VISUAL OBJECT TRACKING VIA GABOR-BASED SALIENT FEATURES EXTRACTION

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 appearance-based method for visual object tracking of rigid objects with pose variations and small scale and 2-dimensional rotation changes is proposed. The algorithm employs a bank of Gabor filters for computing the salient object features, which represent the object model. In each frame, candidate objects of a search region are extracted randomly, following a 2-dimensional Gaussian distribution. The object in the current frame is the candidate object whose cosine similarity to the detected object in the first frame and the object instance in a previous frame where significant change in the object appearance was last observed is maximal.

Index Terms— visual object tracking, local steering kernels, Gabor filters

1. INTRODUCTION

By visual object tracking we define the challenging task of extracting the trajectory of a moving object in a video, by exploiting information obtained from the video content without the use of any sensor data. Tracking algorithms should be able to handle a number of factors which affect the tracking performance, such as changes in the lighting conditions of the video, rapid and non-smooth object movements, noise, etc. Depending on the object representation method, the tracking algorithms can be divided into four broad categories:

- Model-based algorithms [1], which use 3-dimensional description models of the object,
- Appearance-based algorithms [2], which employ 2-dimensional description models of the object texture,
- Contour-based algorithms [3], which perform object tracking by identifying the object contour, and
- Feature-based algorithms [4], which identify and track the object salient features.

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.

Furthermore, there exist hybrid tracking algorithms which combine more than one object representation methods [5]. Most of the tracking algorithms, including the proposed one, employ appearance-based representations of the object, as they are more simple and require less computations than other methods.

Visual object tracking is a fundamental tool in video content analysis, as it enables the study of the motion of the entities (i.e., objects, humans) which appear in a video which, consequently, leads to the extraction of high level descriptions for the content of the video. For example, human activity recognition can be performed by analyzing the trajectories of human body parts, the trajectories of auxiliary objects which take part in an activity, or the relevant position between more than one objects of interest which characterize an activity. In this notion, eating and drinking activity recognition may be performed by analyzing the trajectories of the human hands during food intake, the trajectories of kitchen utensils used for eating and drinking (e.g. glass, fork, etc.), or the relevant position between the hands and the face.

In this paper, a novel algorithm is introduced for tracking rigid objects in videos. The objective of the proposed algorithm is to be used in an automatic nutrition support system for eating and drinking activity recognition. Therefore, the algorithm was tested in videos depicting meal sessions of humans. Experimental results showed that the proposed method is successful in tracking rigid objects which perform smooth movements with changes in the view angle, 2-dimensional rotations and small changes in scale.

2. PROBLEM STATEMENT

Dementia is a syndrome which affects a high percentage of the geriatric population over the age of 65, which causes either a static or a progressive loss of the patient's cognitive ability. At an early stage, dementia may cause deterioration of the nerves, apraxia (i.e. loss of the patient's ability to use tools) and agnosia (i.e. loss of the patient's ability to identify other persons, objects, smells, sounds, or shapes). As a result, the patients lose the ability of executing activities of day

living, such as eating and drinking, by themselves. In order to prevent dehydration and underfeeding of patients suffering from early stage of dementia, the development of a monitoring system is required, which detects the time instances when the patient eats or drinks and measures their duration. If the system detects that the patient hasn't eaten or drunk anything in a certain period of time, a robotic unit reminds him to eat or drink. The monitoring system should process only visual data obtained by surveillance cameras, as body warn sensors or markers on the patient's hands and face may cause disturbance to the patient.

3. LOCAL STEERING KERNEL DESCRIPTORS

Locals Steering Kernels (LSKs) [6] are descriptors of the salient features of an image, which represent how similar a pixel is with its surrounding pixels in a locally defined $P \times P$ window, by taking into account both their illumination difference (their pixel value) and the distance between the neighboring pixels:

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

where $\mathbf{p}$ denotes the image pixel coordinates, $\mathbf{p}_l$ denotes the neighboring pixels coordinates, and $\mathbf{C}_l$ is a covariance matrix, which is estimated from the matrix $\mathbf{J}_l$:

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

which consists of the gradient vectors of the image in a $P \times P$ window around $\mathbf{p}_l$, by applying SVD according to equations (3)-(5) [7]:

$$\mathbf{J}_l = \mathbf{U}_l \cdot \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix}_l. \tag{3}$$

$$\mathbf{C}_l = \gamma \sum_{q=1}^{2} a_q^2 \mathbf{v}_q \mathbf{v}_q^T, \tag{4}$$

$$a_1 = \frac{s_1+1}{s_2+1}, \quad a_2 = \frac{s_2+1}{s_1+1}, \quad \gamma = \left(\frac{s_1 s_2 + 10^{-7}}{P^2}\right)^a. \tag{5}$$

In equation (2), $\mathbf{z}(\mathbf{p}) = [z_x(\mathbf{p}), z_y(\mathbf{p})]^T$ denotes the image gradient vector along x and y axes at the position $\mathbf{p}$, while in equations (5), a is a parameter that restricts γ. In our experiments a takes the value 0.008.

4. LSK COMPUTATION VIA A BANK OF GABOR FILTERS

Gabor filters are band-pass filters, widely used in image processing for edge detection, therefore they can be used for estimating the gradient vector of an image along some direction φ. In the 2-D space, a Gabor filter is defined as a complex sinusoid $s(x, y)$ (i.e. the carrier) modulated by a Gaussian kernel function $g(x, y)$ (i.e. the envelope) [8]:

$$f(x, y) = s(x, y) \cdot g(x, y), \tag{6}$$

where the Gaussian kernel function is defined as:

$$g(x, y) = \frac{K^2}{\sigma^2} \exp\left(-\frac{K^2(x^2+y^2)}{2\sigma^2}\right) \tag{7}$$

and the complex sinusoid is given by:

$$s(x, y) = \exp\left(jK(x\cos\varphi + y\sin\varphi)\right) - \exp\left(-\sigma^2/2\right). \tag{8}$$

In equations (7), (8) $K/2$ denotes the magnitude and φ denotes the direction of the spatial frequency, while σ^2 is a scaling parameter.

A bank of 12 Gabor filters with 4 orientations (0, 45, 90 and 135 degrees) and 3 scales may be used in order to compute the LSK descriptors given by (1) as follows. At first, we average the responses of the Gabor filters with the same orientation and different scales. Let us consider the averaged filter responses at 0 and 90 degrees as the image gradient vectors of (2) along x and y-axes, respectively. In the same notion, the averaged filter responses at φ and $\varphi + 90$ degrees (in our case 45 and 135 degrees) can be considered as the image gradient vectors $\mathbf{z}$ in the rotated by φ degrees coordinate system. Therefore, the LSK descriptors (1) in the rotated coordinate system will be given by:

$$K_\varphi(\mathbf{p}_l - \mathbf{p}) = \frac{\sqrt{\det(\mathbf{C}_l)}}{2\pi} \exp\left\{-\frac{(\mathbf{p}_l-\mathbf{p})^T \mathbf{R}_\varphi^T \mathbf{C}_l \mathbf{R}_\varphi (\mathbf{p}_l-\mathbf{p})}{2}\right\}, \quad l = 1, \ldots, P^2, \tag{9}$$

where $\mathbf{R}_\varphi$ is the rotation matrix

$$\mathbf{R}_\varphi = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}, \tag{10}$$

and $\mathbf{C}_l$ is given by equations (2)-(4). It is straightforward to show that equation (1) is derived by equation (9) for $\varphi = 0$.

5. THE PROPOSED TRACKING FRAMEWORK

The proposed method performs object tracking in a video by employing the modified locally adaptive regression kernel descriptors first introduced in [6] for object representation. The algorithm computes the similarity of candidate objects (patches) in a search region $\mathbf{T} \in \Re^{M_x \times M_y}$ of the target frame

with the object instance in the first frame (initial query image $\mathbf{I} \in \Re^{N_x \times N_y}$) and the object instance in a previous frame (query image $\mathbf{Q} \in \Re^{N_x \times N_y}$), where significant change in the object appearance was last observed. The size of the candidate objects is equal to the size of the query object and the initial query object. The proposed algorithm starts by initialization of the position of the object at the initial video frame. The object initialization can be achieved in two ways: automatically, by using an object detection algorithm, or manually, by inserting the object coordinates in the initial frame. Then, the algorithm executes the following three iterative steps. In the first step, the 1^{st} order Kalman filter is applied for predicting the new object position, the new search region is determined and the candidate objects are initialized. In the second step, the local steering kernel descriptors of the initial query image, the query image, and the candidate objects are extracted. Finally, the similarities of the candidate objects to the initial query image and the query image are calculated and exploited in order to determine the new position of the object.

5.1. Candidate objects selection

In this step, the object position is predicted through the 1^{st} order Kalman filter. Given that $\mathbf{x}_t = [p_x, p_y, dx, dy]^T$ is the state of the object at frame t, the position $\hat{\mathbf{x}}_{t+1}$ of the object at frame $t+1$ is estimated from the motion state estimation model $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{n}_t$ according to the equations:

$$\hat{\mathbf{x}}_{t+1} = \mathbf{A}\hat{\mathbf{x}}_t, \quad (11)$$

$$\hat{\mathbf{P}}_{t+1} = \mathbf{A}\hat{\mathbf{P}}_t\mathbf{A}^T + \mathbf{Q}_s, \quad (12)$$

and the measurement model $\mathbf{z}_{t+1} = \mathbf{H}\mathbf{x}_{t+1} + \mathbf{v}_{t+1}$ is adjusted according to the equations:

$$\mathbf{K}_{t+1} = \hat{\mathbf{P}}_t\mathbf{H}^T(\mathbf{H}\hat{\mathbf{P}}_t\mathbf{H}^T + \mathbf{Q}_m)^{-1} \quad (13)$$

$$\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t + \mathbf{K}_{t+1}(\mathbf{z}_{t+1} - \mathbf{H}\hat{\mathbf{x}}_{t+1}) \quad (14)$$

$$\mathbf{P}_{t+1} = (\mathbf{I} - \mathbf{K}_{t+1}\mathbf{H})\mathbf{P}_{t+1}, \quad (15)$$

where $\mathbf{A}$ is the transition matrix of the system, $\mathbf{n}_t$ is the process noise with covariance matrix $\mathbf{Q}_s$, $\hat{\mathbf{P}}_t$ is the error covariance matrix, $\mathbf{z}_t = [p_x, p_y]^T$ is the system measurement, $\mathbf{H}$ is the measurement matrix, $\mathbf{v}_t$ is the measurement noise with covariance matrix $\mathbf{Q}_m$, and $\mathbf{K}_t$ is the Kalman gain.

The search region in frame t is then defined around $\hat{\mathbf{x}}_{t+1}$ with size $M_x \times M_y = fN_x \times fN_y$, where f is a factor which determines the search region size. The value of f 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. Finally, a candidate objects $\mathbf{Y}_{t+1}$ are selected randomly, according to:

$$\mathbf{Y}_{t+1} = \left\{\mathbf{y}^1_{t+1}, \ldots, \mathbf{y}^a_{t+1}\right\} \sim N(\hat{\mathbf{x}}_{t+1}, \sigma), \quad (16)$$

$\sigma = \text{diag}[M_x/4, M_y/4]$. In out experiments we set $a = 150$.

5.2. Salient features extraction

The salient feature of the initial query image, the query image, and the search region are extracted from equation (9) for φ equal to 0 and 45 degrees. For an image pixel $\mathbf{p}$, equation (9) is computed for each neighboring pixel $\mathbf{p}_l$, $l = 1, \ldots, P^2$, meaning that for each image pixel we export two LSK feature vectors $\mathbf{K}_0(\mathbf{p}), \mathbf{K}_{45}(\mathbf{p}) \in \Re^{P^2 \times 1}$. The final feature vector is extracted by concatenating the two feature vectors:

$$\mathbf{K}(\mathbf{p}) = [\mathbf{K}_0(\mathbf{p})^T \mathbf{K}_{45}(\mathbf{p})^T]^T \in \Re^{2P^2 \times 1}. \quad (17)$$

The resulting LSK feature vector becomes invariant to brightness and contrast changes by using L-1 normalization:

$$\mathbf{N}(\mathbf{p}) = \frac{\mathbf{K}(\mathbf{p})}{\sum_{l=1}^{2P^2} |K(\mathbf{p}_l - \mathbf{p})|} \in \Re^{2P^2 \times 1}. \quad (18)$$

Finally, the LSK feature vectors of the $n = N_x N_y$ pixels of the query image, the initial query image and the candidate objects are ordered column-wise to form the LSK feature matrices $\mathbf{N}_Q \in \Re^{2P^2 \times n}$, $\mathbf{N}_I \in \Re^{2P^2 \times n}$ and $\mathbf{N}_{\mathbf{y}^i} \in \Re^{2P^2 \times n}$, $i = 1, \ldots, a$, respectively.

5.3. Similarity measure and decision extraction

After extracting the LSK feature matrices $\mathbf{N}_Q$, $\mathbf{N}_I$, $\mathbf{N}_{\mathbf{y}^i} \in \Re^{2P^2 \times n}$, $i = 1, \ldots, a$, we measure the similarity of the candidate objects to the query image and the initial query image. At first, we proceed to dimensionality reduction of the initial query image by PCA, producing the matrix $\mathbf{F}_I = \mathbf{A}_I\mathbf{N}_I \in \Re^{d \times n}$, where $\mathbf{A}_I \in \Re^{d \times 2P^2}$ is the projection matrix. In our experiments we set $d = 3$. The LSK feature matrices $\mathbf{N}_Q$, $\mathbf{N}_{\mathbf{y}^i}$, $i = 1, \ldots, a$ of the query image and the candidate objects are then projected to the space created by the projection matrix as follows:

$$\mathbf{F}_Q = \mathbf{A}_I\mathbf{N}_Q \in \Re^{d \times n}, \quad \mathbf{F}_{\mathbf{y}^i} = \mathbf{A}_I\mathbf{N}_{\mathbf{y}^i} \in \Re^{d \times n}. \quad (19)$$

Then, the similarity of the candidate objects to the query image and the initial query image is estimated by the cosine similarity:

$$s_{Qi} = s(\mathbf{F}_Q, \mathbf{F}_{\mathbf{y}^i}), \quad s_{Ii} = s(\mathbf{F}_I, \mathbf{F}_{\mathbf{y}^i}), \quad i = 1, \ldots, a, \quad (20)$$

where

$$s(\mathbf{F}_1, \mathbf{F}_2) = \sum_{l=1,j=1}^{n,d} \frac{F_1(l,j)F_2(l,j)}{\sqrt{\sum_{l=1,j=1}^{n,d} |F_1(l,j)|^2 \sum_{l=1,j=1}^{n,d} |F_2(l,j)|^2}}, \quad (21)$$

and $F_1(l,j)$, $F_2(l,j)$ denote the (l,j) elements of matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ respectively. The final decision for the new object position in frame $t+1$ is taken by:

$$\mathbf{p}_{t+1} = \text{argmax}_{\mathbf{y}^i}\left\{s_{Qi}, s_{Ii}\right\}. \quad (22)$$

If the maximum similarity of the detected object at frame $t+1$ is smaller that the 80% of the maximum similarity at frame t, then a significant change in the object appearance is detected, and the detected object at frame $t+1$ is considered the new query image Q.

6. EXPERIMENTAL RESULTS

The performance of the proposed tracking algorithm was tested on videos depicting eating and drinking activities from the MOBISERV/AIIA eating and drinking activity recognition database [9]. The database consists of videos depicting 12 subjects, 6 male and 6 female, during 4 meal sessions, recorded in 4 different days. In each session, the subject eats and drinks in all possible ways: he eats with a spoon, or a fork, or knife and fork, or with one hand, or with both hands; he drinks from a cup, or a glass, or a glass with a straw. The ultimate goal is to recognize the motion patterns which take part in eating and drinking activities, so that later they will be exploited in a nutrition support system. More precisely, eating activity can be identified by the up and down movement of the hands during bites and/or their relevant distance from the person's face. In the same notion, drinking activity can be characterized by the trajectory of the auxiliary utensil which takes part in the activity (i.e. the glass) and/or the relevant distance between the glass and the head. Moreover, the motion patterns of the head during eating and drinking can also be examined through profile-face tracking or, equivalently, ear tracking. Therefore the algorithm performance was tested on tracking the glass and the face during drinking activity, and the hand, the face and the ear during a meal. The glass-face tracking experiment was performed in a video captured by the frontal camera, where the view of the objects of interest is optimal. The hand-face tracking experiment was performed in a video captured by the upper-frontal camera, where there is no hand occlusion. Finally, the ear tracking experiment was performed in the video captured by the profile camera, which best captures the vertical head movement.

Experimental results are shown in Figure 1. The proposed algorithm is compared to the state-of-the-art appearance-based tracking method [10] (called FT tracker) which is based on integral histograms. The tracking results of the proposed algorithm and the FT tracker are depicted with green (first row) and yellow (second row) bounding boxes, respectively. In Figure 1a) we notice that the proposed framework tracks successfully the transparent rigid object (the glass) and the tracking performance is more stable than the one of the FT tracker. In the case of face tracking, the performance of both methods is equivalent. Figure 1b) shows that, the proposed method tracks successfully the hand during the meal, despite the changes in the hand appearance. On the other hand, the FT tracker looses track of the hand during the drink-up activity. The face tracking in the upper-frontal camera is successful in both methods, however the proposed algorithm performance is more stable. Finally, in Figure 1c) we notice that, the proposed tracker is able to track rigid objects with color similar to the background, in contrast to the FT tracker, which is based on color information and, therefore, looses track of the human ear.

7. CONCLUSION

In this paper we presented a novel appearance-based method for visual object tracking which employs local steering kernels estimated from a bank of Gabor filters for image representation. Experimental results showed the effectiveness of the proposed tracking scheme in tracking successfully any rigid object under pose variations and small changes in scale and 2-d angle and its superiority against a state-of-the-art method. The objective of the proposed tracking scheme is to be used in an automatic nutrition assistance framework.

8. 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.

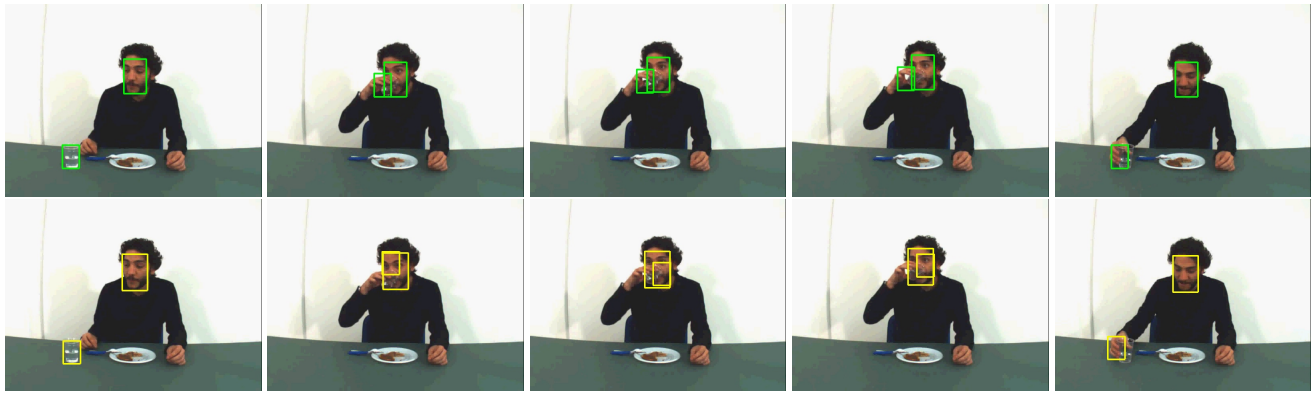
[6] Hae Jong Seo and Peyman Milanfar, "Training-free, generic object detection using locally adaptive regression kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1688–1704, September 2010.

[7] 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.

[8] J. R. Movellan, "Tutorial on Gabor Filters," *Tutorial paper http://mplab.ucsd.edu/tutorials/pdfs/gabor.pdf*, 2008.

[9] http://www.aiia.csd.auth.gr/MOBISERV AIIA/index.htm, ," .

[10] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 798–805.

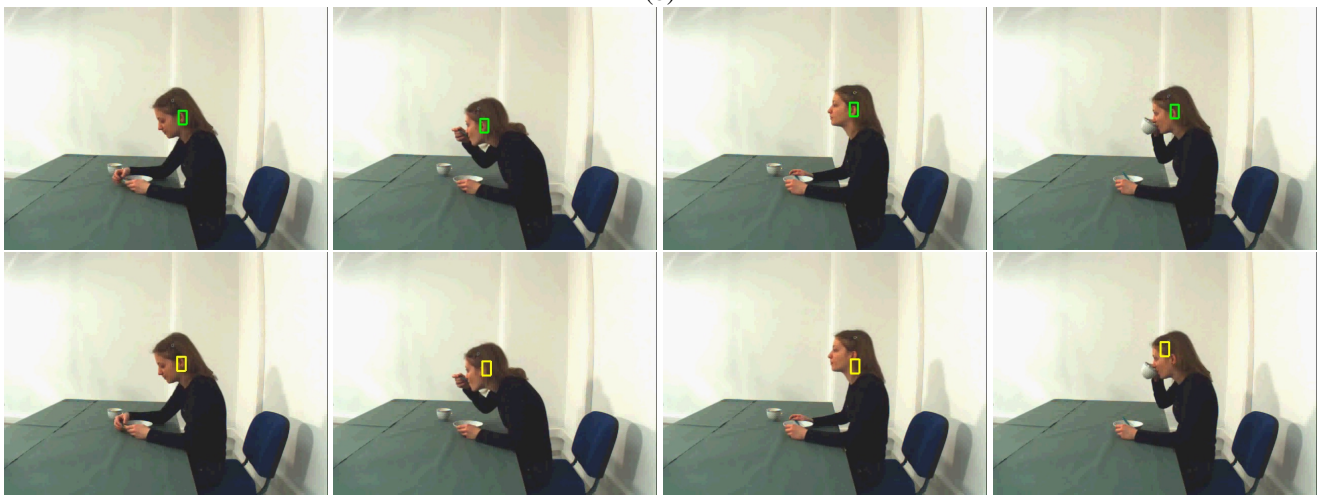


frame 0 frame 38 frame 43 frame 50 frame 81

(a)

frame 0 frame 75 frame 213 frame 265

(b)



frame 0 frame 60 frame 121 frame 229

(c)

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