EATING AND DRINKING ACTIVITY RECOGNITION BASED ON DISCRIMINANT ANALYSIS OF FUZZY DISTANCES AND ACTIVITY VOLUMES

Alexandros Iosifidis, Ermioni Marami, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

Eating and drinking activity recognition can be considered a solitary research field in activity recognition area. The development of an application capable to identify human eating and drinking activity can be really useful in a smart home environment targeting to extend independent living of older persons in the early stages of dementia. In this paper a novel method aiming at eating and drinking activity recognition is presented. Activities are considered as a sequence of human body poses forming 3D volumes, in which the third dimension refers to time. Fuzzy Vector Quantization is performed to associate the 3D volume representation of an activity video with 3D volume prototypes and Linear Discriminant Analysis is used to map activity representations in a low dimensional discriminant feature space. In this space a simple Nearest Centroid classification procedure leads to very satisfactory classification results.

Index Terms— eating activity and drinking activity recognition, fuzzy vector quantization, discriminant analysis, activity volumes

1. INTRODUCTION

In the human behavior analysis field, activity recognition is considered crucial for various applications such as visual surveillance in security systems, video-content annotation, human-computer interaction etc. The term 'activity' corresponds to a middle-level motion pattern performed by a single person and refers to sequences of movements or states where the only real knowledge required is the statistics of the sequence [1]. Regarding the most popular everyday activities that an activity recognition algorithm is required to recognize, besides of walk, run, jump, bend, sit and wave [2, 3, 4, 5], eating and drinking could also be considered as activities of major interest [6]. However, due to its importance, the recognition of eating and drinking activities could be combined in a human meal intake procedure and considered as a solitary research field.

In order to identify food consumption, Cadavid and Abdel-Mottaleb [7] present an algorithm that automatically detects chewing events in surveillance video of a person. Based on Support Vector Machines (SVMs), on an Active Appearance Model (AAM) and on the distinct periodicity of chewing events, they discriminate chewing and non-chewing facial activities such as talking. Another work that deals with human events detection in a video sequence and can be extended in eating and drinking activity recognition is presented in [8]. A color-based ratio histogram analysis, the Gaussian Mixture Models (GMMs) technique and Markov models are used to automatically detect smoking events.

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.

A different approach to the eating and drinking activity detection problem is the elaboration of data acquired by ambient [9] or body-worn (more invasive) sensors [10]. The methods related to the latter case require various body-worn sensors in order to gather vital information for a person's activities such as arm movements, chewing or swallowing that reveal food or liquid intake activity. Ambient sensors mainly provide information about person's location (kitchen, dining room etc.) and could be used in combination with other, more precise, techniques.

In this paper we propose a novel method specialized for eating and drinking activity recognition using information captured by a camera. According to our knowledge no other method focused on this field has been previously proposed in the literature. A privacy preserving human body representation is obtained by describing the human body poses using binary images denoting the Regions of Interest (ROIs) of the depicted person, i.e. his/her head and hands [11]. Activities are described by a number of successive video frames forming 3D volumes, which will be called activity volumes (AVs) hereafter, in which the third dimension refers to time. Volumetric human body representation is combined with Fuzzy Vector Quantization (FVQ) and Linear Discriminant Analysis (LDA) in order to represent video segments depicting activities in a low-dimensional discriminant subspace in which activity classes are linearly separable. This algorithm can be applied for eating and drinking activity recognition in order to identify nutrition abnormalities of older persons in the early stages of dementia. However, the proposed scheme can be trained to recognize other activities as well.

The rest of the paper is organized as follows: Section 2 outlines the problem statement, Section 3 describes the proposed method and its main phases. Section 4 includes the experiments conducted and the corresponding results. Finally, Section 5 draws the conclusion of this work.

2. PROBLEM STATEMENT

The importance of eating and drinking activity recognition can be appreciated considering its usefulness. A system that automatically recognizes eating and drinking activity, using video processing techniques, would greatly contribute to prolonging independent living of older persons aiming at patients in the early stages of dementia. Dementia is a serious cognitive disorder which affects the sufferer's memory, attention, language, and problem solving abilities [9]. This means that mild dementia (early stages of dementia) can be confronted at home by assisting older persons, primarily, in their daily nutrition needs.

The development of a smart home environment using up-to-date technologies, like monitoring cameras or a robot carrying an IP camera, targets in supporting independent living of older persons as

long as possible in their own homes in a non-invasive way. Specialized applications can identify nutrition abnormalities of the supervised person such as underfeeding and dehydration by observing his/her eating behaviors and notifying the person of interest if it is needed, e.g., by a stimulation to eat if he/she has skipped a meal.

In order to preserve the anonymity not only of the persons participated in training data recordings but final application users as well, privacy preserving human body representations are used. More precisely, binary images (white silhouettes in black background) are used in the proposed method. Moreover, in order to further enhance the privacy preserving nature of the proposed approach no video-storage or video transmission outside the house is required since data are processed in real time.

3. PROPOSED METHOD

In the preprocessing phase, videos capturing a person during a meal intake are manually segmented to smaller ones depicting elementary activities, e.g., an eating sequence, producing the so-called activity videos. In the case of continuous activity recognition, i.e., recognition in videos containing multiple elementary activities, smaller videos are automatically produced by using a sliding window consisting of N_{t_w} video frames, in both training and test phases. Thus, the resulted activity videos are produced by, probably, overlapping video segments. Binary images depicting the image locations that belong to the person's ROIs, i.e., his/her head and hands, in white and the remaining locations in black are obtained by applying a skin color segmentation algorithm at each activity video frame. This is a reasonable approach, as the human body parts that are related with the eating and drinking activity are mainly the person's head and the hands. Snapshots of the binary activity videos, along with the initial captured data, are shown in Figure 1. Binary images preserve the human skin ROIs of persons wearing blouses with long sleeves (head and palms) for each one of the following activities, namely: eating, drinking and apraxia. Although apraxia specifies a disorder, here we use this term to define the class that includes activities other than eating or drinking, like chewing, slicing food etc. The last two raws in Figure 1 present snapshots of a person while slicing food (apraxia class).

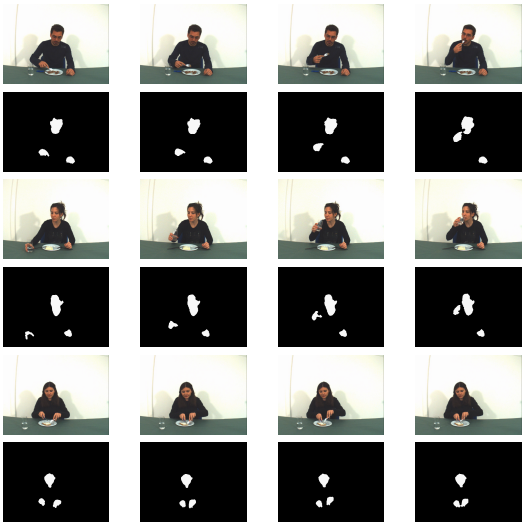

Fig. 1. Successive snapshots extracted from activity videos and the corresponding binary images.

After the extraction of the person's ROIs, each elementary activity video, consisting of N_{t_a} video frames, is converted to a binary activity video consisting of N_{t_a} binary images. Each binary image is centered to the person's center of mass and bounding boxes (BBs) of size equal to the maximum BB that encloses person's ROIs in each activity video are extracted and resized using linear interpolation to produce binary images of predefined size ($L_x \times L_y$ pixels). These binary images are subsequently concatenated to produce 3D AVs of $N_L = L_x \times L_y \times N_{t_a}$ voxels. Since activities vary in duration, the number of binary images N_{t_a} consisting AVs varies for activity videos belonging to different activity classes. Finally, AVs are resized in time in order to produce AVs of fixed size ($N_L = L_x \times L_y \times L_z$ voxels) as illustrated in Figure 2.

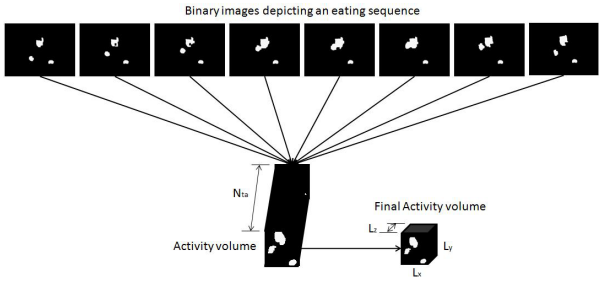

Fig. 2. Activity volume creation.

AVs are vectorized in order to produce activity vectors $\mathbf{p}_i \in \mathcal{R}^{N_L},\ i = 1, ..., N_v$. This is done by vectorizing column-wise each of the L_z binary images consisting the AVs and concatenating the resulted vectors, $\mathbf{p}_i = [\mathbf{p}_{i,1}^T, \mathbf{p}_{i,2}^T, ..., \mathbf{p}_{i,L_z}^T]^T$, where $\mathbf{p}_{i,j},\ j = 1, ..., L_z$ denotes the vector produced by vectorizing the j-th binary image of the i-th AV. To ignore any temporal information concerning the starting frame of the activity videos, e.g., in order to represent two activity videos depicting the activity class 'eat' starting from different human body poses, we exploit the circular invariance property of the Discrete Fourier Transform (DFT):

$$P_i(k) = |\sum_{n=0}^{N_L-1} p_i(n) e^{-i\frac{2\pi k}{N_P}n}|, \quad k = 1, ..., N_L - 1. \qquad (1)$$

That is, each activity video is represented by a vector $\mathbf{P}_i$ which contains the magnitudes of the DFT representation of the corresponding activity vector $\mathbf{p}_i$.

3.1. Training Phase

Let $\mathcal{U}$ be an annotated activity video database, containing N_v training activity videos of N_A activity classes. Each activity video is described by its DFT activity vector representation $\mathbf{P}_i,\ i = 1, ..., N_v$. AV prototypes $\mathbf{v}_d \in \mathcal{R}^{N_L},\ d = 1, ..., N_D$ are calculated using a K-Means clustering algorithm [12] without using the labeling information available in the training phase by minimizing the between cluster sum of squares:

$$\sum_{d=1}^{N_D} \sum_{i=1}^{N_v} \alpha_{id} \|\mathbf{P}_i - \mathbf{v}_d\|^2, \qquad (2)$$

where $\alpha_{ij} = 1$ if $\mathbf{P}_i$ is assigned to the cluster j and zero otherwise. AV prototypes $\mathbf{v}_d,\ d = 1, ..., N_D$ are defined as the cluster centers

(consisting of $n_d = \sum \alpha_{id}$ activity vectors each):

$$\mathbf{v}_d = \frac{1}{n_j} \sum_{i=1}^{N_v} \alpha_{id} \mathbf{P}_i. \quad (3)$$

The optimal number of AV prototypes $N_{D_{opt}}$ is determined by applying the leave-one-person-out (LOPO) cross-validation procedure. This procedure excludes the whole set of patterns belonging to a specific person from the training set and uses this set as validation data. This is done multiple times (folds), one for each person used for validation. That is, in every fold, activity videos depicting all but one persons are used as training samples and activity videos of the remaining person are used for testing.

After the determination of the AV prototypes, fuzzy distances from the DFT activity vector representations $\mathbf{P}_i,\ \ i = 1, ..., N_v$ to all the AV prototypes $\mathbf{v}_d,\ \ d = 1, ..., N_D$ are calculated and the membership vectors $\mathbf{u}_i \in \mathcal{R}^{N_D},\ \ i = 1, ..., N_v$ are obtained:

$$\mathbf{u}_i = \frac{(\parallel \mathbf{P}_i - \mathbf{v}_d \parallel_2)^{-\frac{2}{m-1}}}{\sum_{d=1}^{N_D} (\parallel \mathbf{P}_i - \mathbf{v}_d \parallel_2)^{-\frac{2}{m-1}}}. \quad (4)$$

That is, each AV is mapped to the corresponding membership vector $\mathbf{u}_i$, which represents the activity. Membership vectors denote the similarity of AVs with the AV prototypes. As the number of AVs prototypes N_D is smaller than the dimensionality of the AVs ($N_D \ll N_L$), the dimensionality of the activity representation is reduced. Furthermore, this is a better activity representation, in terms of classification, as it is expected that AVs representing an activity will be quite similar with AV prototypes coming from training AVs representing the same activity and dissimilar with AV prototypes coming from training AVs representing different activities.

Using the known labeling information available in the training phase, a discriminant technique can be exploited in order to discriminate the activity classes. Linear Discriminant Analysis (LDA) is used to map the training membership vectors $\mathbf{u}_i,\ \ i = 1, ..., N_v$ in an optimal discriminant subspace in which the activity classes are linearly separable. The optimal projection matrix $\mathbf{W}_{opt}$ is calculated by:

$$\mathbf{W}_{opt} = \arg\min_{\mathbf{W}} \frac{trace\{\mathbf{W}^T \mathbf{S}_w \mathbf{W}\}}{trace\{\mathbf{W}^T \mathbf{S}_b \mathbf{W}\}}. \quad (5)$$

$\mathbf{S}_w$ and $\mathbf{S}_b$ are the within and between scatter matrices of the training membership vectors respectively:

$$\mathbf{S}_w = \sum_{i=1}^{N_A} \sum_{j=1}^{N_i} \frac{(\mathbf{u}_{ij} - \boldsymbol{\mu}_i)(\mathbf{u}_{ij} - \boldsymbol{\mu}_i)^T}{N_i} \quad (6)$$

$$\mathbf{S}_b = \sum_{i=1}^{N_A} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T}{N_i} \quad (7)$$

where $\mathbf{u}_{ij}$ is the j-th membership vector belonging to activity class i, N_i is the number of membership vectors belonging to activity class i, $\boldsymbol{\mu}_i$ is the mean membership vector of activity class i and $\boldsymbol{\mu}$ is the mean vector of all the training membership vectors.

Discriminant activity vectors, $\mathbf{z}_i \in \mathcal{R}^{N_A - 1}, i = 1, ..., N_v$, are obtained by mapping the membership vectors $\mathbf{u}_i$ to the LDA space by $\mathbf{z}_i = \mathbf{W}_{opt}^T \mathbf{u}_i$. In this space, each activity class is represented by its mean discriminant activity vector $\mathbf{z}_a \in \mathcal{R}^{N_A - 1},\ \ a = 1, ..., N_A$:

$$\mathbf{z}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} \mathbf{z}_{ai}. \quad (8)$$

3.2. Classification Phase

In the classification phase, an activity video is preprocessed using the procedure described in Subsection 3 in order to provide an AV of predefined size ($L_x \times L_y \times L_z$ voxels). This AV is vectorized to produce the activity vector $\mathbf{p}_{test}$. The DFT representation of the activity vector $\mathbf{P}_{test}$ is obtained by applying the DFT transform to $\mathbf{p}_{test}$. Fuzzy distances from the activity vector representing the test activity video $\mathbf{P}_{test}$ to all the AV prototypes $\mathbf{v}_d,\ \ d = 1, ..., N_D$ are calculated and the membership vector $\mathbf{u}_{test} \in \mathcal{R}^{N_D}$ is obtained. The discriminant activity vector $\mathbf{z}_{test} \in \mathcal{R}^{N_A - 1}$ is obtained by mapping $\mathbf{u}_{test}$ to the LDA space. In that space, the activity video is classified to the nearest class center using the Euclidean distance:

$$d(\mathbf{z}_{test}, \mathbf{z}_{a_i}) = \|\mathbf{z}_{test} - \mathbf{z}_{a_i}\|_2. \quad (9)$$

The procedure applied in the classification phase is illustrated in Figure 3.

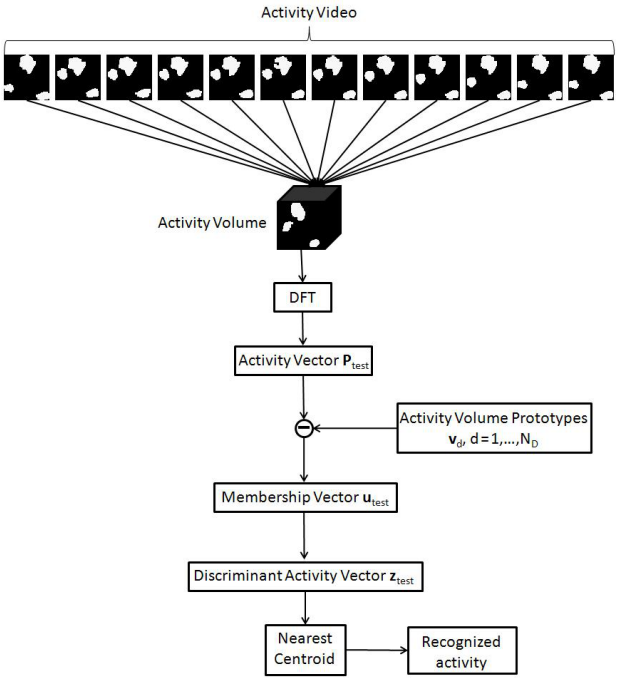

Fig. 3. Activity video classification.

4. EXPERIMENTAL RESULTS

Since there is no publicly available database for eating and drinking activity recognition, in order to assess the performance of the proposed method we created such a specialized data collection. Four persons, wearing clothes with long sleeves, were captured during a meal intake, considering various types of food (sandwich, cereals etc.) and liquids, in order to capture variations of the performed activities. The camera was placed at a distance of 2 meters in front of the participants and recording sessions were performed during three different days. In the preprocessing phase, these videos were manually annotated and segmented and 333 activity videos depicting activity classes 'eat', 'drink' and 'apraxia' were obtained. A color-based image segmentation algorithm was applied to the resulted activity videos in order to obtain binary images illustrating the head and the

hands of the depicted person in white and the remaining locations in black. Specifically, images were converted to HSV color-space and image locations with hue, saturation and value values between pre-defined thresholds were denoted as ROIs. These binary images were further preprocessed as described in Subsection 3 in order to produce AVs consisting of $(64 \times 64 \times 10)$ or $(64 \times 64 \times 20)$ voxels.

Multiple experiments were conducted in order to evaluate the performance of the proposed method to correctly recognize activities performed by different persons. The LOPO cross-validation procedure was performed for different number of AV prototypes. In each fold, the activity videos of three persons of the database were used to train the algorithm, while the activity videos of the remaining person were used for evaluation. Four folds, one for each test person, were performed in order to complete an experiment. The mean classification accuracy rate was calculated for each experiment and the number of AV prototypes corresponding to the experiment providing the best mean classification rate denoted the optimal number of AV prototypes.

The classification accuracy rates achieved in these experiments for different numbers of AV prototypes are illustrated in Figure 4. An overall correct classification rate of 93.3% evaluated by averaging all the per class correct classification rates was attained for AVs consisting of $(64 \times 64 \times 10)$ voxels and 24 AV prototypes. The corresponding confusion matrix can be seen in Table 1. In this table a row represents the actual label of each testing activity video, while a column represents the activity class label provided by the algorithm. The expected variations in style between different persons resulted in small classification errors. However, activity videos are almost perfectly classified. Furthermore, it can be seen that classification rates for different AV sizes are similar. As smaller size of AVs decreases the computational cost, the use of small L_z is suggested.

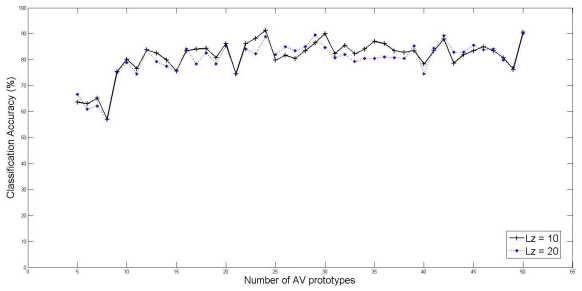

Fig. 4. Activity accuracy rates versus the number of AV prototypes.

Table 1. Confusion matrix containing classification accuracy rates in the eating and drinking activity recognition task performing the LOPO cross-validation procedure.

	drink	eat	apraxia
drink	1		
eat	0.05	0.87	0.08
apraxia	0.01	0.06	0.93

5. CONCLUSION

In this paper we presented a novel method that addresses the eating and drinking activity recognition task. A camera placed in front of a person during a meal is used to gather the necessary information. A color segmentation algorithm is used in order to provide a privacy preserving human body representation. Volumetric representation of activities is achieved by concatenating successive binary images representing human body poses. The circular invariance property of the magnitudes of the DFT transform is exploited to provide time invariant activity video representation. FVQ and LDA are involved in the classification process, whilst the use of an activity representation in a low dimensional discriminant feature space reduces the computational cost and leads to high correct classification rates.

6. REFERENCES

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

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

[3] 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*, 2011.

[4] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks*, 2012.

[5] A. Iosifidis, A. Tefas, and I. Pitas, "Person specific activity recognition using fuzzy learning and discriminant analysis," *Euripean Signal Processing Conference (EUSIPCO)*, 2011.

[6] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *IEEE Conference on Computer Vision and Pattern Recognition 2009, CVPR 2009*, pp. 2929–2936, 2009.

[7] S. Cadavid and M. Abdel-Mottaleb, "Exploiting visual quasi-periodicity for automated chewing event detection using active appearance models and support vector machines," *20th International Conference on Pattern Recognition (ICPR) 2010*, pp. 1714–1717, 2010.

[8] Pin Wu, Jun-Wei Hsieh, Jiun-Cheng Cheng, Shyi-Chyi Cheng, and Shau-Yin Tseng, "Human smoking event detection using visual interaction clues," *20th International Conference on Pattern Recognition (ICPR) 2010*, pp. 4344–4347, 2010.

[9] K. Sim, Ghim-Eng Yap, C. Phua, J. Biswas, Aung Aung Phyo Wai, A. Tolstikov, Weimin Huang, and P. Yap, "Improving the accuracy of erroneous-plan recognition system for activities of daily living," *12th IEEE International Conference on e-Health Networking Applications and Services (Healthcom) 2010*, pp. 28–35, 2010.

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

[11] 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.

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