PERSON SPECIFIC ACTIVITY RECOGNITION USING FUZZY LEARNING AND DISCRIMINANT ANALYSIS

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

One of the major issues that activity recognition methods should be able to face is the style variations observed in the execution of activities performed by different humans. In order to address this issue we propose a person-specific activity recognition framework in which human identification proceeds activity recognition. After recognizing the ID of the human depicted in a video stream, a person-specific activity classifier is responsible to recognize the activity performed by the human. Exploiting the enriched human body information captured by a multi-camera setup, view-invariant person and activity representations are obtained. The classification procedure involves Fuzzy Vector Quantization and Linear Discriminant Analysis. The proposed method is applied on drinking and eating activity recognition as well as on other activity recognition tasks. Experiments show that the person-specific approach outperforms the person-independent one.

1. INTRODUCTION

Human activity recognition is an active research field due to its importance in a wide range of applications. It can be considered as the main preprocessing step in human behavior analysis applications, such as visual surveillance. Furthermore, it can be used in entertainment industry in order to provide 3D actor reconstruction for digital cinema and interactive games. The term activity can be used in several ways. That is why several taxonomies have been proposed by researchers in order to describe human motion hierarchies. In this paper activities are referred as the middle level human motion patterns, i.e., the term activity refers to a simple human motion pattern such as a walking step.

In order to describe activities the global human body information, in terms of binary images that denote the image locations occupied by the human regions of interest, can be used. This human body representation provides the human body configurations related with the poses consisting activities. By using image segmentation methods, such as background subtraction or other color-based image segmentation techniques, binary images denoting the regions of interest can be extracted effectively.

It is evident that the observation angle, that activities are captured from, plays a crucial role in the effectiveness for most of the methods proposed in the literature, as they exploit information captured by one camera. In order to overcome this problem, researchers have provided view-invariant human body representations [8], [14]. However, they have been proved to be of moderate invariance and applicability in real applications. An other option is the use of multi-camera setups [1], [13]. By capturing the human body by different viewing-angles enriched human body representation is achieved. Thus the viewing-angle effect is eliminated. For a detailed description of recent work on the action recognition field the reader is referred to [7], [11].

Two important issues that an activity recognition method should be able to face is the fact that body proportions differ between different humans and that humans perform activities in different styles. That is, the body silhouettes representing different humans captured by the same observation angle performing the same activity will probable differ. Sometimes it is possible that body poses of a human performing an activity are similar with the body poses of another human performing a different activity. This is due to the fact that there is not a formal description of activities. By intuition, the use of the human ID should address these issues and increase the recognition rates.

Having these in mind, we investigate the person-specific activity recognition task. We propose a unified framework in which human identification proceeds human action recognition. That is, after recognizing the ID of a person depicted in a video stream, a person-specific activity recognition classifier is responsible to recognize the activity performed by this person. This approach is motivated by relative work in human face verification [6], [15] in which the use of the human ID leads to the production of a more discriminant data representation and increases the recognition rates. To the best of our knowledge this is the first time that this approach is applied to the activity recognition field.

Depending on the application, an activity recognition method is required to recognize different types of activities. A usual case is the recognition of everyday activities such as walk, run, jump or bend as they can be used to describe the behavior of people in public places. Another interesting case is the recognition of drinking and eating activity that can be used in automatic nutrition assistance systems integrated in smart homes environments. Their aim is to prolong independent living of older persons targeting to patients in the early stages of dementia that have the risk of underfeeding or dehydration. The nutrition assistance system may remind or encourage the patients to eat or drink something when lack of eating/drinking is detected. Several methods have been proposed to this end, most of them exploiting information provided by sensors [2], [10]. The use of video information seems to be more suitable for this task, as it provides a non-invasive recognition procedure.

Experiments performed in two activity recognition databases, using either one camera or a multi-camera setup show that the person-specific approach outperforms the person-independent one on the activity recognition task.

2. PROPOSED METHOD

The proposed method utilizes a converging N-camera setup, as the one shown in Figure 1 (for $N = 8$). The place that can be seen from all the N cameras is referred as the camera setup capture volume. A person inside the capture volume is captured by all the N cameras by a different viewing-angle. The person can freely move inside the capture volume. This affects the viewing angle of all the cameras. For example, if at a time instance camera #1 captures the pesron's frontal view, a change in his/her motion direction may result this camera to capture his/her side view. This is the so-called camera viewpoint identification problem and should be solved in order to perform view-invariant human activity recognition. This problem is addressed properly in our method by exploiting the circular shift invariance property of the Discrete Fourier Transform (DFT). As one can observe the use of one camera is a sub-case of the N-camera setup ($N = 1$).

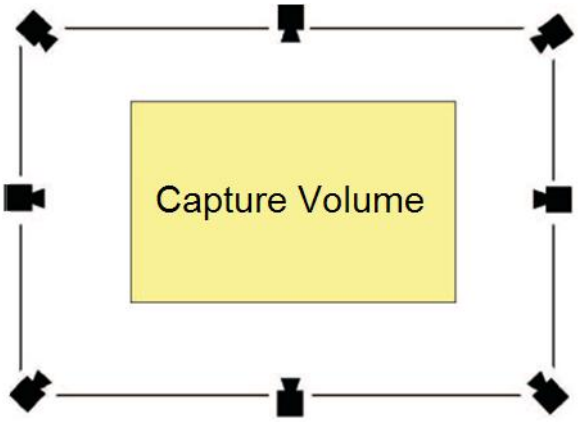

Figure 1: *An eight-view converging camera setup.*

Activities are described by a number of consecutive human body poses in terms of binary images depicting the human regions of interest (ROIs) in white and the background in black. Videos containing multiple activities are segmented in smaller videos depicting single activities, thus producing the so-called activity videos. Synchronized activity videos from all the N cameras are referred as N-view activity videos.

2.1 Preprocessing

In order to extract binary images denoting the person's ROIs, image segmentation techniques [9], [3] are applied to the frames of the action videos captured by all the N cameras. These images are centered to the ROIs' center of mass and bounding boxes of size equal to the maximum bounding box that encloses the person's ROIs in each activity video are extracted and rescaled in order to produce binary posture images with fixed size ($L_x \times L_y$) pixels. In our experiments $L_x = L_y = 64$. Five binary posture images depicting a human performing five activities (walk, run, jump in place, jump forward and bend) captured by the side (90^o) viewing-angle are shown in Figure 2.

The N binary posture images from all the N cameras corresponding to the same time instance are concatenated, using the camera ordering, to produce the so-called N-view binary posture images as shown in Figure 3.

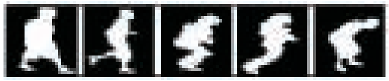

Figure 2: *Five single-view binary posture images.*

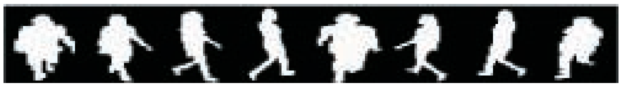

Figure 3: *An* 8*-view binary posture image.*

The N-view binary posture images are scanned column-wise and produce the so called posture vectors, $\mathbf{p}_i \in \mathscr{R}^{N_s}$, $N_s = L_x \times L_y \times N$. In order to solve the camera viewpoint identification problem the circular invariance property of the magnitudes of the DFT is exploited. Each posture vector is transformed to a vector containing the magnitudes of its DFT transform:

$$P_{ij}(k) = |\sum_{n=0}^{N_s-1} p_{ij}(n)e^{-i\frac{2\pi k}{N_s}n}|, \quad k = 1,...,N_s-1. \tag{1}$$

The use of Fast Fourier Transform (FFT) can fasten the procedure. That is, each activity is represented by a set of posture vectors $\mathscr{P}_i = \{\mathbf{P}_{i1}, \mathbf{P}_{i2}, ..., \mathbf{P}_{iN_i}\}$, where N_i is the number of N-view binary posture images consisting each activity video.

2.2 Classifier

In the training phase all the N_T training posture vectors P_{ij} representing all the N_v training activity videos are clustered to a fixed number of clusters using a K-Means algorithm [12], without exploiting the available labeling information, producing D posture prototype vectors $\mathbf{v}_k \in \mathscr{R}^{N_s}$, $k = 1,...,D$. After the computation of the posture prototypes, the fuzzy membership of each posture vector to all the posture prototypes is calculated:

$$u_{k,ij} = \frac{(\parallel \mathbf{P}_{ij} - \mathbf{v}_k \parallel_2)^{-2(m-1)^{-1}}}{\sum_{j=1}^{D}(\parallel \mathbf{P}_{ij} - \mathbf{v}_k \parallel_2))^{-2(m-1)^{-1}}}, \tag{2}$$

where m is the fuzzification parameter, $m = 1.1$ in our experiments. Each posture vector is mapped to the corresponding membership vector $\mathbf{u}_{ij} \in \mathscr{R}^D, \mathbf{u}_{ij} = [u_{1,ij}, u_{2,ij}, ..., u_{D,ij}]^T$. Finally, each N-view activity video is represented by the mean membership vector the so-called activity vector, $\mathbf{s}_i \in \mathscr{R}^D$:

$$\mathbf{s}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{u}_{ij}\,. \tag{3}$$

Exploiting the labeling information available in the training phase, Linear Discriminant Analysis (LDA) [4] is used to map the training activity vectors $\mathbf{s}_i$ to an optimal discriminant subspace $\mathscr{R}^d$, $d < D$, in which the activity classes are linearly separable. This is approximated by minimizing the following criterion:

$$\mathbf{J}_{opt} = \underset{\mathbf{J}}{\operatorname{argmax}}\frac{|\,\mathbf{J}^T\mathbf{S}_b\mathbf{J}\,|}{|\,\mathbf{J}^T\mathbf{S}_w\mathbf{J}\,|} \tag{4}$$

In Equation 4, $\mathbf{S}_w$ is the within class scatter matrix and $\mathbf{S}_b$ is the between class scatter matrix of the training activity vectors $\mathbf{s}_i$:

$$\mathbf{S}_w = \sum_{i=1}^{N_A} \sum_{j=1}^{N_i} \frac{(\mathbf{s}_{ij} - \mathbf{m}_i)(\mathbf{s}_{ij} - \mathbf{m}_i)^T}{N_i} \qquad (5)$$

$$\mathbf{S}_b = \sum_{i=1}^{N_A} \frac{(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T}{N_i} \qquad (6)$$

where N_A is the number of activity classes, $\mathbf{s}_{ij}$ denotes the j-th activity vector belonging to the i-th activity class, N_i is the number of activity vectors of activity class i, $\mathbf{m}_i$ is the mean activity vector of the i-th activity class and $\mathbf{m}$ is the total mean activity vector of the training set.

After calculating $\mathbf{J}_{opt}$ activity vectors $\mathbf{s}_i$ are mapped to the corresponding discriminant activity vectors $\mathbf{z}_i \in \mathscr{R}_d$ by:

$$\mathbf{z}_i = \mathbf{J}_{opt}^T \mathbf{s}_i \qquad (7)$$

In the classification phase the testing discriminant activity vector is classified to the Nearest class Centroid using the Euclidean distance.

2.3 Classification Schemes

As already mentioned, in this paper we are interested to investigate the impact of the human ID in the activity classification task. For this reason we compare the two classification schemes presented in Figures 4 and 5.

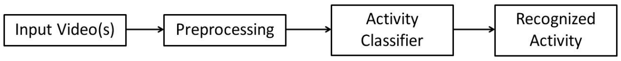

Figure 4: *Person-independent activity classification scheme.*

Both of them use classifiers as the one presented in Subsection 2.2. In the first one, only the activity labeling information of the training set is exploited. That is, in the training phase the training activity videos are accompanied with their activity class labels in order to train a person-independent activity classifier. In the second one, both activity and human ID labeling information of the training set are exploited. In both classification steps the same human body representation is used. That is, in the first classification step an activity based person identification classifier is trained using the person ID labeling information of the training activity videos. In the second classification step M person-specific activity classifiers are trained using the activity labeling information of the training activity videos depicting each of the M persons consisting the training set.

3. EXPERIMENTS

In order to compare the ability of the two classification schemes in activity recognition, we conducted experiments in two databases coming from different applications.

3.1 Multi-view activity recognition

The first one is an online available multi-view activity recognition database, in which eight persons perform eight activities (walk (wk), run (rn), jump in place (jp), jump forward (jf), bend (bd), sit down (st), fall (fl) and wave one hand (wo). The camera setup was an 8-view converging camera setup, as the one shown in Figure 1 with capture volume dimensions approximately $4 \times 3 \times 2$ meters. The studio background was of uniform blue color. More details can be found in [5]. From the eight activities we selected five (wk, rn, jp, jf and wo) as the rest were performed once from each person in the database. Binary images were extracted by thresholding on the blue color using HSV color-space. Videos were manually segmented in activity videos and preprocessed as described in Subsection 2.1.

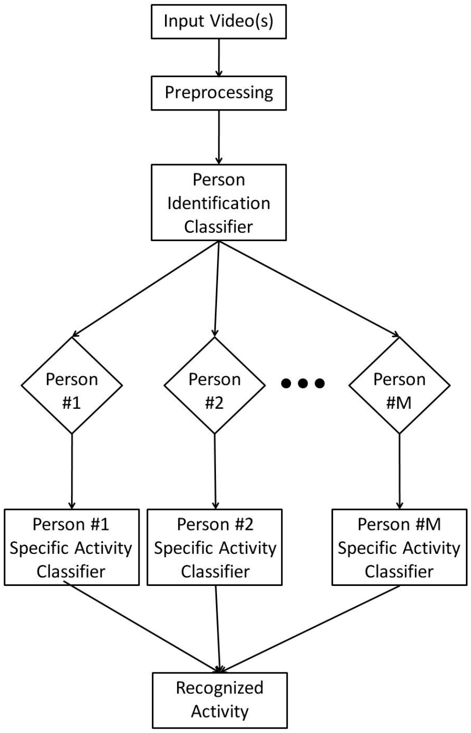
Input Video(s)
Preprocessing
Person Identification Classifier
Person #1
Person #2
Person #M
Person #1 Specific Activity Classifier
Person #2 Specific Activity Classifier
Person #M Specific Activity Classifier
Recognized Activity

Figure 5: *Person-specific activity classification scheme.*

The leave-one-out cross-validation procedure was performed in order to determine the optimal number of posture vector prototypes for both of the classification schemes. This procedure is used in order to estimate the ability of a classification method to correctly classify data that it was not trained on. It consists of several steps (folds). In each step, some of the data are used to train the algorithm, while the rest are used for evaluation. In our case at every fold we kept activity videos depicting one iteration of each activity class performed by all the persons for testing and the remaining activity videos were used for training. This procedure was applied four times, since we had four different instances of each activity. Experiments for different number of posture vector prototypes were performed for both the person-independent and the person-specific activity classification schemes. Classification accuracy equal to 92% was achieved for forty posture vector prototypes in the person-independent case. In the person-specific case the optimal parameters were found to be twenty posture vector prototypes for the person identification classifier and thirty posture vector prototypes for the activity classifiers. For these parameters classification accuracy equal to 96.4% was achieved. Confusion matrices for both schemes using the optimal parameters are presented in

Tables 1 and 2.

Table 1: *Confusion matrix for five activities using person-independent activity classification scheme. A row represents the actual activity class and a column the activity recognized by the algorithm.*

	wk	rn	jp	jf	wo
wk	1				
rn		1			
jp		0.15	0.65	0.2	
jf			0.05	0.95	
wo					1

Table 2: *Confusion matrix for five activities using person-specific activity classification scheme. A row represents the actual activity class and a column the activity recognized by the algorithm.*

	wk	rn	jp	jf	wo
wk	1				
rn	0.03	0.94	0.03		
jp			0.94	0.06	
jf	0.06			0.94	
wo					1

3.2 Eating and drinking activity recognition

In order to compare the two classification schemes in a single-view camera setup ($N = 1$), we created an eating/drinking activity recognition database. Four persons were captured by one camera in a distance of two meters in front of them during a meal. This was performed for three different days. Activity videos depicting activity classes ”eat”, ”drink” and ”apraxia” were manually segmented and binary images denoting the person’s head and hands were extracted by performing color-based image segmentation in HSV color-space. Three posture images depicting instances of a person having a meal are shown in Figure 6.

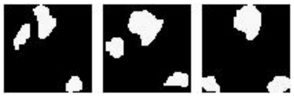

Figure 6: *Binary posture images depicting a person having a meal. From left to right: drink, eat and apraxia.*

The leave-one-iteration-out cross-validation procedure was applied again for different number of posture vector prototypes for both of the classification schemes. The optimal number of posture vector prototypes for the person-independent activity classification scheme was 24 and resulted to a classification accuracy equal to $96,6\%$. In the human-specific activity recognition case the optimal parameters were 12 and 13 posture vector prototypes for the person identification and the activity classifiers respectively. Using these parameters a classification accuracy equal to $98,3\%$ was achieved. The corresponding to the optimal parameters confusion matrices for both schemes are presented in Tables 3 and 4.

Table 3: *Confusion matrix for three activities using person-independent activity classification scheme. A row represents the actual activity class and a column the activity recognized by the algorithm.*

	dr	ea	ap
dr	0.95		0.05
ea		1	
ap	0.05		0.95

Table 4: *Confusion matrix for three activities using person-specific activity classification scheme. A row represents the actual activity class and a column the activity recognized by the algorithm.*

	dr	ea	ap
dr	1		
ea		1	
ap	0.05		0.95

As can be seen in both cases the person-specific activity recognition scheme outperforms the person-independent one. This confirms our intuition that the use of human ID helps the action classification, as style variations do not affect the recognition results.

4. CONCLUSIONS

In this paper we presented a view-invariant person-specific framework aiming to activity recognition. Information captured by different observation angles produces a view-invariant human body representation. The use of Fuzzy Vector Quantization and Linear Discriminant Analysis leads to an activity representation in a low dimensional discriminant subspace. The use of human ID increases the classification rates, as variations in style do not affect the recognition results.

Acknowledgment

This work has been funded by the Collaborative European Project MOBISERV FP7-248434 (http://www.mobiserv.eu), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.

REFERENCES

[1] M. Ahmad and S. Lee. Human action recognition using shape and clg-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7):2237–2252, July 2008.

[2] O. Amft, H. Junker, and G. Troster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 160–163. IEEE, 2005.

[3] S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff, and O. Schreer. Vision-based skin-colour segmentation of moving hands for real-time applications. In *Proc. of 1st European Conf. on Visual Media Production (CVMP)*, pages 524–529, 2004.

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

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

[6] G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas. Class-specific kernel discriminant analysis for face verification. *IEEE Transactions on Information Forensics and Security*, 2(3):570–587, September 2007.

[7] X. Ji and H. Liu. Advances in view-invariant human motion analysis: A review. *IEEE Transactions on Systems, Man and Cybernetics Part–C*, 40(1):13–24, Jan. 2010.

[8] J. Niebles and L. Fei-Fei. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, Nov. 2002.

[9] M. Piccardi. Background subtraction techniques: a review. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 4, pages 3099–3104. Ieee, 2005.

[10] K. Sim, G. Yap, C. Phua, J. Biswas, P. Wai, A. Aung, A. Tolstikov, W. Huang, and P. Yap. Improving the accuracy of erroneous-plan recognition system for Activities of Daily Living. In *e-Health Networking Applications and Services (Healthcom), 2010 12th IEEE International Conference on*, pages 28–35. IEEE, 2010.

[11] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, Nov. 2008.

[12] A. Webb. *Statistical pattern recognition*. A Hodder Arnold Publication, 1999.

[13] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2–3):249–257, Nov./Dec. 2006.

[14] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 984–989, June 2005.

[15] S. Zafeiriou, A. Tefas, and I. Pitas. Learning discriminant person-specific facial models using expandable graphs. *IEEE Transactions on Information Forensics and Security*, 2(1):55–68, March 2007.