Journal of Visual Communication and Image Representation 51 (2018) 131–143

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Video saliency detection via bagging-based prediction and spatiotemporal propagation☆

Xiaofei Zhou[a,b], Zhi Liu[a,b,*], Kai Li[a,b], Guangling Sun[b]

[a] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[b] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Keywords:
Spatiotemporal saliency
Unconstrained video
Bagging
Prediction
Propagation

ABSTRACT

The task of spatiotemporal saliency detection is to distinguish the salient objects from background across all the frames in the video. Although many spatiotemporal models have been designed from various aspects, it is still a very challenging task for handing the unconstrained videos with complicated motions and complex scenes. Therefore, in this paper we propose a novel spatiotemporal saliency model to estimate salient objects in unconstrained videos. Specifically, a bagging-based saliency prediction model, *i.e.* an ensembling regressor, which is the combination of random forest regressors learned from undersampled training sets, is first used to perform saliency prediction for each current frame. Then, both forward and backward propagation within a local temporal window are deployed on each current frame to make a complement to the predicted saliency map and yield the temporal saliency map, in which the backward propagation is constructed based on the temporary saliency estimation of the following frames. Finally, by building the appearance and motion based graphs in a parallel way, spatial propagation is employed over the temporal saliency map to generate the final spatiotemporal saliency map. Through experiments on two challenging datasets, the proposed model consistently outperforms the state-of-the-art models for popping out salient objects in unconstrained videos.

1. Introduction

The human visual system (HVS) can effortlessly capture visually salient objects in the complicated static and dynamic scenes using the inherent visual attention mechanism. The traditional saliency model was based on biologically plausible architecture [1] and feature integration theory [2], and was exploited to predict human fixations [3,4]. Afterwards, saliency models have been extended to estimate salient objects and a number of models based on different theories have been proposed in the past decades. Meanwhile, saliency models have benefited a wide range of applications such as salient object detection and segmentation [5–10], content-aware image/video retargeting [11–13], content-based image/video compression [14–16], image/video quality assessment [17,18,77], and visual scanpath prediction [19,20].

In recent years, saliency model for images is a booming topic, and a lot of efforts have been made and some recent benchmarks have been reported [21,22]. Meanwhile, compared to images, the temporal information in videos is a crucial cue for spatiotemporal saliency model. Recently, the research on spatiotemporal saliency models for videos also received increasing attention. Certainly, many prior efforts have been made from various aspects such as the center-surround scheme [23–29], information theory [30–32], control theory [33,34], frequency domain analysis [35,36], machine learning [5,37–40], sparse representation [41–44], information fusion [45–56], and regional saliency computation [58–62]. Although the aforementioned progress can achieve the decent effect to a certain degree, their performances will degrade when handling a variety of unconstrained videos with complicated motion and complex scenes such as nonlinear deformation, fast motion, dynamic background, and occlusion. Specifically, the spatiotemporal saliency maps generated using these models are insufficient to highlight salient objects uniformly and suppress background effectively.

With the aim to improve the performance of saliency detection in unconstrained videos, this paper proposes a novel spatiotemporal saliency model to effectively separate salient objects from background. Our model proceeds on a per-frame basis, operates within a local temporal window centered on each current frame, and is guided by the outputs of previous frames towards the salient objects in the following frames. Besides, in contrast to previous models, the salient object mask

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China.
E-mail addresses: liuzhisjtu@163.com, liuzhi@staff.shu.edu.cn (Z. Liu).

https://doi.org/10.1016/j.jvcir.2018.01.014
Received 7 September 2017; Received in revised form 24 December 2017; Accepted 20 January 2018

of the first frame is given as a prior in our model, which serves as a saliency prior and indicates the potential salient regions. The advantages of our model lie in the following three aspects:

Firstly, inspired by visual tracking [63,64], we extend online learning to saliency detection and learn prediction models using the random forest regressor based on previous frames. However, the class imbalance problem often occurs on the training data. It has a negative effect on the performance of prediction model, since the model tends to be overwhelmed by the majority class and ignores the minority class. Some efforts try to solve the class imbalance problem based on an ensemble based method [72] or a cost-sensitive learning strategy [73]. To mitigate the aforementioned issues, we propose a bagging-based prediction model, which ensembles the regressors built on multiple undersampled training sets, to perform saliency prediction. The predicted saliency map can indicate most part of salient objects on each current frame.

Secondly, motivated by previous works [58–62], a novel bidirectional temporal propagation method, operated within a local temporal window, is constructed to enhance temporal consistency and complement to saliency prediction. Specifically, the forward propagation, which is constructed on the obtained final spatiotemporal saliency maps of previous frames, is first performed for each current frame, and then the backward propagation is executed in a particular way, which is constructed on the temporary saliency estimation of the following frames. The generated temporal saliency map discovers more salient object regions compared to the predicted saliency map.

Thirdly, reviewing previous works such as graph based image saliency models [65] and two-phase spatial propagation in [60], it can be seen that the graph for saliency detection is usually constructed using appearance information (color, texture, etc.) only. However, the motion information between frames is also an important cue for video saliency detection. Driven by this point, we propose a novel spatial propagation method, which brings the two complementary sources of information together in a unified manner. Concretely, two graphs are first constructed using color and motion features. The propagation is performed over the temporal saliency map in a parallel way on the two graphs. The output generally shows substantially stronger results with the better highlighted salient objects and suppressed background regions.

Overall, the main contributions of this paper are threefold:

(1) Driven by the online learning and the class imbalance problem, we propose a bagging-based saliency prediction model, *i.e.* an ensembling regressor, which is the integration of random forest regressors learned from multiple undersampled training sets.
(2) To enhance temporal consistency and complement to saliency prediction, we propose a novel bidirectional temporal propagation method. In particular, the backward propagation to each current frame is constructed based on the temporary saliency estimation of the following frames.
(3) To fully exploit the complementary effect between appearance and motion information in a unified manner, we propose a novel spatial propagation method, which is performed via two graphs based on appearance and motion, respectively, in a parallel way.

The rest of this paper is organized as follows. The related work is reviewed in Section 2, and the proposed model is described in Section 3. Experimental results and analyses are presented in Section 4, and conclusions are given in Section 5.

2. Related work

The research on saliency detection for still images has continued for decades and amounts of effective models have been proposed as mentioned in [21,22]. For example, the incorporation of low-level and high-level prior learning is employed by [78] to compute the visual saliency. In [79], the manifold ranking-based matrix factorization model is proposed to incorporate the features extracted from each superpixel. In [82], a saliency integration approach via the use of similar images is proposed to elevate the saliency detection performance. Besides, the deep-learning based models [75,80] and multiple-instance learning based models [81] are proposed to improve the saliency detection performance. Generally speaking, the aforementioned efforts major in image saliency detection, thus they are inappropriate to perform video saliency detection. According to the issue of this paper, in this section, we will mainly review the state-of-the-art spatiotemporal saliency models designed from various aspects in the recent years.

As a pioneering work, Itti et al. [3] proposed a well-known center-surround scheme, which exploits luminance, color and orientation across different scales to generate the saliency map. Subsequently, in [23], a surprise model was further designed, in which a set of features including color, luminance, orientation, flicker and motion energy are exploited. Along this way, there are also some other models which estimate the difference of each patch/volume and its spatiotemporal surroundings. Based on the discriminant center-surround hypothesis [24,25], the Kullback-Leibler divergence on dynamic texture feature is exploited in [26]. In [27], local regression kernels based self-resemblance is used to measure saliency. In [18], a multiscale background model represented by Gaussian pyramid is used to detect objects undergoing salient motion. In [28], the earth mover's distance (EMD) is used for computing the center-surround difference in the spatiotemporal receptive field. The contrast of luminance and directional coherence is adopted to measure spatiotemporal saliency in [29].

Besides the classical center-surround scheme based models, lots of spatiotemporal saliency models, which are based on various mechanisms including information theory, control theory, frequency domain analysis, machine learning, sparse representation and the fusion scheme of spatial and temporal saliency, are also proposed in recent years.

The knowledge of information theory can be used for saliency measurement. For instance, self-information in [30] is used to represent the object saliency. The minimum conditional entropy in [31] follows a local approach to estimate saliency. The incremental coding length in [32] is exploited to measure the perspective entropy gain of each feature and achieve attention selectivity. On the basis of control theory, some effective works first model the video sequence as a linear dynamic system, and then use the observability of output [33] and the controllability of states [34] to measure salient motion, respectively.

The frequency domain analysis method is also exploited for video saliency detection. Inspired by spectral residual for image saliency detection [35], the temporal spectral residual on video slices along X-T and Y-T planes is exploited to perform motion saliency detection in [36]. Except for spectral residual, a spatiotemporal saliency model based on phase spectrum of quaternion Fourier transforms is proposed in [14].

With the increasing attention on machine learning, some efforts have also been attempted for video saliency detection. In [37], stimulus-driven and task-related factors are exploited to model video saliency under the framework of probabilistic multi-task learning. A standard soft margin support vector machine with Gaussian kernels is adopted to predict interesting locations in [38]. In [39], both low-level and high-level features are used to predict visual attention from videos by using support vector regression. In [40], the one-class support vector machine is used to remove the consistent trajectories in motion, yielding salient regions in videos. Besides, the conditional random field (CRF) is exploited to integrate multiscale contrast, center-surround histogram and spatial distribution of color and motion vector field in [5].

The sparse representation method is also adopted in spatiotemporal saliency models. In [41], a sparse feature selection model, which incorporates temporal consistence and temporal difference, is used to generate saliency maps for videos. Besides, the regularized feature reconstruction [42] and the sparse low-rank decomposition [43,44] are employed in video saliency measurement.