Computer Vision and Image Understanding 135 (2015) 126–140

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Kernel regression in mixed feature spaces for spatio-temporal saliency detection ☆

Yansheng Li, Yihua Tan *, Jin-Gang Yu, Shengxiang Qi, Jinwen Tian

School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 21 May 2014
Accepted 30 January 2015
Available online 7 February 2015

Keywords:
Spatio-temporal saliency
Kernel regression
Mixed feature spaces
Hybrid fusion strategy

ABSTRACT

Spatio-temporal saliency detection has attracted lots of research interests due to its competitive performance on wide multimedia applications. For spatio-temporal saliency detection, existing bottom-up algorithms often over-simplify the fusion strategy, which results in the inferior performance than the human vision system. In this paper, a novel bottom-up spatio-temporal saliency model is proposed to improve the accuracy of attentional region estimation in videos through fully exploiting the merit of fusion. In order to represent the space constructed by several types of features such as location, appearance and temporal cues extracted from video, kernel regression in mixed feature spaces (KR-MFS) including three approximation entity-models is proposed. Using KR-MFS, a hybrid fusion strategy which considers the combination of spatial and temporal saliency of each individual unit and incorporates the impacts from the neighboring units is presented and embedded into the spatio-temporal saliency model. The proposed model has been evaluated on the publicly available dataset. Experimental results show that the proposed spatio-temporal saliency model can achieve better performance than the state-of-the-art approaches.

1. Introduction

Visual saliency modeling has been widely considered to be a promising approach to automatically and efficiently localizing the "important" content in images or videos [1–3]. While saliency detection from still images depends only on spatial information (i.e., the current scene), for the task of saliency estimation from videos, temporal information (i.e., the accumulative information along time) also plays an important role, which is commonly termed as *spatio-temporal saliency* in the literature [4]. As highlighted in [4], spatio-temporal saliency modeling provides a natural way to identify important regions from dynamic scenes, which can benefit a wide range of computer vision applications, such as action recognition [5], moving object detection in the stationary or dynamic background [6–8], and image/video compression [9].

A lot of spatio-temporal saliency models have recently been proposed, which in general fall into two categories, namely the data-driven *bottom-up* models [2,3,5–15] and the task-oriented *top-down* models [16–18]. The later category of models usually resorts to supervised learning, with a pre-determined task and the corresponding training data, while the former does not rely on specific tasks. In this paper, we are mainly focused on bottom-up saliency modeling for the purpose of detecting salient moving object (region) from videos, which desirably can facilitate the computer vision tasks such as moving object detection [6–8] and visual tracking [19].

The vast majority of existing computational models for spatio-temporal saliency detection [2,3,6,7,10,12–15] follow the paradigm that the spatial saliency and the temporal saliency are first measured separately and then fused to obtain the final saliency map. Neurobiological evidence supporting this paradigm can be found in [20], which suggests that visual information processing in the brain goes along two parallel and concurrent streams, namely the ventral stream and the dorsal stream. The former has high spatial sensitivity and mainly processes appearance information, while the latter has high temporal sensitivity and mainly processes motion information. Consequently, the hypothetical segregation of the two streams naturally raises a question, i.e., the so-called visual integration problem: how does the cortex combine the information from the two streams? In spatio-temporal saliency models, in addition to the spatial and temporal saliency calculation modules which simulate the perceptual functionality of the ventral and the dorsal streams, there should also be a fusion strategy for simulating the visual integration functionality. In previous works [2,3,6,7,10,12–15], the fusion problem is typically

☆ This paper has been recommended for acceptance by John K. Tsotsos.

* Corresponding author. Fax: +86 027 87556301.
E-mail address: yhtan@hust.edu.cn (Y. Tan).

http://dx.doi.org/10.1016/j.cviu.2015.01.011
1077-3142/

addressed by performing multiplicative, additive, or maximum operations over the corresponding perceptual units independently, which totally ignores the connection among the neighboring units. While these fusion strategies might be appropriate for eye fixation prediction as discussed in many previous works, they are disadvantageous for those tasks which require highlighting foreground regions uniformly. In addition, the over-simplicity of existing fusion strategies may be one of the key reasons why existing computational models underperform the human vision system in predicting salient moving objects. These facts motivate us to develop a more effective approach for fusing the spatial and the temporal saliency.

Our proposed spatio-temporal saliency model follows the aforementioned paradigm while being mainly focused on the fusion strategy. Due to its computational efficiency and perceptual superiority [21–23], superpixel is taken as the basic computational unit in our model. Similar to [24], each unit is represented by a feature vector defined in the mixed feature space, that is, the Cartesian product of the location feature subspace, the appearance feature subspace and the temporal feature subspace. Spatio-temporal saliency can then be treated as a function value of the mixed feature, which can be learnt by means of regression. In order to accommodate the locality and multi-modality of the mixed feature space, we first propose the model of *kernel regression in mixed feature spaces (KR-MFS)*, which includes three variants: kernel regression using local constant approximation in mixed feature space (KR-LC-MFS), kernel regression using local regularized linear approximation in mixed feature space (KR-LRL-MFS) and kernel regression using local regularized kernel approximation in mixed feature space (KR-LRK-MFS). In addition, the relationship between KR-MFS and the classical kernel regression [25–28] is discussed. Then, based on KR-MFS, a hybrid fusion strategy is proposed. Different from the existing fusion strategies treat each unit individually, the proposed fusion strategy is able to exploit the spatial structure of the neighboring units. Finally, by integrating the spatial and temporal saliency calculation modules and the hybrid fusion module, we propose a unified spatio-temporal saliency model, which can outperform the state-of-the-art approaches on the benchmark datasets. To sum up, the contributions of this paper are as follows:

- A generic model, i.e., kernel regression in mixed feature spaces (KR-MFS), and its three variants, i.e., KR-LC-MFS, KR-LRL-MFS and KR-LRK-MFS, are proposed, which can be potentially applied to many real-world problems besides the spatio-temporal saliency detection as concerned in this paper.
- Based on KR-MFS, a hybrid fusion strategy is proposed. The key property of the proposed fusion strategy is that it is able to take into account the influences from the neighboring units.
- By the aide of the hybrid fusion strategy, our proposed spatio-temporal saliency model can outperform the state-of-the-art approaches. Additionally, our model is well modularized and therefore has good extendibility.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 details the proposed KR-MFS model and the variants. Section 4 describes the spatio-temporal saliency model using KR-MFS. In Section 5, our saliency model is comprehensively evaluated and compared against the state-of-the-art algorithms. Section 6 gives the conclusion of this paper.

2. Related work

In this section, we briefly review the previous works including the saliency measure and fusion strategies for spatio-temporal saliency modeling, and kernel regression and its applications in computer vision. Note that top-down saliency modeling is excluded from our literature review here since we only concentrate on bottom-up saliency in this paper.

2.1. Spatio-temporal saliency measurement

As mentioned before, a majority of existing spatio-temporal saliency models [2,3,6,7,10,12–15] have adopted the paradigm, that the spatial and the temporal saliency are first computed separately and then fused. In [1,2], the Difference of Gaussian (DoG) filter with center-surround mechanism is utilized to measure saliency. Itti et al. deploy the spatial and temporal surprise to define saliency and detect salient events in video, where surprise is mathematically computed by Kullback–Liebler (KL) divergence [3]. Liu et al. utilize a spatio-temporal volume to integrate the spatial and temporal saliency, where the spatial and temporal saliency is measured by Shannon's self-information [6]. In [7], spatial saliency is measured by self-resemblance of the texture descriptors, while the temporal saliency is computed by the sum of absolute difference between temporal gradients of the center and the surrounding regions. Zhai and Shah define temporal saliency as motion contrast based on the estimated point correspondences using RANSAC and spatial saliency as global color histogram contrast [10]. In [12], cortical-like filters corresponding to the two separate outputs of the retina, namely parvocellular and magnocellular, are utilized to compute the spatial and temporal saliency respectively. Ren et al. apply feature reconstruction error to measure spatial and temporal saliency [13]. In [14], temporal saliency is computed based on dynamic consistent optical flow and spatial saliency is computed by integrating the activation maps from multiple low-level feature channels. Recently, Fang et al. exploit a psychological model for speed perception which is further utilized to estimate temporal saliency; meanwhile spatial saliency is computed by averaging multiple saliency maps corresponding to multiple spatial features [15].

There also exists spatio-temporal saliency models which do not follow the aforementioned paradigm, like in [8,9,11]. In these methods, the spatial and temporal saliency calculation and the fusion modules are integrated as a whole. Mahadevan and Vasconcelos extend the discriminant center-surround hypothesis from image to video by using dynamic texture to model the spatio-temporal characteristics of the visual stimuli [8]. In [9], the spatial and temporal features are first extracted, and then the spatio-temporal saliency is measured by Quaternion Fourier Transform. Seo and Milanfar [11] adopt the local regression kernel to represent each pixel/voxel, where the center-surround contrast based on the resemblance of a pixel/voxel to its surroundings is utilized to measure spatio-temporal saliency.

Generally, the overwhelming majority of spatio-temporal saliency models follow the aforementioned paradigm. Our proposed model also follows the same paradigm.

2.2. Spatio and temporal saliency fusion

In [2], the motion channel is simply treated as an ordinary one like the other spatial feature channels such as color, intensity, and orientation, and then all these channels are normalized and combined into the final saliency map. By fitting the neural data, the additive fusion scheme is recommended in [3]. Liu et al. propose a weighted additive fusion approach, where weights are estimated from the spatio-temporal volumes [6]. By considering the influence of scale, Kim et al. propose a multi-scale additive fusion method [7]. In [12,13], some common fusion approaches such as multiplicative, additive, and maximum fusion, are evaluated and discussed. A maximum fusion strategy is also introduced in [14]. The work of