. Fuzzy Vector Quantization and Linear Discriminant Analysis are exploited in order to provide a discriminant activity representation. Person identification, activity recognition and viewing angle specification results are obtained for all the available cameras independently. By properly combining these results, a view-invariant activity-independent person identification method is obtained. The proposed approach has been tested in challenging problem setups, simulating real application situations. Experimental results are very promising.

I. INTRODUCTION

Persons identification hereafter, has been heavily researched in the past few decades, due to its importance in security applications. Most of the existing person identification approaches, such as the ones based on face, iris and fingerprint recognition, assume restricted identification setups and person cooperation. For example, the person
should stand at a standard distance in front of a camera and look at a specific point, or have physical contact with sensors. In order to overcome these restrictions, non-invasive biometrics have been exploited, refering to the anatomical or behavioral traits associated with a specific person that can be used for automatic recognition [1]. Gait recognition, i.e., the identification of individuals by the way they walk, has been widely used for this purpose, as it provides a non-invasive way to recognize persons at a distance. Gait recognition has gained researchers’ attention in the last decade and numerous such methods have been proposed in the literature [2], [3], [4], [5].

One disadvantage of gait recognition is the assumption that the person under investigation walks, which is not always the case. Most methods proposed in the literature would probably fail in the case where the person performs a different activity, for example if he/she bends. Thus, the activity information should be taken into account in order to provide the correct person ID. Despite the fact that gait recognition has been studied a lot, the use of other activities has not been exploited yet for person identification. In fact, ‘walk’ can be seen as a special case of a wide range of human activities, such as ‘run’, ‘bend’, ‘jump’, ‘eat’ or ‘drink’, which can be used in order to reveal the identity of a person depicted in a video stream. Similar to gait, the global human body information, in the manner of human body proportions and shape, is conserved while observing a person performing other activities. In addition, dynamics observed in different activities may be very distinctive. That is, although people walk in quite a similar way, they may perform other activities, like eating, quite differently. This means that it is more probable to achieve good identification performance if we exploit several, possibly all, different activities a person performs. Indeed, other activities, besides walk, may contain more discriminant information for person identification, as execution style of one activity may uniquely describe a person.

We are interested in non-invasive person identification using multiple cameras that observe the person under investigation, while he/she performs several every day activities. It is evident that the viewing angle plays a significant role in the person identification accuracy [6], [7]. This is due to the fact that the shape of the human body, as well as the body dynamics observed during the execution of activities, differ a lot when the person is observed by different viewing angles. This is the so-called viewing angle effect. Most person identification methods proposed in the literature use one camera and assume the same viewing angle during the training and recognition phases. This renders them inappropriate for applications aiming to view-invariant person identification. In order
to provide a non-restrictive person identification setup, the identification accuracy method should ideally not be affected by the viewing angle.

As previously mentioned, the activity-based person identification approach is relatively new. Researchers have focused their efforts on the identification of persons through gait. For this reason, the problem addressed by most methods can be regarded as a special case of the problem addressed by the proposed method. Next, we review a few methods that exploit activity information for person identification. Furthermore, because gait can be seen as a special case of the proposed approach, we review several state-of-the-art gait recognition methods.

A method that extends gait recognition to include running activity is presented in [8]. Temporal template matching is applied to extract the angles of lower leg rotation during the entire gait cycle. The magnitude of the Fourier transform of these rotation signals over time provides a gait representation. Classification is achieved using a k-nearest neighbor framework. While this method can incorporate running in person identification, the use of other activities in such a framework is not straightforward. This is because most activities, such as ‘wave one hand’, ‘eat’ or ‘drink’, cannot be suitably described in this approach. Furthermore, it assumes recognition of the lower part of the human body. The gait representation is based on edge detection followed by template matching and is prone to false positives. In addition, this method uses only one camera and its performance is closely related to the viewing angle. An activity-based person identification method is presented in [9]. Human body poses, described by binary body masks, are used to identify persons while executing activities. Fuzzy Vector Quantization (FVQ) is used to associate body poses forming an activity with reference body poses obtained in the training phase. Linear Discriminant Analysis (LDA) is used to map the activity representation in a low-dimensional discriminant feature space. In this space, the human body activity representation in a test sequence is classified to the nearest class centroid using cosine similarity. This method exploits a robust human body representation. Binary images denoting human body poses can effectively be obtained by using background subtraction techniques. Activities used for person identification besides ‘walk’ are ‘run’, ‘skip’, ‘gallop sideways’, ‘jump jack’, ‘jump forward’ and ‘jump in place’. It should be noted that experiments have proven that the most distinctive activity was found to be ‘skip’, while ‘walk’ was ranked sixth in discriminant capacity (out of 8 activities). The main problem of this method is that it uses one camera only. Thus the viewing angle should be the same in training and test phases.
Regarding gait recognition methods, the main effort has been focused on finding a convenient representation of the human body, able to highlight the differences between individuals during a walk cycle. One widely used gait representation is the Gait Energy Image (GEI) [10], which is a grayscale image obtained by averaging the silhouettes extracted over a gait cycle. A similar representation is Motion Silhouette Image (MSI) [11]. In MSI, the intensity of each pixel is computed as a function of motion in the temporal dimension over a complete gait cycle. Gait Entropy Image (GEnI) is proposed in [12]. GEnI is a gray-scale image, whose pixels are determined by calculating their entropy in a complete gait cycle. While GEI and MSI capture the shape information of the person during a gait cycle, GEnI highlights the dynamic body areas. After determining the gait representation, most of the proposed methods exploit subspace techniques in order to map this representation into a more discriminant space, where the classification is performed either to the nearest class centroid or to the k-nearest neighbors. GEnI is followed by Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in [12] and classification is achieved using a nearest neighbor procedure.

A fused classifier is proposed in [10]. GEIs produced by real training data and GEIs produced by synthetic data are mapped to discriminant subspaces by applying PCA and LDA. In these spaces, real and synthetic GEIs are classified to the nearest class centroid and identification results are combined in order to provide the ID of the person depicted in a test sequence. A gait feature selection technique that combines GEI and GEnI is described in [12]. Adaptive Component and Discriminant Analysis (ACDA) is proposed in order to map the resulting gait features in a discriminant subspace. ACDA is based on Component Discriminant Analysis (CDA) and incorporates the gait feature selection procedure in the discriminant analysis. General Tensor Discriminant Analysis (GTDA) is proposed in [13]. A Gabor-based gait representation is obtained by decomposing GEI by Gabor filters. GTDA is introduced as a discriminant analysis technique, which reduces the data dimensionality, while preserving the discriminant information.

The aforementioned methods exploit information captured by only one camera. Thus, their performance is closely related with the viewing angle. Furthermore, they use a compact human body representation, i.e., the human body is represented by a single image. Similar human body representations have been proposed for activity recognition. Motion Energy Image (MEI) and Motion History Image (MHI) have been proposed in [14] as compact activity
representations. MEI denotes image locations occupied by the human body during activity execution, while pixel values of MHI are a function of motion. Although these representations provide good classification rates for simple activity patterns, different activity types provide the same MEI and MHI human body representation. This affects the performance of the classification method. Thus, temporal information should be taken into account in order to distinguish similar activity patterns.

Several methods that exploit the temporal information of a gait sequence have been proposed. Consecutive human body poses forming a gait cycle are used in [15]. These poses are normalized using a population Hidden Markov Model (pHMM), in order to emphasize the shape aspects of gait. PCA followed by LDA is performed to the normalized poses, in order to reduce feature dimensionality and increase class discrimination. Finally, the similarity between two normalized sequences is defined as the sum of Euclidean distances between human body pose representations in the LDA space. Consecutive human body poses are also used in [16]. Radon transform is performed to these poses and a 3D volume, whose third dimension is time, is obtained, by exploiting the linear property of the Radon transform. Classification is achieved by mapping this representation to a discriminant subspace using LDA and using a nearest class centroid approach. These methods preserve the information contained in the consecutive body poses, but they utilize one camera only. Therefore, their performance depends on the viewing angle.

In order to overcome the viewing angle effect, three main approaches have been proposed so far: a) the introduction of view-invariant gait representations, b) mapping of the gait representation from an arbitrary viewing angle to a reference one and performing identification in this viewing angle and c) using multi-camera setups. In [17], one camera was used to capture the human body from an arbitrary viewing angle. Human body parts were detected and tracked during a gait cycle. 2D trajectories of head and feet of the depicted person were normalized and a convenient view-invariant gait representation was achieved. A method that projects images from an arbitrary view to a reference one, using perspective projection was proposed in [18]. Both these methods do not provide a fully view-invariant gait representation, as self occlusions, observed during a gait cycle, significantly affect the resulting gait representation. Methods belonging to the second category use a View Transformation Model (VTM), in order to map sequences from an arbitrary viewing angle to a reference one. This is done through a factorization
process, usually performed by Singular Value Decomposition (SVD). A Fourier-based gait representation was obtained in [19], by applying Discrete Fourier Transform (DFT) to human poses consisting a gait cycle, while an optimized GEI is proposed in [20]. After calculating the gait representations, a SVD-based VTM was applied. While these methods can handle the viewing angle effect, the final gait representation is not robust to noise. The use of multi-camera setups has been proposed in [21], [22]. In [21], a 3D human body representation was obtained, by applying image-based rendering techniques. Combining information coming from multiple cameras, a synthetic human body representation viewed by the same angle as the one in the training phase was created. The same idea was implemented by calculating the visual hull of the person in [22]. These techniques exploit the enriched human body information captured by multiple cameras, but they have a large computational cost. Furthermore, they require a calibrated camera setup and that the person must be visible from all cameras. In other cases, the human body representation may be affected leading to reduced recognition performance or failure.

In this paper, we propose a novel method for activity-based person identification. Walk is assumed to be one, but not the only, activity, which will be exploited for person identification. We use an uncalibrated multi-camera setup, in order to capture the human body from different viewing angles. The person is represented by his/her body poses during activity execution. This is a low computational cost representation, as it does not involve a 3D reconstruction of the human body. In order to exploit the static body information, provided by the human body poses forming an activity, we utilize dynemes, which were previously proposed for activity recognition [23], [24]. By calculating the similarity of each test human body pose with all dynemes, the temporal information is preserved. The classification procedure involves the projection of the final activity representation in a low dimensional discriminant subspace using LDA. For each of the cameras, person identification, activity recognition and viewing angle specification is performed independently. By properly combining the results produced for all the available cameras, a view-invariant and activity-independent person identification method is obtained. Furthermore, the proposed approach addresses most of the open issues appearing in person identification, that will be discussed in the subsection II-A.

The main contributions of this paper are: 1) The use of activities (besides ‘walk’) in person identification. 2) The use of dynemes for person identification. In [23], it has been shown that dynemes can be used for activity recognition. Here, we show that dynemes can also be used for activity-based person identification. Thus, dynemes can
be considered as similar human body poses representing similar persons. 3) The use of cumulative fuzzy distances from the dynemes to achieve time-invariant person identity representation. 4) The use of a Bayesian framework to combine the person identification, activity recognition and viewing angle specification results produced for all the available cameras in the identification phase.

In the following, each step of the proposed method is described in detail. The procedure followed in order to obtain the human body poses is presented in subsection II-B. Dynemes determination and activity representation are described in subsections II-C and II-D. LDA projection and the calculation of discriminant human ID, activity class and viewing angle representations are described in subsection II-E. The procedure followed in order to combine person identification, activity classification and viewing angle specification from all the available cameras is presented in subsection II-F. Finally, subsection II-G presents the person identification procedure followed in the test phase.

II. DYNEME-BASED PERSON IDENTIFICATION

A. Problem Statement

Let \( \mathcal{U} \) be an annotated person identification database consisting of \( N_P \) persons, in which each person performs multiple instances of \( N_A \) activities. In this paper, the term activity refers to simple human motion patterns (cycles), which may be periodic, e.g., in a 'walk' sequence, or not, e.g., in a 'bend' sequence, shown in Figure 1. Each person in the database is captured by \( N_C \) cameras. \( t \) frames \( f_i, i = 1, ..., t \) from a specific camera form a single-view video \( \mathbf{F} = [f_1^T, f_2^T, ..., f_t^T]^T \). A single-view video depicting one activity instance, e.g., a walk step, is referred to as activity video. Let \( \mathcal{A} = \{a_j\}, j = 1, ..., N_A \) be an activity set denoting \( N_A \) activity classes, such as walk, jump, run, etc. An activity instance performed by one person is depicted in \( N_C \) activity videos, each corresponding to a different viewing angle. Since activities differ in duration, the number \( t \) of video frames of activity videos belonging to different activity classes may differ. For example, a walk step may consist of only 10 video frames \((t = 10)\), whereas a bend sequence of 40 video frames \((t = 40)\) in a 25 fps video. Let a person of the database under investigation perform a novel (not included in the person identification database) activity instance, belonging to an activity existing in the activity set \( \mathcal{A} \). Person identification refers to the association of a person depicted in one or more test video sequences with a known person included in the training database.
This task is not trivial, as there are many challenges that a person identification method should be able to face. The assumption that the person under investigation performs one activity only in a test video is not always valid. For example, while he/she walks, he/she may start running. In addition, activities differ in duration. This is observed even in different instances of the same activity performed by different persons, or in different instances of the same activity performed by the same person.

It is possible that different persons are similar, in terms of body volume. In this case, style variations observed during the execution of some activities may provide person discriminant information. However, this does not necessarily happen for all the viewing angles. Thus, it is possible that one viewing angle has more person-discriminant capacity.

In the cases of multiple cameras, the identification of the position of each camera with respect to the human body must be solved, in order to perform view-invariant person identification. Moreover, each camera may capture the person from an arbitrary distance, affecting the area occupied by the projection of the human body in each camera plane. The person may be visible from $N \leq N_C$ cameras. This can be observed in cases where the person is outside of some cameras field of view, or in the case of occlusions. Cameras forming the camera setup may differ in resolution and frame rate. Synchronization errors are usual in real multi-camera setups, which means that there might be a delay between the captured frames by different cameras. Differences in resolution, frame rate and synchronization can affect person identification accuracy.

Continuous person identification over time should be allowed that be easily applied in different applications, such as home surveillance and sports monitoring. Finally, the overall computational load should be small, so that (near) real-time person identification can be performed. In order to perform person identification in a non restrictive manner, all the aforementioned issues have to be addressed properly. The proposed method handles all the above
mentioned issues.

B. Preprocessing Phase

A person performing an activity is captured by $N$ cameras to create $N$ activity videos. In the training phase, the number of cameras is fixed and is equal to $N_C$. In the test phase, the number $N$ of cameras is not necessary predefined. That is, we may use 8 cameras for training and we should be able to perform identification using, for example, the videos captured by 3 cameras. Activity videos consist of $t$ video frames. As activities vary in duration, the number of video frames $t$ differs for different activity classes. This is observed even for different instances of the same activity class performed either by the same or by different persons. To simplify notation, we assume that activity videos of the same activity class $a_j$ consist of the same number of $t_j$ video frames, although this is not necessary true, as explained in the previous section. However, as it will be explained in next subsections, the proposed method can cope with activity classes of varying duration. During training, multi-period videos, i.e., videos depicting multiple activity instances, are manually split to produce $T$ activity videos, which are subsequently used for training. In the case of continuous person identification (test phase) in videos depicting a person performing multiple instances of multiple activity classes, a sliding window of overlapping video segments consisting of $t_w$ video frames is used and person identification is performed at each window position.

Video segmentation techniques, such as background subtraction [26], [27] or color-based image segmentation methods [28] can be applied to activity video frames to produce binary images (masks) denoting the human body regions of interest (ROIs). These images are centered to center of mass of the binary mask and bounding boxes of size equal to the size of the maximum bounding box that encloses the person’s masks in each activity video are extracted and rescaled to $H \times W$ pixels to produce binary images of fixed size to obtain a scale-invariant human body image representation. For the experiments presented in this paper, we chose the size of the resulted binary images to be equal to $32 \times 32$ pixels, which has been found experimentally to be a good compromise between computational cost and person identification accuracy. Seven binary images denoting body poses during the execution of seven activities (‘walk’, ‘run’, ‘jump in place’, ‘jump forward’, ‘wave one hand’, ‘eat’ and ‘drink’) captured from various viewing angles are shown in Figure 2.

The resulting binary images are vectorized to produce the so-called posture vectors $p_{ik} \in \mathbb{R}^C$, $i = 1, ..., T$, $k =$
Fig. 2. Posture frames of human poses during the execution of seven activities depicted from various viewing angles. From left to right 'walk', 'run', 'jump in place', 'jump forward', 'wave right hand', 'eat' and 'drink'. Posture vectors \( p_{ik} \) are produced by vectorizing posture frames forming an activity video.

\[ 1, \ldots, t_j, C = H \times W, \] where the first index refers to the activity video and the second one the video frame number associated with this vector. That is, every activity video \( i \) consisting of \( t_j \) video frames is represented by a set of posture vectors \( p_{ik} \). In our experiments we have scanned the resulted binary images in a column-wise manner. However, one could produce \( p_{ik} \) by scanning the binary masks row-wise.

C. Dynemes Calculation

In the training phase, posture vectors \( p_{ik}, \ i = 1, \ldots, T, \ k = 1, \ldots, t_j \), having \( t_j \) video frames each, are used, without exploiting the available labeling information (person ID, activity class and viewing angle), to produce the dynemes. This is achieved by applying \( K \)-Means clustering \([29]\), which clusters the training posture vectors \( p_{ik} \) to \( D \) clusters, so as to minimize the within-cluster sum of squares:

\[
\sum_{d=1}^{D} \sum_{i=1}^{T} \alpha_{id} \left\| p_{ik} - v_d \right\|^2,
\]

where \( \alpha_{id} = 1 \), if \( p_{ik} \) is assigned to the cluster \( d \) and \( \alpha_{id} = 0 \), otherwise. Dynemes \( v_d, \ d = 1, \ldots, D \) are defined as the cluster centers, each cluster consisting of \( n_d = \sum \alpha_{id} \) posture vectors:

\[
v_d = \frac{1}{n_d} \sum_{i=1}^{T} \alpha_{id} p_{ik}.
\]

Other clustering techniques, such as Spectral Clustering \([30]\), Self Organizing Maps \([31]\), Fuzzy \( C \)-Means \([32]\), etc, can be used for dyneme calculation. The choice of clustering algorithm was proven to have only minor impact in person identification accuracy. We found experimentally that \( K \)-Means performs well for the specific task and is faster than other clustering algorithms. This is why we chose \( K \)-Means for dynemes calculation. The optimal number of dynemes is determined using the cross-validation procedure. It is a procedure that determines the ability of a learning algorithm to have the best person identification accuracy on data that it was not trained on. The learning algorithm is trained multiple times (folds) using all but some data, which are subsequently used for evaluation.
Depending on the evaluation data, there are several cross-validation procedures. In our case, the activity videos depicting one activity instance of each activity class for all the persons were used for evaluation. A set of 110 dynemes calculated over activity videos depicting eight persons performing five activities (‘walk’, ‘run’, ‘jump in place’, ‘jump forward’ and ‘wave one hand’), captured by eight viewing angles ($0^\circ$, $45^\circ$, $90^\circ$, $135^\circ$, $180^\circ$, $225^\circ$, $270^\circ$ and $315^\circ$, with respect to the frontal person’s view in a clock-wise manner) is illustrated in Figure 3.

![Figure 3](image.png)

**Fig. 3.** 110 dynemes obtained by clustering the posture vectors coming from activity videos from 8 persons performing 5 activities captured from 8 viewing angles. Dynemes are vectorized to produce $v_d$.

By observing Figure 3, it can be seen that dynemes preserve the human body shape information, i.e., they depict similar body poses of similar persons captured by the same viewing angle. For example, one can observe that dynemes $5b$, $8d$ and $6g$ correspond to poses of females, while dynemes $1f$, $5h$, $9i$ and $10j$ correspond to poses of males. It can also be seen that the characteristics of the persons corresponding to dynemes $8d$ and $6g$ are quite similar and different from those corresponding to the dyneme $5b$. Furthermore, dynemes contain activity information. For example, dynemes $1c$, $1d$, $1f$, $3h$, $3i$ and $9c$ correspond to activity ‘wave one hand’ depicted from different viewing angles. Finally, it can be seen that dynemes can provide the viewing angle information. For example dynemes $1g$, $2a$, $4b$, $9a$ and $9i$ depict a $90^\circ$ side view, while dynemes $1e$, $2e$, $5c$, $7a$ and $8d$ depict a $270^\circ$ side view.

**D. Activity Representation**

Let $t_j$, $j = 1, ..., N_A$ posture vectors $p_{ik}$, $k = 1, ..., t_j$ describing the $i$-th activity video. We can form an alternative human body representation for poses found in activity video $i$, by mapping all posture vectors $p_{ik}$ to the dynemes $v_d$, $d = 1, ..., D$. Fuzzy distances of every $p_{ik}$ to all the dynemes $v_d$, $d = 1, ..., D$ are calculated to
determine the similarity of every posture vector with every dyneme:

\[ d_{ikd} = \left( \| p_{ik} - v_d \|_2 \right)^{-\frac{2}{m-1}} \]  

(3)

where \( m \) is the fuzzification parameter (\( m > 1 \)), which is set equal to 1.1 for all the experiments presented in this paper. Fuzzy distances allow for a smooth distance representation between posture vectors and dynemes [23]. The use of other distance metrics, such as \( L_1 \), \( L_2 \) or Mahalanobis distances, can be utilized to this end. However, it has been experimentally found that the use of fuzzy distances outperform these alternative choices.

After the calculation of fuzzy distances, each posture vector is mapped to the following distance vector \( d_{ik} = [d_{ik1}, d_{ik2}, ..., d_{ikD}]^T \). Distance vectors \( d_{ik}, k = 1, ..., t_j \) are normalized to produce membership vectors \( u_{ik} \in \mathbb{R}^D \), which form the final representations of the posture vectors in the dyneme space:

\[ u_{ik} = \frac{d_{ik}}{\| d_{ik} \|}. \]  

(4)

The above described procedure is illustrated in figure 4. The mean vector \( s_i \in \mathbb{R}^D \) of all the \( t_j \) membership vectors comprising the \( i \)-the activity video is called activity vector and represents the \( i \)-th activity video:

\[ s_i = \frac{1}{t_j} \sum_{k=1}^{t_j} u_{ik}. \]  

(5)

![Fig. 4. Membership vector derivation procedure.](image)

The use of the mean vector leads to a duration-invariant activity representation. That is, we expect the normalized cumulative membership of a specific human body representation to be invariant to the duration of the activity. Thus, the activity representation can handle intra-class duration variations. Finally, the activity vectors representing all the \( T \) training activity videos \( s_i, i = 1, ..., T \) are normalized to have zero mean and unit standard deviation. In the test phase, all the \( N \) activity vectors \( s_i, i = 1, ..., N \) that correspond to \( N \) test activity videos depicting the person from different viewing angles are normalized accordingly.
E. LDA Projection

Since the activity vectors $s_i$ have high dimensionality (equal to the number of dynemes), we can use LDA in order to project activity vectors in a low-dimensional discriminant subspace exploiting the labeling information available in the training phase. LDA is a technique which is used to obtain the optimal projection matrix $W_{opt}$ which minimizes Fisher criterion:

$$W_{opt} = \arg \min_W \text{trace}\{W^T S_w W\} \over \text{trace}\{W^T S_b W\}. \quad (6)$$

$S_w$ and $S_b$ are the within class and between class scatter matrices, respectively:

$$S_w = \sum_{n=1}^{C} \sum_{i=1}^{T} r_n^i (s_i - \mu_n)(s_i - \mu_n)^T \quad (7)$$

$$S_b = \sum_{n=1}^{C} \frac{(\mu_n - \mu)(\mu_n - \mu)^T}{N_n} \quad (8)$$

where $C$ is the number of classes, $r_n^i = 1$ for training activity vectors $s_i$ belonging to class $n$ and $r_n^i = 0$ otherwise, $\mu_n$ is the mean vector of class $n$ having $N_n = \sum_{i=1}^{T} r_n^i$ training activity vectors and $\mu$ is the total mean vector of the training set. After calculating $W_{opt}$, the activity representation in the LDA space, the so-called discriminant activity vector $h_i$, is obtained by:

$$h_i = W_{opt}^T s_i. \quad (9)$$

Each training activity vector $s_i$ is followed by its human ID, activity class and viewing angle labels. We use this information in order to determine discriminant subspaces, in which the corresponding discriminant activity vectors of different classes are linearly separable. That is, we apply LDA to the activity vectors using the human ID labels, in order to determine the optimal discriminant subspace represented by the projection matrix $W_A$, in which the discriminant activity vectors representing different persons, noted as $h_i$, are linearly separable. Similarly, activity class and viewing angle labels are used in order to determine discriminant subspaces represented by the corresponding projection matrices $W_I$ and $W_V$, respectively, in which the discriminant activity vectors of different activity and viewing angle classes, noted as $a_i$ and $v_i$ respectively, are linearly separable. Thus, by using the same activity representation combined with different labeling information, we obtain appropriate human ID, activity class and viewing angle representations (Figure 5). In these spaces, the test discriminant vectors are classified to the nearest class centroid.
Fig. 5. Activity vector $s_i$ is used to produce discriminant vectors in different subspaces for human ID, activity class and viewing angle classification.

F. Person Identification based on a Bayesian Framework

During testing, a person performs an activity instance captured by $N \leq N_C$ cameras. This results in the creation of $N$ activity videos, each depicting the person performing the same activity from a different viewing angle. Each activity video is represented by three discriminant activity vectors $h_i$, $a_i$ and $v_i$, $i = 1, ..., N$, as described in subsection II-E. These discriminant activity vectors can be labeled according to the nearest class centroid. Thus, for each test activity video $i = 1, ..., N$, the human ID, activity class and viewing angle labels can be obtained. The person depicted in these $N$ activity videos can be identified using a probabilistic framework for decision fusion, which exploits classification results obtained for all $N$ available cameras, as described subsequently.

Let $\hat{h}_i$, $\hat{a}_i$ and $\hat{v}_i$ denote the human ID, activity class and viewing angle labels as estimated using the activity video captured by camera $i$. Let $P(\hat{h}_i)$, $P(\hat{a}_i)$ and $P(\hat{v}_i)$ denote the a priori probabilities corresponding to these recognition results and $P(j)$ be the a priori probability of occurrence of the $j$-th person. Let $P(\hat{h}_1, \hat{a}_1, \hat{v}_1, ..., \hat{h}_N, \hat{a}_N, \hat{v}_N)$ be the joint probability of all the recognition results provided for all the $N$ cameras and $P(\hat{h}_1, \hat{a}_1, \hat{v}_1, ..., \hat{h}_N, \hat{a}_N, \hat{v}_N|j)$ be the conditional probability of all the recognition results for the $j$-th person in the database. The a posteriori probability $P(j|\hat{h}_1, \hat{a}_1, \hat{v}_1, ..., \hat{h}_N, \hat{a}_N, \hat{v}_N)$ of the $j$-th person in the database given the recognition results $\hat{h}_i$, $\hat{a}_i$, $\hat{v}_i$, $i = 1, ..., N$ can be estimated using the formula:

$$P(j|\hat{h}_1, \hat{a}_1, \hat{v}_1, ..., \hat{h}_N, \hat{a}_N, \hat{v}_N) = \frac{P(\hat{h}_1, \hat{a}_1, \hat{v}_1, ..., \hat{h}_N, \hat{a}_N, \hat{v}_N|j) \cdot P(j)}{\sum_{l=1}^{N_P} P(\hat{h}_1, \hat{a}_1, \hat{v}_1, ..., \hat{h}_N, \hat{a}_N, \hat{v}_N|l) \cdot P(l)}$$

coming from a Bayesian framework.
Assuming that the training and evaluation data come from the same distributions, \( P(\hat{h}_1, \hat{a}_1, \hat{v}_1, \ldots, \hat{h}_N, \hat{a}_N, \hat{v}_N | j) \) can be estimated during the training phase. \( P(j) \) can be estimated by the frequency of person appearance in the training database. In the case of equiprobable human appearance, we can choose \( P(j) = \frac{1}{N_P} \). The main disadvantage of this approach is that the number of combinations of all the \( N_C \) cameras providing \( N_P \) human ID, \( N_A \) activity class and \( N_C \) viewing angle results is prohibitive. For example, in the case of \( N_C = 8, N_P = 8 \) and \( N_A = 5 \) the number of all possible combinations is equal to \( N_P^{N_C} + N_A + N_C = 9.2234 \cdot 10^{18} \). Thus, in order to estimate the probabilities \( P(\hat{h}_1, \hat{a}_1, \hat{v}_1, \ldots, \hat{h}_N, \hat{a}_N, \hat{v}_N | j) \) an enormous training database should be used.

To overcome this problem, the person identification task can be applied to each of the \( N \) cameras independently and the \( N \) identification results could subsequently be fused, in order to recognize the depicted person. That is, activity videos corresponding to each of the \( N \) cameras available in the test phase will be classified to one person ID class and, subsequently, a classifier fusion strategy will be used in order to combine these classification results [33]. Thus, for camera \( i \), the probability \( P(j | \hat{h}_i, \hat{a}_i, \hat{v}_i) \) of person \( j \) given the recognition results \( \hat{h}_i, \hat{a}_i, \hat{v}_i \) can be estimated by:

\[
P(j | \hat{h}_i, \hat{a}_i, \hat{v}_i) = \frac{P(\hat{h}_i, \hat{a}_i, \hat{v}_i | j) \cdot P(j)}{\sum_{l=1}^{N_P} P(\hat{h}_i, \hat{a}_i, \hat{v}_i | l) \cdot P(l)}. \tag{11}
\]

Probabilities \( P(j | \hat{h}_i, \hat{a}_i, \hat{v}_i) \) are estimated during the training phase. That is, after determining the optimal discriminant subspaces corresponding to the human ID, activity class and viewing angle discriminant vectors, the training activity vectors are mapped to these spaces and classified to the nearest class centroid. Thus, for each training activity video, a human ID, an activity class and a viewing angle label are obtained. Using these results and the human ID labels available in the training phase, probabilities \( P(\hat{h}_i, \hat{a}_i, \hat{v}_i | j) \) are estimated. \( P(j) \) is estimated by the frequency of person \( j \) appearance in the training database. Finally, probabilities \( P(j | \hat{h}_i, \hat{a}_i, \hat{v}_i) \) are estimated using (11).

The denominator \( \sum_{j=1}^{N_P} P(\hat{h}_i, \hat{a}_i, \hat{v}_i | j) \cdot P(j) = P(\hat{h}_i, \hat{a}_i, \hat{v}_i) \) in (11) refers to the probability that activity video \( i \) depicts activity \( \hat{a}_i \), is captured from viewing angle \( \hat{v}_i \) and depicts person \( \hat{h}_i \). This probability indicates the discriminatory power of each activity, when it is observed from the recognized viewing angle, to correctly identify a person. It is related with the fact that different activities and different viewing angles carry different discriminant information in terms of person identification. Thus, some activity-viewing angle combinations may be more dis-
criminant, in terms of person identification. In the case where the combination \((\hat{a}_i, \hat{v}_i)\) is capable to distinguish a person, \(P(\hat{h}_i, \hat{a}_i, \hat{v}_i)\) is expected to be equal to its true value, as estimated by the frequency of person appearance in the database. In a different case, \(P(\hat{h}_i, \hat{a}_i, \hat{v}_i)\) will have a different value. If a person is confused with others, meaning that activity videos of different persons are identified to belong to the person at hand, this value will be higher, resulting to a lower conditional probability \(P(j|\hat{h}_i, \hat{a}_i, \hat{v}_i)\) according to (11). On the other hand, in the case where a person is well distinguished from all the others, but some activity videos depicting him/her are assigned to the ID of another person, this value will be lower, resulting to a higher conditional probability \(P(j|\hat{h}_i, \hat{a}_i, \hat{v}_i)\).

Thus, in the first case there is a doubt about the person identification result, while in the second case we are almost sure that the person has been correctly identified.

The probability \(P(\hat{h}_i, \hat{a}_i, \hat{v}_i|j)\) is the joint probability of the recognized labels \(\hat{h}_i, \hat{a}_i, \hat{v}_i\) given that the \(j\)-th person is the test one. It is indicative to the similarity between different persons when they are observed by the viewing angle \(\hat{v}_i\) while performing activity \(\hat{a}_i\). In the case where an activity class-viewing angle combination characterizes a person, the corresponding probability \(P(\hat{h}_i = j, \hat{a}_i, \hat{v}_i|j)\) will have a high value. In a different case, \(P(\hat{h}_i = j, \hat{a}_i, \hat{v}_i|j)\) will be low. These probabilities will play a significant role in the final decision, as different activities captured from different viewing angles will have different influence on the final identification result.

After the estimation of the conditional probabilities \(P(j|\hat{h}_i, \hat{a}_i, \hat{v}_i), j = 1, ..., N_P, i = 1, ..., N\) corresponding to the activity videos coming from all the \(N\) cameras using (11), the classifiers outcomes should be combined in order to identify the depicted person. Several classifiers combination strategies can be utilized to this end, including Maximum A Posteriori (MAP) and Maximum Likelihood (ML) estimation [34]. In [33] it was shown that, the Sum and Product combination rules, can be derived by using the MAP estimation approach under different assumptions. Through experimentation, it was found that the Sum rule outperforms the Product one. Furthermore, by presenting an error sensitivity study, it was shown that the Sum rule is much more resilient to estimation errors. In order to chose our classifiers combination strategy, we have conducted experiments, presented in Subsection III-A, indicating that the Sum rule outperforms both Product combination rule and ML estimation, confirming the results presented in [33]. Thus, by using the conditional probabilities \(P(j|\hat{h}_i, \hat{a}_i, \hat{v}_i)\), the person is identified as the one that provides
the maximum mean probability sum, i.e., by using the Sum combination rule:

\[ j^* = \arg\max_j \frac{1}{N} \sum_{i=1}^{N} P(j|\hat{h}_i, \hat{a}_i, \hat{v}_i), \ j = 1, \ldots, N_p. \] (12)

G. Person Identification (test phase)

Let a person perform an activity captured by \( N \leq N_C \) cameras, resulting in \( N \) activity videos. In the case of continuous person identification, a sliding window consisted of \( t_{w} \) video frames is used to create the \( N \) videos that depict the activity. Person identification is performed at every window location. These videos are preprocessed as discussed in Section II-B in order to produce \( N \times t \) posture vectors \( p_{ij}, i = 1, \ldots, N, \ j = 1, \ldots, t \), where \( t = t_j \) or \( t = t_{w} \). Fuzzy distances \( d_{ijk} \) from all the posture vectors \( p_{ij} \) to every dyneme \( v_d, d = 1, \ldots, D \), are calculated and \( N \) activity vectors \( s_i, i = 1, \ldots, N \), are obtained. These activity vectors are projected to the discriminant subspaces specified in the training phase and discriminant activity vectors \( h_i, a_i \) and \( v_i \) are obtained. Each discriminant activity vector is classified to the nearest class centroid and \( N \) human ID, activity class and viewing angle labels are obtained. The person depicted in the activity videos is identified as the person that provides the highest probability sum according to the Bayesian Framework (12). Figure 6 illustrates the procedure followed in the identification phase.

III. EXPERIMENTS

In this section, we present experiments conducted in order to evaluate the proposed method. We used a multi-view activity recognition database, in order to demonstrate the effectiveness of the proposed method in dealing with all open issues discussed in Subsection II-A. Furthermore, we used a gait database, in order to compare our approach with other methods proposed in the literature for gait recognition. Finally, we created an eating and drinking activity database and conducted experiments, in order to present how the proposed method can be adapted to other applications.

A. Experiments on the i3DPost multi-view database

The i3DPost multi-view activity database [25] consists of 64 high-resolution image sequences depicting eight persons (six males and two females) performing eight activities. Eight cameras having a wide 45\% baseline to
provide 360° coverage of the capture volume were spaced in a ring having diameter 8m at a height of 2m above the studio floor. The studio was covered by a uniform blue background. The activities appearing in the database are: 'walk', 'run', 'jump in place', 'jump forward', 'bend', 'fall down', 'sit on a chair' and 'wave one hand'. The persons appearing in the database are Chris, Haidi, Han, Jean, Joe, John, Natali and Nick. Binary body image masks were obtained by discarding the blue color in the HSV color space. 8 synchronized binary images depicting the same person performing the activity 'jump forward' is illustrated in Figure 7.

The cross-validation procedure has been applied to the i3DPost eight-view database. We have used the image sequences depicting the activities 'walk', 'run', 'jump in place', 'jump forward' and 'wave one hand' in our experiments, as the remaining activities have been performed only once by each person and could not be used for training. Multi-period videos were manually split to activity videos depicting one activity instance. Four instances of each activity class were used. In each fold of the cross-validation procedure, activity videos depicting three instances of each activity class performed by all the persons in the database have been used for training. Activity
videos depicting the remaining activity instances were used for evaluation. Figure 8 illustrates the mean person identification rates obtained for different numbers of dynemes and classifiers combination strategies. It can be seen that, the Sum rule outperforms both Product rule and ML estimation. Furthermore, it can be seen that, by increasing the number of dynemes, the identification rate increases. This is reasonable, as a higher number of dynemes can describe more different human body shapes and poses. Finally, it worths mentioning the stability observed in the person identification rates when the Sum rule is applied. The optimal number of dynemes was found to be equal to 110, providing a mean identification rate equal to 94.37%. By using a higher number of dynemes, the identification rates obtained remain high, near 90%. As we are interested in the best identification performance, we report the identification results obtained by using the Sum rule and the optimal number of dynemes hereafter. The confusion matrix corresponding to the optimal parameters is illustrated in Table I.

This experiment illustrates the ability of the proposed method to correctly identify a person performing an arbitrary activity. That is, given a set of activity videos, 8 in this experiment, coming from all the cameras forming the camera
TABLE I

CONFUSION MATRIX FOR THE EIGHT PERSONS APPEARING IN THE i3DPost multi-view activity recognition database. A row represents the actual human ID and a column the person identified by the algorithm.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Recognized</th>
<th>chr</th>
<th>hai</th>
<th>han</th>
<th>jea</th>
<th>joh</th>
<th>nat</th>
<th>nik</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hai</td>
<td>0.95</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>han</td>
<td>0.1</td>
<td>0.85</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jea</td>
<td>0.05</td>
<td>0.9</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>joh</td>
<td>0.05</td>
<td></td>
<td>0.9</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nat</td>
<td>0.05</td>
<td></td>
<td></td>
<td>0.9</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nik</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

setup, it is expected that the recognized ID of the person under investigation is 94.37% likely to correspond to his/her actual ID. However, as it was previously discussed, several activities contain more discriminant information that others. This means that several activities may provide higher identification accuracy. This is investigated in Subsection III-E.

B. Person identification in different camera setups

In this subsection, a comparative study between different test camera setups is presented. The algorithm has been trained using videos from all the 8 cameras forming the database camera setup and has been tested it using fewer cameras. That is, the test camera setup was a subset of the one used to train the algorithm. Using the optimal number of dynemes (110), the cross-validation procedure has been applied using the i3DPost database for different test camera setups. In each fold of the cross-validation procedure, the activity videos depicting three instances of each activity class performed by all the persons in the database captured by all the \( N_C \) cameras forming the training camera setup have been used for training. The activity videos depicting the remaining activity instances of each activity class captured by the \( N < N_C \) cameras forming the test camera setup have been used for evaluation. Figure 9 illustrates the person identification rates achieved, by applying this procedure for four different camera setups. It should be noted that, as the direction of motion of persons differs, the cameras forming the test camera setups do not correspond to a specific viewing angle. This means that camera \#1 could correspond to any view of the depicted person. An experiment that investigates the ability of each viewing angle to correctly identify the
depicted person is presented in subsection III-E. As can be seen in Figure 9, an identification rate equal to 91.2% was achieved by using 6 cameras placed on opposite sides of the capture volume. It can be seen that, even for 3 cameras, a person identification rate equal to 83.06% was obtained.

Fig. 9. Person identification rates for four different camera setups.

C. Discriminant ability of different viewing angles

A comparative study that specifies the ability of different viewing angles to correctly identify the person depicted in a sequence is presented in this subsection. Using the optimal number of dynemes (110), the cross-validation procedure has been applied using the i3DPost database for the activity videos depicting the persons in each of the eight viewing angles. That is, eight single-view person identification procedures have been performed. In each fold of the cross-validation procedure, the activity videos depicting three instances of each activity class performed by all persons in the database captured by all $N_C$ cameras have been used for training. The activity videos depicting the remaining activity instances of each activity class corresponding to one specific viewing angle have been used for evaluation. Figure 10 illustrates the person identification rates achieved for each viewing angle. The probability to correctly identify a person performing one of the five activities captured from the frontal view is equal to 61.86%. The best viewing angle was found to be the side $45^\circ$ view, which provided an identification rate equal to 73.91%. It should be noted that the combination of the person identification results using the Bayesian framework increases the person identification rates to 94.37%, when all the eight cameras are used in the test phase.

D. Person identification in the presence of total person occlusion

In real applications, it is probable that a person performing an activity is visible from fewer ($N < N_C$) cameras. This happens when the person is out of the field of view of some cameras, or he/she is occluded. A person identification method should be able cope with such situations.
Identification rates obtained by training the algorithm using activity videos depicting the training persons from all the eight viewing angles and testing it using activity videos depicting the test person from a specific view angle.

To simulate person identification using an arbitrary number of cameras, an experiment was set as follows: the cross-validation procedure has been applied using the I3DPost database for different number of cameras in the identification phase using 110 dynemes. That is, in the training phase, the activity videos depicting the persons performing three instances of each activity class captured from all the eight cameras have been used to train the algorithm. In the evaluation phase, the number and the capturing view of the test activity videos were randomly chosen. This experiment was applied for different number of cameras in the evaluation phase. The identification rates achieved in these experiments can be seen in Figure 11. Using one arbitrary camera, a recognition rate equal to 71.68% was observed, while using six arbitrary cameras the identification rate was increased to 92.5%. Identification rates equal to 94.34% were observed for seven and eight cameras. This experiment illustrates the ability of the proposed approach to perform person identification at high accuracy in the case of using an arbitrary number of cameras that depict the person from arbitrary viewing angles.

Person identification rates versus the number of cameras used in the test phase.
E. Discriminant ability of different activities

As was previously mentioned, it is expected that the discriminant information of different activities is not the same. Depending on the performed activity, the identification accuracy may differ. This is critical in some applications. For example, in street surveillance it is expected that most people walk. In sports video analysis, it is expected that several activities besides walk, such as jump and run, are performed. Therefore, we investigated the discriminant ability of different activities in person identification using the optimal number of dynemes (110) and the cross-validation procedure applied on the i3DPost database for the activity videos belonging to different activity classes. That is, eight activity-specific person identification procedures were performed. In each fold of the cross-validation procedure, the activity videos depicting three instances of each activity class performed by all the persons in the database captured by all the 8 cameras have been used for training. The activity videos depicting the remaining activity instances of one activity class captured by all the 8 cameras have been used for evaluation. Table II presents the person identification rates achieved for each of the activity classes. As can be seen, the most discriminant activities were found to be 'jump in place' and 'wave one hand', providing person identification rates equal to 100%, while 'walk' was rated fifth, providing only 84.37% mean person identification rate.

TABLE II

<table>
<thead>
<tr>
<th>Activity</th>
<th>walk</th>
<th>run</th>
<th>jump in place</th>
<th>jump forward</th>
<th>wave one hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean identification rate</td>
<td>84.37%</td>
<td>90.62%</td>
<td>100%</td>
<td>96.87%</td>
<td>100%</td>
</tr>
</tbody>
</table>

F. Person identification at different video frame rates

To simulate the situation of person identification using cameras having different frame rates, compared to the ones used in the training phase, an experiment was set as follows. The cross-validation procedure using 110 dynemes has been applied using the I3DPost database for varying camera frame rate in the test phase. That is, in the training phase the employed activity videos contained their actual number of frames. In the test phase, the number of the activity video frames were fewer, in order to achieve identification at lower frame rate. Thus, for person identification at half frame rate, the test activity videos were formed using only the even frames, whose time index $n_t$ satisfies $n_t$
mod 2 = 0, where mod refers to the modulo operator. In the general case of training to testing frame rate ratio $R = \frac{1}{K}$, the test activity videos consisted of the frames whose time index $n_t$ satisfies $n_t \mod K = 0$. Figure 12 illustrates the results for various $K$. It can be seen that the frame rate ratio between training and test phases does not influence significantly the performance of the proposed method.

![Graph](image)

Fig. 12. Person identification rates for various video frame rate ratios between training and test phase.

G. Experiments on the MOBISERV eating and drinking recognition database

We created an eating and drinking activity recognition database depicting twelve persons (six females and six males) taking a meal. A camera was placed at a distance of 2 meters in front of them. Four meals were recorded each for a different day. The activities appearing in the database are: ‘eat’, ‘drink’ and ‘apraxia’. Each person performed multiple instances of these activities in each meal. A color-based image segmentation technique was applied to the video frames in order to create binary images depicting the head and the hands of the depicted person. Specifically, the video frames were converted to HSV color-space and pixels with values similar to the human skin were set to one and the remaining were set to zero. Two binary images depicting a person during a meal are illustrated in Figure 13.

![Images](image)

Fig. 13. Video frames depicting a person of the MOBISERV eating and drinking recognition database a) eating and b) drinking and the corresponding binary body images c) and d).
Videos depicting each meal were manually split to activity videos, depicting activity instances of 'eat' and 'drink'. We used the activity videos depicting the persons eating with fork and drinking with cup, resulting to a total number of 954 activity videos. The cross-validation procedure has been applied for different numbers of dynemes. In each fold, activity videos corresponding to three days have been used to train the algorithm. The activity videos corresponding to the fourth day have been subsequently used for testing. This procedure was applied four times, one for each test day, in order to complete an experiment. The optimal number of dynemes was found to be 230 that provided a mean person identification rate equal to 87.83%. The confusion matrix of this experiment is illustrated in Table III. In this table, a row represents the actual human ID and a column the person identified by the algorithm.

### TABLE III

**CONFUSION MATRIX FOR PERSON IDENTIFICATION IN THE EATING AND DRINKING ACTIVITY RECOGNITION DATABASE**

<table>
<thead>
<tr>
<th></th>
<th>p01</th>
<th>p02</th>
<th>p03</th>
<th>p04</th>
<th>p05</th>
<th>p06</th>
<th>p07</th>
<th>p08</th>
<th>p09</th>
<th>p10</th>
<th>p11</th>
<th>p12</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>0.87</td>
<td>0.06</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p02</td>
<td>0.17</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p03</td>
<td>0.02</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p04</td>
<td>0.82</td>
<td>0.16</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p05</td>
<td>0.83</td>
<td>0.09</td>
<td>0.08</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p06</td>
<td>0.05</td>
<td>0.90</td>
<td>0.04</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p07</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p08</td>
<td>0.04</td>
<td>0.10</td>
<td>0.72</td>
<td>0.07</td>
<td>0.02</td>
<td>0.01</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p09</td>
<td>0.07</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p10</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.82</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p11</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p12</td>
<td>0.04</td>
<td>0.01</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**H. Experiments on the CASIA gait multi-view database**

CASIA multi-view gait database (Dataset B) [7] consists of 744 image sequences depicting 124 persons (31 females and 93 males). Each person is depicted in 6 sequences at 11 different viewing angles (0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162° and 180°) with respect to the person frontal view, in an counter-clockwise manner. Thus, it contains a total of $6 \times 124 \times 11 = 8184$ gait sequences. Binary body image masks are provided in the database. 11 binary images depicting a person in the database from each viewing angle are illustrated in Figure 14.

In order to compare the proposed method with existing person identification methods proposed in the literature,
the cross-validation procedure was applied using the CASIA (dataset B) eleven-view database. Multi-period videos were used as activity videos. That is, image sequences depicting multiple walk cycles were used to train and test the algorithm. In each fold of the cross-validation procedure, activity videos depicting five walk sequences performed by all the persons in the database were used for training. Sequences depicting the remaining walking sequences were used for evaluation. An identification rate equal to 93.27% was achieved using 500 dynemes. In Table IV, we compare our method with several methods proposed in the literature for view-invariant gait recognition. As can be seen, the proposed method outperforms all of them. Furthermore, our approach can incorporate different activities in the identification procedure, while other methods will probable fail in such a case.

Fig. 14.  
(a) 11 video frames depicting a person of the CASIA (Dataset B) multi-view gait recognition database from different viewing angles and (b) the corresponding binary body images.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification rate</td>
<td>90%</td>
<td>90.72%</td>
<td>87.5%</td>
<td>85.5%</td>
</tr>
</tbody>
</table>

I. Single-view person identification in CASIA multi-view gait database

In this experiment, we compare the proposed method with several single-view gait person identification methods proposed in the literature. To do so, we performed the cross-validation procedure using the CASIA database (dataset B). We used the image sequences depicting frontal person views. Multi-period videos were used to train and test the algorithm. At each fold of the cross-validation procedure, activity videos depicting five walk sequences performed by all the persons in the database were used for training. Sequences depicting the remaining walk sequences were
used for evaluation. An identification rate equal to 97.04% was achieved using 340 dynemes. Table V illustrates comparison results with two methods proposed in the literature in a cross validation procedure. The proposed method has person identification rate that is very close to the state of the art.

TABLE V

<table>
<thead>
<tr>
<th>Person identification rates for single-view gait recognition in the CASIA gait database performing the cross-validation procedure.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method [38]</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Identification rate</td>
</tr>
</tbody>
</table>

In order to compare our method with other methods proposed for gait recognition using the CASIA database (dataset B), which do not follow the cross-validation procedure for evaluation, four sequences for each person depicting his/her frontal view were used to train the algorithm, while the remaining sequences were used for evaluation. An identification rate equal to 100% was achieved using 340 dynemes. In Table VI we compare the proposed method with state of the art methods proposed for view-dependent gait recognition.

TABLE VI

<table>
<thead>
<tr>
<th>Person identification rates for single-view gait recognition in the CASIA gait database using four sequences depicting each person for training and the remaining two sequences for evaluation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Identification rate</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

In this paper, we presented a unified framework aiming at activity-based person identification. A generic view-invariant person identification method has been proposed, by exploiting the information provided by a multi-camera setup and incorporating several activity classes in the person identification procedure. The adopted activity representation scheme exploits the global human body information, in the sense of binary body masks. Primitive human body poses, the so-called dynemes, are determined in the training phase. Test activity videos, comprised of human body poses, are represented by their similarity with the dynemes. By applying LDA to this activity video
representation, three optimal discriminant subspaces are determined, in order to use the activity video for human identity, activity class and viewing angle classification. In these spaces, classification is achieved by using a nearest class centroid procedure. By properly combining classification results coming from all the available cameras in the person identification (test) phase using a Bayesian approach, the proposed method achieves high identification rates. The effectiveness of the proposed method in challenging problem setups has been demonstrated through experimentation and extensive comparison with other approaches. Furthermore, it has been shown that the same framework can be applied in different applications, without any modification.

ACKNOWLEDGMENT

This work has been funded by 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.

REFERENCES


**Alexandros Iosifidis** received the Diploma in Electrical and Computer Engineering in 2008 and the M.S. degree in Electrical and Computer Engineering in 2010, both from the Democritus University of Thrace, Xanthi, Greece. From 2008 to 2010, he was a researcher in the Department of Production and Management Engineering, Democritus University of Thrace. He is currently pursuing the Ph.D. degree in the Artificial Intelligence and Information Analysis Lab of the Department of Informatics in the Aristotle University of Thessaloniki, Greece. He has participated in 3 research projects financed by European funds. His research interests include image and video processing, pattern recognition, computational intelligence and computer vision.

**Anastasios Tefas** received the B.Sc. in informatics in 1997 and the Ph.D. degree in informatics in 2002, both from the Aristotle University of Thessaloniki, Greece. Since 2008, he has been a Lecturer at the Department of Informatics, Aristotle University of Thessaloniki. From 2006 to 2008, he was an Assistant Professor at the Department of Information Management, Technological Institute of Kavala. From 2003 to 2004, he was a temporary lecturer in the Department of Informatics, University of Thessaloniki. From 1997 to 2002, he was a researcher and teaching assistant in the Department of Informatics, University of Thessaloniki. Dr. Tefas participated in 10 research projects financed by national and European funds. He has co-authored 25 journal papers, 80 papers in international conferences and contributed 7 chapters to edited books in his area of expertise. Over 1400 citations have been recorded to his publications and his H-index is 19 according to Google scholar. His current research interests include computational intelligence, pattern recognition, statistical machine learning, digital signal and image processing and computer vision.
Ioannis Pitas, fellow IEEE, received the Diploma and PhD degree in Electrical Engineering, both from the Aristotle University of Thessaloniki, Greece. Since 1994, he has been a Professor at the Department of Informatics of the same University. He served as a Visiting Research Associate or Visiting Assistant Professor at several Universities. His current interests are in the areas of intelligent digital media, image/video processing (2D/3D) and human-centered interfaces. He has published over 670 papers, contributed in 39 books in his areas of interest and edited or (co-)authored another 8 books. He has also been an invited speaker and/or member of the program committee of several scientific conferences and workshops. In the past he served as Associate Editor or co-Editor of eight international journals and General or Technical Chair of four international conferences (including ICIP2001). He participated in 63 R&D projects, primarily funded by the European Union and is/was principal investigator/researcher in 39 such projects. He has 12600+ citations to his work and H-index 55+. 