Signal Processing: *Image Communication* 39 (2015) 418–431

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos

Vincent Buso [a,*], Iván González-Díaz [b], Jenny Benois-Pineau [a]

[a] *Laboratoire Bordelais de Recherche en Informatique, Université Bordeaux, 33405 Talence, France*
[b] *Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés 28911, Madrid, Spain*

ARTICLE INFO

Available online 3 June 2015

Keywords:
Saliency maps
Egocentric vision
Object recognition
Vision modelling
Image processing
Video processing

ABSTRACT

We propose a new top down probabilistic saliency model for egocentric video content. It aims to predict top-down visual attention maps focused on manipulated objects, that are then used for psycho-visual weighting of features in the problem of manipulated object recognition. The model is probabilistically defined using both global and local appearance features extracted from automatically segmented arm areas and objects. A psycho-visual experiment has been conducted in a guided framework that compares our proposal and other popular state-of-the-art models with respect to human gaze fixations. The obtained results show that our approach outperforms several popular bottom-up saliency approaches in a well-known egocentric dataset. Furthermore, an additional task-driven assessment for object recognition in egocentric video reveals that the proposed method improves the performance of several state-of-the-art techniques for object detection.

1. Introduction and motivation

The rationale and application of this research is in the objective assessment and life-logging of Alzheimer patients in their Instrumental Activities of Daily Living (IADLs). For this particular task, egocentric video analysis has gained a lot of interest. Indeed this kind of video content is recorded by cameras worn by a person, representing a cheap and effective way to record users' activity, and offering a unique point of view on the manipulated objects (see Fig. 1). Recent studies have demonstrated how crucial is the recognition of manipulated objects for activity recognition under this scenario [1,2]. Considering methods for object recognition, two kinds of approaches can be identified in the literature: those relying on sliding windows, and those ones that first try to segment the foreground area containing the object of interest.

Concerning the first type, the authors of [1] applied the well-known Discriminatively Trained Deformable Part-Based (DPM) Models [3] to egocentric video. The second kind of approaches follows the well-known paradigm of foreground object segmentation to guide the object recognition process. The authors of [4] proposed a method that firstly segments the foreground areas from the background of each frame. Once the segmentation is made, the method detects and labels regions associated with the hands and the object being manipulated, respectively, and finally assigns an object label to the frame.

Concerning the second line of research, to drive the recognition process to relevant areas in the images, saliency or visual attention modelling was incorporated to the object recognition paradigms, showing an increase in the system performances [5,6]. In this paper, we will follow this line of research, as we are particularly interested in the modelling of human visual attention based on task-oriented top-down cues.

* Corresponding author.
E-mail addresses: vbuso@labri.fr (V. Buso), igonzalez@tsc.uc3m.es (I. González-Díaz), benois-p@labri.fr (J. Benois-Pineau).

http://dx.doi.org/10.1016/j.image.2015.05.006
0923-5965/© 2015 Elsevier B.V. All rights reserved.

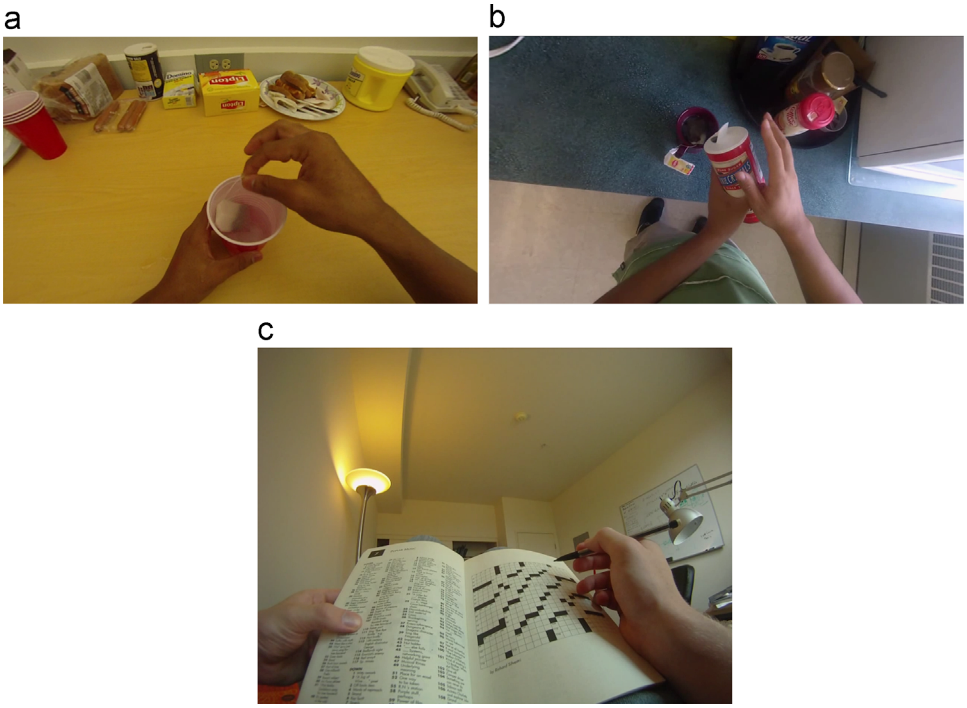

Fig. 1. Examples of egocentric datasets illustrating the unique point of view on manipulated objects. (a) GTEA dataset [4], (b) EDSHK dataset [7], and (c) ADL dataset [1].

Generally speaking, two types of attention are commonly distinguished in the literature: bottom-up or stimulus-driven and top-down attention or goal-driven. [8,9]. The authors of [8] define the top-down attention as the voluntary allocation of attention to certain features, objects, or regions in space. They also state that attention is not only voluntary directed as low-level salient stimuli can also attract attention, even though the subject had no intention to attend these stimuli. A recent study [10] about how saliency maps are created in the human brain shows that an object captures our attention depending both on its bottom-up saliency and top-down control.

Modelling of human visual attention has been an intensively explored research subject since the last quarter of the 20th century and nowadays the majority of saliency computation methods are designed from a bottom-up perspective [11]. Bottom-up models are stimulus-driven, mainly based on low-level properties of the scene such as colour, gradients orientation, motion or even depth. Consequently, bottom-up attention is fast, involuntary and, most likely feed-forward [11].

One of the first complete models of visual attention was proposed as a fusion of features based on the Human Visual System modelling (HVS) by Itti [12]. Since then, much work has been made in this domain [13–15]. The reader is referred to a recent benchmark of several saliency models for more details [16].

However, although the literature concerning models of top-down attention is clearly less extensive, the introduction of top-down factors (e.g., face, speech and music, camera motion) into the modelling of visual attention has provided impressive results in previous works [17,18]. In addition, some attempts in the literature have been made to model both kinds of attention for scene understanding in a rather “generic” way. In [19] the authors claim that the top-down factor can be well explained by the focus in image, as the producer of visual content always focuses his camera on the object of interest. Nevertheless, it is difficult to admit this hypothesis for expressing the top-down attention of the observer of the content: it is always task-driven [11].

More recent works using machine learning approaches to learn top-down behaviours based on eye-fixation or annotated salient regions have proven also to be very useful for static images [20–22] as well as for videos [23,24]. Furthermore, with advent of Deep Learning Networks (DNN), some novel approaches have been designed in the field of object recognition, which build class-agnostic object detectors to generate candidate salient bounding-boxes which are then labelled by later class-specific object classifiers [25,26]. However, it seems impossible for us to propose a universal method for prediction of the top-down visual attention component, as it is voluntary directed attention and therefore it is specific for the task of each visual search. Nevertheless, the prior knowledge about the task the observer is supposed to perform allows extracting semantic clues from the video content which would ease such a prediction.

The current state-of the art in computer vision allows detection of some categories of objects with a high confidence. A variety of face or skin detectors have been proposed since the last two decades [27]. Hence, when modelling a top-down attention in a specific visual search task, we can use such “easily recognizable” semantic