Neural representation and learning for multi-view human action recognition

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas
Aristotle University of Thessaloniki, Depeartment of Informatics
Box 451, 54124 Thessaloniki, Greece
Email: {aiosif,tefas,pitas}@aiia.csd.auth.gr

***Abstract*—In this paper we propose a novel method aiming at view-independent multi-view action recognition. Instead of combining the information provided by all the cameras forming the camera setup, for action representation and classification, we perform single-view action representation and classification to all the available videos depicting the person under consideration independently. Action representation involves a self organizing neural network training followed by fuzzy vector quantization. Action classification is performed by a feedforward neural network which is trained for view-invariant action recognition. Multiple action classification results combination based on Bayesian learning, in the recognition phase, results to high action recognition accuracy. The performance of the proposed action recognition method is evaluated on two publicly available databases, aiming at different application scenarios.**

I. Introduction

Human action recognition is an important task finding applications in many fields. It can be considered as the main pre-processing step in high-level semantic video analysis applications, including visual surveillance [1], human-computer interaction and games [2] and video content annotation [3]. Due to its importance, it has enjoyed considerable research study in the last two decades. Since the early 90's, hundreds of human action recognition methods have been proposed in the literature aiming at action recognition in several application scenarios. Depending on the application scenario, each method utilizes one or multiple cameras to capture the visual information that is needed in order to describe actions. Methods using one camera are referred as single-view methods, since they exploit information captured by one viewing angle, while methods using multiple cameras, i.e., a multi-camera setup, are referred as multi-view methods.

Multi-view methods have been recently proposed in the literature in order to address the so-called viewing angle effect [4]. It is evident that the human body during action execution can be considered as a high level deformable object. Thus, human actions, when captured by different viewing angles are quite different. Single-view methods [5], usually, assume that actions are captured by the same viewing angle during both the training and recognition phases. However, this is a strong assumption, which is not met in many applications. Multi-view methods can effectively address this issue, since they exploit visual information coming from multiple viewing angles [6], [7]. That is, by capturing the human body during action execution from many viewing angles, actions can be better described. This advantage of multi-view methods is moderated by their higher computational cost, since multiple video streams need to be processed.

Real-time operation is very important in many human action recognition systems. This is why simple human action descriptions have been adopted by most methods. The most widely adopted action description represents actions as series of human body configurations (poses), in the sense of binary images denoting the human body. For example, action 'walk' can be described as consecutive human body poses like those illustrated in Figure 1.

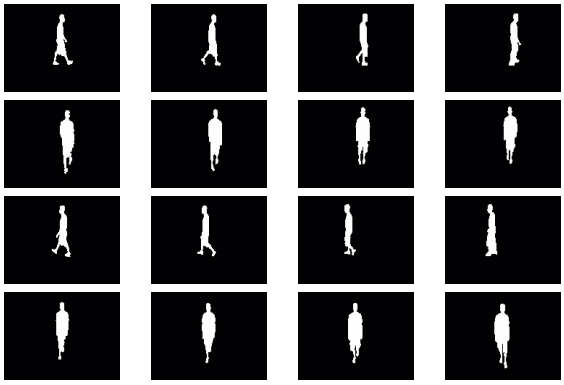

Fig. 1. *Action 'walk' as series of human body poses when captured by different viewing angles. From up to down:* 90^o, 0^o, 270^o *and* 180^o.

Most multi-view methods proposed in the literature combine the human body poses before action description and classification. Weinland et. al. [8] employ the binary body images in order to compute visual hulls which are, subsequently, accumulated over a time period to produce the so-called Motion History Volumes (MHVs). Actions are described by calculating view-invariant features in Fourier space, which are obtained after the transformation of MHVs into cylindrical coordinates around the human body vertical axes. This method can perform view-independent action recognition employing a 3D human body representation. However, it is computational expensive and, thus, it can not be used in applications where the real-time operation is important. Furthermore, visual hulls calculation requires calibrated and synchronized camera setups. This involves a configuration process before the method can operate using a different camera setup. In order to address these issues, two low-computational cost multi-view human

body representations, obtained by an un-calibrated multi-camera setup have been proposed in [9], [10]. Binary body images coming from different viewing angles are combined in order to produce the so-called multi-view postures. View-invariant human body representation is obtained either by re-arranging the binary body images with respect to the human body orientation by using morphological operations and human body proportions [9], or by calculating view-invariant multi-view postures in the DFT space [10]. Both these human body representations lead to view-independent action recognition setting several assumptions. Specifically, the person under consideration should be visible from all the cameras forming the camera setup. The camera setups used in both the training and recognition phases must consist of the same number of cameras, placed in the same positions. Finally, the cameras forming both the training and recognition camera setups need to be synchronized.

In real applications [11], the cameras forming the recognition setup may have different properties, such as video frame resolution and frame rate, and synchronization errors are usual in multi-camera setups. The number of cameras forming the training and recognition camera setups may not be the same. Finally, and most important, the person under consideration may not be visible from all the cameras in the recognition phase. Having these in mind, it can be seen that the assumptions that most multi-view methods set are very restrictive. We believe that these assumptions are mainly due to the adopted information combination strategy. Action description based on human body representation that exploits all the available information coming from all the cameras at once, is not flexible.

In this paper we propose a combination strategy that can address all the above mentioned issues appearing in real application scenarios involving action recognition using multi-camera setups. We use a camera setup consisting of $N_C \geq 1$ cameras in order to capture the human body during action execution from multiple viewing angles in the training phase. This results to the creation of multiple videos depicting the training action instances from various viewing angles. Videos depicting action instances will be called action videos hereafter. Using all the training action videos and the corresponding action labels we describe actions following a single-view strategy and we, subsequently, train a classifier that will be used for action video classification in the recognition phase. We, finally, employ an action video classification results combination strategy based on Bayesian learning. In the recognition phase, $N \leq N_C$ action videos depicting the person under consideration performing an action instance from all the N available cameras are classified independently, producing N action classification results. These classification results are, subsequently, combined in order to recognize the test action instance.

The paper is structured as follows. We describe the proposed action classification method in Section II. Specifically, the adopted action video representation scheme is described in Subsection II-A. Classification of action videos is described in Subsection II-B. The proposed action classification results combination strategy is described in Subsection II-C. Experimental results conducted in order to assess the performance of the proposed method are illustrated in Section III. Finally, conclusions are drawn in Section IV.

II. Proposed Method

The proposed method operates on binary action videos. In the case of color action videos, moving object segmentation [12], [13], or color-based image segmentation techniques [14] are applied to the color action videos frames in order to produce binary action videos denoting the human body poses.

A. Action Representation

Let a binary action video i consist of N_i video frames. These frames are centered with respect to the human body center of mass. The size of the maximum bounding box that encloses the human body in the entire video is determined and the centered binary video frames are cropped using this bounding box size. The resulting images are rescaled to fixed size ($N_H \times N_W$ pixels) images in order to produce binary pose images, like those illustrated in Figure 2.

Fig. 2. *Binary human pose images of eight actions taken from various viewing angles.*

Binary pose images are vectorized in order to produce the so-called posture vectors $\mathbf{p}_{ij} \in \mathbb{R}^D$, $j = 1, ..., N_i$, $D = N_H \cdot N_W$.

Let us assume that the above described procedure has been applied to all the N_T training action videos, resulting to $N_P = \sum_{i=1}^{N_T} N_i$ posture vectors $\mathbf{p}_{ij}$. We employ all these N_P posture vectors in order to calculate K posture prototypes $\mathbf{v}_k$, $k = 1, ..., K$. In this work we train a Self Organizing Map (SOM) [15], resulting to a topographic map of the training posture vectors. SOM neurons can be considered to be representative posture vectors, corresponding to representative human body poses during action execution. An example SOM produced by using action videos belonging to eight actions captured by an eight-view camera setup is illustrated in figure 3.

The SOM neurons $\mathbf{w}_k \in \mathbb{R}^D$ are randomly initialized and updated by introducing the training posture vectors $\mathbf{p}_{ij}$ multiple times (epoches) in a random order. At each update iteration, the involved posture vector $\mathbf{p}_{ij}$ is compared with all the SOM neurons $\mathbf{w}_k$ by calculating the corresponding Euclidean distance:

$$d_{ijk} = \| \mathbf{p}_{ij} - \mathbf{w}_k \|_2, \quad (1)$$

The SOM neuron that provides the smallest Euclidean distance is used to determine a topographical neighborhood, based on which all the SOM neurons $\mathbf{w}_k$, $k = 1, ..., K$ are updated, using the following update rule:

$$\mathbf{w}_k(n+1) = \mathbf{w}_k(n) + \eta h_k(n)(\mathbf{p}_{ij} - \mathbf{w}_k(n)). \quad (2)$$

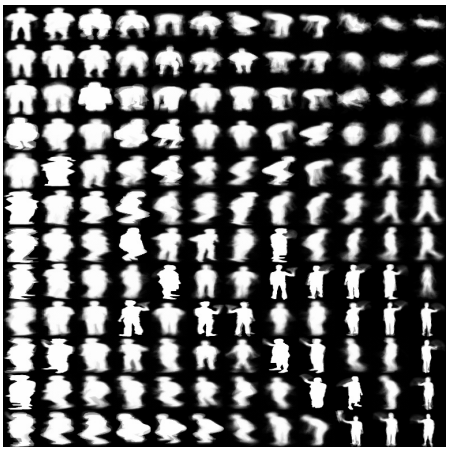

Fig. 3. *A 12×12 SOM obtained by using action videos depicting eight persons performing eight actions (walk, run, jump in place, jump forward, bend, sit, fall down and wave one hand).*

where η is the update rate parameter and $h_k(n) = exp(\frac{r_k^2}{2\sigma^2(n)})$. r_k is the lateral distance between the winning neuron and neuron k.

After SOM calculation, we map training posture vectors $\mathbf{p}_{ij}$ to the so-called fuzzy membership vectors $\mathbf{u}_{ij} = [u_{ij1}, ..., u_{ijK}]^T$, where:

$$u_{ijk} = \frac{(\parallel \mathbf{p}_{ij} - \mathbf{w}_k \parallel_2)^{-\frac{2}{m-1}}}{\sum_{n=1}^{K}(\parallel \mathbf{p}_{ij} - \mathbf{w}_n \parallel_2)^{-\frac{2}{m-1}}}, \quad (3)$$

$m > 1$ is the fuzzification parameter.

Mean membership vectors $\mathbf{u}_{ij}$ describe the similarity of each posture vector $\mathbf{p}_{ij}$ with all the SOM neurons $\mathbf{w}_k$. The mean membership vectors $\mathbf{u}_i$ are, subsequently, used to represent the training action videos, i.e.:

$$\mathbf{u}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{u}_{ij}. \quad (4)$$

Mean membership vectors $\mathbf{u}_i$ are, finally, normalized to have zero mean and unit variance producing the so-called action vectors $\mathbf{s}_i$, $i = 1, ..., N_T$. Mean membership vectors representing test action videos are normalized accordingly.

B. Action Video Classification

After obtaining the action vectors $\mathbf{s}_i$ representing all the training action videos, we train a classifier exploiting the action labels available for the training action videos. In this work, we train a feedforward neural network, using the Backpropagation training algorithm [16]. For an action recognition problem aiming to distinguish N_A action classes consisting an action class set $\mathcal{A} = \{\alpha_1, ..., \alpha_{N_A}\}$, the adopted network topology consists of K input neurons and N_A output neurons. For each of the training action vectors $\mathbf{s}_i$, $i = 1, ..., N_T$, the corresponding output vector $\mathbf{o}_i \in \mathbb{R}^{N_A}$ is determined to have values equal to $o_{ik} = 1$ for action vectors belonging to action class k and $o_{ik} = -1$ otherwise.

Backpropagation algorithm is an iterative procedure aiming to determine the network's weights that minimize the Mean Square Error (MSE) between the actual network outputs $\hat{\mathbf{o}}_i$ and the desired network outputs $\mathbf{o}_i$:

$$E[\frac{1}{N_A}(\hat{\mathbf{o}}_i - \mathbf{o}_i)^2] < \varepsilon. \quad (5)$$

For each of the training action vectors $\mathbf{s}_i$, network responses $\hat{o}_{ij}$ are calculated by:

$$\hat{o}_{ik} = f_s(\mathbf{s}_i^T\mathbf{W}_k),\ k = 1, ..., N_A, \quad (6)$$

where $\mathbf{W}_k$ is a vector that contains the network weights corresponding to output k and f_s is the sigmoid function. Network weights $\mathbf{W}_{kl}, k = 1, ..., N_A, l = 1, ..., K$ are updated by using the following update rule:

$$\Delta W_{kl}(n+1) = c\Delta W_{kl}(n) + \eta\delta_k(n)\mathbf{s}_{il}(n), \quad (7)$$

where $\delta_k(n)$ is the local gradient for the k-th neuron, η is the learning-rate, c is the momentum constant and n is the iteration number. Alternatively, the network's weight values can be calculated by applying the Levenberg - Marquadrt algorithm [17] for error back-propagation.

After completing the network's training procedure, a test action vector $\mathbf{s}_{test}$ can be introduced to the network and be classified to the action class that corresponds to the highest network's output, i.e.:

$$l_{test} = \underset{i}{argmax}\ \hat{o}_{test,i}. \quad (8)$$

C. Action Classification Results Combination

Let us assume that a person performs an action instance and that he/she is captured by N cameras. This results to the creation of N action videos, which can be classified to one of the N_A action classes by introducing the corresponding action videos to the feedforward network as described in Subsections II-A and II-B. Since all these N action videos depict the same action instance performed by the same person, we would like to combine all these N action classification results in order to decide in which action class the action instance belongs to. One intuitive approach would be the classification of the action instance by performing a majority voting algorithm on these N action classification results. However, this may not be the optimal combination approach. In the cases where the action classes appearing in the classification problem are quite similar, a simple majority voting procedure may result in classification errors or unrecognized action instances. Consider the case of distinguishing actions 'walk' and 'run' by using a multi-camera setup consisting of an even number of cameras. Since it is difficult to distinguish these two actions, it is possible that half of the action videos will be classified to action class 'walk' and the remaining action videos will be classified to action class 'run'. In this case, a majority voting combination approach will not be able to provide the correct recognition result.

In order to avoid such situations, we formulate an action classification results combination strategy based on Bayesian

learning. Assuming that both training and test action vectors follow the same distributions, we introduce all the N_T training action vectors to the trained feedforward network and we obtain its outputs $\hat{\mathbf{o}}_i$. Training action vectors are, subsequently, classified to the action classes corresponding to the highest network outputs, i.e.:

$$l_i = \underset{j}{argmax}\ \hat{o}_{ij},\ i = 1, ..., N_T. \tag{9}$$

Given these classification results and the available training action video labels, we can calculate the probabilities $P(l_i|\alpha_j)$, i.e., the probabilities that action video i belonging to action class α_j was classified to action class l_i. Furthermore, by assuming equiprobable action classes, we can assume that the a priori probability of action class α_j, $j = 1, ..., N_A$ is equal to $P(\alpha_j) = \frac{1}{N_A}$. If this assumption is not met, $P(\alpha_j)$ can be calculated as $P(\alpha_j) = \frac{N_{\alpha,j}}{N_T}$, where $N_{\alpha,j}$ is the number of training action videos belonging to action class α_j.

Let us now assume that the N action vectors representing a test action instance are introduced to the feedforward network and N action class labels $l_{test,i}$ are obtained. Given probabilities $P(\alpha_j)$ and $P(l_i|\alpha_j)$ determined in the training phase, the test action instance can be classified to the action class that provides the maximum a posteriori probability sum [18], i.e.:

$$l_{test} = \underset{j}{argmax} \sum_{i=1}^{N} P(a_j|l_{test,i}). \tag{10}$$

Probabilities $P(a_j|l_{test,i})$ are calculated by using the Bayes formula:

$$P(a_j|l_{test,i}) = \frac{P(l_{test,i}|a_j) \cdot P(a_j)}{\sum_{n=1}^{N_A} P(l_{test,i}|a_n) \cdot P(a_n)}. \tag{11}$$

As can be observed, by following the above described combination strategy, a flexible multi-view action recognition method is obtained. The camera setups used in the training and recognition phases do not need to be calibrated. The number of cameras forming the training and test camera setups may differ, while synchronization errors between the cameras do not affect its performance, since each action video is classified independently. Finally, by adopting a low-computational cost human body representation, the proposed multi-view action recognition method can operate in high frame rates, compared to other multi-view methods employing 3D human body representation.

III. Experimental Results

In this Section we present experimental results assessing the performance of the proposed action recognition method. Since the proposed approach can operate using one or multiple cameras, we present experiments conducted on two publicly available action recognition databases, aiming at different application scenarios. The first one, is a multi-view action recognition database aiming at recognition of daily actions. The second database, is a single-view action recognition database aiming at recognition of action appearing in meal intakes. Comprehensive descriptions of these databases followed by experimental results obtained by applying the proposed method on them are provided in the following.

Regarding the parameters used by the proposed method, the following values have been used: $N_H = 32$, $N_W = 32$, $m = 1.1$, $n = 0.1$ and $c = 0.1$. For the determination of the optimal values of the remaining parameters, the Leave-One-Person-Out (LOPO) cross-validation procedure has been used. LOPO involves training the method by using the action videos depicting all but one persons in the database and testing it by using the action videos depicting the remaining ones. This procedure is performed multiple times, equal to the number of persons in the database, in order to complete one experiment. Multiple experiments have been performed by using different parameter values, and the optimal ones were determined to be those providing the highest action recognition rate.

A. The i3DPost multi-view database

The i3DPost multi-view database [19] contains high resolution image sequences depicting eight persons performing eight daily actions. Eight cameras were used in order to capture the persons while they performed one or multiple instances of eight actions classes: 'walk', 'run', 'jump in place' (jump1), 'jump forward' (jump2), 'bend', 'fall', 'sit on a chair' (sit) and 'wave one hand' (wave). Example action video frames depicting a person of the database from different viewing angles are illustrated in Figure 4. Binary action videos have been produced by applying an image segmentation technique to the action video frames exploiting the properties of the HSV color-space.

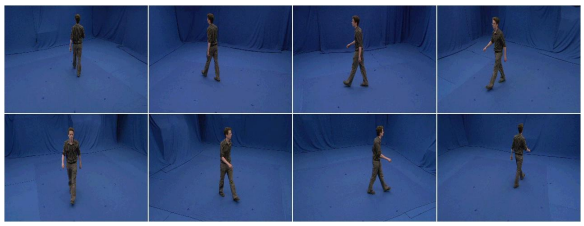

Fig. 4. *Action video frames of the i3DPost database depicting a person walking from multiple viewing angles.*

The LOPO cross-validation procedure has been performed multiple times by using the produced binary action videos. In order to assess the performance of the proposed action classification results combination strategy, we have performed the LOPO procedure by using the same parameter values and combining the action classification results using a majority voting algorithm. Figure 5 illustrates the action classification rates obtained by using the optimal parameter values for both combination approaches and various SOM topologies. An action classification rate equal to 94.37% has been obtained by using a 12×12 SOM for action video representation and the proposed combination strategy. The corresponding action classification rate for the majority voting approach is equal to 92.65%. The confusion matrices of these experiments are illustrated in Figure 6.

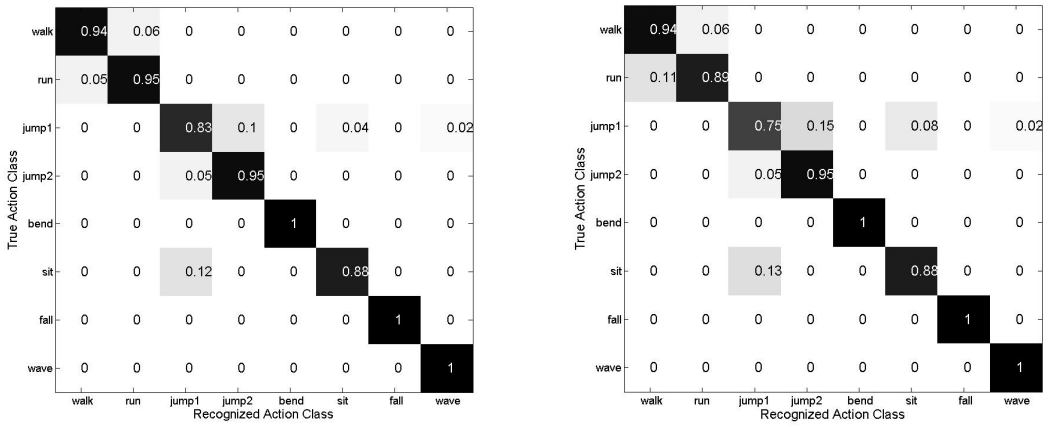

Fig. 6. *Confusion matrices in i3DPost database following the a) Bayesian and b) majority voting action classification results combination strategies.*

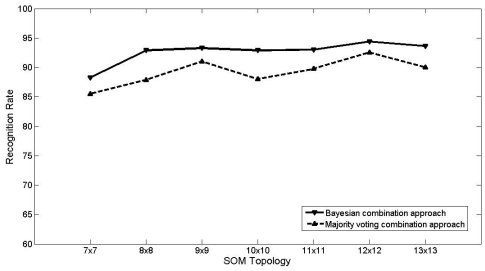

Fig. 5. *Action classification rates on the i3DPost database.*

As can be seen, high action classification rates have been obtained in all these experiments. Most actions are well distinguished from the others. Action classes 'jump in place', 'jump forward' and 'sit' are more difficult to be recognized. This is reasonable, since they contain a high number of common human body poses and variations in execution style between different persons may result that action instances performed by one person belonging to one action class are more similar to action instances belonging to another action class performed by another person. Finally, it can be seen that the proposed combination strategy outperforms the majority voting one in all these experiments.

Comparing the proposed method with other methods evaluating their performance in the i3DPost database using the same experimental protocol, we can see that the proposed method clearly outperform the one presented in [9], where the best reported action classification rate is equal to 90.88%. The method presented in [10], reports a 94.37% action classification rate, equal to the one obtained by the proposed method. However, the method in [10] can operate only in the cases where the person under consideration is visible from all the cameras forming the recognition camera setup, which need to be synchronized. Furthermore, the training and recognition camera setups need to be consisted of the same number of cameras, placed at the same positions. In different cases, its performance will decrease. Thus, the proposed method can successfully operate in the cases where the method in [10] will fail.

B. The AIIA-MOBISERV single-view database

The AIIA-MOBISERV single-view database [20], [21] contains low resolution image sequences depicting twelve persons having a meal. One camera was placed at a distance of 2m in front of a table. Each person was recorded multiple times, each for a different day. The persons eat using a fork, a cutlery or a spoon and drink from a cup or a glass. Furthermore, the persons perform other actions appearing in meal intakes, such as slicing their food and chewing it. The human body ROIs used to describe the persons' poses were determined to be their heads and hands. Binary action videos have been produced by applying a color-based image segmentation technique [14]. Example binary human pose images are illustrated in Figure 7.

Fig. 7. *Binary action video frames of the AIIA-MOBISERV database depicting a person a) eating, b) drinking and c) slicing his food.*

In our experiments we formulated a three-class classification problem. That is, the videos appearing in the database were accompanied by action class labels belonging to action classes 'eat' and 'drink'. Videos not belonging to these two action classes, i.e., the videos depicting the persons slicing their food or chewing it, formed a third action class, which we named as 'apraxia'. The choice of three, instead of seven, 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 case of the above described three-class classification problem are high, since due to human body proportions 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 than from an action video depicting a sequence belonging to an other eating subclass, e.g. 'eat with fork'.

The LOPO cross-validation procedure has been performed multiple times by using the binary action videos of the AIIA-MOBISERV database. The action classification rates obtained for various SOM topologies are illustrated in Figure 8. The optimal SOM topologies were found to be equal to 12×12 providing an action classification rate equal to 89.66%. The confusion matrix of this experiment is illustrated in Figure 9.

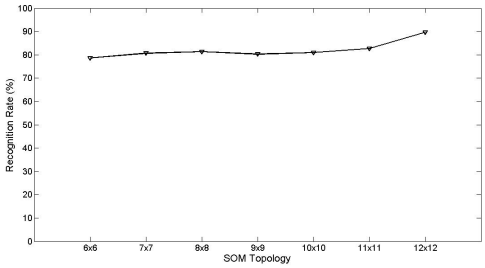

Fig. 8. *Action classification rates on the i3DPost database.*

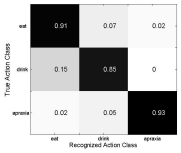

Fig. 9. *Action classification rates on the i3DPost database.*

As can be seen the action classification rates obtained for all three classes are high. There is a small confusion between action classes 'eat' and 'drink' which is reasonable, since motion variations appearing in these two classes are very few.

IV. Conclusion

In this paper we presented a novel multi-view method aiming at view-independent action recognition utilizing multi-camera setups. Instead of combining the information provided by all the cameras forming the camera setup, for action representation and classification, we perform single-view action representation and classification to all the available videos depicting the person under consideration independently. Action representation involves a self organizing neural network training followed by fuzzy vector quantization. Action classification is performed by a feedforward neural network which is trained for view-invariant action recognition. Multiple classification results combination based on Bayesian learning, in the recognition phase, results to high action recognition accuracy. The proposed multi-view approach can operate in situations appearing in real-application scenarios, such as total person occlusion in some of the cameras forming the recognition camera setup and synchronization errors between the cameras.

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] L. Weilun, H. Jungong, and P. With, "Flexible human behavior analysis framework for video surveillance applications," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, Article ID 920121, 9 pages, 2010.

[2] P. Barr, J. Noble, and R. Biddle, "Video game values: Human-computer interaction and games," *Interacting with Computers*, vol. 19, no. 2, pp. 180–195, Mar. 2007.

[3] B. Song, E. Tuncel, and A. Chowdhury, "Towards a multi-terminal video compression algorithm by integrating distributed source coding with geometrical constraints," *Journal of Multimedia*, vol. 2, no. 3, pp. 9–16, 2007.

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

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[6] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *Transactions on Systems, Man and Cybernetics Part–C*, vol. 40, no. 1, pp. 13–24, Jan. 2010.

[7] M. Holte, C. Tran, M. Trivedi, and T. Moeslund, "Human action recognition using multiple views: a comparative perspective on recent developments," *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pp. 47–52, 2011.

[8] 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.

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

[10] 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.

[11] F. Qureshi and D. Terzopoulos, "Surveillance camera scheduling: A virtual vision approach," *Proceedings Third ACM International Workshop on Video Surveillance and Sensor Networks*, vol. 12, pp. 269–283, Nov. 2005.

[12] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," *19th International Conference on Pattern Recognition ICPR*, pp. 1–4, 2009.

[13] M. Piccardi, "Background subtraction techniques: a review," *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3099–3104, 2005.

[14] E. Marami, A. Tefas, and I. Pitas, "Nutrition assistance based on skin color segmentation and support vector machines," *Man-Machine Interactions 2*, pp. 179–187, 2011.

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

[16] P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," 1974.

[17] J. More, "The levenberg-marquardt algorithm: implementation and theory," *Numerical analysis*, pp. 105–116, 1978.

[18] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[19] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," *6th Conference on Visual Media Production*, pp. 159–168, Nov. 2009.

[20] "http://poseidon.csd.auth.gr/mobiserv-aiia/index.html."

[21] A. Iosifidis, A. Tefas, and I. Pitas, "Activity based person identification using fuzzy representation and discriminant learning," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, 2012.