VISUAL OBJECT TRACKING BASED ON THE OBJECT'S SALIENT FEATURES WITH APPLICATION IN AUTOMATIC NUTRITION ASSISTANCE

O. Zoidi, A. Tefas and I. Pitas

Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 540 06, GREECE
{ozoidi,tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

A novel method for object tracking in videos which can find application in eating and drinking activity recognition is proposed. The query object is detected in the first video frame, extracting a new query image. The initial query image along with the obtained query image are then compared with patches within a determined search region around the position of the detected object in the previous frame. For each image, the local steering kernels are extracted and the similarity between a query image and the patches of the video frame is measured by calculating the cosine similarity. The proposed method finds application in eating and drinking activity recognition.

Index Terms— visual object tracking, eating activity recognition, drinking activity recognition, local steering kernels

1. INTRODUCTION

Visual object tracking is the challenging task of tracking the trajectory of a moving object in a video. Ideal tracking algorithms should track successfully any object under any conditions in real time. However, in practice this is not possible as numerous environmental parameters affect the tracking performance. Such parameters include lighting conditions variations, partial/total/self occlusion of the object, rapid and complicated object movements, noise, etc. The existing tracking algorithms are categorized into model-based methods [1], which employ a 3-dimensional model of the object, appearance-based methods [2], which use texture information, contour-based methods [3], which track the object contour, feature-based methods [4], which employ a feature-based representation of the object, and hybrid methods [5], which comprise combinations of the above. Due to their simplicity and low computational cost, appearance-based tracking methods, like the one introduced in this paper, are the most widely used.

Visual object tracking finds numerous applications in human computer interaction systems, surveillance systems, e-health, etc. The method introduced in this paper performs tracking of rigid objects which perform smooth movements under small pose changes, 2-dimensional rotations, and small scale changes. The tracking performance will be tested in videos depicting eating and drinking activities. Ulterior motive is to use the objects' trajectories in an activity recognition framework.

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 [6]. The majority of research in activity recognition focuses on identifying the "verbs" which characterize an activity [7], and only a few of them target the problem of recognizing the objects which take part in them [6]. 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. In this paper we present a novel method for object tracking which finds application in eating and drinking activity recognition.

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 the 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 eating and drinking activity can prevent the patient's dehydration by analyzing the patient's eating and drinking behavior and, if necessary, reminding him to eat or drink. In order to cause minimum disturbance to the patient, the system should detect eating and 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 hands and face. When the patient spends a long time without eating and drinking, a robotic unit stimulates his feeling of hunger and thirst, e.g., by asking whether he wants something to eat or 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 object of interest in the first video frame (initial query image $\mathbf{I} \in \Re^{N_x \times N_y}$) and the denoted object in the previous video frame (query image $\mathbf{Q} \in \Re^{N_x \times N_y}$) with equally sized patches of the following frame in a search region $\mathbf{T} \in \Re^{M_x \times M_y}$ around the predicted position of the object, which is based on its position in the previous two frames. 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 search region is determined. Then, the salient features of the initial query image, the query image, and the search region are extracted. Finally, the similarities of the search region patches with the initial query image and the query image are computed and used for determining the new object position.

3.1. Object position prediction and search region 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:

$$\hat{\mathbf{p}}_t = \mathbf{p}_{t-1} + \mathbf{v}_{t-1}, \tag{1}$$

where $\mathbf{v}_{t-1} = \mathbf{p}_{t-1} - \mathbf{p}_{t-2}$ is the object velocity in frame $t-1$ and $\mathbf{p}_{t-1}$, $\mathbf{p}_{t-2}$ are the object coordinates in frames $t-1$ and $t-2$, respectively. The search region in frame t is then defined around the position $\hat{\mathbf{p}}_t$. The search region size is determined to be equal with the size of the detected object plus a margin of m pixels. In our experiments, we set $m = 15$ 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 search region.

3.2. Local Steering Kernel feature extraction

The salient feature of the initial query image, the query image, and the search region are extracted according to the following procedure. Initially, the image is transformed to the La^*b^* color space and, subsequently, it is split to its three channels. For each color channel, the local steering kernel descriptors (LSK) [8] are extracted in a locally defined $P \times P$ window. LSKs take into account both the illumination (pixel value) difference and the distance between neighboring pixels:

$$K(\mathbf{p}_l - \mathbf{p}) = \frac{\sqrt{\det(\mathbf{C}_l)}}{2\pi} \cdot \exp\left\{ -\frac{(\mathbf{p}_l - \mathbf{p})^T \mathbf{C}_l (\mathbf{p}_l - \mathbf{p})}{2} \right\}, \quad l = 1, \ldots, P^2, \tag{2}$$

where $\mathbf{p}$ are the coordinates of the image pixel, $\mathbf{p}_l$ are the coordinates of the neighboring pixels, and $\mathbf{C}_l$ is a covariance matrix, estimated from the matrix $\mathbf{J}_l$ of the gradient vectors of the image in a $P \times P$ window around $\mathbf{p}_l$:

$$\mathbf{J}_l = \begin{bmatrix} z_x(\mathbf{p}_1) & z_y(\mathbf{p}_1) \\ \vdots & \vdots \\ z_x(\mathbf{p}_{P^2}) & z_y(\mathbf{p}_{P^2}) \end{bmatrix}, \tag{3}$$

where $\mathbf{z}(\mathbf{p}) = [z_x(\mathbf{p}), z_y(\mathbf{p})]^T$ is the image gradient vector along x and y axes at the position $\mathbf{p}$, by applying SVD according to equations (4)-(6) [9]:

$$\mathbf{J}_l = \mathbf{U}_l \cdot \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix}_l. \tag{4}$$

$$\mathbf{C}_l = \gamma \sum_{q=1}^{2} a_q^2 \mathbf{v}_q \mathbf{v}_q^T, \tag{5}$$

$$a_1 = \frac{s_1 + 1}{s_2 + 1}, \quad a_2 = \frac{s_2 + 1}{s_1 + 1}, \quad \gamma = \left(\frac{s_1 s_2 + 10^{-7}}{P^2} \right)^a. \tag{6}$$

In equations (6), a is a parameter that restricts γ and in our experiments takes the value 0.008.

For an image pixel $\mathbf{p}$, equation (2) is computed for each neighboring pixel $\mathbf{p}_l$, $l = 1, \ldots, P^2$, meaning that for each image pixel we export an LSK feature vector $\mathbf{K}(\mathbf{p}) \in \Re^{P^2 \times 1}$. The produced LSK feature vector becomes invariant to brightness and contrast changes by using normalization according to:

$$\mathbf{N}(\mathbf{p}) = \frac{\mathbf{K}(\mathbf{p})}{\sum_{l=1}^{P^2} |K(\mathbf{p}_l - \mathbf{p})|} \in \Re^{P^2 \times 1}. \tag{7}$$

The LSK feature vectors of the $n = N_x N_y$ pixels of the query image are ordered column-wise to form the LSK feature matrix $\mathbf{N}_Q \in \Re^{P^2 \times n}$. The LSK feature matrices $\mathbf{N}_I \in \Re^{P^2 \times n}$ and $\mathbf{N}_T \in \Re^{P^2 \times n_T}$, $n_T = M_x M_y$, for the initial query image and the search region, respectively, are formed accordingly.

3.3. Similarity measure and decision extraction

After extracting the LSK feature matrices $\mathbf{N}_Q$, $\mathbf{N}_I$, for the query image and the initial query image respectively, we measure the similarity of the search region patches to the query image and the initial query image. At first, we proceed to dimensionality reduction by PCA keeping 80% of the image information, producing matrices $\mathbf{F}_Q \in \Re^{d\times n}$, $\mathbf{F}_I \in \Re^{d\times n}$:

$$\mathbf{F}_Q = \mathbf{A}_Q\mathbf{N}_Q, \qquad \mathbf{F}_I = \mathbf{A}_I\mathbf{N}_I, \tag{8}$$

We compute two projection matrices, one for the query image $\mathbf{A}_Q \in \Re^{d\times P^2}$ and one for the initial query image $\mathbf{A}_I \in \Re^{d\times P^2}$. The LSK feature matrix $\mathbf{N}_T$ of the search region is then projected to the spaces created by the two projection matrices as follows:

$$\mathbf{F}_{T_Q} = \mathbf{A}_Q\mathbf{N}_T \in \Re^{d\times n_T}, \qquad \mathbf{F}_{T_I} = \mathbf{A}_I\mathbf{N}_T \in \Re^{d\times n_T}, \tag{9}$$

The search region of size $M_x \times M_y$ is then divided into patches $\mathbf{T}_{ij}$, $i = 1,\ldots,m_x = M_x - N_x + 1$, $j = 1,\ldots,m_y = M_y - N_y + 1$, of size $N_x \times N_y$. For each patch $\mathbf{T}_{ij}$ the corresponding LSK feature matrices $\mathbf{F}_{T_{Qij}} \in \Re^{d\times n}$, $\mathbf{F}_{T_{Iij}} \in \Re^{d\times n}$ contain only the columns of $\mathbf{F}_{T_Q}$, $\mathbf{F}_{T_I}$ which correspond to the pixels of the patch $\mathbf{T}_{ij}$. The similarity of the search region patches to the query image and the initial query image is then computed by the cosine similarity:

$$s_{Qij} = s(\mathbf{F}_Q, \mathbf{F}_{T_{Qij}}), \qquad s_{Iij} = s(\mathbf{F}_I, \mathbf{F}_{T_{Iij}}), \tag{10}$$

where

$$s(\mathbf{F}_1,\mathbf{F}_2) = \sum_{l=1,j=1}^{n,d} \frac{F_1(l,j)F_2(l,j)}{\sqrt{\sum_{l=1,j=1}^{n,d}|F_1(l,j)|^2 \sum_{l=1,j=1}^{n,d}|F_2(l,j)|^2}}, \tag{11}$$

and $F_1(l,j)$, $F_2(l,j)$ are the (l,j) elements of matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ respectively. The cosine similarity values s_{Qij}, s_{Iij} of each color channel are grouped in the resemblance maps $\mathbf{R}_{Qc}, \mathbf{R}_{Ic} \in \Re^{m_x\times m_y}$, $c = 1,2,3$. Before we continue to the next step, we add the resemblance maps of each color channel to produce the total resemblance maps $\mathbf{R}_Q, \mathbf{R}_I \in \Re^{m_x\times m_y}$:

$$\mathbf{R}_Q = \sum_{c=1}^{3}\mathbf{R}_{Qc}, \qquad \mathbf{R}_I = \sum_{c=1}^{3}\mathbf{R}_{Ic}. \tag{12}$$

Finally, the average resemblance map $\mathbf{R} \in \Re^{m_x\times m_y}$ is computed:

$$\mathbf{R} = \frac{1}{2}\left(\mathbf{R}_Q + \mathbf{R}_I\right). \tag{13}$$

The final decision for the new object position in frame t is taken by:

$$\mathbf{p}_t = argmax_{i,j}\left\{\mathbf{R} + \max_{i,j}\{\mathbf{R}(ij)\}\cdot\mathbf{W}\right\}, \tag{14}$$

where $\mathbf{W} \in \Re^{m_x\times m_y}$ is a matrix of weights which correspond to the probability of the image patch to be the object according to the prediction method of subsection 3.1. The elements of $\mathbf{W}$ follow the normal distribution:

$$w_{ij} = \frac{1}{2\pi\sigma^2}exp\left\{-\frac{(i-\bar{i})^2+(j-\bar{j})^2}{2\sigma^2}\right\}, \tag{15}$$

where $i = 1,\ldots,m_x$, $j = 1,\ldots,m_y$, $\bar{i} = m_x/2$, $\bar{j} = m_y/2$.

4. EXPERIMENTAL RESULTS

In this section we demonstrate the performance of the proposed tracking scheme in videos depicting eating and drinking activities. The videos were recorded in the AIIA laboratory. The ultimate purpose is to exploit the tracking results in an eating and drinking activity recognition framework by recognizing motion patterns which take part in the activities. More precisely, eating activity recognition can be performed by computing the relevant distance in pixels between the human hand holding the knife or spoon and the face, while drinking activity can be detected by calculating the distance in pixels between the glass and the head. Therefore, in our experiments, we test the performance of the proposed tracking scheme in tracking the glass and the head during drinking activity and the hand and the head during eating activity.

Experimental results have been performed in several videos depicting eating and drinking activities by different persons. One example is shown in Figure 1. We notice that the proposed tracker is able to handle the changes in the viewing angle of the glass (Figure 1a) in the duration of the drinking activity and the rotation of the hand during eating activity (Figure 1c). Therefore, the proposed tracker is robust in view and rotation changes of rigid objects. At this point we note that, in general, the hand is an articulated object which constantly changes shape and the proposed tracker cannot handle the deformations of the shape of the object. However, during eating activity, the hand holding the fork or spoon remains folded in an fist-like pose, therefore we can consider it as being rigid. Furthermore, Figures 1b, and 1d show that the tracker is able to track the human face during both eating and drinking activities. Moreover, from Figure 1b it is derived that the tracker keeps track of the face even when it is occluded by the glass. This is achieved because the tracker takes into account the similarity with the initial query object. Further experiments in videos depicting other activities showed the robustness of the proposed tracker in tracking successfully rigid objects.

5. CONCLUSION

In this paper we presented a novel appearance-based method for visual object tracking which employs local steering kernels for image representation. The objective of the proposed tracking scheme is to be used in an automatic nutrition assistance framework, however it can be used in tracking any rigid object.

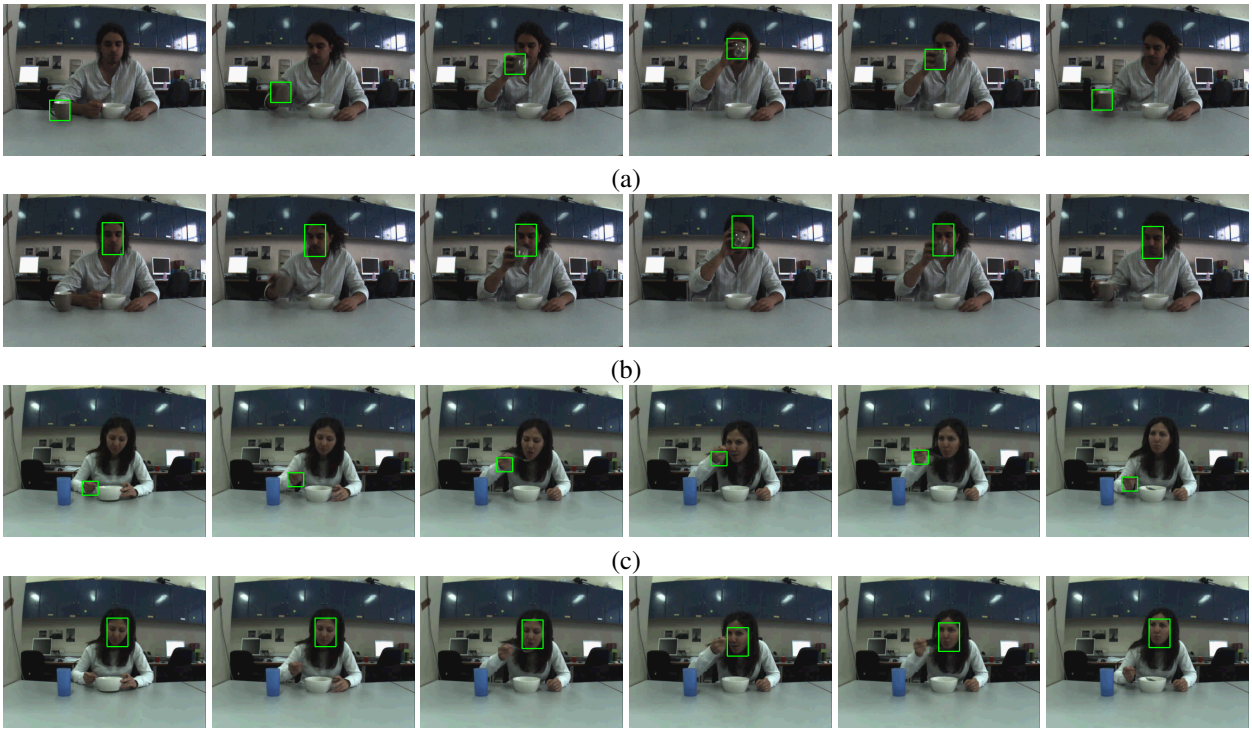

Fig. 1. Tracking results in videos depicting eating and drinking activities

6. ACKNOWLEDGEMENT

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.

7. REFERENCES

[1] D. Roller, K.Daniilidis, and H. H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *International Journal of Computer Vision*, vol. 10, pp. 257–281, 1993.

[2] C. Yang, R. Duraiswam, and L. Davis, "Efficient mean-shift tracking via a new similarity measure," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, June 2005, vol. 1, pp. 176 – 183.

[3] A. Yilmaz, X. Li, and M. Shah, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1531 –1536, nov. 2004.

[4] L. Fan, M. Riihimaki, and I.Kunttu, "A feature-based object tracking approach for realtime image processing on mobile devices," in *17th IEEE International Conference on Image Processing (ICIP)*, sept. 2010, pp. 3921 –3924.

[5] Li-Qun Xu and Pere Puig, "A hybrid blob- and appearance-based framework for multi-object tracking through complex occlusions," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.*, oct. 2005, pp. 73 – 80.

[6] Jianxin Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J.M. Rehg, "A scalable approach to activity recognition based on object use," in *IEEE 11th International Conference on Computer Vision (ICCV)*, oct. 2007, pp. 1 –8.

[7] M. Singh, A. Basu, and 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.

[8] Hae Jong Seo and Peyman Milanfar, "Training-free, generic object detection using locally adaptive regression kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1688–1704, September 2010.

[9] Hae Jong J. Seo and Peyman Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, vol. 9, no. 12, 2009.