Learning sparse representations for view-independent human action recognition based on fuzzy distances

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, Greece Tel,Fax: +30-2310996304

{*aiosif,tefas,pitas*}*@aiia.csd.auth.gr*

Abstract

In this paper, a method aiming at view-independent human action recognition is presented. Actions are described as series of successive human body poses. Action videos representation is based on fuzzy vector quantization, while action classification is performed by a novel classification algorithm, the so-called Sparsity-based Learning Machine (SbLM), involving two optimization steps. The first one determines a non-linear data mapping to a high-dimensional feature space determined by an l1-minimization process exploiting an overcomplete dictionary formed by the training samples. The second one, involves a training process in order to determine the optimal separating hyperplanes in the resulted high-dimensional feature space. The performance of the proposed human action recognition method is evaluated on two publicly available action recognition databases aiming at different application scenarios.

Keywords: Activity recognition; Fuzzy Vector Quantization; Sparse Data Representation; Action classification

1. Introduction

Human action recognition is one of the most active research fields in computer vision with an increasing number of real-world applications, including intelligent visual surveillance, human-computer interaction, content-based video compression and retrieval, semantic video annotation and augmented reality. Due to its importance, it has been heavily researched in the last two decades and a plurality of action recognition methods have been proposed in the literature, each taking into account several aspects of the action recognition problem, depending on the application scenario.

Most of the proposed methods utilize one camera, in order to obtain the necessary visual information. However, due to the fact that the human body during action execution, when it is observed by arbitrary viewing angles, can be considered to as a high level deformable subject, such methods usually set the assumption of fixed and a priori known observation angle. This observation angle is defined to be the one that provides the maximal action distinction [1]. In order to overcome this limitation, multi-view action recognition methods have been proposed. These methods operate by using multi-camera systems. By capturing the human body from multiple viewing angles, multi-view methods, exploit the enriched visual information in order to create a view-independent human body description leading to view-independent action representation. After obtaining the available visual information, a pre-processing step is, usually, performed on the video data depicting action instances in order to obtain a convenient human body representation. Such descriptions include features based on motion information and optical flow [2] and features devised mainly for action representation [3, 4]. However, neurobiological studies [5] have concluded that humans can perceive actions by observing only the human body configurations during action execution. Thus, actions can be described as sequences of successive human body poses [6]. Popular choices for multi-view human body representation include visual hulls of the 3D human body [7, 8], multi-view postures [9] and skeletal and super-quadratic body models [10]. Such human body representations, including the one adopted by the proposed method, set the assumption that the entire human body is visible from each camera and require video frame segmentation in order to determine binary human body silhouettes on each video frame. This is why they have been, mainly, adopted for action recognition in applications involving a fairly controlled environment. Finally, pattern recognition techniques are applied in order to obtain a convenient action class representation and to classify new, unknown, action instances. Discriminant Analysis techniques [11], Support Vector Machines (SVMs) [12] and Artificial Neural Networks (ANNs) [13], have been widely used to this end providing satisfactory action classification rates. Among them, Single-hidden Layer Feedforward Neural networks (SLFNs) have been widely used due to their ability to approximate any target continuous function and classify any disjoint regions. Recently, the concept of sparsity-based data representation has attracted attention in computer vision due to its powerful discrimination ability and the ability of sparse representations to uncover semantic information.

Although most multi-view methods proposed in the literature can effectively perform view-independent human action recognition exploiting multi-view visual information, they inherently set several assumptions. That is, the person under consideration should be visible from all the cameras forming the recognition camera setup and the recognition camera setup should be synchronized and/or calibrated. In different cases, the action recognition performance will decrease, or they will fail to operate properly. In order to address this issue, we follow a different approach. We choose to perform single-view view-independent action recognition on each video, coming from different cameras, depicting the person under consideration performing an action. Action classification results are, subsequently, fused in order to recognize the performed action. By adopting this approach, the proposed method can effectively address several issues that may appear in real application scenarios involving multi-camera setups [14], like total occlusion of the human body in some of the cameras, synchronization errors between the cameras and different properties of the camera setups used in training and recognition phases. This is due to the fact that each video coming from a different camera is processed independently and, thus, delays on the video frames corresponding to different cameras do not affect the adopted action representation. Furthermore, by including only the cameras capturing the performed action on the final fusion process, the cases of total human body occlusion in some of the cameras and different camera setups in training and test phases are easily addressed. However, multi-view action recognition is challenging in this setting due to the well known viewing angle effect [1, 15]. In order to successfully operate an action recognition method should be able to first identify the view angle that the person is captured from each camera and, subsequently, perform view-dependent action recognition. Thus, view angle determination should proceed action recognition. In this paper we follow a different approach by adopting a sparsity-based video representation, which is shown to be discriminative and able to uncover semantic information concerning both action recognition and view angle determination. Thus, view angle determination is not required as a pre-processing step, since the obtained video representation encodes information relating to both the action class and view angle. Based on this fact and inspired by the efficient Extreme Learning Machine (ELM) algorithm for SLFN network raining [16], we propose a new learning algorithm, the so-called Sparsity-based Learning Machine (SbLM). The motivation behind SbLM is the fact that, since ELM can achieve high classification performance by adopting a random data mapping in a high dimensional feature space, the adoption of a meaningful discriminant high dimensional data representation should lead to increased action classification performance. Since sparsity based data representation has proven its discriminative power in classification problems, such an action representation would be an excellent choice. Furthermore, as it is shown, the adoption of sparsity-based data representation is able to uncover semantic information regarding both action class and view angle, which is of interest for view-independent action recognition exploiting multi-view information.

The main contributions of this paper are: 1) the proposal of a new learning algorithm exploiting sparsity-based data representation using an overcomplete basis formed by the training data, 2) the proposal of two dictionary variants for the SbLM algorithm formed by the mean cumulative fuzzy distances of posture vectors to all the dynemes and 3) the evaluation of SbLM in single-view and multi-view action recognition on two publicly available databases.

The rest of this paper is structured as follows. Section 2 presents a literature review of the action recognition methods adopting sparsity-based action representation. Section 3 provides an overview of the recognition framework used in the proposed approach and a small discussion concerning the action recognition task. Section 4 describes the proposed action recognition method and experimental results evaluating its performance are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. Related Work

Human action recognition methods employing sparsity-based action representation for classification have been, recently, proposed in the literature, inspired by the effectiveness of such representations in image classification. Researchers have, mainly, focused their attention on finding a convenient action description, while classification is, usually, performed by following the smallest residual error classification rule [17]. In [18], actions are described as sequences of consecutive human body poses, in terms of binary images denoting the human body regions. Each pose is represented by a feature vector and feature vectors representing all the poses forming the training action videos are stored in a matrix to produce an overcomplete dictionary. Human body poses forming a test action video are represented by sparse linear combinations of the training poses and classified to the action class providing the smallest residual error. Finally, the test action video is classified by following the majority rule. In [19, 20] actions are described by histograms of interest points detected in action videos. Histograms representing training action videos form a dictionary and the histogram representing a test action video is classified to the action class that provides the smallest residual error. In [21], action videos are described by bags of features such as optical flow, velocity, gradient and divergence. Covariance matrices of these features representing all the training action videos are vectorized in order to produce a dictionary. Finally, test action videos are represented by the corresponding covariance matrices and classified to the action class that provides the smallest residual error.

Instead of employing the entire training set for dictionary creation, several methods have been proposed in order to learn a representative dictionary. Such methods employ the training samples in order to learn an overcomplete dictionary minimizing an objective function denoting the appearance information between dictionary items. In [22], actions are described as vectors obtained by calculating central moments of image patches corresponding to interest points neighborhood. Random projections are, subsequently, used for dimensionality reduction. The reduced dimensionality feature vectors representing training action videos are employed for dictionary learning, while test action videos are classified to the action class that provides the smallest residual error employing the learned dictionary. Three dictionary types are evaluated: class independent, class-specific and a shared dictionary, which is produced by concatenating the class independent and class-specific dictionaries. In [23], the class label information associated with dictionary items is exploited in order to learn a compact and discriminant dictionary for human action representation.

All the above described methods aim at single-view action recognition and, thus, they set the assumption of known viewing angle during training and recognition phases. A method that does not set this assumption is proposed in [24]. A multi-camera setup is employed in order to describe actions as sequences of 3D human body poses. View-independent features obtained by applying the Discrete Fourier Transform on 3D Motion History Volumes [8], resulted by accumulating the 3D human body poses with respect to time. Sparsity-based action video classification is performed by following the smallest residual error classification rule.

3. Problem Statement

Let $\mathcal{A}$ be a set of N_A action classes, such as walk, run, bend, etc. Let $\mathcal{U}$ be a video database containing videos depicting N_P persons performing instances of actions belonging to the action class set $\mathcal{A}$. Such videos will be called action videos hereafter. In the case where the database camera setup is formed by $N_C > 1$ cameras, each action instance in the video database $\mathcal{U}$ is captured from N_C viewing angles. The number of video frames consisting action videos may vary, since actions differ in duration. For example, a walking step is, usually, depicted in 10 - 15 video frames of a 25 fps video, while a bend sequence is depicted in 40 - 60 video frames. Such duration variations may be observed even in two different realizations of the same action, due to execution style variations between different persons, or mood variations of the same person. Let us, now, assume that a person, who may or may not appear in the video database $\mathcal{U}$, performs an instance of an action included in the action class set $\mathcal{A}$, and that he/she is captured by $N \geq 1$ cameras. This results to the creation of N action videos, each depicting the same action instance from a different viewing angle. Action recognition is the task of assigning the new, unknown, action instance to one of the N_A action classes.

Action recognition is a difficult task in this setup. Since we aim to perform view-independent human action recognition, the person is free to move and thus, his/her position and orientation with respect to the camera setup coordinate system may vary. Furthermore, actions highly overlap in the video frame space, since the same human body poses appear in multiple actions. Taking into consideration the body size variations between different persons and the action execution style variations that appear in different realizations of actions, it is possible that an action instance belonging to an action class performed by a person will be more similar to an action instance belonging to a different action class performed by another person, than to an action instance belonging to its actual action class. Finally, in the recognition phase, the human body may be visible from an arbitrary number of cameras, since the person under consideration may perform the action outside some cameras' field of view, or he/she may be occluded in some of the cameras.

4. Proposed Method

In this Section, each step of the proposed action recognition method is described in detail. The adopted action representation is presented in Subsection 4.1. The proposed SbLM classification algorithm is described in Subsection 4.2. Finally, the procedure followed in the recognition phase is described in Subsection 4.3.

4.1. Action Representation

Let us assume that the video database $\mathcal{U}$ contains N_T action instances, depicted in $N_V = N_T \cdot N_C$ action videos. Image segmentation techniques [25] are applied to the video frames of these action videos in order to create binary images depicting the image locations belonging to the human body in white and the background in black. These binary images are centered to the human body Regions Of Interest (ROIs) center of mass, cropped to the ROIs region and resized to fixed size $H \times W$ posture images. Posture images are represented as matrices, which are, subsequently, vectorized column-wise in order to create the so-called posture vectors $\mathbf{p}_{ij} \in \mathbb{R}^{H \cdot W}$, where $i = 1, ..., N_V$ denotes the action video index and $j = 1, ..., N_i$ runs along the video frames of action video i. An example of posture vector creation is illustrated in Figure 2.

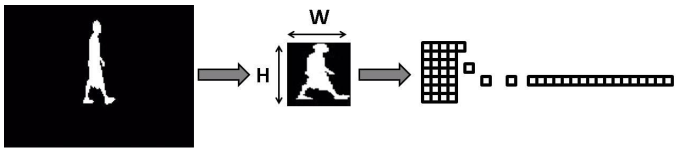


Figure 1: *Posture vectors creation.*

In the training phase, all the posture vectors of the training action videos are used in order to produce D representative posture vectors $\mathbf{z}_d \in \mathbb{R}^{H \cdot W}$, $d = 1, ..., D$, the so-called dynemes [26]. Dynemes calculation is performed by clustering the training posture vectors $\mathbf{p}_{ij}$ without exploiting the known action labels of the training action videos. The choice of clustering algorithm was proven to have only minor impact in action recognition accuracy. We found experimentally that K-Means [11] performs well for the specific task and is faster than other clustering algorithms. Thus, we adopt K-Means for dynemes calculation, minimizing the within-cluster scatter, i.e.:

$$\sum_{d=1}^{D} \sum_{i=1}^{N_V} \sum_{j=1}^{N_i} \alpha_{ijd} \|\mathbf{p}_{ij} - \mathbf{z}_d\|^2, \tag{1}$$

where α_{ijd} is an index denoting if posture vector $\mathbf{p}_{ij}$ belongs to cluster d. That is, $\alpha_{ijd} = 1$ if $\mathbf{p}_{ij}$ belongs to cluster d and $\alpha_{ijd} = 0$ otherwise. Dynemes are defined to be the cluster mean vectors:

$$\mathbf{z}_d = \frac{1}{n_d} \sum_{i=1}^{N_V} \sum_{j=1}^{N_i} \alpha_{ijd} \mathbf{p}_{ij}. \tag{2}$$

$n_d = \sum_{i=1}^{N_V} \sum_{j=1}^{N_i} \alpha_{ijd}$ is the number of posture vectors belonging to cluster d. Figure 2 illustrates 60 dynemes calculated by using the action videos of the i3DPost action recognition database, that has been used in the first set of our experiments.

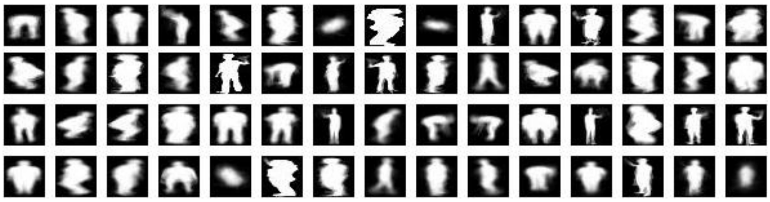


Figure 2: 60 *dynemes produced by using action videos depicting eight persons performing eight actions from eight viewing angles.*

As can be seen, dynemes correspond to representative human body poses captured by arbitrary viewing angles.

After dynemes calculation, each posture vector is mapped to the so-called membership vector $\mathbf{u}_{ij} \in \mathbb{R}^D$ encoding the similarity of $\mathbf{p}_{ij}$ with all the dynemes $\mathbf{v}_d$. This is done by calculating the fuzzy distances between $\mathbf{p}_{ij}$ and all the dynemes $\mathbf{z}_d$, according to a fuzzification parameter $m > 1.0$:

$$u_{ijd} = \frac{(\| \mathbf{p}_{ij} - \mathbf{z}_d \|_2))^{-\frac{2}{m-1}}}{\sum_{k=1}^{D}(\| \mathbf{p}_{ij} - \mathbf{z}_k \|_2))^{-\frac{2}{m-1}}},\ d = 1, ..., D. \tag{3}$$

Membership vectors corresponding to each action video are averaged in order to produce the so-called action vector $\mathbf{s}_i \in \mathbb{R}^D$, which is used in order to represent action videos:

$$\mathbf{s}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{u}_{ij}. \tag{4}$$

Finally, the action vectors representing all the N_V training action videos are normalized to have unit l_2 norm, zero mean and unit standard deviation. In the recognition phase, all the action vectors $\mathbf{s}_{test,j},\ j = 1, ..., N$ representing the N test action videos depicting an action instance from all the N cameras involved in the recognition process are normalized accordingly. We will employ the training action vectors $\mathbf{s}_i,\ i = 1, \ldots, N_V$ in order to train the proposed SbLM algorithm, as we will describe in the following Section.

4.2. Sparsity-based Learning Machine

In this Section we briefly describe the ELM algorithm [16] for SLFNs training and we subsequently describe the proposed SbLM algorithm. Let $\mathbf{s}_i$, $i = 1, ..., N_V$ be the set of training action vectors, accompanied with the corresponding action class labels $c_i \in \mathcal{A}$. In ELM, the network's input weights $\mathbf{W}_{in}$ are randomly chosen, while the output weights $\mathbf{W}_{out}$ are analytically calculated. The network's target vector corresponding to action vector $\mathbf{s}_i$, $\mathbf{t}_i = [t_{i1}, ..., t_{iN_A}]^T$, is set to $t_{ik} = 1$ when action vector $\mathbf{s}_i$ belongs to class k and $t_{ik} = -1$ otherwise.

Let us assume that the network's hidden layer consists of L neurons and that $\mathbf{b} \in \mathbb{R}^L$ is a vector containing the hidden layer neurons bias values, which are randomly chosen as well. Let $\mathbf{v}_j$ and $\mathbf{w}_k$ denote the j-th column of input weights matrix $\mathbf{W}_{in} \in \mathbb{R}^{D \times L}$ and the k-th column of output weights matrix $\mathbf{W}_{out} \in \mathbb{R}^{L \times N_A}$, respectively. The output vector $\mathbf{o}_i$ of the ELM network corresponding to training action vector $\mathbf{s}_i$ is calculated by $o_{ik} = \mathbf{w}_k^T \mathbf{g}_i,\ k = 1, ..., N_A$, where $\mathbf{g}_i$ is the $\mathbf{s}_i$ representation in the ELM space. By storing the action vectors representation in the ELM space in a matrix $G = [\mathbf{g}_1, ..., \mathbf{g}_{N_V}]$, the network's outputs corresponding to the training action vectors can be written in a matrix form as $\mathbf{O} = \mathbf{W}_{out}^T \mathbf{G}$. Finally, by assuming that the network's predicted outputs $\mathbf{O}$ are equal to the network's targets, i.e., $\mathbf{o}_i = \mathbf{t}_i,\ i = 1, \ldots, N_v$, and using linear activation function for the output neurons, $\mathbf{W}_{out}$ can be analytically calculated, i.e., $\mathbf{W}_{out} = \mathbf{G}^\dagger \mathbf{T}^T$, where $\mathbf{G}^\dagger = (\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{G}$ and $\mathbf{T}$ a matrix containing the network's target vectors.

In standard ELM algorithm the action vectors representation in the ELM space is obtained by performing a non-linear mapping using randomly chosen input weights $\mathbf{W}_{in}$. Furthermore, the dimensionality of the ELM space L must be empirically chosen. In order to find the optimal ELM space dimensionality several methods have been proposed [27, 28]. Such methods either start by using a large number of hidden neurons and iteratively decrease it as long as the classification residual error remains above a pre-defined threshold, or start by using a small number of hidden neurons and iteratively increase it. These methods depend on user pre-specified parameter values, like the maximal number of hidden neurons and residual error threshold. Furthermore, the determined optimal number of hidden neurons depends on the training data at hand. For example, if some of the training data are replaced by others, the optimal number of hidden layer neurons should be determined again.

In the following we describe the proposed SbLM algorithm which has been formulated inspired by the above described learning process. Intuitively, the use of an action vector representation in a high dimensional feature space that highlights action class discrimination should lead to an increase in the action classification performance. We investigate the discriminative ability of sparsity based data representations to this end. Specifically, the proposed SbLM algorithm maps action vectors $\mathbf{s}_i$ to a high dimensional feature space determined by the reconstruction weights determined by an l1-minimization optimization problem. Subsequently, a learning process is performed in order to determine the optimal separating hyperplanes in the resulted high dimensional space for action classes discrimination.

Let us denote with $\mathbf{A} \in \mathbb{R}^{D \times L}$ a dictionary that will be used in order to reconstruct a training action vector $\mathbf{s}_i$ using weights appearing in a vector $\mathbf{g}_i \in \mathbb{R}^L$. $\mathbf{g}_i$ can be determined by solving the following l1-minimization problem:

$$\arg\min_{\mathbf{g}_i} \|\mathbf{A}\mathbf{g}_i - \mathbf{s}_i\|_2^2 + \lambda\|\mathbf{g}_i\|_1, \tag{5}$$

where λ is a parameter that penalizes the non-zero elements of $\mathbf{g}_i$. Such minimization problems can be efficiently solved by using existing software packages [29, 30]. Following [18, 19, 20, 21], $\mathbf{A}$ can be an overcomplete dictionary formed by all the training data, in our case the training action vectors, i.e., $\mathbf{A} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{N_V}]$. One may think that the obvious solution in this problem is $\mathbf{G} = \mathbf{I}_{N_V}$, since each training action vector $\mathbf{s}_i$ can be perfectly reconstructed by a single column of $\mathbf{A}$ and specifically by the i-th column. In order to avoid this situation and since it is expected that test action vectors will not be perfectly reconstructed by using a single training action vector, we employ the following two alternatives for $\mathbf{A}$ determination:

- **Action video independent:** In this case we assume that the training action vector $\mathbf{s}_i$ can be reconstructed by using the action vectors representing all the remaining action videos in the video database $\mathcal{U}$. That is, $\mathbf{A}$ is formed by the training action vectors, $\mathbf{s}_j,\ j = 1, ..., N_V,\ j \neq i$. However, since test action vectors will be reconstructed by using all the training action vectors, $\mathbf{s}_i$ is reconstructed by using the following dictionary:

$$\mathbf{A}^i = [\mathbf{s}_1, ..., \mathbf{s}_{i-1}, \mathbf{0}_1, \mathbf{s}_{i+1}, ..., \mathbf{s}_{N_V}], \tag{6}$$

  where $\mathbf{0}_1 \in \mathbb{R}^D$ is a vector of zeros. By solving the minimization problem in (5) using $\mathbf{A}^i$ and $\mathbf{s}_i$, a reconstruction vector $\mathbf{g}_i^i$ is obtained. Finally, $\mathbf{g}_i$ is calculated by:

$$\mathbf{g}_i = (1 - \mu)\mathbf{g}_i^i + \mu\mathbf{1}^i, \tag{7}$$

  where $\mathbf{1}^i \in \mathbb{R}^L$ is a vector having all its elements equal to zero, except of element i, i.e., $\mathbf{1}^i = [0, ..., 0, 1, 0, ..., 0]^T$.

- **Person independent:** In this case we assume that the training action vectors representing action videos depicting each person in the video database $\mathcal{U}$, can be reconstructed by using the action vectors of the remaining persons. That is, we assume that training action vectors $\mathbf{s}_i,\ i = 1, ..., N_V$ are accompanied by the corresponding action class and person ID labels, c_i and

q_i, respectively. The action vectors representing the action videos depicting the k-th person are stored in a matrix $\mathbf{A}_k = \{\mathbf{s}_j,\ q_j = k\}$ and each training action vector $\mathbf{s}_i,\ i = 1, ..., N_V$ is reconstructed by using the following dictionary:

$$\mathbf{A}^i = [\mathbf{A}_1, ..., \mathbf{A}_{q_i-1}, \mathbf{0}_{N_{q_i}}, \mathbf{A}_{q_i+1}, ..., \mathbf{A}_{N_P}], \tag{8}$$

where $\mathbf{0}_{N_{q_i}} \in \mathbb{R}^{D\times N_{q_i}}$ is a matrix of zeros and N_{q_i} is the number of action vectors belonging to person q_i.

By following the above described approach, the dimensionality of the new action vector representation is determined by the number of training action vectors, i.e., $L = N_V$. In the cases where the training set size is very large, dictionary learning techniques, like those proposed in [22, 23], can be employed for representative dictionary learning. However, the adoption of dictionary learning techniques has the following disadvantages: 1) it depends on user pre-specified parameter values, 2) the determined optimal dictionary size depends on the training data at hand and 3) the dictionary learning process is time consuming. Taking into account that the dictionary learning process should be performed multiple times, for different parameter values, in a try-and-error sense, it can be seen that the adoption of such dictionary learning techniques is computationally expensive. In the cases where the number of training samples is not very high (like in our case where the number of action videos appearing in current action databases is between one and two thousand) the use of the training data for dictionary creation is computationally inexpensive and exploits all the available information.

The proposed action vector representation exploits the discriminant ability of sparsity-based data representation. In Figure 3 we present the representations of an action video depicting the first person in the i3DPost multi-view action database walking, for the ELM and the proposed SbLM approaches. As can be seen in Figure 3a, the ELM based action vector representation does not seem to contain discriminant information, since it is a result of non-linear mapping using randomly chosen weights. In both SbLM approaches, as illustrated in Figures 3b,c, the obtained action vector representations contain discriminant information, since the highlighted coefficients correspond to training action belonging to the correct action class in most cases. Specifically, it can be seen that the action vector under investigation is more similar to action vectors belonging to the first, fourth and seventh persons in the database representing action videos depicting them walking. In addition, let us consider that each training action vector corresponds to an action video that has been recorded by a specific viewing angle.

Thus, the weights determined for a test action vector contain information regarding the test action video observation angle [1]. In this way, we can see that the viewing angle effect is inherently addressed by the adopted sparsity based action vector representation.

After calculating $\mathbf{g}_i,\ i = 1, ..., N_V$, the available action labels for the training action videos can be utilized in order to train a classification algorithm for action class representation and classification of new, unknown, action videos. Expecting that training and test action vectors follow the same distributions, the assumption of zero training error is a reasonable choice. However, in the cases where the training data set contains outliers, this assumption may reduce the classifiers generalization ability. By assuming small training errors, we can formulate a regularization based optimization problem having the following form:

$$\textbf{Minimize: } L_P = \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{c}{2}\sum_{i=1}^{N_V}\|\boldsymbol{\xi}_i\|_2^2$$

$$\textbf{Subject to: } \mathbf{W}^T\mathbf{g}_i = \mathbf{t}_i - \boldsymbol{\xi}_i,\ \ i = 1, ..., N_V,$$

where $\mathbf{W} \in \mathbb{R}^{L\times N_A}$ is a matrix formed by vectors denoting the optimal hyperplanes separating action classes in the above described multi-class action classification problem in an One-Versus-All (OVA) manner, $\boldsymbol{\xi}_i \in \mathbb{R}^{N_A}$ is the training error vector corresponding to training action vector $\mathbf{s}_i$ and c is a parameter denoting the importance of the training error in the optimization problem. $\mathbf{W}$ can be calculated by:

$$\mathbf{W} = \left(\mathbf{G}\mathbf{G}^T + \frac{1}{c}\mathbf{I}\right)^{-1}\mathbf{G}\mathbf{T}^T. \tag{9}$$

where $\mathbf{I} \in \mathbb{R}^{N_V\times N_V}$ is the identity matrix. The derivation of (10) by using the primal minimization problem L_P is shown in the Appendix.

Here we should note that similar classification schemes have been employed in ELM as well [31, 32]. These methods investigate multi-class classification problems as well and the proposed formulas for output weights $\mathbf{W}_{out}$ calculation look similar to (10). However, in both these methods, data representation in the decision space is based on randomly chosen weights, contrary to the proposed SbLM, employing sparsity based data representation.

4.3. Action Recognition (Test Phase)

Let us assume that a person, who is free to move in a place that is monitored by a camera setup consisting of $N_C \geq 1$ cameras, performs an instance of an

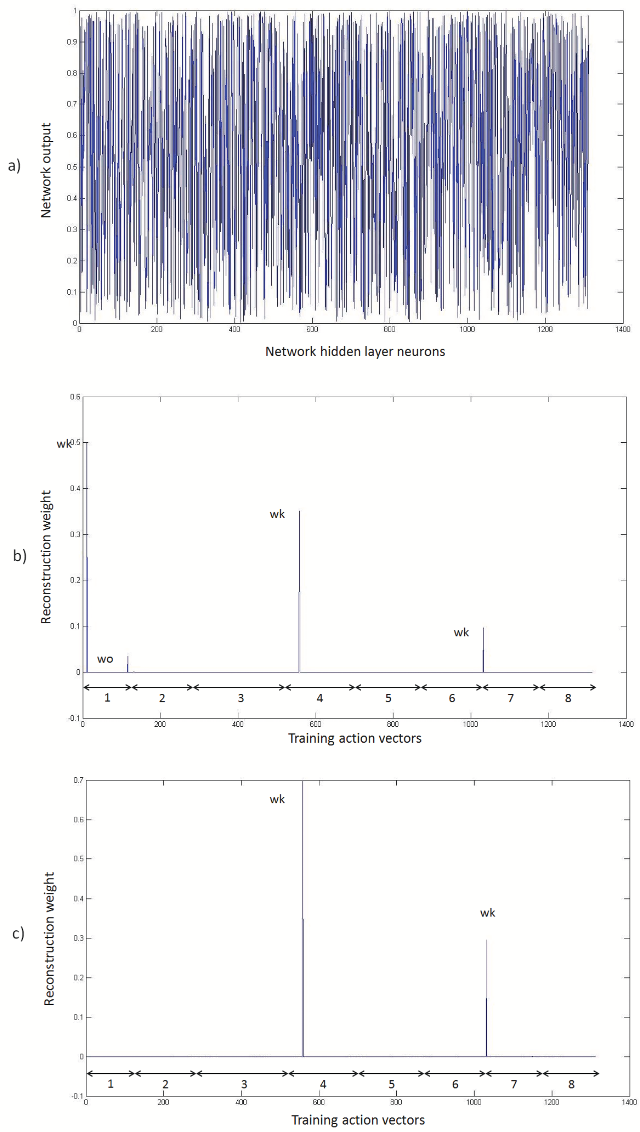


Figure 3: *Action vector representation of an action video depicting a person walking for a) standard ELM algorithm, b) action video independent SbLM and c) person independent SbLM. In b) and c) we illustrate the person ID and action class labels of the corresponding training action vectors (wk = walk, wo = wave).*

action appearing in the action class set $\mathcal{A}$. Visual human body tracking techniques [33] can be used in order to determine the cameras in which the person is visible from. Let us assume that he/she is visible from $1 \leq N \leq N_C$ cameras. This results to the creation of N action videos, each depicting the same action instance from a different viewing angle. These action videos are preprocessed following the procedure described in Section 4.1 resulting to N test action vectors $\mathbf{s}_{t,i},\ i = 1, ..., N$. $\mathbf{s}_{t,i}$ are represented as sparse combinations of the training action vectors $\mathbf{s}_i$ and N vectors containing the reconstruction weights $\mathbf{g}_{t,i}$ are obtained. $\mathbf{g}_{t,i}$ are, subsequently, classified to the action class corresponding to the highest output, i.e.:

$$c_{t,i} = arg\max_k \mathbf{w}_k^T \mathbf{g}_{t,i},\ k = 1, ..., N_A. \tag{10}$$

Finally, the test action instance is classified to an action class by following the majority voting rule, i.e.:

$$c_t = arg\max_j \sum_{i=1}^{N} \beta_{ij}, \tag{11}$$

where is an index denoting if action vector $\mathbf{s}_{ti}$ has been classified to action class j, i.e., $\beta_{ij} = 1$ if $c_{t,i} = j$ and $\beta_{ij} = 0$ otherwise. As can be seen, by adopting such an action recognition scheme, the number of cameras used in the recognition phase is not assumed to be a priori known. As will be shown in the experimental evaluation presented in the next Section, the proposed method can, successfully, operate in the cases where the number of cameras N involved in the recognition process is not equal to the number of cameras N_C forming the recognition camera setup. This may result in the cases where the person under consideration is not visible from some of the cameras due to total human body occlusion, or due to the fact that he/she moves outside of some cameras field of view. Furthermore, synchronization errors, i.e., delays on the obtained video frames, between the cameras forming the recognition camera setup do not affect the action classification performance. This is due to the fact that each action video coming from a different camera is processed independently, contrary to multi-view methods combining visual information coming from multiple views in order to obtain a view-independent human body representation and subsequently perform classification. In both cases, i.e., total human body occlusion in some of the cameras and synchronization errors between cameras, such methods will, probably, fail to operate well, since the obtained human body representation will be incorrect.

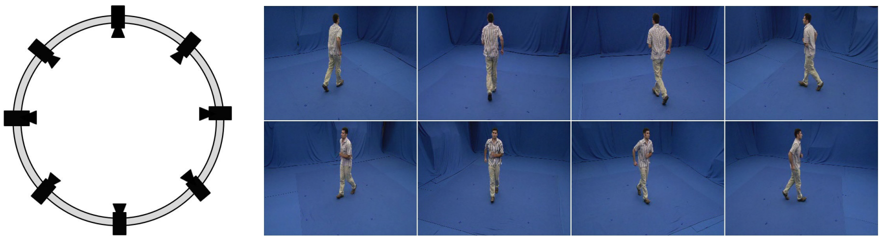
5. Experimental Results

In this Section we present experiments conducted in order to evaluate the performance of the proposed action recognition method. We perform multi-view action recognition on the i3DPost action database aiming at recognition of daily actions and single-view action recognition (as a special case of multi-view action recognition for $N_C = 1$) on the AIIA-MOBISERV eating and drinking database containing actions appearing in meal intakes. In order to evaluate the discriminant information captured by the proposed sparsity based action vector representation, we compare the performance of the proposed SbLM algorithm with that of the ELM variants proposed in [16, 31, 32] in all the presented experiments, since these methods employ a similar optimization-based classification scheme. Since most action recognition methods employing sparse representation perform action classification based on the smallest residual error [17], as discussed in section 2, we compare the smallest residual error classification rule with the proposed SbLM in all the presented experiments. We, also, compare the performance of the proposed SbLM with that of nearest class centroid classification using the Euclidean distance, which has been used as a baseline. Furthermore, we compare the performance of the proposed action classification scheme with that of the View Transformation Model (VTM) [34] based on Singular Value Decomposition (SVD), and that of sparse coding-based action vectors creation combined with the proposed SbLM algorithm. Finally, we compare the performance of the proposed action recognition method with that of other methods, recently proposed in the literature.

Regarding the algorithms' parameters, the following values have been used: $W = H = 32$ and $\mu = 0.5$. Following [32], the number of neurons forming the network's hidden layer for the ELM algorithms proposed in [16, 31, 32] was chosen to be equal to $L = 1000$. The optimal number of dynemes, the fuzzification parameter m and the optimal value for the regularization parameter c have been determined by performing the leave-one-person-out cross-validation (LOPOCV) procedure and following a grid search strategy. That is, the algorithms have been trained multiple times (folds), equal to the number of the persons appearing in the database, by using the action videos depicting all but one persons and tested on the action videos depicting the remaining one. The LOPOCV procedure has been performed for different numbers of dynemes $D = 10d,\ d = 1, ..., 20$, fuzzification parameter values $m = 1.0 + 0.05n,\ n = 1, ..., 10$ and values of regularization parameter $\lambda = 10^r,\ r = -4, ..., 4$. The mean action classification rate over all the folds has been used to measure the performance of each algorithm in an experiment.

5.1. *Experiments on the i3DPost database*

The i3DPost multi-view database [35] contains high resolution (1080×1920 pixels) image sequences depicting eight persons performing eight actions. The database camera setup consists of eight cameras placed in the perimeter of a ring at a height of 2 meters above the studio floor (Figure 4a). The actions appearing in the database are: 'walk' (wk), 'run' (rn), 'jump in place' (jp), 'jump forward' (jf), 'bend' (bd), 'fall down' (fl), 'sit on a chair' (st) and 'wave one hand' (wo). Example images depicting a person in the database walking when viewed from different viewing angles are illustrated in Figure 4b.



(a) (b)

Figure 4: *a) The i3DPost action recognition database camera setup and b) images depicting a person of the database from different viewing angles running.*

In our first set of experiments we have performed the LOPOCV procedure by using all the action videos of the database. That is, in each fold of an experiment, the algorithms have been trained by using the action videos depicting seven of the persons performing actions from all the available eight viewing angles and tested by using all the action videos depicting the remaining person. The obtained mean action classification rates for different regularization parameter r values are illustrated in Figure 5. As can be seen in this Figure, high action classification rates have been obtained. The best action classification rates obtained for the ELM algorithms proposed in [31] and [32] are equal to 95.08% and 95.5%, respectively. SbLM algorithm outperforms both [31] and [32] providing best action rates equal to 96.12% and 97.69% for the person independent, noted as SbLM PI, and action video independent, noted as SbLM VI, cases respectively. Furthermore, it can be seen that the SbLM VI algorithm is the overall winner in a wide range of regularization parameter r values, i.e., for r taking values between -3 and 2. The confusion matrices corresponding to the optimal parameter values for the SbLM algorithms are illustrated in Figure 6. It is worth noting here that by performing the LOPOCV procedure using the standard ELM algorithm [16] an action classification rate equal to 68.57% was obtained. This can be explained by the fact that the standard ELM algorithm assumes zero training error and, thus, variations appearing in action videos depicting different persons affect its generalization performance. Action vectors classification based on the smallest residual error resulted to an action classification rate equal to 94.15%. The VTM-based and the sparsity codding-based classification schemes resulted to action classification rates equal to 95.77% and 95.5%, respectively. Finally, action vectors classification based on the smallest Euclidean distance from the mean action class vectors resulted to action classification rates equal to 12.5%. The small action classification rate in the case of action vectors classification in the input space was expected due to the viewing-angle effect [1]. Finally, in Table 1 we compare the performance of the proposed action recognition method with that of other methods employing silhouette-based action representations and evaluating their performance on the i3DPost database.

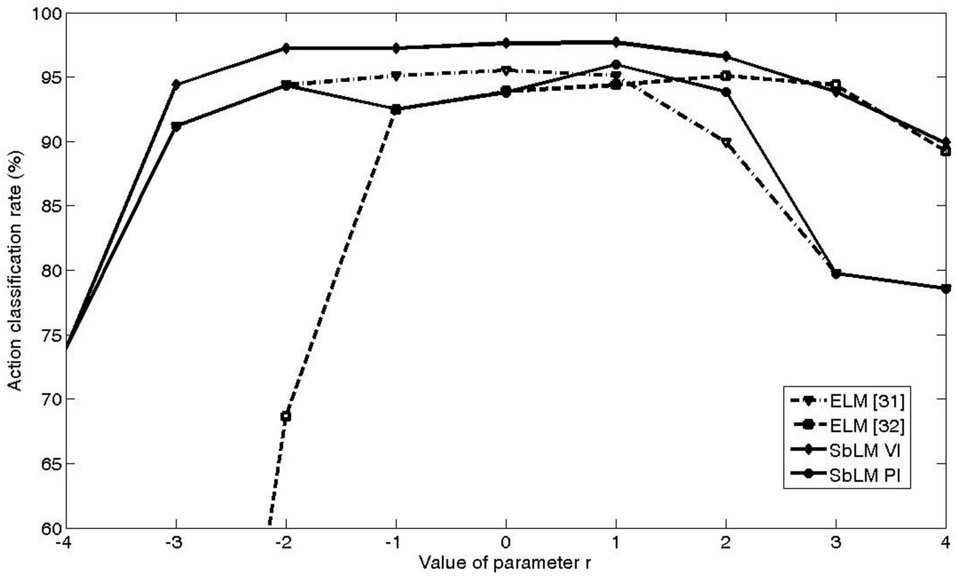
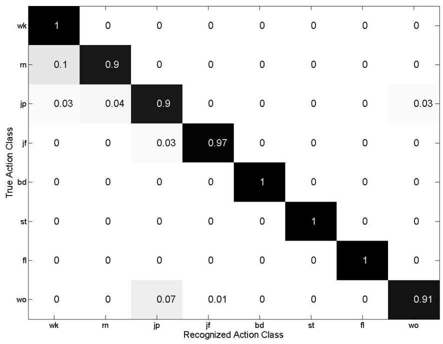
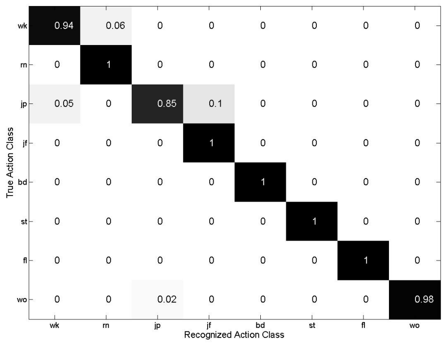


Figure 5: *Action classification rates on the i3DPost database by using all the available cameras ($N_C = 8$, $N = 8$).*

In order to simulate the situation of performing human action recognition under total human body occlusion in some of the cameras used in the recognition phase an experiment has been set as follows. We have performed the LOPOCV procedure multiple times by using different numbers of recognition cameras N.



(a)



(b)

Figure 6: *Confusion matrices on the i3DPost database: a) person independent SbLM and b) action video independent SbLM.*

Table 1: Comparison results in the i3DPost multi-view action recognition database.

Method [36]	Method [9]	Method [37]	Method [38]	**SbLM PI**	**SbLM VI**
$90.\%$	94.34%	94.87%	96.34%	**96.12%**	**97.69%**

That is, in each fold of the cross validation procedure we have trained the algorithms by using all the action videos depicting seven persons in the database and tested them by using action videos of the remaining person coming from N randomly chosen cameras. This means that each test action instance was classified by using N action videos depicting the person under consideration from N randomly chosen viewing angles. The action classification rates obtained for these experiments are illustrated in Figure 7.

As can be seen in Figure 7, the action video independent SbLM algorithm clearly outperforms all the competing algorithms. By using only $N = 3$ randomly chosen cameras in the recognition phase an action classification rate equal to 96.73% has been obtained, while by using $N = 6$ cameras the obtained action classification rate is equal to 97.69%. We should note here that the decrease of the action classification rates in the case of $N = 2$ is observed due to the fact that the majority voting combination scheme can not provide a classification result in the cases where the two action video classification results are not the same. In these cases, the corresponding test action instance is characterized as "unknown", and, thus, it is regarded as incorrect. In table 2 we compare the performance of the proposed method with that of the method proposed in [37] for different numbers

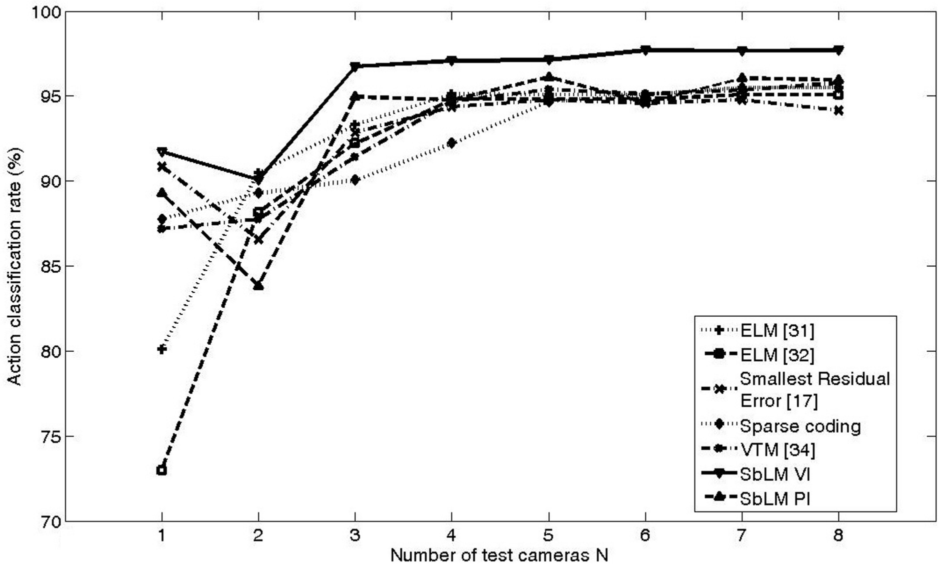


Figure 7: *Action classification rates on the i3DPost database by using different numbers of randomly chosen cameras N ($N_C = 8$).*

of test cameras N. As can be seen, the proposed method clearly outperforms the method in [37] in all cases.

Table 2: Comparison results in the i3DPost multi-view action recognition database for different N ($N_C = 8$).

Number of cameras N	1	3	4	5
Method [37]	79%	84.9%	90%	94.85%
SbLM PI	**89.28%**	**94.95%**	**94.78%**	**96.1%**
SbLM VI	**91.72%**	**96.73%**	**97.06%**	**97.12%**

Finally, in order to simulate the situation of multi-view human action recognition based on view-independent single-view training, an experiment has been set as follows. We have performed the LOPOCV procedure multiple times by training the algorithms with one, randomly chosen camera, for each training action instance and using different numbers of recognition cameras N. That is, in each fold of the cross validation procedure we have trained the algorithms by using one camera depicting the seven persons in the database and tested them by using action videos of the remaining person coming from N randomly chosen cameras. The action classification rates obtained for these experiments are illustrated in Figure 8.

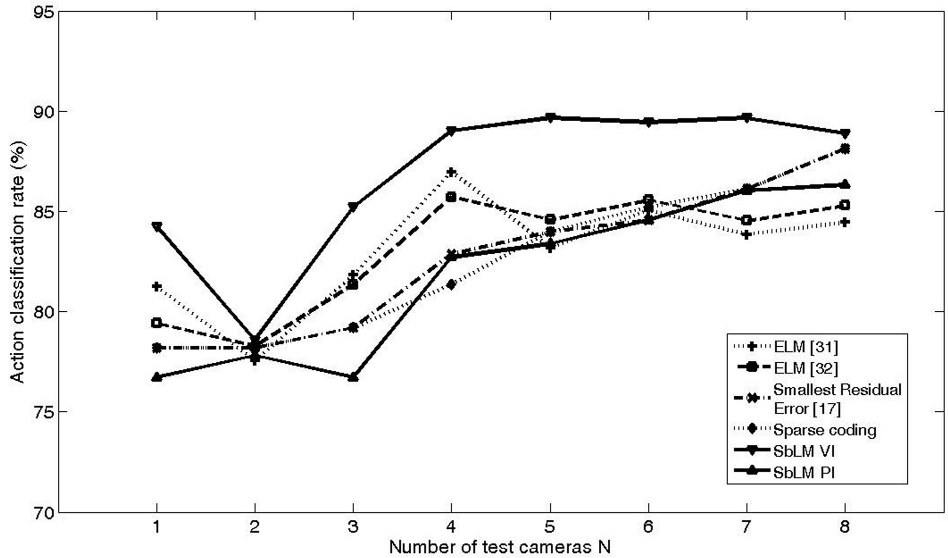


Figure 8: *Action classification rates on the i3DPost database by using different numbers of randomly chosen cameras N ($N_C = 1$).*

As can be seen in Figure 8, the SbLM VI algorithm clearly outperforms all the competing algorithms. By using only $N = 4$ randomly chosen cameras in the recognition phase an action classification rate equal to 89.01% has been obtained, while by using $N = 7$ cameras the obtained action classification rate is equal to 89.67%. Overall, it can be seen that the proposed action classification scheme can be applied in multi-view action recognition based on either multi-view or view-independent single-view training procedures.

5.2. Experiments on the AIIA-MOBISERV database

The AIIA-MOBISERV eating and drinking database [39, 6] contains low resolution (480×640 pixels) videos depicting twelve persons (six females and six males) during a meal. Four meals have been recorded for each person. The persons eat using spoon, cutlery and fork and drink from a cup or a glass. Between eating and drinking periods, the persons perform several actions, such as slicing their food, chewing it and resting. In our experiments we have formulated a three class classification problem containing three action classes: ’eat’, ’drink’ and ’apraxia’. Action class ’apraxia’ contains action videos depicting the persons slicing their food, chewing it and resting. Example video frames depicting an instance of each action class are illustrated in Figure 9.

(a) (b) (c)

Figure 9: *Video frames of the AIIA-MOBISERV eating and drinking database.*

The choice of three, instead of eight, classes was made in order to evaluate the performance of the proposed method in the case of high intra-class variations. Indeed, as it is expected, intra-class variations in the above described three-class classification problem are high, since due to different human body proportions among individuals and action execution style variations it is possible that an action video depicting a person eating with spoon is more similar to an action video depicting another person drinking from a cup and at the same time is quite different from an action video depicting a sequence belonging to an other eating subclass, e.g. eat with fork. In the presented experiments the image ROIs denoting human body poses were determined to be the persons' heads and hands, i.e., the skin-like video frame regions.

We have performed the LOPOCV procedure by using the action videos of the AIIA-MOBISERV database. The mean action classification rates obtained for different regularization parameter r values are illustrated in Figure 10. As can be seen high action classification rates have been obtained in these experiments. The use of the two ELM algorithms proposed in [31] and [32] resulted to best action classification rates equal to 90.2% and 90%, respectively. Standard ELM algorithm [16] provided an action classification rate equal to 68.65%. Action vectors classification based on the smallest residual error rule and the sparsity codding-based classification scheme resulted to action classification rates equal to 90.3% and 90%, respectively. Action vectors classification based on the smallest Euclidean distance from the mean action class vectors resulted to an action classification rate equal to 86.06%. Finally, the best action classification rates obtained by applying the two variants of the proposed SbLM algorithms are equal to 91.64% and 92.36%, for the person independent and action video independent cases, respectively. As can be seen in Figure 10, the two SbLM algorithms clearly outperform the other three ELM variants, since in most cases they provide higher action classification rates. The confusion matrices obtained by using the SbLM algorithms and the optimal regularization parameter r values are illustrated in Figure 11.

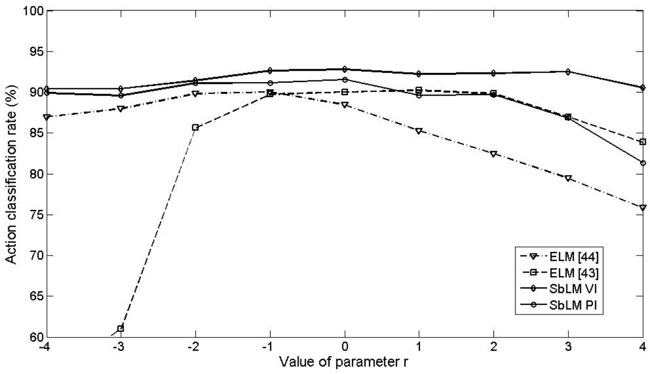


Figure 10: *Action classification rates on the AIIA-MOBISERV database.*

(a) (b)

Figure 11: *Confusion matrices on the AIIA-MOBISERV database: a) person independent SbLM and b) action video independent SbLM.*

6. Discussion and Conclusion

In this paper we presented method aiming at view-independent human action recognition. Actions are described as sequences of successive human body poses. Action representation involves fuzzy vector quantization based on representative human body poses determined from the human body poses of the training action videos. Inspired by the effectiveness of the sparsity-based data representation for classification and the efficient Extreme Learning Machine algorithm, the Sparsity-based Learning Machine (SbLM) has been proposed for view-independent action video classification. In the case of multi-view human action recognition, action classification results combination leads to high action classification rates. The proposed method can successfully handle the situations of total human body occlusion in some of the cameras forming the recognition camera setup, as well as synchronization errors between the cameras. Experimental results have shown that the adoption of a high dimensional discriminant action representation can increase the action classification performance in the proposed setting.

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.

Appendix A.

Here we describe the solution of the following optimization problem:

$$\textbf{Minimize: } L_P = \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{c}{2}\sum_{i=1}^{N_V}\|\boldsymbol{\xi}_i\|_2^2$$

$$\textbf{Subject to: } \mathbf{W}^T\mathbf{g}_i = \mathbf{t}_i - \boldsymbol{\xi}_i,\;\; i = 1, ..., N_V,$$

By substituting the condition in the primal problem we obtain:

$$L_P = \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{c}{2}\sum_{i=1}^{N_V}\|\mathbf{t}_i - \mathbf{W}^T\mathbf{g}_i\|_2^2. \tag{A.1}$$

Solving for $\frac{\vartheta L_P}{\vartheta \mathbf{W}} = 0$, $\mathbf{W}$ can be obtained by:

$$\mathbf{W} = \frac{c}{2}\left(2\mathbf{G}\mathbf{T}^T - 2\mathbf{G}\mathbf{G}^T\mathbf{W}\right) \Rightarrow \left(\mathbf{I} + c\mathbf{G}\mathbf{G}^T\right) = c\mathbf{G}\mathbf{T}^T \Rightarrow \left(\mathbf{G}\mathbf{G}^T + \frac{1}{c}\mathbf{I}\right)\mathbf{W} = \mathbf{G}\mathbf{T}^T$$

$$\mathbf{W} = \left(\mathbf{G}\mathbf{G}^T + \frac{1}{c}\mathbf{I}\right)^{-1}\mathbf{G}\mathbf{T}^T. \tag{A.2}$$

References

[1] S. Yu, D. Tan, T. Tan, Modelling the effect of view angle variation on appearance-based gait recognition, Asian Conference on Computer Vision (2006) 807–816.

[2] S. Ali, M. Shah, Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2) (2010) 288–303.

[3] R. Minhas, A. Baradarani, S. Seifzadeh, Q. Wu, Human action recognition using extreme learning machine based on visual vocabularies, Neurocomputing 73 (2010) 1906–1917.

[4] L. Shao, R. Gao, Y. Liu, H. Zhang, Transform Based Spatio-Temporal Descriptors for Human Action Recognition, Neurocomputing 74 (2011) 962–973.

[5] M. Giese, T. Poggio, Neural mechanisms for the recognition of biological movements, Nature Reviews Neuroscience 4 (3) (2003) 179–192.

[6] A. Iosifidis, A. Tefas, I. Pitas, Activity based Person Identification using Fuzzy Representation and Discriminant Learning, IEEE Transactions on Information Forensics and Security 7 (2) (2012) 530–542.

[7] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2006) 249–257.

[8] D. Weinland, R. Ronfard, E. Boyer, Automatic discovery of action taxonomies from multiple views, IEEE Conference on Conference Computer Vision and Pattern Recognition 2 (2006) 1639–1645.

[9] A. Iosifidis, A. Tefas, N. Nikolaidis, I. Pitas, Multi-view human movement recognition based on Fuzzy distances and Linear Discriminant Analysis, Computer Vision and Image Understanding 116 (2012) 347–360.

[10] C. Tran, M. Trivedi, Human body modelling and tracking using volumetric representation: Selected recent studies and possibilities for extensions, ACM/IEEE International Conference on Distributed Smart Cameras (2008) 1–9.

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

[12] C. Burges, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery 2 (2) (1998) 121–167.

[13] S. Haykin, Neural networks and learning machines, Upper Saddle River, New Jersey .

[14] F. Qureshi, D. Terzopoulos, Surveillance camera scheduling: A virtual vision approach, ACM International Workshop on Video Surveillance and Sensor Networks 12 (2005) 131–140.

[15] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Multi-view gait recognition based on motion regression using multiplayer perceptron, IEEE International Conference on Pattern Recognition (2010) 2186–2189.

[16] G. Huang, Q. Zhu, C. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks, IEEE International Joint Conference on Neural Networks 2 (2004) 985–990.

[17] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.

[18] C. Liu, Y. Yang, Y. Chen, Human action recognition using sparse representation, International Conference on Intelligent Computing and Intelligent Systems 4 (2009) 184–188.

[19] A. Liu, D. Han, Spatiotemporal Sparsity Induced Similarity Measure for Human Action Recognition, International Journal of Digital Content Technology and its Applications 4 (8) (2010) 143–149.

[20] J. Zhang, Y. Wang, J. Chen, Q. Li, Sparse representation for action recognition, International Congress on Image and Signal Processing 1 (2010) 372–376.

[21] K. Guo, P. Ishwar, J. Konrad, Action recognition using sparse representation on covariance manifolds of optical flow, International Conference on Advanced Video and Signal Based Surveillance (2010) 188–195.

[22] T. Guha, R. Ward, Learning Sparse Representations for Human Action Recognition, Transactions on Pattern Analysis and Machine Intelligence 34 (2011) 1576–1588.

[23] Q. Qiu, Z. Jiang, R. Chellappa, Sparse Dictionary-based Representation and Recognition of Action Attributes, International Conference on Computer Vision (2011) 707–714.

[24] S. Cosar, M. Cetin, A group sparsity-driven approach to 3-D action recognition, International Conference on Computer Vision Workshops (2011) 1904 –1911.

[25] M. Piccardi, Background subtraction techniques: a review, International Conference on Systems, Man and Cybernetics 4 (2005) 3099–3104.

[26] N. Gkalelis, A. Tefas, I. Pitas, Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition, IEEE Transactions on Circuits and Systems for Video Technology 18 (11) (2008) 1511–1521.

[27] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, A. Lendasse, OP-ELM: optimally pruned extreme learning machine, IEEE Transactions on Neural Networks 21 (1) (2010) 158–162.

[28] G. Huang, L. Chen, Convex incremental extreme learning machine, Neurocomputing 70 (16-18) (2007) 3056–3062.

[29] A. Yang, S. Sastry, A. Ganesh, Y. Ma, Fast l1-minimization algorithms and an application in robust face recognition: A review, IEEE International Conference on Image Processing (2010) 1849–1852.

[30] E. Candes, J. Romberg, l1-magic: Recovery of sparse signals via convex programming, URL: www.acm.caltech.edu/l1magic/downloads/l1magic 4.

[31] W. Deng, Q. Zheng, L. Chen, Regularized extreme learning machine, Symposium on Computational Intelligence and Data Mining (2009) 389–395.

[32] G. Huang, H. Zhou, X. Ding, R. Zhang, Extreme Learning Machine for Regression and Multiclass Classification, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42 (2012) 512–529.

[33] J. Pantrigo, A. Sanchez, J. Mira, Representation spaces in a visual-based human action recognition system, Neurocomputing 72 (2009) 901–915.

[34] S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, Robust View Transformation Model for Gait Recognition, IEEE International Conference on Image Processing (2011) 2073–2076.

[35] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, I. Pitas, The i3DPost multi-view and 3D human action/interaction database, Conference on Visual Media Production (2009) 159–168.

[36] N. Gkalelis, N. Nikolaidis, I. Pitas, View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation, IEEE International Conference on Multimedia and Expo (2009) 394–397.

[37] A. Iosifidis, A. Tefas, I. Pitas, View-Invariant Action Recognition Based on Artificial Neural Networks, IEEE Transactions on Neural Networks and Learning Systems 23 (3) (2012) 412–424.

[38] A. Iosifidis, A. Tefas, I. Pitas, Multi-view Action Recognition Based on Action Volumes, Fuzzy Distances and Cluster Discriminant Analysis, Pattern Recognition Letters 93 (6) (2013) 1445–1457.

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