Pattern Recognition 56 (2016) 129–141

Contents lists available at ScienceDirect

Pattern Recognition

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

Perceptual modeling in the problem of active object recognition in visual scenes

Iván González-Díaz [a,*,1], Vincent Buso [b], Jenny Benois-Pineau [b]

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

ARTICLE INFO

Article history:
Received 27 March 2015
Received in revised form
19 January 2016
Accepted 4 March 2016
Available online 15 March 2016

Keywords:
Perceptual modeling
Visual saliency
Active object recognition
Foveal and peripheral pathways

ABSTRACT

Incorporating models of human perception into the process of scene interpretation and object recognition in visual content is a strong trend in computer vision. In this paper we tackle the modeling of visual perception via automatic visual saliency maps for object recognition. Visual saliency represents an efficient way to drive the scene analysis towards particular areas considered 'of interest' for a viewer and an efficient alternative to computationally intensive sliding window methods for object recognition. Using saliency maps, we consider biologically inspired independent paths of central and peripheral vision and apply them to fundamental steps of the so-called Bag-of-Words (BoW) paradigm, such as features sampling, pooling and encoding. Our proposal has been evaluated addressing the challenging task of active object recognition, and the results show that our method not only improves the baselines, but also achieves state-of-the-art performance in various datasets at very competitive computational times.

1. Introduction

Object recognition is a very active research field for the computer vision community. For such a task, the Bag-of-Words (BoW) model [1,2] is still one of the most prevalent approaches due to its simplicity. However, its performance is greatly limited in case of occlusions or small objects in cluttered backgrounds. In contrast, sliding window methods have turned out to be more robust against these problems. They perform a window-based scanning process that searches for objects in several locations and scales in the image, thus addressing both the detection and accurate localization of objects even when they are small. Examples of these methods can be found in the literature for detecting faces [3], pedestrians [4], more generic objects [5], and even mixed with the BoW [6]. Nevertheless, these methods still suffer from several drawbacks: (a) although efficient implementations exist, the computational complexity due to the computation of features within each candidate window, and the evaluation of the objective function cannot be neglected; (b) they require a strong human effort to manually annotate bounding boxes in the training data; (c) an exhaustive scanning might cause more false detections; or (d) unless explicitly incorporated, context information around the object that might become a valuable cue of its presence is usually discarded.

Alternatively, modeling the selective process of human perception of visual scenes represents an efficient way to drive the scene analysis towards particular areas considered 'of interest' or 'salient'. This is why it has become a very active trend in computer vision [7]. Due to the use of saliency maps, the search for objects in images is more focused, thus improving the recognition performance and additionally reducing the computational burden. Even more, saliency methods can be naturally applied to both BoW [8] and sliding window approaches [9,10].

Models of visual attention, such as the one proposed by Itti and Koch [11] or Harel's graph implementation [12] are frequently used in literature for computing saliency maps. Various authors have shown how driving the processing to those particular areas with high values in the saliency maps improves the system performance in various computer vision tasks, such as image retrieval [13], object recognition [14,15], object tracking [16,17], or action recognition [18,19]. However, although much fundamental work has been done to generate good representations of visual saliency from still images or video content, their application to object recognition has not been yet explored in-depth. Indeed, it is still commonly restricted to a pre-processing stage that filters out non-relevant areas from the process [8].

In this paper, therefore, we provide a systematic study of the application of saliency to the challenging task of active object recognition. In a given scene, active objects are those objects

* Corresponding author. Tel.: +34 916246262; fax: +34 916248749.
E-mail addresses: igonzalez@tsc.uc3m.es (I. González-Díaz), vbuso@labri.fr (V. Buso), benois-p@labri.fr (J. Benois-Pineau).
[1] This research was carried out when he was working at the LaBRI.

http://dx.doi.org/10.1016/j.patcog.2016.03.007
0031-3203/

which are interacted (manipulated, observed) by the users and, therefore, play a key role to understand the semantics of the scene. Furthermore, we claim that, in many scenarios in which humans perform activities by manipulating objects, an action can be effectively defined as a sequence of 'active' objects [20]. Hence, we do not aim to detect every object in the scene, but only those ones considered as active. This problem fits well with the nature of saliency since it aims to drive the recognition process to the areas of interest in the image, therefore preventing from the detection of non-active objects that belong to the background of the scene.

Our saliency-based approach aims to model the retina in the Human Visual System (HVS), and consider biologically inspired independent foveal and peripheral visual paths. By plugging our contributions in the BoW paradigm, we investigate how visual attention modeling can be applied to various steps in the processing pipeline. To the best of our knowledge, this is the first in-depth study about the application of visual saliency to object recognition with BoW approach at several stages, as: (i) we extend the state-of-the-art on *Saliency-sensitive non-uniform feature sampling* in a new *Saliency-sensitive variable-resolution feature space*, (ii) we introduce a completely new *Saliency-sensitive Coding of features* and use the (iii) *Saliency-based feature pooling* which has been shown to be efficient in referenced research [20,13].

The benefits of our approach are multiple: (i) the computation of saliency maps is category-independent and a common step for any object detector, (ii) compared to sliding window methods, by looking at the salient area we can avoid much of the computational overhead caused by an exhaustive scanning process, (iii) our automatic saliency maps not only focus on the object of interest of a scene but usually contain some context around the object, (iv) an object recognition method working with saliency maps does not need ground-truth bounding boxes for training, which dramatically reduces the human resources devoted to the database annotation. In contrast, a known limitation of the use of saliency is that, as it focuses on the objects/area of interest of the scene, it may prevent systems from detecting those objects located outside these areas and that belong to the background of the scene.

In order to assess these benefits, we have selected an experimental benchmark composed of both video and image datasets containing scenes in which just a few objects are considered and have been manually labeled as active. The video datasets are 1st-person camera view (egocentric videos), which have recently gained a lot of attention due to the emerging end-user applications involving the use of wearable cameras in scenarios such as robotics, telemedicine or life-logging [21]. Furthermore, as this kind of content fits well with the problem being addressed, we can find previous works in the literature that have previously applied visual saliency to egocentric video analysis [22,8,23,24]. On the contrary, the image datasets are 3rd-person camera view and demonstrates that our method is not restricted to egocentric contents.

The remainder of the paper is organized as follows: in Section 2 we discuss the work related to the application of perceptual modeling to computer vision and, particularly, to object recognition. Next, in Section 3, and just for the sake of completeness, we provide a brief description of the method used to compute saliency in video. Section 4 describes in detail our saliency-based approach for active object recognition. In Section 5 an in-depth evaluation is provided that assesses our model under the various scenarios, and compares it to other state-of-the-art approaches. Finally, Section 6 summarizes our conclusions and gives perspectives.

2. Related work in saliency-based object recognition

Modeling visual perception in the problem of object recognition consists in the automatic prediction of the areas in the scene which, by their spatial, luminance, color and motion properties, would attract human gaze [12]. This is the so-called 'bottom-up' visual attention prediction. The rationale of using such low-level prediction is in the hypothesis that objects are characterized by peculiarities in these description channels. There exist different predictors for visual attention: e.g. those that predict the dynamics, that is saccadic motion [25], or those which focus on fixations [26]. Furthermore, the predicted visual attention is often expressed in the form of 'saliency maps' [11,8].

In any case, this paper does not focus on the particular method to compute saliency but, alternatively, studies how this valuable information can be plugged into an object recognition pipeline. In general, previous approaches tackling this problem can be broadly divided into three categories: methods using *binary segmentation masks*, *Saliency-based Pooling*, and *saliency-based sampling*.

Traditionally, most works have relied on binary saliency maps, also known as foreground masks, as a way to delimit the particular area of the image to be processed. This is the case of [27], where object matching is improved by filtering out the local descriptors located in non-salient areas, or the more recent proposal [8], where the authors incorporated foreground masks to the BoW paradigm by restricting the detection of local features to particular salient areas of the image. A similar approach is followed in [22], where a method for object recognition in egocentric video firstly identifies foreground areas in each frame, and consequently detects and labels regions associated with the hands and the object being manipulated.

Following the second strategy, the works in [13,15] substitute these binary masks by a soft-pooling scheme over real-valued saliency maps. In particular, both works build over the BoW paradigm, and consider the continuous values of a saliency map to weigh the contribution of each visual word. In addition, in [13] two complementary image signatures are considered: one associated with the foreground, and another modeling the background. These signatures enable foreground and background-based object recognition, or even combined recognition in which both the object of interest and the context are considered. In [14], a discriminative approach for pooling visual features is proposed that integrates within a unified framework the computation of saliency maps and the learning of SVM-based classifiers. In this case, saliency maps are category-dependent functions that learn the spatial distribution of visual words associated with particular object categories. This approach has been successfully applied to various computer vision tasks, such as action recognition or scene classification.

Concerning the third category of methods, other works have used saliency to perform non-uniform sampling of local features in images, so that more information is gathered on those areas considered as salient. In [28] the authors propose a classification method based on the use of decision trees over randomly sampled square patches of different sizes. To improve this random sampling process, category-specific saliency maps store the most likely locations and scales of positive patches of each class. The works in [18,19] also explore the same idea in the BoW paradigm, so that local descriptors are computed over regions randomly sampled using saliency maps. Finally, [9,10] are yet other examples of this kind of approach, where saliency maps drive the search process of sliding-window object detectors, thus drastically reducing the number of windows being evaluated. Finally, there exist other approaches that follow the so-called 'fixation point strategy', whereby they sequentially analyze a set of image representations or 'glimpses' from each visual fixation a human would perform on