J. Vis. Commun. Image R. 38 (2016) 45–54

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Finding spatio-temporal salient paths for video objects discovery ☆

Ye Luo [a,*], Junsong Yuan [b], Jianwei Lu [a]

[a] School of Software Engineering, Tongji University, Shanghai 201804, China
[b] Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 10 September 2015
Accepted 2 February 2016
Available online 20 February 2016

Keywords:
Spatio-temporal path
Salient object detection
Temporal coherence
Video saliency estimation
Saliency density
Dynamic programming
Salient path discovery
Computer vision

ABSTRACT

Many videos capture and follow salient objects in a scene. Detecting such salient objects is thus of great interests to video analytics and search. However, the discovery of salient objects in an unsupervised way is a challenging problem as there is no prior knowledge of the salient objects provided. Different from existing salient object detection methods, we propose to detect and track salient object by finding a spatio-temporal path which has the largest accumulated saliency density in the video. Inspired by the observation that salient video objects usually appear in consecutive frames, we leverage the motion coherence of videos into the path discovery and make the salient object detection more robust. Without any prior knowledge of the salient objects, our method can detect salient objects of various shapes and sizes, and is able to handle noisy saliency maps and moving cameras. Experimental results on two public datasets validate the effectiveness of the proposed method in both qualitative and quantitative terms. Comparisons with the state-of-the-art methods further demonstrate the superiority of our method on salient object detection in videos.

1. Introduction

As the proliferation of visual saliency estimation methods, salient object detection either from images or videos becomes an important topic. Moreover, due to the exploding number of internet videos (e.g. Youtube), salient video object detection has seen increasing demands and has extensive applications in object recognition, human action detection, video summarization, video retargeting, etc.

Although various methods have been proposed to estimate saliency, accurate detection of the salient object in a saliency map remains a challenging problem. It is not easy to accurately locate the salient object and crop it out from a noisy saliency map which is usually caused by the distractions from the cluttered background. Moreover, without any prior knowledge of the salient objects (e.g. location, size, category, shape, etc.), it is very difficult to detect them just based on the confidence of the local saliency. To locate the salient object, Luo et al. [13] proposed to find a bounding box of the maximum saliency density (MSD) based on the observation that the average density of the salient object region is much larger than that of any other regions on saliency maps. Although this method works well for object discovery in images, the major limitations of extending it to salient video object detection are that the saliency density is the sole factor to detect objects and the temporal correlation among consecutive frames is ignored. Some examples can be seen in Fig. 1. A referee who is breaking into the view is wrongly detected as the salient object by the saliency alone. On the contrary, the football player and the flying ball are more likely to be the concerns of audience. How to accurately detect salient objects in noisy saliency maps and leverage the temporal coherence is the problem we study in this paper.

Despite many methods proposed to detect salient objects, few of them focus on locating salient objects in videos. There are methods proposed to locate a specific object in videos, such as the slide window search [7] and the optimal path discovery (OPD) [17], but neither of them targets to find salient objects. Moreover, instead of using a bounding box to locate objects, OPD targets to find a path with various sizes of bounding boxes as the windows indicate the existence of an object or an event. However, they did not consider the temporal correlations among each spatial windows.

To address the aforementioned problems, we propose a novel unsupervised method to discover salient objects from a given video clip and a simple version of our previous work can be found in [12], where salient object detection is formulated as an optimal salient path discovery problem. Suppose any 2D sliding window in a frame is a node and has a local detection score (i.e. saliency density) assigned by the estimated saliency map. By connecting nodes

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail addresses: luoy0009@e.ntu.edu.sg (Y. Luo), jsyuan@ntu.edu.sg (J. Yuan), jwlu33@tongji.edu.cn (J. Lu).

http://dx.doi.org/10.1016/j.jvcir.2016.02.001
1047-3203/

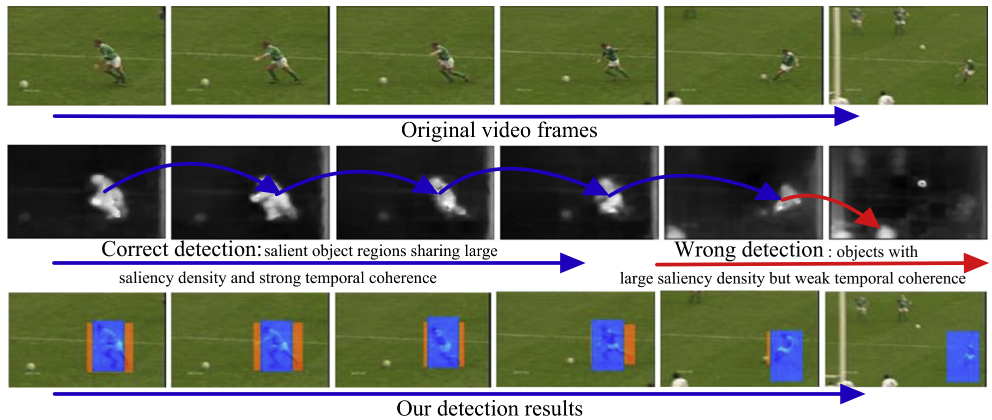

Fig. 1. Salient objects in a video can be detected by discovering a salient path which has the maximum accumulated saliency density. By adding the temporal coherence constraints to the most dense regions, our method accurately detect and track the salient objects. Better to view in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and fusing the local evidence of them, we build a spatio-temporal salient path. After adjusting the score by adding the temporal coherence constraints among nodes, the accumulated score for all nodes along the path shows the likelihood of the occurrences of salient objects. Thus, the path which has the maximum accumulated score corresponds to the target salient objects. Our salient object discovery method can maximize the likelihood of the occurrences of the salient objects and rectify the detection results by enhancing the temporal coherence among salient object regions. Regarding the solution, a dynamic programming algorithm is proposed and a global optimal solution is obtained afterward. Different from OPD, our solution is not performed on a discriminative score map and the temporal correlations among nodes in the path are taken into consideration. Last but not the least, in order to better characterize the role of motion in video saliency estimation, we propose our own way to estimate motion contrast. Extensive experiments on hundreds of videos from two public datasets demonstrate the effectiveness of the proposed method on salient video object detections.

To summarize, our major contribution is twofold:

(1) We first formulate the video salient object detection as an optimal salient path discovery problem, which makes a more robust detection via incorporating temporal coherence into a salient path.
(2) To obtain the global optimal solution of the salient path discovery problem, we propose an efficient dynamic programming algorithm, which is based on the saliency density rather than the traditional classification scores. The experiments also show its efficiency on salient video object detection.

The rest of paper is arranged as follows. Section 2 gives an overview of the related work. Sections 3 and 4 present our estimation method for video saliency and detection method for salient video objects. The experimental results are presented in Section 5 and we conclude in Section 6.

2. Related work

Object detection either from images or videos is an important topic in computer vision [19]. Many existing methods have been proposed to detect either objects from a specific category or generic objects without clear category labels from videos [10,8]. Considering the complexity of video data, efficient object localization methods rather than object detection methods also draw a lot of interests, such as sliding window search [7] and optimal path discovery (OPD) [17,18]. Both methods locate objects of a specific category either in images or in videos. On the contrary, to our best knowledge, few methods focus on locating category free objects in videos. In this paper, we target to automatically locate generic objects which draw our interests across a video.

Salient object detection has been broadly explored [2,20]. In order to automatically segment objects from videos, a Markov Random Field model with estimated saliency as priors is used for salient object segmentation [3]. However, the computation complexity is relatively high. An efficient and unsupervised method has been proposed to detect salient video objects by finding maximum saliency density regions [13]. However, it ignores the temporal correlations among the salient objects. Liu et al. proposed to train a Conditional Random Field (CRF) model and then use the model to detect the salient objects [11]. Benefited from the training, they may achieve better performance but lost the generality to handle real data as an unsupervised method does.

Object tracking is also closely related to our work [6]. Instead of tracking a specific object, saliency based object tracking aims to track objects which are salient in a video [5,21,14,1]. Similar to traditional tracking methods, salient objects are first detected by the trained CRF model and then tracked by the Particle Filters [21]. Mahadevan et al. proposed to track salient objects by first learning discriminative features to describe foreground objects in previous frame and then detecting the salient objects by the responses of the learned features in current frame [14]. As a further improvement of the aforementioned two methods, the descriptors of the background and the objects are first learned and particle filters are used for tracking in [1]. Different from them, our method targets to first detect and then locate the salient objects in every frame in an unsupervised way.

3. From video saliency map to salient video object

In this section, we first introduce our video saliency estimation method, then the candidate regions of the salient objects on the obtained video saliency map are accordingly defined and presented.

3.1. Video saliency estimation

Considering the achievement on image saliency estimation and the main difference among images and videos relying on the