Multi-view Action Recognition Based on Action Volumes, Fuzzy Distances and Cluster Discriminant Analysis

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

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

Abstract

In this paper, we present a view-independent action recognition method exploiting a low computational-cost volumetric action representation. Binary images depicting the human body during action execution are accumulated in order to produce the so-called action volumes. A novel time-invariant action representation is obtained by exploiting the circular shift invariance property of the magnitudes of the Discrete Fourier Transform coefficients. The similarity of an action volume with representative action volumes is exploited in order to map it to a lower-dimensional feature space that preserves the action class properties. Discriminant learning is, subsequently, employed for further dimensionality reduction and action class discrimination. By using such an action representation, the proposed approach performs fast action recognition. By combining action recognition results coming from different view angles, high recognition rates are obtained. The proposed method is extended to interaction recognition, i.e., to human action recognition involving two persons. The proposed approach is evaluated on a publicly available action recognition database using experimental settings simulating situations that may appear in real-life applications, as well as on a new nutrition support action recognition database.

Keywords: Action Recognition; Action Volumes; Fuzzy Vector Quantization; Cluster Discriminant Analysis

1. Introduction

Human action recognition is a very active research topic in computer vision, finding applications in many important tasks, including visual surveillance, human–computer interaction and games, model-based compression, augmented reality and semantic video annotation. Human motion can be graded in several levels, depending on its complexity. In this paper, we adopt the taxonomy used in [1], where an action refers to a middle-level human motion pattern, such as walk or run. Depending on the number of cameras used to obtain the visual information, action recognition techniques can be categorized in single-view and multi-view ones [2]. Single-view methods utilize one camera, while multi-view ones utilize a multiple camera setup. Most methods proposed in the literature belong to the single-view category. Their main disadvantage is

the assumption of the same view angle during both training and recognition phases, which leads to a restrictive action recognition framework. If this assumption is not met, their performance decreases. This is the well known view angle effect [3]. Using multiple cameras, the human body is captured from multiple view angles, and, thus, view-independent action recognition can be achieved. This advantage is mitigated by the higher computational cost, as a result of the increased data need to be processed.

Action recognition is not a trivial task. Inter- and/or intra-class variations are usual in action classes. An action recognition method should be able to take into account variations in human body proportions, action speed and execution style, observed in different people, when they execute the same action. At the same time, it should be able to discriminate different action classes. Sometimes, such a discrimination is quite difficult, as execution style variations may result in one action, performed by one person, being similar to another one performed by another person. Furthermore, the human body may be (partially) occluded, which will result in poor human body visualization. The position of the person, as well as his/her orientation should not be assumed a priori known. That is, the human body representation should be view-invariant and not affected by the distance between the camera(s) and the person. Finally, the specification of the adopted camera setup should not be restricted. Sometimes, the camera(s) used in training and recognition phases may differ in resolution and frame rate. In the cases where a multi-camera setup is used, synchronization errors might occur. Furthermore, a multiple camera setup may require camera calibration, which means that the algorithms should be re-trained for different camera setups.

Actions are usually described by using either features based on motion information and optical flow [4, 5], or features devised mainly for action representation [6, 7, 8, 9]. Although the use of such features leads to satisfactory action recognition results, their computation is expensive. Thus, in order to perform action recognition at high frame rates, the use of simpler action representations is required. Neurobiological studies [10] have concluded that the human brain can perceive actions by observing only the human body poses (postures) during action execution. Thus, actions can be described as sequences of consecutive human body poses, in terms of human body silhouettes [11, 12]. Such human body representation has also been widely used in the relating task of human gait recognition [13, 14, 15, 16]. Such poses are obtained either by applying image segmentation techniques, such as background substraction [17], or by performing human body pose estimation on video frames. The result of this preprocessing is the generation of binary human body images (*human body silhouettes*), such as the one illustrated in Figure 1b. These images can either be seen as 2D human body representations or in a multi-view setting, they can be be combined in order to produce 3D human body representations. By combining human body silhouettes depicting the same body pose from different view angles, a 3D human body representation is obtained, e.g. by calculating the corresponding visual hull. This approach is known as *shape from silhouettes* [18] and requires a synchronized and calibrated camera setup. An alternative approach is the use of 3D human pose estimation techniques [19]. A 3D human body pose is illustrated in Figure 1c. By simply concatenating the human body silhouettes coming from different view angles, a *multi-view human body pose* is obtained [20], like the one illustrated in Figure 2. This is a low-computational cost human body representation, which does not require a calibrated camera setup. However,

the cameras forming the camera setup should be synchronized.

By combining 2D human body postures coming from the same view angle, a 3D action representation, in which the third dimension refers to time, can be obtained. When accumulating 2D human body poses, *Motion History Images (MHIs)* [21] and *Action Volumes (AVs)* [12] are obtained. MHIs are 2D grayscale images, in which the pixel intensity is a function of time. AVs are 3D action representations, in which consecutive human body poses are placed in consecutive volume sections. Finally, by combining 3D human body poses corresponding to subsequent time instances, a 4D action representation can be obtained. By accumulating 3D human body poses *Motion History Volumes (MHVs)* [22] are obtained, which is the 4D extension of MHIs. Figure 1d illustrates a MHI depicting a person running, while Figure 1e illustrates a MHV depicting a person bending. A major disadvantage of volumetric action representations is their dimensionality. For example, an action depicted in a 20 frame video consisting 240×320 pixels video frames is represented by a $240 \times 320 \times 20 = 1536000$-dimensional vector. Because the size all available action recognition databases is small, compared to the dimensionality of volumetric action representations, these representations become sparse, posing serious problems in action recognition accuracy.

After describing actions, most action recognition methods employ dimensionality reduction techniques in order to reduce the action representation dimensionality and enhance the action classes discrimination. Popular choices to this end include Linear Discriminant Analysis (LDA) [22, 23, 24] and Non-Negative Matrix Factorization (NMF) [25, 26]. In the cases where multiple descriptions are provided, for example in the case where actions are described by using two different action representations, two approaches have been adopted. In the first one, the available action representations are concatenated in order to form a new, enhanced action representation, while in the second one, the available action representations are employed in order to determine a new, more discriminant, feature space for data projection and classification [27, 28]. Finally, action class representation and classification of new, unknown, action instances is performed in the reduced dimensionality feature space by employing supervised classification schemes, such as Support Vector Machines (SVMs) [29, 30, 31], Artificial Neural Networks (ANNs) [32, 33], or Nearest Class Centroid based classification [11, 20, 22].

In this paper, a novel view-independent method that exploits information captured by a multi-camera setup is proposed. Binary human body silhouettes corresponding to the same view angle are accumulated, in order to produce an AV-based action representation. Each AV is mapped to a vector denoting its similarity with representative AVs determined in the training phase. This non-linear mapping results in a low-dimensional action representation, which is further mapped to a lower-dimensional feature space, in which action classes are more discriminant. We employ CDA to this end, in order to handle the multimodality usually appearing in action classes. The proposed discriminant action representation takes into account inter- and intra-class variations. In the cases of high intra-class variations, the number of clusters forming an action class can be high for better action class representation. At the same time, discriminant action vectors belonging to different action classes are quite different in the decision space. The use of a low-dimensional discriminant action representation results in fast and accurate action recognition.

The proposed method provides view-independent action recognition. It can operate using one camera, or an uncalibrated multi-camera setup. In the case of multiple cameras, action recognition results obtained from different view angles are combined to form the final action recognition decision. By following this approach, the person under consideration is not assumed to be visible from all the cameras forming the recognition camera setup. The method assumes that the person can freely move at an arbitrary distance from the camera(s) and that he/she is observed by an arbitrary number of cameras. The cameras forming the multi-camera setup may differ in resolution and synchronization errors are proven not to affect its performance.

The main novel contributions of this paper are the following ones: a) we propose a novel time-invariant volumetric action representation. We represent an action by a fixed size AV. By exploiting the circular shift invariance property of the magnitude of the Discrete Fourier Transform (DFT), we obtain a time-invariant action representation. This means that the resulting action representation is independent of the action duration and the initial human body pose. b) We propose a non-linear mapping for volumetric action representation in a reduced dimensionality space, called AV space, that preserves the action class properties. This space is determined by the training AVs. c) We propose the use of CDA for dimensionality reduction in the AV space. By exploiting the CDA properties, we obtain a discriminant low-dimensional action representation, which can handle the action class multimodality. d) We show that the proposed method can, easily, be extended in order to take into account interactions among two persons.

Compared with our previous work [34], the proposed multi-view action recognition method has the following advantages: $1)$ in the proposed method each action video is processed independently and, thus, the number of action videos used in the recognition phase may vary. Furthermore, the number of training videos are much more than those used in [34], avoiding the Small Sample Size (SSS) problem [35] that is related to statistical learning techniques, such as LDA and CDA. $2)$ The adopted camera setup needs not to be synchronized. $3)$ The proposed method is extended to interaction recognition, i.e., to human action recognition involving two persons. $4)$ An experimental study of discriminant ability of different view angles in the proposed multi-view framework is provided.

The remainder of this paper is structured as follows. Section 2 provides an overview of the recognition framework used in the proposed approach. Section 3 presents technical details that clarify the processing steps performed in the proposed method. Section 4 presents experiments conducted for assessing the performance of the proposed method. Finally, conclusions are drawn in Section 5.

2. Problem Statement

Let an arbitrary number of $N_C \geq 1$ cameras form a multi-view camera setup. Let $\mathcal{A} = \{\alpha_1, \alpha_1, ..., \alpha_{N_A}\}$ be an action class set consisting of N_A action classes, such as 'walk', 'run', 'jump', etc. Let a person, captured by $1 \leq N \leq N_C$ cameras, perform an action belonging to the action class set $\mathcal{A}$. This results to the production of N videos depicting this person performing the same action from different view angles. These videos are referred as *action videos*. A complete action period, e.g., a walking step, is referred as *elementary action*. Action videos are split to smaller videos, each

depicting one elementary action. These videos are referred as *elementary action videos*. Elementary actions differ in duration, e.g. a walk step is depicted in $9-15$ video frames in a 25 fps video depending on walk speed, whereas a bend sequence requires at least 30 video frames per bend cycle. This is also observed in elementary action videos depicting the same action performed by different persons, as well as different elementary action videos depicting the same action performed by the same person in different time periods. Thus, elementary action videos may differ in duration.

Action recognition is the recognition of an elementary action video set $\mathcal{V} = \{\mathbf{u}_1, ..., \mathbf{u}_N\}$, consisting of N elementary action videos depicting a person performing an elementary action from N different view angles, to one of the action classes $\alpha_j,\ j = 1, ..., N_A$. Needless to say that the case where $N_C = 1$ refers to single-view action recognition, whereas $N_C > 1$ refers to multi-view action recognition. It should be noted that multi-view action recognition contains also the special case, when the algorithm is trained using multiple cameras and recognition is performed using only one camera, i.e., $N_C > 1$ and $N = 1$, e.g. when during testing the person is visible only by one camera.

3. Proposed Method

3.1. Preprocessing

During the training phase, action videos coming from all N_C cameras depicting a number of consecutive action periods are manually split in elementary action videos, which are subsequently used to train and test the algorithm. In the case of continuous action recognition, i.e., in action videos depicting multiple action instances, we adopt the approach followed in [12]. That is, action videos are, automatically, temporally split in smaller videos using a sliding window of an appropriately chosen length N_{t_w}. Thus, overlapping action video segments are used to train the algorithm. In the test phase, elementary action videos are obtained by using a sliding window consisting of the same number of video frames N_{t_w} and action recognition is performed at each sliding window position.

Image segmentation techniques are applied to the elementary action video frames to create binary human body images (silhouettes). This is a low computational cost human body representation which has been widely adopted by action, as well as gait, recognition methods aiming at fast operation. In the cases where the obtained binary masks are of low quality due to noise or inaccurate body segmentation, post-processing techniques, such as morphological operations (erosion/dilation), filtering, or more advanced ones, can be applied in order to improve their quality. In cases where image segmentation techniques cannot provide binary images of adequate quality, human body posture/pose estimation techniques can be used in order to provide the human silhouettes. These images are centered at the center of mass of the person mask. Then the size of the maximal bounding box (ROI) that encloses the person body mask in the various video frames of each elementary action video is determined. Binary images of size equal to this maximal bounding box corresponding to each elementary action video are created and rescaled accordingly to $H \times W$ pixels to produce binary posture images of fixed size. Binary posture images corresponding to one elementary action

video are accumulated in order to produce an AV which represents this video. As AVs may differ in size, they are rescaled to fixed size volumes containing $H \times W \times T$ pixels. Essentially, this step is a low pass filtering procedure, which reduces the effect of the local shape errors in the human body images. In the experiments presented in this paper, we have used the values $H = W = 32$ and $T = 16$. However, these values are not restricted and one can create AVs of different size. We have experimentally found that these values provide satisfactory action classification results, requiring a relatively low computational cost. Figure 3a-h illustrate eight AVs representing elementary action videos depicting the same person performing eight actions.

AVs are vectorized in order to produce action vectors $\mathbf{p}_i \in \mathcal{R}^Q$, $Q = W \times H \times T$. This is done by representing the T binary images forming the AVs as matrices. These matrices are column-wise vectorized and concatenated to produce the action vector $\mathbf{p}_i = [\mathbf{p}_{i1}^T, \mathbf{p}_{i2}^T, ..., \mathbf{p}_{iZ}^T]^T$, where $\mathbf{p}_{ij},\ j = 1, ..., T$ denotes the vector produced by vectorizing the j-th binary image belonging to the i-th AV. After vectorizing the AVs, elementary action videos are represented by the corresponding action vectors $\mathbf{p}_i$. However, by using the above described approach, elementary videos depicting action instances of the same action class starting from different human body poses result in different action vectors representation.

To ignore any temporal information concerning the starting frame of the elementary action videos, we exploit the circular sift invariance property of the magnitudes of Discrete Fourier Transform (DFT):

$$\tilde{p}_i(k) = |\sum_{n=0}^{Q-1} p_i(n) e^{-i\frac{2\pi k}{Z}n}|, \quad k = 1, ..., Q-1. \tag{1}$$

That is, each AV is represented by a vector $\tilde{\mathbf{p}}_i$ which contains the magnitudes of the DFT of the corresponding action vector $\mathbf{p}_i$, which is referred as action vector hereafter. It should be noted here that, by selecting the dimensionality of $\mathbf{p}_i$ being power of 2, DFT-based AV representation can be efficiently calculated using FFT. Furthermore, by exploiting the symmetric property of the DFT transform for real-valued discrete-time signals, the resulted $\tilde{\mathbf{p}}_i$ can be reduced in dimensionality by exploiting only the first $D = (Q/2) + 1$ DFT coefficients obtained by (1).

The dimensionality of this AV representation is high. For example, in our experiments $D = 8193$. This will affect any procedure that will be used for action recognition using this action representation [36]. In order to avoid this, we propose to perform a non-linear mapping of the resulting action vectors to a lower-dimensional space, determined by the training AVs. This involves the determination of representative AVs and the calculation of the similarity between action vectors and these representative AVs. These procedures are described in subsections 3.2 and 3.3, respectively.

3.2. Calculation of Representative AVs

In the training phase, the AVs representing all N_T training elementary action videos are clustered to K clusters, without using the available action and view angle labels. This can be done by applying any clustering algorithm on the training AVs [37, 38, 39, 40]. Spectral Clustering [41] has proven to be a clustering algorithm which preserves the intrinsic data structure. Training action vectors $\tilde{\mathbf{p}}_i,\ i = 1, ..., N_T$ are

assumed to form the nodes of a similarity graph $(\mathcal{P}, \mathcal{E})$, where $\mathcal{P} = \{\tilde{\mathbf{p}}_1, \tilde{\mathbf{p}}_2, ..., \tilde{\mathbf{p}}_{N_T}\}$ is the set of training action vectors and $\mathcal{E}$ is a set of edges between the graph nodes. The weights of the edges are calculated using a parametric monotonically decreasing function $f_{ij}(\sigma) = f(d(\tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_j), \sigma)$, which measures the similarity between every pair of action vectors. In our experiments, we used the Gaussian similarity function, i.e., $f_{ij}(\sigma) = e^{\frac{\|\tilde{\mathbf{p}}_i - \tilde{\mathbf{p}}_j\|_2}{\sigma^2}}$, where the value of σ was empirically set to $\sigma = 0.25 \cdot \|\mathbf{p}_i - \tilde{\mathbf{p}}_j\|_2$. The weights of the graph edges constitute a symmetric, non-negative matrix $\mathbf{W}$. The degree matrix $\mathbf{D}$ is defined as the diagonal matrix having entries $\mathbf{D}_{ii} = \sum_{i=1}^{N_T} f_{ij}(\sigma)$. Using $\mathbf{W}$ and $\mathbf{D}$, the affinity matrix $\mathbf{Q}$ is calculated by $\mathbf{Q} = \mathbf{D}^{-1}\mathbf{W}$. $\mathbf{Q}$ is a $N_T \times N_T$ matrix, which contains the neighboring information of the action vectors $\tilde{\mathbf{p}}_i$ and can be interpreted as the transition table of a random walk among the set of vertices $\mathcal{P}$. Spectral clustering performs eigenanalysis on $\mathbf{Q}$ and constructs a new $N_T \times K$ matrix consisting of the K eigenvectors corresponding to the highest eigenvalues of $\mathbf{Q}$. It has been shown [41] that the rows of this matrix form a new representation of the training AVs, which is more useful for clustering. After determining this representation, a K-Means algorithm [42] is applied in order to assign each of the action vectors to one of the K clusters. Finally, the K representative AV vectors $\mathbf{v}_k,\ \ k = 1, ..., K$ are chosen to be the mean vectors of each cluster.

The optimal number K of representative AVs is determined by applying the cross-validation procedure [43]. This is a procedure used to determine the ability of a learning algorithm to generalize upon data it was not trained on. During its operation, the algorithm is trained on all but some data, which are subsequently used for testing. An experiment consists of multiple training and test cycles (folds). Depending on the data used for testing, several cross-validation procedures exist. In order to determine the ability of the proposed method to generalize upon action videos depicting different persons, we perform the leave-one-person-out cross-validation. That is, at each fold of the cross-validation procedure we used the elementary action videos depicting all but one persons in the database. The remaining elementary action videos were subsequently used for testing. This procedure was applied multiple times, equal to the number of the persons appearing in the database, in order to complete an experiment. The cross-validation procedure is performed for multiple values of K and the optimal is the one providing the best action recognition accuracy.

3.3. Action Representation

After the representative AVs calculation, the fuzzy distances between the action vectors $\tilde{\mathbf{p}}_i$ and all the representative AVs $\mathbf{v}_k$ are calculated:

$$d_{ik} = (\| \tilde{\mathbf{p}}_i - \mathbf{v}_k \|_2))^{-\frac{2}{q-1}}. \tag{2}$$

q is the fuzzification parameter ($q > 1$). Its value is also determined by applying the cross-validation procedure on the training action vectors $\tilde{p}_i$. Each action vector is mapped to a low-dimensional distance vector $\mathbf{d}_i = [d_{i1}, d_{i2}, ..., d_{iK}]^T \in \mathbb{R}^K$, typically $K \ll H \times W \times T$. The resulting vector is subsequently normalized to produce the membership vector $\mathbf{u}_i \in \mathbb{R}^K$, i.e., $\mathbf{u}_i = \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|}$.

Finally, the membership vectors $\mathbf{u}_i$ representing the training elementary action videos are normalized to have zero mean and unit standard deviation. Membership vectors of test elementary action videos are normalized accordingly.

3.4. CDA Projection

In order to discriminate action classes, the action class label information available in the training phase is exploited. The dimensionality of training membership vectors $\mathbf{u}_i \in \mathbb{R}^K$ can be reduced to $C < K$ dimensions, using a discriminant subspace method. As it was previously discussed, we expect that AVs belonging to the same action class captured by different view angles will be quite different. Thus, the adopted dimensionality reduction method should not set the assumption of unimodal action classes. Cluster Discriminant Analysis (CDA) looks for a linear transform $\mathbf{\Psi}_{opt}$, which maps the membership vectors $\mathbf{u}_i$ to a low-dimensional space, where the action classes are linearly separable. The dimensionality of the resulting space is equal to $C \leq b-1$, where b is the total number of clusters of all action classes. By applying CDA, membership vectors are mapped to the so-called discriminant action vectors $\mathbf{s}_i = \mathbf{\Psi}_{opt}^T \mathbf{u}_i \in \mathbb{R}^C$. $\mathbf{\Psi}_{opt}$ is determined as the linear transform minimizing Fisher criterion:

$$\mathbf{\Psi}_{opt} = \arg\min_{\mathbf{\Psi}} \frac{trace\{\mathbf{\Psi}^T \mathbf{S}_w \mathbf{\Psi}\}}{trace\{\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi}\}}, \tag{3}$$

where $\mathbf{S}_w$ is the within-class and $\mathbf{S}_b$ is the between-class scatter matrices. Let us denote by b_j the number of clusters inside the action class α_j, by $\boldsymbol{\mu}^{jk}$ the mean vector for the k-th cluster of action class α_j and set $a_i^{jk} = 1$, if the i-th membership vector $\mathbf{u}_i$ belongs to the k-th cluster of α_j, or zero otherwise. Then:

$$\mathbf{S}_w = \sum_{j=1}^{N_A} \sum_{k=1}^{b_j} \sum_{i=1}^{N_T} a_i^{jk} (\mathbf{u}_i - \boldsymbol{\mu}^{jk})(\mathbf{u}_i - \boldsymbol{\mu}^{jk})^T, \; \mathbf{S}_b = \sum_{j=1}^{N_A} \sum_{l \neq i} \sum_{k=1}^{b_j} \sum_{h=1}^{b_l} (\boldsymbol{\mu}^{jk} - \boldsymbol{\mu}^{lh})(\boldsymbol{\mu}^{jk} - \boldsymbol{\mu}^{lh})^T \tag{4}$$

The solution of (3) is given by the generalized eigenvalue decomposition of $\mathbf{S}_w^{-1}\mathbf{S}_b$. $\mathbf{\Psi}_{opt}$ consists of the eigenvectors which correspond to the largest eigenvalues. In our experiments, we have used the eigenvectors corresponding to the $b-1$ largest eigenvalues. However, one can use fewer eigenvectors for the discriminant action representation. After obtaining $\mathbf{\Psi}_{opt}$, each membership vector is mapped to the CDA space by applying $\mathbf{z}_i = \mathbf{\Psi}_{opt}^T \mathbf{u}_i$. In this space, action class j is represented by the b_j cluster centers $\mathbf{m}^{jk} = \mathbf{\Psi}_{opt}^T \boldsymbol{\mu}^{jk}, \; j = 1, ..., N_A, \; k = 1, ..., b_j$.

Obviously, the number of clusters b_j forming the action classes is not a-priori known and, probably, differs among action classes. In order to automatically define b_j, several criteria have been proposed in the literature. In our experiments we followed the procedure described in [44].

3.5. Action Recognition

To classify an unknown elementary action video containing $N_{t_{test}}$ binary images, each binary image is centered to the person mask center of mass and binary posture images of size equal to that of the maximum bounding box that encloses the person body mask in all video frames are created and rescaled accordingly to $H \times W$ pixels

to produce binary posture images of fixed size. These $N_{t_{test}}$ binary posture images are accumulated to produce the corresponding AV, which is rescaled to a fixed size AV consisting of $H \times W \times T$ pixels. The action vector $\tilde{\mathbf{p}}_{test}$ is produced by vectorizing the resulting AV and calculating the magnitudes of its DFT. The similarity of $\tilde{\mathbf{p}}_{test}$ with all the representative AVs is calculated and $\tilde{\mathbf{p}}_{test}$ is mapped to the corresponding distance vector $\mathbf{d}_{test}$, which is subsequently normalized to produce the membership vector $\mathbf{u}_{test}$. $\mathbf{u}_{test}$ is projected to the CDA space and the discriminant action vector $\mathbf{s}_{test}$ is obtained. In this space, action classification can be performed by applying any classification algorithm, such as SVM, ANN, Nearest Neighbor based classification, etc. However, by assuming that the two previously performed dimensionality reduction steps, i.e., FVQ and CDA projection, have successfully revealed the action classes structure, $\mathbf{s}_{test}$ is assigned to the action class label of the nearest cluster centroid using the Euclidean distance, i.e.,:

$$\alpha_{test} = \arg\min_{j} \|\mathbf{m}^{jk} - \mathbf{z}_{test}\|_2, \;\; j = 1, ..., N_A, \; k = 1, ..., b_j. \tag{5}$$

When multiple cameras are used in the action recognition procedure, multiple elementary action videos are produced, each capturing the elementary action from a different view angle. We perform action recognition to each elementary action video independently and we obtain multiple action recognition results. Finally, the recognition results obtained for each view angle are combined, in order to classify the elementary action to one of the action classes belonging to action class set $\mathcal{A}$. To this end, we perform a majority voting procedure over the recognition results provided by all the available cameras. Figure 4 illustrates the procedure followed in the recognition phase. It should be noted here that video frame resolution differences and synchronization errors between the cameras forming the camera setup do not influence the performance of the proposed method, as each elementary action video is processed independently. Furthermore, the proposed multi-view approach does not require a calibrated camera setup.

4. Experimental Results

In this section, we present experiments conducted on two action recognition databases, in order to evaluate the performance of the proposed method. We perform multi-view action recognition on the i3DPost eight-view action recognition database [45], by using different experimental settings in order to demonstrate the ability of the proposed approach to address several issues that may appear in real applications. We show that the proposed approach can easily be extended in order to recognize person interactions, i.e., actions that involve two persons. Finally, we conducted experiments on a new single-view nutrition support action recognition database, which has been created for the needs of the European R&D project MOBISERV.

4.1. Experiments on the i3DPost multi-view database

The i3DPost multi-view action recognition database contains 80 high-resolution, 1920×1080 pixel, image sequences depicting eight persons (six males and two females) performing eight actions and two person interactions. The actions appearing

in the database are: 'walk' (wk), 'run' (rn), 'jump in place' (jp), 'jump forward' (jf), 'bend' (bd), 'fall' (fl), 'sit' (st) and 'wave one hand' (wo), while the two interactions are: 'hand shaking' (hs) and 'pull' (pl). An 8-view converging camera setup which provides a 360^o coverage of the capture volume, was used to capture each action sequence. The studio background was of uniform blue color. Binary body images were extracted by thresholding the blue color in the HSV color space. Figure 5 illustrates an example of eight synchronized frames of the i3DPost database.

4.2. Cross-validation on i3DPost multi-view database

The leave-one-person-out cross-validation procedure has been performed using the elementary action videos depicting the eight persons performing the previously mentioned actions. Multiple experiments have been conducted for varying number of representative AVs. Figures 6a,b illustrate the obtained mean action recognition accuracies for different numbers of representative AVs and fuzzification parameter q values, respectively. As can be seen in Figure 6b, the performance of the proposed method is quite stable concerning the value of q. Since we are interested to the best action classification performance, we use the value $q = 1.1$ in all the experiments illustrated hereafter. In this Figure, we also present the mean action recognition accuracies obtained by applying PCA to the action vectors $\tilde{\mathbf{p}}_i$ for unsupervised dimensionality reduction and LDA or CDA followed by nearest class, or cluster, centroid for recognition. The horizontal axis, refers to the dimensionality of the action representation after applying unsupervised dimensionality reduction. In the PCA case, it refers to the dimensionality of the projected action vectors in the PCA space, while in the case of AV-based unsupervised dimensionality reduction procedure, it refers to the number of representative AVs. As can be seen, the PCA-based dimensionality reduction approach outperforms the AV-based one in the cases of low-dimensionality feature space. A mean action recognition accuracy equal to 91.67% was observed by mapping action vectors $\tilde{\mathbf{p}}_i$ to a 20-dimensional feature space by applying PCA and by performing action recognition using CDA and 2 clusters per action class. The corresponding mean action recognition accuracy for the LDA approach was equal to 89.95%, while the proposed approach achieves recognition accuracy equal to 76.21%. In the same Figure, it can be observed that the performance of AV-based unsupervised dimensionality reduction approach increases for higher AV space dimensionality. The best mean action recognition accuracy, equal to 96.34%, was observed for 120 representative AVs and one cluster per action class. This was expected, as by using a higher number of representative AVs, actions can be described in more detail. This means that differences appearing in human body proportions and action execution style are better preserved by using a high number of representative AVs. Overall, the proposed approach outperforms the competing ones (PCA+LDA, PCA+CDA).

Figure 7a illustrates the confusion matrix corresponding to the optimal parameters of the proposed method. As can be seen, actions 'walk', 'run', 'bend', 'fall' and 'wave one hand' are perfectly recognized, while the actions 'jump in place', 'jump forward' and 'sit' are confused. This is reasonable, as these three actions share a large number of human body poses. In Table 1, we compare the performance of the proposed method with other state-of-the-art methods aiming at multi-view action recognition on

the i3DPost database. As can be seen, the proposed approach achieves state-of-the-art performance in this camera setting, providing 6 to 8% higher recognition accuracy compared to other competing methods.

4.3. Recognition using different camera setups

In this experiment, we investigate the effect of using different camera setups in the training and recognition phases. We have performed the leave-one-person-out cross-validation procedure on the i3DPost database using the elementary action videos depicting the eight persons in the database for 120 representative AVs. The training camera setup consisted of all eight available cameras. In the test phase, the employed camera setup consisted of fewer cameras. The test camera setups and the mean action recognition rates obtained for each experiment are illustrated in Figure 8a. As can be seen, even with two cameras placed in arbitrary positions, mean action recognition rates greater than 82.9% were observed. By using three cameras, this accuracy increased to 90.24%, when using four cameras having viewpoint angle at multiples of 90^o a mean action recognition accuracy equal to 91.46% was obtained.

4.4. Recognition at different video frame rates

To simulate the situation of recognizing actions using cameras at different video frame rates between training and test phases, an experiment was set up as follows. The leave-one-person-out cross-validation procedure has been performed on the i3DPost database using the elementary action videos depicting the eight persons in the database for 120 AVs. During one fold of this procedure, the elementary action videos depicting seven persons were used to train the algorithm using their actual number of video frames. In the test phase, the elementary action videos depicting the eighth person consisted of fewer video frames, in order to achieve action recognition at lower frame rate. We performed this experiment for various numbers of frame rate ratio (R) between the training and testing elementary action videos. That is, the test to training video frame rate was equal to $\frac{1}{R}$. This means that the test elementary videos were consisted by the video frames satisfying $n_t \mod R = 0$, where n_t denotes the frame number and mod the modulo operator. Figure 6c illustrates the results of this experiment. As can be seen, a frame rate ratio equal to $R = 2$, i.e., when performing recognition using cameras operating at 13 fps, while performing training using cameras operating at 25 fps, causes only a moderate drop on the performance of the proposed method, as a mean action recognition accuracy equal to 90.24% was observed. However, a frame rate ratio equal to $\frac{1}{3}$, i.e., when performing recognition at 8 fps, reduces the recognition accuracy to 70.7%. This is reasonable, as when $R > 2$ the AVs representing the test elementary action videos differ significantly from the ones used to train the algorithm.

4.5. Discriminant ability of different view angles

This experiment was performed in order to evaluate the action recognition ability at different view angles. We performed the leave-one-person-out cross-validation procedure on the i3DPost database, using the elementary action videos depicting the eight persons in the database for 120 representative AVs. At every fold of this procedure, the elementary videos depicting seven persons of the database coming from all

the eight cameras were used to train the algorithm. In the test phase, the elementary videos depicting the eighth person from a specific view angle were used. That is, the experimental setup $N_C = 8$ and $N = 1$ has been used. Figure 8b illustrates the results of this experiment. As can be seen, the front and back view angles provide the poorest recognition accuracies (equal to 70.12% and 71.95%, respectively), while the side views provide better recognition performance. The optimal view angle was found to be the 350^o side-view, which provided a recognition accuracy equal to 84.75%. It should be noted though, that a simple majority voting procedure on these recognition results increases the recognition accuracy to 96.34%, as mentioned in subsection 4.2.

4.6. Recognition of person interactions

The leave-one-person-out cross-validation procedure has been performed on the i3DPost database using the elementary action videos depicting actions and interactions between two persons. A mean recognition accuracy equal to 94.44% has been observed for 160 representative AVs and one cluster per action class. The corresponding confusion matrix is illustrated in Figure 7b. As can be observed, actions are not confused with interactions. This is due to the fact that the AVs resulting from human interactions are very different from the ones resulting from actions, as the first involve two human body silhouettes. Figures 3i,j illustrates AVs describing two human interactions. Moreover, it can be observed that there is a confusion between the two interactions. However, high recognition rates were observed even for these cases. As can be seen, the proposed approach can be directly applied to recognize human actions as well as interactions.

4.7. Experiments on eating and drinking action recognition

Eating and drinking action recognition is a very important task, e.g. for monitoring the status of the elderly people in the early stages of dementia, while still living independently, to proven dehydration. In the framework of EU R&D project MOBISERV, we created an eating and drinking action 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. Two meals have been recorded, each for a different day. The actions appearing in the database are: 'eat', 'drink' and 'apraxia'. Each action class contains several subclasses. That is, action class 'eat' contains the cases where the persons eats using a cutlery, a fork, or takes a bite, while action class 'drink' contains the cases where the persons drinks using a cup, a glass, or a straw. Finally, action class 'apraxia' contains the cases where the person is slicing food or chewing it. This is why we named this action class 'apraxia', instead of 'rest'.

Each person performed multiple instances of these actions at each meal. A color-based image segmentation technique has been 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 the HSV color-space and pixels with HSV values similar to the human skin denoted the human skin regions. Example video frames and binary skin-colored regions are illustrated in Figure 9a. Videos depicting each meal were manually split to produce elementary action videos. This procedure resulted in the creation of 1288 elementary action videos. Figure 9a also illustrates sample AVs resulting from elementary action videos in the database.

The leave-one-person-out cross-validation procedure has been performed for different numbers of representative AVs. Figure 9b illustrates the mean action recognition rates obtained for these experiments. As can be seen, by using few representative AVs, the performance of the proposed method is moderate. For example by using only one representative AV, a mean action recognition accuracy equal to 41.45% was obtained. By using 10 representative AVs, the action classes were better described in the AV space resulting to an mean action recognition accuracy equal to 69.95%. The optimal number of representative AVs was found to be 200 providing a mean action recognition accuracy equal to 89.37%. The confusion matrix of this experiment is illustrated in Figure 9c. As can be seen, eating and drinking action recognition is quite difficult. This is due to the fact that the differences between the body movements involved during such action execution belonging to different classes are small. However, even for this case, high action recognition rates have been obtained.

5. Conclusion

In this paper, a novel view-independent human action recognition method that exploits a 3D action representation was presented. The human body poses, in the manner of binary human body images, are temporally concatenated in order to produce AVs which represent the action. The similarity of an AV to representative AVs, determined in the training phase, is exploited in order to map a test AV to a low-dimensional space. CDA is employed in order to further project the action representation on a discriminant subspace. By allowing multiple clusters per action class, intra-class variations are taken into account, leading to a convenient action class representation. The use of a multi-camera setup allows the action description from different view angles, leading to view-independent action representation and recognition. The use of low-computational 3D action representation combined with dimensionality reduction results in fast and accurate action recognition. Experimental results show that the proposed approach can properly address several issues that may appear in action recognition.

Acknowledgment

The research leading to these results has received funding from the Collaborative European Project MOBISERV FP7-248434 (http://www.mobiserv.eu), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.

References

[1] A. Bobick, "Movement, activity and action: the role of knowledge in the perception of motion," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 352, no. 1358, pp. 1257–1265, 1997.

[2] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

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

[4] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288–303, feb. 2010.

[5] J. Hoey and J. Little, "Representation and recognition of complex human motion," *Proceedings of IEEE Conference on Computer Vision*, vol. 1, pp. 752–759, 2000.

[6] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1310–1323, July 2011.

[7] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 867–882, May 2011.

[8] Z. Shang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 436–450, March 2012.

[9] W. Bian, D. Tao, and Y. Rui, "Cross-domain human action recognition," *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics*, vol. 42, no. 2, pp. 298–307, April 2012.

[10] M. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 179–192, Mar. 2003.

[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 and Systems for Video Technology*, vol. 18, no. 11, pp. 1511–1521, Nov. 2008.

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[13] I. Orovic, S. Stankovic, and M. Amin, "A new approach for classification of human gait based on time-frequency feature representations," *Signal Processing*, vol. 91, no. 6, pp. 1448–1456, 2011.

[14] J. Cheng, M. Qiao, W. Bian, and D. Tao, "3d human posture segmentation by spectral clustering with surface normal constraint," *Signal Processing*, vol. 91, no. 9, p. 2204.

[15] A. Roy, S. Sural, and J. Mukherjee, "Gait recognition using pose kinematics and pose energy image," *Signal Processing*, vol. 92, no. 3, pp. 780–792, 2011.

[16] D. Truong Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray, "People re-identification by spectral classification of silhouettes," *Signal Processing*, vol. 90, no. 8, pp. 2362–2374, 2010.

[17] M. Seki, T. Wada, H. Fujiwara, and K. Sumi, "Background subtraction based on the cooccurrence of image variations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2003, pp. 65–72.

[18] W. Martin and J. Aggarwal, "Volumetric descriptions of objects from multiple views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 150–158, 1983.

[19] M. Lee and I. Cohen, "A model-based approach for estimating human 3D poses in static images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 905–916, 2006.

[20] A. Iosifidis, N. Nikolaidis, and I. Pitas, "Movement recognition exploiting multi-view information," in *IEEE International Workshop on Multimedia Signal Processing*. IEEE, 2010, pp. 427–431.

[21] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analalysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[22] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2–3, pp. 249–257, Nov./Dec. 2006.

[23] Y. Yılmaz and A. Cemgil, "Algorithms for probabilistic latent tensor factorization," *Signal Processing*, vol. 92, no. 8, pp. 1853–1863, 2011.

[24] J. Driesen *et al.*, "Supervised input space scaling for non-negative matrix factorization," *Signal Processing*, vol. 92, no. 8, pp. 1864–1874, 2011.

[25] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Non-negative patch alignment framework," *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1218–1230, 2011.

[26] N. Guan, D. Tao, and B. Yuan, "Nenmf: An optimal gradient method for non-negative matrix factorization," *IEEE Transactions on Signal Processing*.

[27] B. Long, P. Yu, and Z. Zhang, "A general model for multiple view unsupervised learning," *International Conference on Data Mining - SIAM*, pp. 822–833, 2008.

[28] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 6, pp. 1438–1446, 2010.

[29] C. Figuera, J. Rojo-Álvarez, M. Wilby, I. Mora-Jiménez, and A. Caamaño, "Advanced support vector machines for 802.11 indoor location," *Signal Processing*, vol. 92, no. 9, pp. 2126–2136, 2012.

[30] J. Dai and F. Mai, "On the svmpath initialization," *Signal Processing*, vol. 92, no. 5, p. 1258.

[31] J. Chai, H. Liu, B. Chen, and Z. Bao, "Large margin nearest local mean classifier," *Signal Processing*, vol. 90, no. 1, pp. 236–248, 2010.

[32] S. Haykin, "Neural networks and learning machines," *Upper Saddle River, New Jersey*, 2008.

[33] Y. Le Cun, "Efficient learning and second order methods," in *Tutorial presented at Neural Information Processing Systems*, vol. 5, 1993.

[34] 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, no. 3, pp. 347–360, 2012.

[35] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face recognition using lda-based algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.

[36] D. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1–32, 2000.

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

[38] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[39] R. López-Sastre, D. Oñoro-Rubio, P. Gil-Jiménez, and S. Maldonado-Bascón, "Fast reciprocal nearest neighbors clustering," *Signal Processing*, vol. 92, no. 1, pp. 270–275, 2011.

[40] L. Galluccio, O. Michel, P. Comon, and A. Hero, "Graph based k-means clustering," *Signal Processing*, vol. 92, no. 9, pp. 1970–1984, 2012.

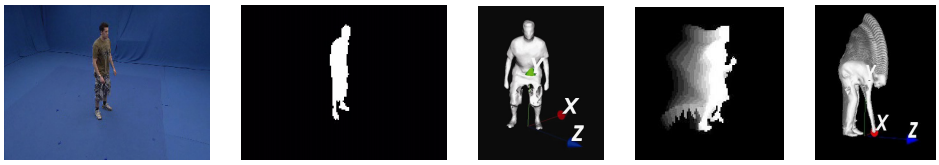
[41] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[42] A. Webb, *Statistical Pattern Recognition, 2nd ed.* Wiley, 2002.

[43] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice-Hall, 1982.

[44] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007. [Online]. Available: www.psi.toronto.edu/affinitypropagation

[45] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3d human action/interaction database," in *6th Conference on Visual Media Production*, Nov. 2009, pp. 159–168.



(a) (b) (c) (d) (e)

Figure 1: *a) A video frame depicting a person, b) a binary human body image, c) a 3D human body pose (visual hull), d) a MHI depicting action 'run' and e) a MHV depicting action 'bend'.*

Figure 2: *A multi-view human body pose.*

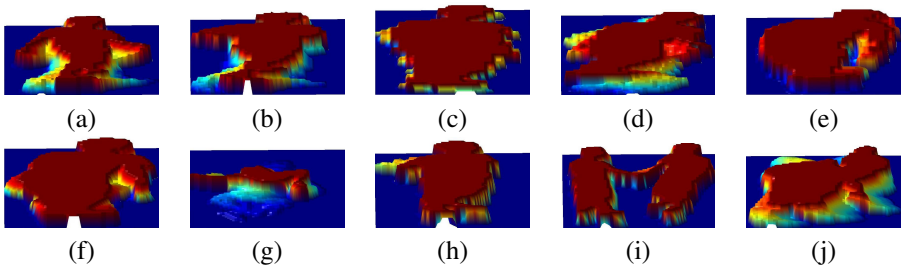


Figure 3: *AVs depicting actions: a) 'walk', b) 'run', c) 'jump in place', d) 'jump forward', e) 'bend', f) 'sit', g) 'fall' and h) 'wave one hand', and interactions i) 'hand shaking' and j) 'pull'.*

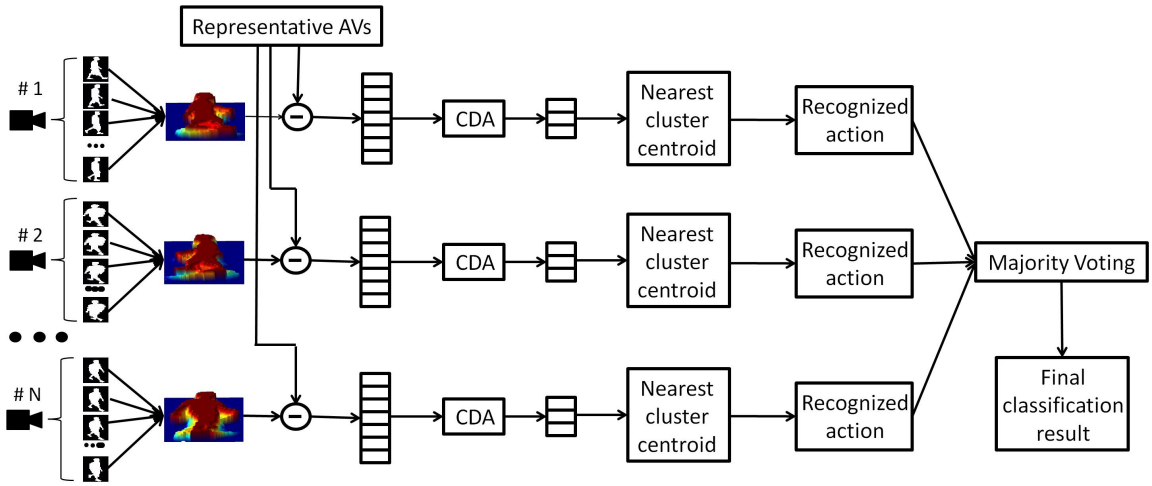


Figure 4: *Action recognition procedure.*

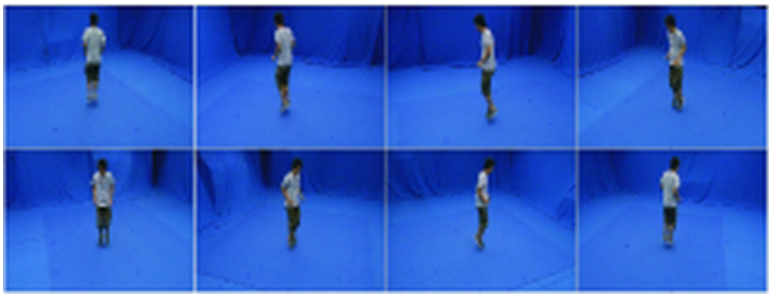


Figure 5: *Eight synchronized video frames of the i3DPost database.*

Table 1: Recognition accuracy on the i3DPost database.

Method	5 actions	8 actions	6 actions
Method [48]	90%	-	-
Method [47]	-	90.88%	-
Method [29]	-	94.37%	-
Method [18]	-	-	89.58%
Proposed method	**97.8%**	**96.34%**	**98.16%**

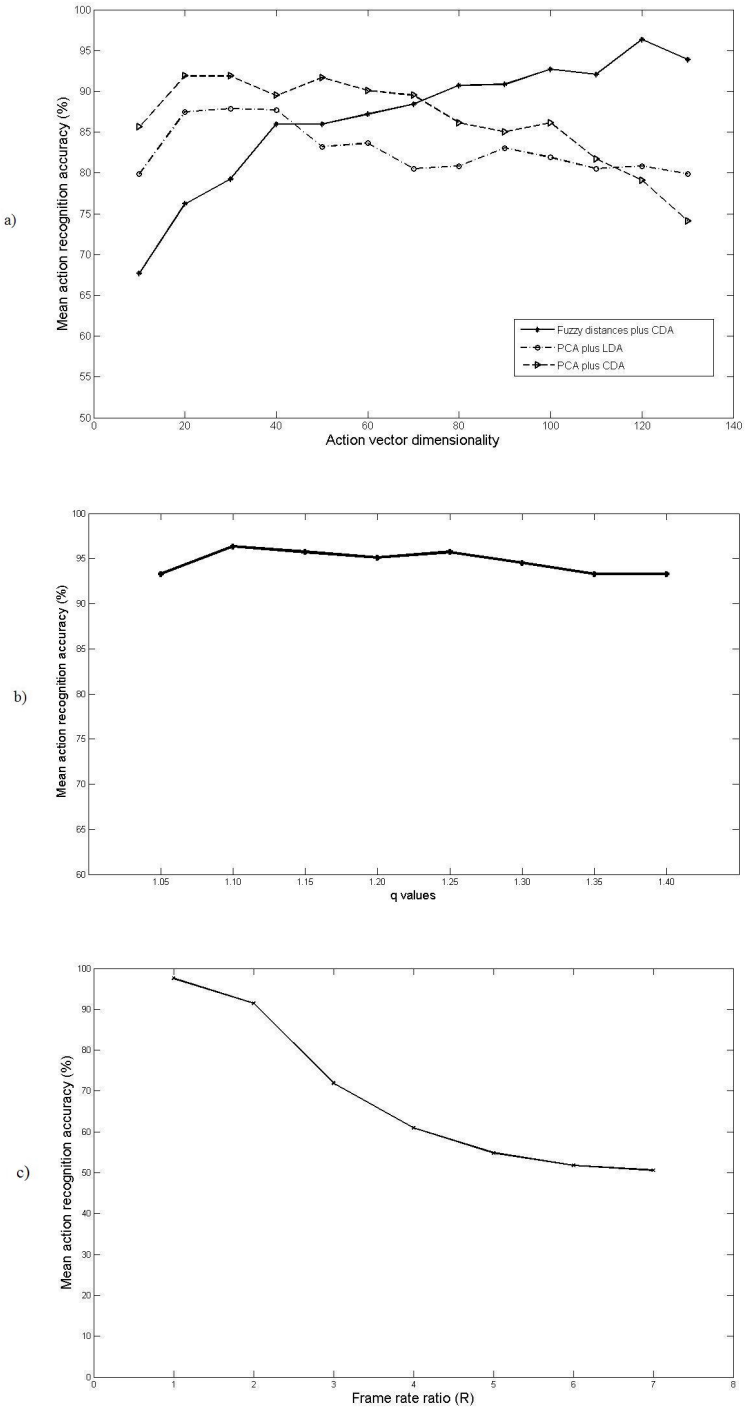


Figure 6: *a) Mean recognition accuracy on i3DPost database vs number of representative AVs, b) Mean recognition accuracy vs the fuzzification parameter q value and c) action recognition using different frame rate ratio (R) between training and test phases.*

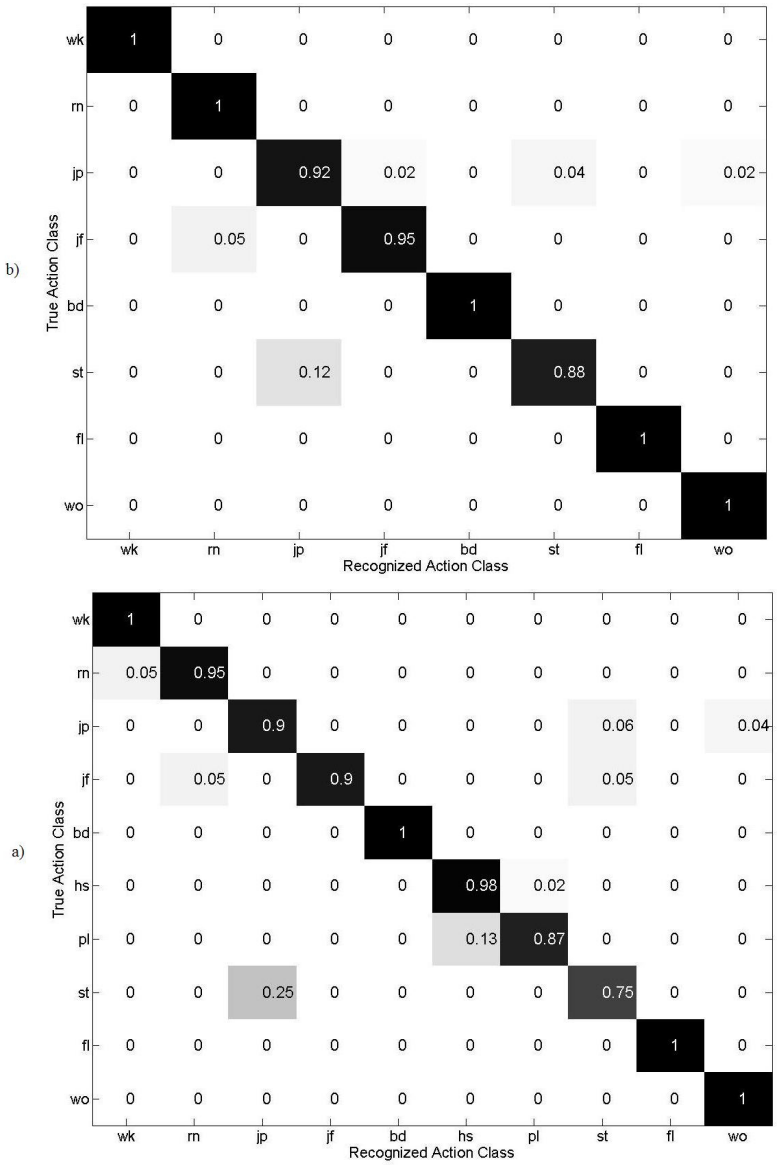


Figure 7: *a) Confusion matrices containing recognition accuracy for eight actions and b) for eight actions and two interactions on the i3DPost database.*

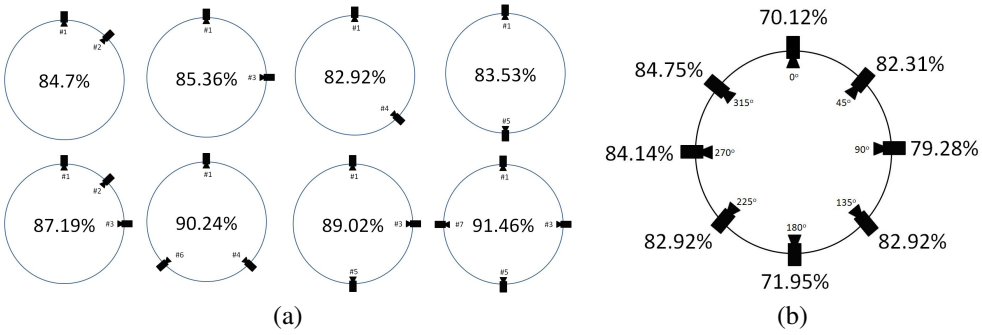


Figure 8: *a) Action recognition results on i3DPost database using different test camera setups and b) discriminant ability of different view angles.*

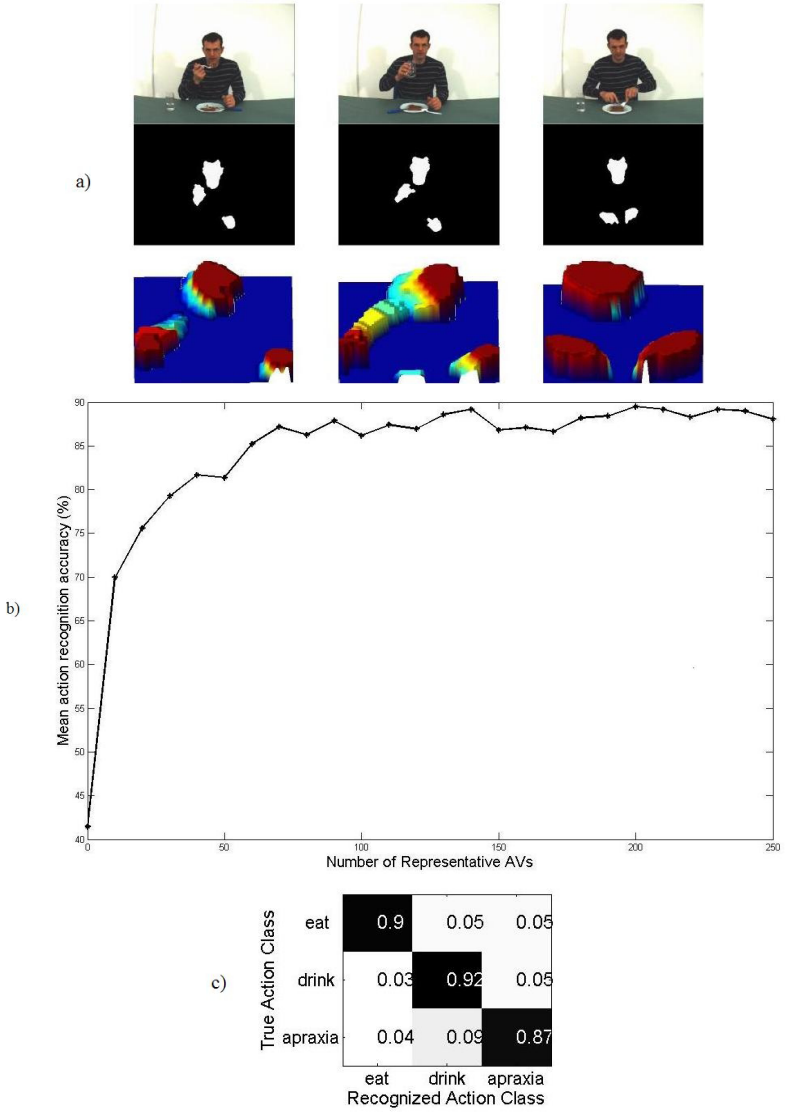


Figure 9: *a) Example video frames depicting a person having a meal accompanied with the corresponding binary body images and the produced AVs, b) recognition accuracy on eating and drinking action recognition vs the number of representative AVs and c) confusion matrix containing recognition rates for eating and drinking action recognition.*