Neurocomputing 207 (2016) 178–188

Contents lists available at ScienceDirect

Neurocomputing

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

# Perception-oriented video saliency detection via spatio-temporal attention analysis

Sheng-hua Zhong [a,b], Yan Liu [c,*], To-Yee Ng [c], Yang Liu [d]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shen Zhen, PR China
[b] Department of Psychological & Brain Sciences, The Johns Hopkins University, Baltimore, MD, US
[c] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, PR China
[d] Department of Computer Science, The Hong Kong Baptist University, Hong Kong, PR China

ARTICLE INFO

*Article history:*
Received 7 December 2015
Received in revised form
7 March 2016
Accepted 27 April 2016
Communicated by L. Shao
Available online 11 May 2016

*Keywords:*
Perception-oriented video saliency
Spatio-temporal modeling
orientation inhomogeneous feature map
Dynamic consistency
Visual attention

ABSTRACT

Human visual system actively seeks salient regions and movements in video sequences to reduce the search effort. Computational visual saliency detection model provides important information for semantic understanding in many real world applications. In this paper, we propose a novel perception-oriented video saliency detection model to detect the attended regions for both interesting objects and dominant motions in video sequences. Based on the visual orientation inhomogeneity of human perception, a novel spatial saliency detection technique called visual orientation inhomogeneous saliency model is proposed. In temporal saliency detection, a novel optical flow model is created based on the dynamic consistency of motion. We fused the spatial and the temporal saliency maps together to build the spatio-temporal attention analysis model toward a uniform framework. The proposed model is evaluated on three typical video datasets with six visual saliency detection algorithms and achieves remarkable performance. Empirical validations demonstrate the salient regions detected by the proposed model highlight the dominant and interesting objects effectively and efficiently. More importantly, the saliency regions detected by the proposed model are consistent with human subjective eye tracking data.
© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

Visual perception is an active process to interpret the surrounding environment by processing information contained in visual light [1]. Theories and observations of visual perception have been the main source of inspiration for computer vision and artificial intelligence. Perceptual models play an increasingly significant role in optimizations of various applications, such as perceptual-based video coding [2], quality assessment [3], real-time user authentication [4], sound classification [5], and audio watermarking [6].

Visual attention analysis plays an important role in perception-oriented modeling and attracts interest from a broad range of researchers and scientists. First, attention is the behavioral and cognitive process of selectively concentrating on one aspect of the environment while ignoring other things [7]. In human, attention is facilitated by a retina that has evolved a high-resolution central fovea and a low-resolution periphery [8], which is also closely related to multiple other cognitive processes, such as: perception, memory, and learning. Hence, the research on visual attention analysis provides a window on the human perception and other cognitive processes. Second, based on the eye tracking data recorded by the high-speed eye tracker, attention is more easily to be observed than other cognitive processes. Third, the research on attention analysis provides an effective means to connect human perception and the various applications in further processing [9], including image quality assessment [10], object detection [11], action recognition [12], image retargeting [13], video abstraction [14], removing label ambiguity in image [15], etc.

The strongest attractors of attention are stimuli that pop-out from their neighbors in space or time usually referred to as “saliency” [16]. Visual attention analysis simulates the human visual system behavior by automatically producing saliency maps of the target image or video sequence [17]. The saliency map is proposed to measure the conspicuity and calculate the likelihood of a location in visual data to attract attention [18]. Therefore, the visual saliency detection provides predictions on which regions are likely to attract observers' attention [19]. Although image saliency detection has been long studied, little work has been extended to video sequences due to the data complexity. After the standard real world video datasets with subjects' eye tracking data emerge, such as the video action dataset [20], a more detailed and quantitative research for video saliency detection and analysis will be feasible.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2016.04.048
0925-2312/© 2016 Elsevier B.V. All rights reserved.

The conference version of our preliminary work was published in [21]. This work demonstrates good performance on saliency detection based on dynamic consistency of motion. But in spatial saliency modeling, it inherits the classical bottom-up spatial saliency map. In this paper, we propose a novel perception-oriented video saliency detection method called spatio-temporal attention analysis model (STAM) by referring to the characters of the human visual system. The STAM follows the three-part scheme of video saliency detection, including spatial saliency detection, temporal saliency detection, and the fusion of spatial and temporal saliency maps. In feature extraction stage of spatial saliency detection, multiple low-level visual features including: intensity, color, orientation, and contrast are extracted at multiple scales. Instead of using the original orientation feature map, we propose a novel technique called visual orientation inhomogeneous saliency model (VOIS). In our orientation feature map, the information in cardinal orientations is retained, but the information in oblique orientations is weakened with the inhomogenous weight. Then, the activation maps are built based on multiple low-level feature maps. And the saliency map is finally constructed by a normalized combination of the activation map. In temporal saliency map modeling part, a novel dynamic consistent optical flow model (DCOF) is proposed based on the human visual dynamic continuity. Different from the classical optical flow model which estimates motion between each adjacent frame pair independently, the proposed DCOF takes account of the motion consistency in video sequence. In saliency fusion stage, the "skew-max" fusion method is utilized to fuse the spatial and temporal saliency maps together and construct the final video saliency map.

In the following parts of this paper, we discuss the related work on video saliency detection in Section 2. A novel spatio-temporal video saliency detection technique is introduced in Section 3. In Section 4, we demonstrate the performance of the proposed video saliency detection model on three video sequence datasets. The paper is closed with conclusion and future work in Section 5.

## 2. Related work on saliency detection

Video saliency detection calculates the salient degree of each location by comparing with its neighbors both in spatial and in temporal areas. Previously, most existing computational saliency models depend on the intrinsic bottom-up spatial features of the visual stimuli by referring to the human visual system [22,23]. Neurophysiological experiments have proved that neurons in the middle temporal visual area (MT) compute local motion contrast. Such neurons, which underlie the perception of motion pop-out and figure-ground segmentation, influence the attention allocation [24]. After realizing the importance of motion information in video attention, the motion feature has been added into the saliency models [25,26]. Recently, to simulate two pathways (parvocellular and magnocellular) of the human visual system, the video saliency detection procedure is divided into spatial and temporal pathways [27]. These two pathways (the P and M pathways) correspond to the static and dynamic information of video. In P pathway, parvocellular cell has greater spatial resolution, but lower temporal resolution. Conversely, in M pathway, magnocellular cell has greater temporal resolution, but lower spatial resolution.

Typically, in spatial pathway saliency detection, most of the video saliency techniques follow the classical image saliency architecture including three stages: feature extraction, activation, and normalization. Multiple low-level visual features such as intensity, color, orientation, and contrast are firstly extracted at multiple scales. Then, the activation maps are built based on multiple low-level feature maps. After the activation maps are computed, they are normalized and combined into a spatial saliency map that represents the saliency of each pixel [28]. Almost all of the existing bottom-up models are inspired by the theories from human visual system [29]. Among them, the most famous one was proposed by Itti et al. [30]. They developed the center surround structure akin to on-type and off-type visual receptive field. We denote this model as ITTI in our experiment. In recent years, more proposed work simulated the multi-scale and multi-orientation function of primary visual cortex. Achanta et al. detected the saliency map with a Difference of Gaussians (DOG) model to describe the spatial properties of visual regions [31]. Gabor filters and Log-Gabor wavelets were utilized to explore the salient features such as spatial localization, spatial frequency characteristics in [32] and [29], respectively.

In temporal pathway saliency detection, optical flow is the most widely used method in existing video saliency detection models [20,31,33]. These models rely on the classical optical flow method to extract the motion vector between each frame pair independently as the temporal saliency map. The classical formulation of optical flow was first introduced by Horn and Schunck [34]. They optimized a functional based on residuals from the brightness constancy constraint, and a regularization term expressing the smoothness assumption of the flow field. Black and Anandan further addressed the outlier sensitivity problem of initial optical flow model by replacing the quadratic error function with a robust formulation [35]. Although different efforts have been put into improving the optical flow, the median filtering is the most important source to improve the performance of the classical optical flow model [36]. According to the extensive test by [37], the median filtering makes non-robust methods more robust and improves the accuracy of the optical flow models. Unfortunately, although the optical flow techniques can accurately detect the motion in the direction of intensity gradient, the temporal saliency is not perfectly equal to the amplitude of all the motion between each adjacent frame pair. Indeed, only the continuous motion of the prominent object should be popped out as the indicator of the temporal salient region.

## 3. Spatio-temporal attention model

In this section, we propose a novel spatio-temporal attention analysis model (STAM). The schematic illustration of the proposed STAM is described in Fig. 1.

The whole spatio-temporal attention analysis model can be partitioned into two pathways. In spatial saliency map construction, we follow the three common stages of the classical bottom-up spatial saliency map. A novel spatial saliency map technique called visual orientation inhomogeneous saliency model (VOIS) is proposed in Section 3.1. In VOIS, we will provide a human-like orientation feature map extraction based on the visual orientation inhomogeneity of human perception. In temporal saliency map construction, a novel dynamic consistent optical flow model (DCOF) is proposed in Section 3.2 based on the human visual dynamic continuity. Different from the classical optical flow model estimates motion between each adjacent frame pair independently, DCOF both underlines the consistency of motion saliency in the current frame and between the consecutive frames. In Section 3.3, we simply adopt the "skew-max" fusion method from existing work to obtain the final video saliency map.

### 3.1. Spatial saliency map construction

The leading models of spatial saliency map construction can be divided into three stages:

Download English Version:

https://daneshyari.com/en/article/494449

Download Persian Version:

https://daneshyari.com/article/494449

Daneshyari.com