Object Tracking Based on Local Steering Kernels for Drinking Activity Recognition

Olga Zoidi, Anastasios Tefas, Ioannis Pitas
Department of Informatics, Aristotle University of Thessaloniki
Box 451, 540 06 Thessaloniki, Greece
E-mail: {ozoidi,tefas,pitas}@aiia.csd.auth.gr

Abstract. *A novel method for object tracking in videos for drinking activity recognition is proposed. The query object is detected in the first video frame, extracting a new query image. The obtained query image is then compared with patches within a determined region of interest around the position of the detected object in the previous frame. For each image, the local steering kernels are extracted and the similarity between the query image and the patches of the video frame is measured by calculating the matrix cosine similarity. The proposed method finds application in drinking activity recognition, by tracking the object, i.e., the glass, being used.*

Keywords. object tracking, drinking activity recognition, local steering kernels.

1. Introduction

Activity recognition has received great attention over the years, as it finds numerous everyday life applications, such as video surveillance, human - computer interaction (HCI) systems, computer games, etc. In general, activities can be described by a set of "verbs" which characterize the actions, i.e., the sequence of movements, performed by the human, and a set of "nouns" which determine the objects that take part in the actions [1]. The majority of research in activity recognition focuses on identifying the "verbs" which characterize an activity [2][3], and only a few of them target the problem of recognizing the objects which take part in them [1].

Apart from the recognition of the most common human activities like walking, running, jumping, bending, sitting and waving, eating and drinking activity recognition consist a research area with a major application field, including monitoring of patients with eating disorders. The implemented eating and drinking activity recognition algorithms either use data obtained from ambient or body-worn sensors, or visual information obtained from one or more cameras.

A method for automatic recognition of arm gestures related to eating and drinking activities based on data acquired by sensors attached on the wrist and upper arm is proposed in [4], while in [5], a two-modal eating and drinking activity detection system is introduced, which employs information from both ambient (a Radio-Frequency Identification (RFID) reader) and body-worn (a wrist accelerometer) sensors. The combination of the two modalities was succeeded with a Dynamic Baysean Network (DBN).

In [6], information obtained from a surveillance camera is exploited in order to detect chewing events and discriminate them from non-chewing events, such as talking. The method is based on Active Appearance Models (AAM) in order to detect the periodicity of chewing events, and on Support Vector Machines (SVM) in order to take the classification decision. Smoking events, which are activities very similar to eating and drinking, are detected in [7]. Moreover, visual information from a video is exploited through color-based ratio histogram analysis and Gaussian Mixture Models (GMM), while Markov models are employed for the classification decision.

In this paper we present a novel method for object tracking which finds application in drinking activity recognition. More precisely, the drinking activity is detected by observing the trajectory of the object, i.e., the glass, which takes part in the activity.

The rest of the paper is organized as follows: Section 2 outlines the problem statement. Section 3 provides a detailed description of the proposed object tracking method. Section 4 presents a description of the database used in experiments and experimental results. Finally, Section 5 draws the conclusion of this work.

2. Problem statement

Dementia is a syndrome more frequent to the elderly population which causes a serious loss of the sufferer's cognitive abilities. Patients with early stage of dementia have a high risk of dehydration, as they experience symptoms of deterioration of the nerves, loss of sense of smell, apraxia (loss of the ability to execute or carry out learned purposeful movements), agnosia (loss of ability to recognize objects, persons, sounds, shapes, or smells), etc. Therefore, the development of a central monitoring system which detects and measures the duration of drinking activity can prevent the patient's dehydration by analyzing the patient's drinking behavior and, if necessary, reminding him to drink. In order to cause minimum disturbance to the patient, the system should detect drinking activity using only visual data obtained by a set of surveillance cameras, without the use of any markers on the cup or the patient's face. When the patient spends a long time without drinking, a robotic unit stimulates his feeling of thirst, e.g., by asking wether he wants something to drink.

3. Object tracking

The proposed method embodies the object detection method based on locally adaptive regression kernels first introduced in [8] in an algorithm which performs object tracking in a video sequence, by comparing the denoted object in the previous video frame (query image) with equally sized patches of the following frame in a region of interest (ROI) around the predicted position of the object, which is based on its position in the previous two frames (target frame ROI). The proposed algorithm starts with the initialization of the object position at the first video frame. The initialization can be achieved either by using an object detection algorithm, such as the one proposed in [8], or by manually inserting the coordinates of the object in the first frame. Afterwards, the algorithm executes three iterative steps. In the first step, the query image is extracted from the previous frame, the new object position is predicted and the new target ROI is determined. Then, the salient features of the query image and the target frame ROI are extracted. Finally, the frame patch with the greater similarity to the query image is detected and its coordinates are extracted.

Step1: Prediction of object position and target frame ROI initialization

In this step, the detected object in the previous frame is exported and saved as the new query image. Then, the position $\mathbf{p}_t = [p_x, p_y]^T$ of the object in the current frame t is predicted according to the following equation:

$$\mathbf{p}_t = 2\mathbf{p}_{t-1} - \mathbf{p}_{t-2}, \qquad (1)$$

where $\mathbf{p}_{t-1}$, $\mathbf{p}_{t-2}$ are the object coordinates in frames $t-1$ and $t-2$ respectively. The target ROI in the current frame is then defined around the predicted object coordinates $\mathbf{p}_t$. The ROI size is determined to be equal with the size of the detected object plus a margin of m pixels. In our experiments, we set $m = 10$ pixels. The value of m depends on the maximum velocity of the object and it should be large enough to keep track on the object in the selected ROI.

Step 2: Feature extraction

Initially, the query image and target frame ROI are transformed from the RGB colorspace to the La*b* colorspace and, subsequently, they are divided into their three channels. In order to get a good representation of both the query and the target image channels, we extract their salient features, by comparing how similar each pixel is with its surrounding pixels in a locally defined $P \times P$ window. The salient features are computed using the so-called local steering kernel descriptors (LSK) [8]. These descriptors take into account both the illumination (pixel value) difference and the distance between pixels. Their general form is:

$$K(\mathbf{x}_l - \mathbf{x}) = \frac{K(\mathbf{H}_l^{-1}(\mathbf{x}_l - \mathbf{x}))}{\det(\mathbf{H}_l)}, \quad l = 1, ..., P^2, \qquad (2)$$

where $\mathbf{x}$ are the coordinates of the image pixel, $\mathbf{x}_l$ are the coordinates of the neighboring pixels, and $\mathbf{H}_l$ is the 2×2 steering matrix:

$$\mathbf{H}_l = h\mathbf{C}_l^{-1/2}, \qquad (3)$$

where h is a global smoothing parameter and $\mathbf{C}_l$ is a covariance matrix, estimated from the gradient vectors of the pixels in a region around $\mathbf{x}_l$. In [8] the chosen function for the kernel $K(\cdot)$ is the Gaussian function:

$$K(\mathbf{x}_l - \mathbf{x}) = \frac{\sqrt{\det(\mathbf{C}_l)}}{h^2} \cdot \\ \cdot \exp\left\{-\frac{(\mathbf{x}_l - \mathbf{x})^T \mathbf{C}_l (\mathbf{x}_l - \mathbf{x})}{2h^2}\right\}. \quad (4)$$

This kernel function is chosen because, when normalized, it is invariant to brightness changes and robust to contrast changes. In order to compute the covariance matrix , first we form the $P^2 \times 2$ matrix $\mathbf{J}_l$:

$$\mathbf{J}_l = \left[\begin{array}{cc} z_{x_1}(x_1) & z_{x_2}(x_1) \\ \vdots & \vdots \\ z_{x_1}(x_{P^2}) & z_{x_2}(x_{P^2}) \end{array}\right], \quad (5)$$

with rows the transpose gradient vectors along the axes x_1 and x_2 of the pixels in the $P \times P$ window around the pixel $\mathbf{x}_l$, and then apply SVD:

$$\mathbf{J}_l = \mathbf{U}_l \cdot \left[\begin{array}{cc} s_1 & 0 \\ 0 & s_2 \end{array}\right] \cdot \left[\begin{array}{c} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{array}\right]_l . \quad (6)$$

The covariance matrix $\mathbf{C}_l$ is then calculated as in [9]:

$$\mathbf{C}_l = \gamma \sum_{q=1}^{2} a_q^2 \mathbf{v}_q \mathbf{v}_q^T, \quad (7)$$

where

$$a_1 = \frac{s_1 + 1}{s_2 + 1}, \quad a_2 = \frac{s_2 + 1}{s_1 + 1}, \\ \gamma = \left(\frac{s_1 s_2 + 10^{-7}}{P^2}\right)^a . \quad (8)$$

In equations (8), a is a parameter that restricts γ and in our experiments takes the value 0.008.

The local steering kernel descriptors for the query image and the target frame ROI channels at a pixel indexed by j are produced by normalizing equation (4) in the $P \times P$ window around $\mathbf{x}$:

$$\mathbf{W}_Q(\mathbf{x}_l - \mathbf{x}) = \frac{K_Q^j(\mathbf{x}_l - \mathbf{x})}{\sum_{l=1}^{P^2} K_Q^j(\mathbf{x}_l - \mathbf{x})}, \\ j = 1, ..., n \quad (9)$$

$$\mathbf{W}_T(\mathbf{x}_l - \mathbf{x}) = \frac{K_T^j(\mathbf{x}_l - \mathbf{x})}{\sum_{l=1}^{P^2} K_T^j(\mathbf{x}_l - \mathbf{x})}, \\ j = 1, ..., n_T \quad (10)$$

where n and n_T are the number of pixels of the query and target image channels respectively. Before we continue to the next step, we column-stack matrices $\mathbf{W}_Q^j$ and $\mathbf{W}_T^j$ to vectors $\mathbf{w}_Q^j$ and $\mathbf{w}_T^j$, and produce new matrices $\mathbf{W}_Q \in \mathbb{R}^{P^2 \times n}$ and $\mathbf{W}_T \in \mathbb{R}^{P^2 \times n_T}$:

$$\mathbf{W}_Q = [\mathbf{w}_Q^1, ..., \mathbf{w}_Q^n] \\ \mathbf{W}_T = [\mathbf{w}_T^1, ..., \mathbf{w}_T^{n_T}]. \quad (11)$$

Finally, we apply PCA to $\mathbf{W}_Q$, producing the projection matrix $\mathbf{A}_Q \in \mathbb{R}^{d \times n}$, and use $\mathbf{A}_Q$ to compute the salient feature matrices $\mathbf{F}_Q \in \mathbb{R}^{d \times n}$ and $\mathbf{F}_T \in \mathbb{R}^{d \times n_T}$. Therefore, the salient features that best describe each image channel are extracted from $\mathbf{W}_Q$ and $\mathbf{W}_T$ by keeping their d principal components that contain at least 80% of the information:

$$\mathbf{F}_Q = [\mathbf{f}_Q^1, ..., \mathbf{f}_Q^n] = \mathbf{A}_Q^T \mathbf{W}_Q \\ \mathbf{F}_T = [\mathbf{f}_T^1, ..., \mathbf{f}_T^{n_T}] = \mathbf{A}_Q^T \mathbf{W}_T. \quad (12)$$

Step 3: Similarity measure

As we mentioned earlier, this method detects an object by examining the similarity between the query image channels $\mathbf{F}_Q$ and the equally sized patches $\mathbf{F}_{T_l}$ of the target frame ROI channels. The resemblance between $\mathbf{F}_Q$ and $\mathbf{F}_{T_l}$ is found by applying a generalization of the cosine similarity measure for matrices, called Matrix Cosine Similarity (MCS) [8], given by:

$$\rho_i = \rho(\mathbf{F}_Q, \mathbf{F}_{T_l}) = \frac{\text{trace}\left(\mathbf{F}_G^T \mathbf{F}_{T_l}\right)}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_l}\|_F} = \\ = \sum_{l=1,j=1}^{n,d} \frac{f_Q^{(l,j)} f_{T_l}^{(l,j)}}{\sqrt{\sum_{l=1,j=1}^{n,d} |f_Q^{(l,j)}|^2 \sum_{l=1,j=1}^{n,d} |f_{T_l}^{(l,j)}|^2}}$$

where $\| \cdot \|_F$ is the Frobenius norm and $f_Q^{(l,j)}$, $f_{T_l}^{(l,j)}$ are the (l,j) elements of $\mathbf{F}_Q$ and $\mathbf{F}_{T_l}$ respectively. The MCS values ρ_i of each channel are mapped to $\mathbb{R}^+$ space through the transform:

$$f(\rho_i) = \frac{\rho_i^2}{1 - \rho_i^2} \in [0, \infty) \quad (13)$$

and grouped in a resemblance map (**RM**). Experiments showed that $f(\rho_i)$ generates a resemblance map with better contrast and dynamic range than ρ_i [8].

Before we continue to the next step, we add the **RM**s of each channel to produce a total **RM**, that contains the complete similarity information between the query and target images:

$$\mathbf{RM}_{tot} = \sum_{q=1}^{3} \mathbf{RM}_q. \tag{14}$$

The maximum value of $\mathbf{RM}_{tot}$ is selected and compared with a threshold τ. If the maximum value is greater than the chosen threshold, then the current position of the object is found and the new coordinates are extracted. Otherwise, the query object isn't detected in the chosen ROI and the algorithm stops.

4. Experiments

4.1 Database description

For the experiments, we used the AIIA Eating and Drinking Activity Recognition Database (EDAR - AIIA) created by the AIIA laboratory. It consists of 12 multi-view video sessions. Four still Cameras in different positions captured four views of the eating scene respectively: frontal, upper-frontal, left and upper-right. The hardware used consisted of four Color Sony XCD-V60CR Digital Cameras and a PC containing an Octal Core Intel E5420 CPU of 2.5 GHz and RAM of 3.25 GB. The videos depict eating and drinking activity of 4 different persons with one to four recordings for each of them. This database includes males and females eating with spoon (eating cereals) and drinking from a glass (drinking water).

4.2 Experimental results

In this section we demonstrate the performance of our object tracking method using a video which depicts drinking activity during a meal session. The duration of the video is 1 minute and 20 seconds and the total number of frames is 1200. During the capture and in between bites, the subject takes three sips from the cup performing three corresponding drinking activities. Drinking activity can be split into two actions: drink-up, which is the action of raising the glass to the mouth, and drink-down, which is the action of lowering the glass onto the table.

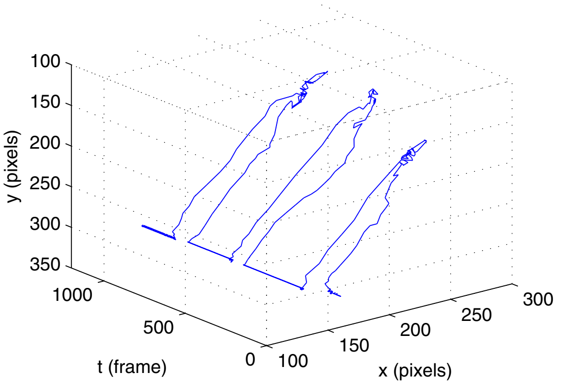
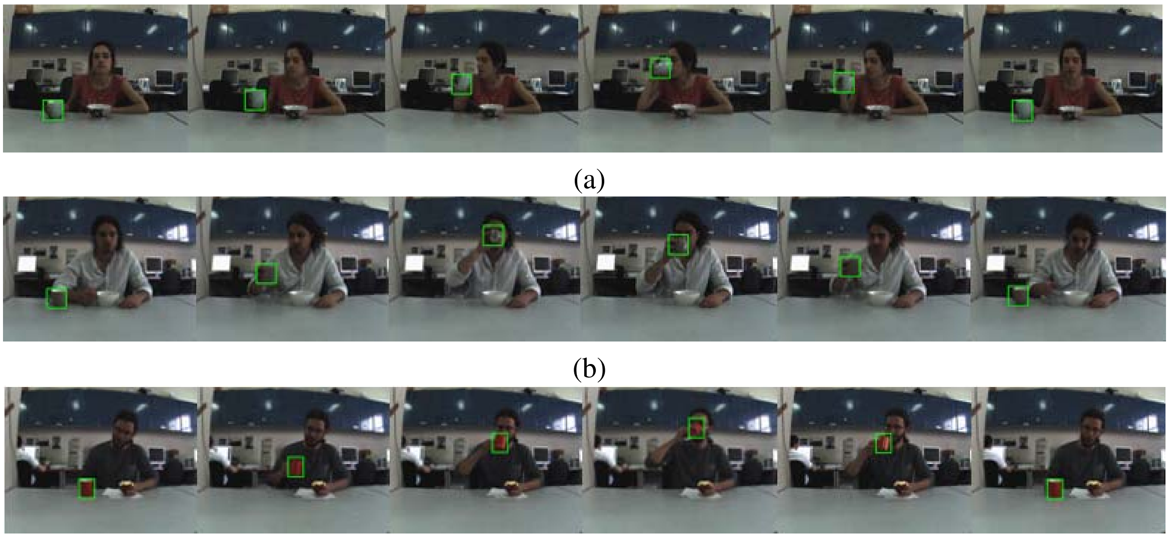


Figure 2: The 3-dimensional trajectory of the glass during the meal session of the video.

The main object which takes part in a drinking activity is the glass. Therefore, we apply our proposed tracking method in order to extract the trajectory of the glass. When the algorithm detects the position of the object, it draws a green rectangle perimetric its area. Snapshots of the algorithm performance during the first sip are demonstrated in Fig. 1a.

Fig. 2 depicts the trajectory of the tracked glass in the 3 dimensional space. The denoted x and y axes represent the coordinates of the glass in the video, therefore they are measured in pixels. Because in image coordinates the y axis increases downwards, in Fig. 2 it is reversed, in order to correspond better to the intuitive notion of the glass rising up during drinking activity. The denoted t axis represents time and it is measured in video frames. Given that the video has 1200 frames in total, time takes values from 0 to 1999. The first sipping takes place between frames 100 and 190, the second between frames 550 and 600, and the third between frames 900 and 1000. We notice that the three drinking activities have different durations. Knowing the camera frame rate, which in our case is 15 fps, we can estimate the duration of each sipping by dividing the number of the frames with the frame rate. Therefore, the duration of the three drinking activities is estimated to be approximately 6 sec, 3.33 sec, and 6.66 sec, respectively.

The projection of the 3-dimensional trajectory of Fig. 2 to the 2-dimensional space generates the trajectories of Fig. 3 and Fig. 4. From Fig. 2-4



(a)

(b)

(c)

Figure 1: Six snapshots of the object detection algorithm performance during the first drinking activity of three video sequences in the EDAR-AIIA database.

we notice that in every sipping the glass trajectory follows a certain pattern. More precisely, in drink-up activity the glass' x-coordinate increases (given that the glass lies on the right side of the subject) and its y-coordinate decreases, while in drink-down activity the x and y coordinates return close to their initial value. On the other hand, when the subject isn't drinking, the x and y coordinates have a constant value. It can be easily derived that in the case of a left handed person the trajectory of the x-coordinate in Fig. 3 is flipped horizontal. Therefore, the activity recognition algorithm should be trained with both right-handed and left-handed persons. Finally, the object trajectory can be normalized with respect to the initial position of the object in the first frame in order to avoid misalignment problems.

Application of the proposed method in the EDAR-AIIA database showed encouraging results. Instances of the algorithm performance in two other video sequences of the database are depicted in Fig. 1b,c.

5. Conclusion

In this paper we presented a novel method for object tracking based on local steering kernels, which finds application in drinking activity recognition. Experimental results showed the

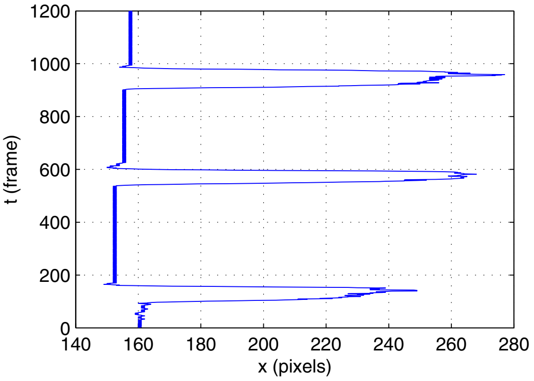


Figure 3: Graph of the glass x-coordinate with respect to time t.

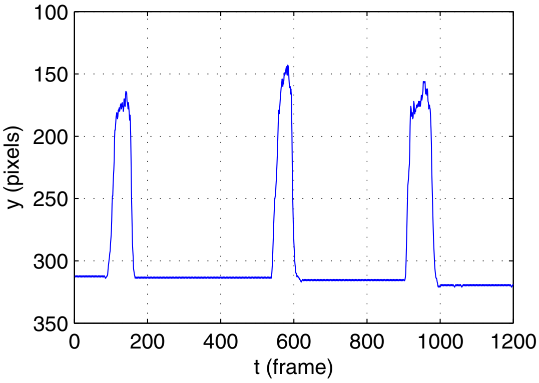


Figure 4: Graph of the glass y-coordinate with respect to time t.

effectiveness of the method in tracking and furthermore analyzing the trajectory of the object which takes part in the activity and extracting significant information about the motion patterns of the activity. The idea of recognizing activities by tracking the objects that are associated with them can be extended in other activities besides drinking, such as eating, where the main utensils which take part in the activity are the spoon, the fork, and/or the knife.

6. Acknowledgement

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] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, J.M. Rehg. "A Scalable Approach to Activity Recognition based on Object Use", IEEE 11th International Conference on Computer Vision (ICCV) 2007. pp.1-8, 14-21 Oct. 2007

[2] J. Parkka, L. Cluitmans, M. Ermes. "Personalization Algorithm for Real-Time Activity Recognition Using PDA, Wireless Motion Bands, and Binary Decision Tree," IEEE Transactions on Information Technology in Biomedicine, vol.14, no.5, pp.1211-1215, Sept. 2010

[3] M. Singh, A. Basu, M.K. Mandal. "Human Activity Recognition Based on Silhouette Directionality," IEEE Transactions on Circuits and Systems for Video Technology, vol.18, no.9, pp.1280-1292, Sept. 2008

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

[5] A. Tolstikov, J. Biswas, Chen-Khong Tham, P. Yap. "Eating Activity Primitives Detection - a Step Towards ADL Recognition", e-health Networking, Applications and Services 2008, 10th International Conference on HealthCom 2008, pages 35-41, 2008

[6] S. Cadavid, 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, pages 1714-1717, 2010

[7] Pin Wu, Jun-Wei Hsieh, Jiun-Cheng Cheng, Shyi-Chyi Cheng, Shau-Yin Tseng. "Human Smoking Event Detection Using Visual Interaction Clues", 20th International Conference on Pattern Recognition (ICPR) 2010, pages 4344-4347, 2010

[8] H.J. Seo, P. Milanfar. "Training-Free, Generic Object Detection Using Locally Adaptive Regression Kernels," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, no.9, pp.1688-1704, Sept. 2010

[9] H.J. Seo, P. Milanfar. "Static and Spacetime Visual Saliency Detection by Self-Resemblance ", Journal of Vision (2009) 9(12):15, 1-27