Optics & Laser Technology 80 (2016) 1–15

Contents lists available at ScienceDirect

Optics & Laser Technology

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

Full length article

Motion saliency detection using a temporal fourier transform

Zhe Chen [a,b], Xin Wang [a], Zhen Sun [a], Zhijian Wang [a,*]

[a] College of Computer and Information Engineering, Hohai University, Nanjing, China
[b] Key Laboratory of Trusted Cloud Computing and Big Data Analysis, Nanjing Xiaozhuang University, Nanjing, China

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form
1 December 2015
Accepted 14 December 2015
Available online 28 December 2015

Keywords:
Motion saliency
Temporal Fourier Transform
Moving object detection

ABSTRACT

Motion saliency detection aims at detecting the dynamic semantic regions in a video sequence. It is very important for many vision tasks. This paper proposes a new type of motion saliency detection method, Temporal Fourier Transform, for fast motion saliency detection. Different from conventional motion saliency detection methods that use complex mathematical models or features, variations in the phase spectrum of consecutive frames are identified and extracted as the key to obtaining the location of salient motion. As all the calculation is made on the temporal frequency spectrum, our model is independent of features, background models, or other forms of prior knowledge about scenes. The benefits of the proposed approach are evaluated for various videos where the number of moving objects, illumination, and background are all different. Compared with some the state of the art methods, our method achieves both good accuracy and fast computation.

1. Introduction

The first row of Fig. 1 shows a temporal sequence of urban images acquired at street level. These images could be described individually using localized identifiers such as “many trees,” “an SUV,” “a pedestrian,” “a street lamp,” “a billboard,” “many traffic lanes,” or “many manhole covers.” Each of these descriptions is correct and represents the major essence of the image-what most people would identify as important or salient.

Detecting these salient regions automatically is a profound challenge in computer vision. In an attempt to solve this problem, most traditional object detectors need to establish either the object feature or the background model [1,2,3]. However, in contrast to the attention mechanism of animal vision, these methods are generally complex computationally. Hence, in order to rapidly focus on generic salient objects, state-of-the-art methods mathematically imitate the attention mechanism of humans or primates. Conventionally, biologically inspired methods focused on identifying fixation points that a human viewer would focus on at first glance. Since fixation point extraction is commonly based on advanced features, these methods are limited to high-quality images [4–9]. Other techniques, which are independent of feature points, work in the frequency domain. These feature-free methods rely on variations in the frequency spectrum which are in line with the “pop-out” display in the image [10,11,12]. Generally, these methods are superior for processing static images and have the ability to extract dominant objects as shown in the third row of Fig. 1.

Upon inspection, it is obvious that the frames in Fig. 1 were captured from a video clip. Hence, if these frames are jointly described in the time scale, the description could be “a pedestrian walks along the sidewalk.” In contrast to the first description of individual images, the last description includes nothing but the moving pedestrian as the salient object, as shown in the bottom row of Fig. 1. Obviously, most of the attention is focused on dynamic (motion) information in the scene while static objects such as trees, the SUV, and the billboard are spontaneously regarded as “background.” This type of motion saliency detection mechanism is more suitable for applications where motion information between frames is dominant, such as object tracking [13] or video compression [14]. However, very little research is conducted in this area.

This paper proposes a novel method for motion saliency detection. The underlying principle is the correspondence observed in temporal sequences between variations in the phase spectrum and the time-varied “pop-out.” Hence, regions containing moving objects are marked salient by calculation of the phase spectrum. Different from the complex background modeling methods and previous works on the frequency spectrum of images and videos, the proposed motion saliency detection method is most computationally efficient and gives better or comparative results.

We demonstrate the utility of our motion saliency detection method by applying it to moving object detection [15,16,17]. We

* Corresponding author.
E-mail address: zhjwang@hhu.edu.cn (Z. Wang).

http://dx.doi.org/10.1016/j.optlastec.2015.12.013
0030-3992/

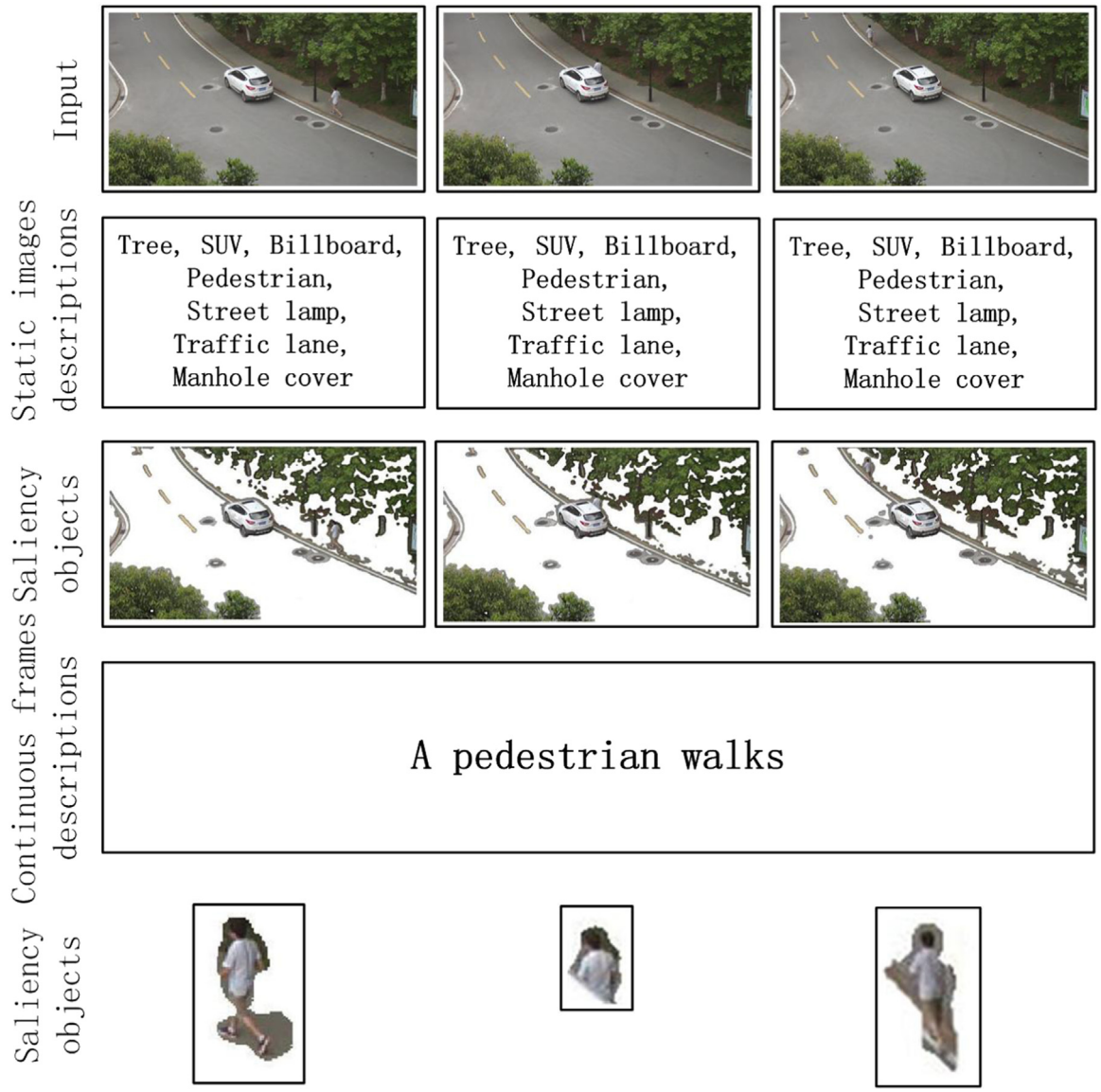

Fig. 1. Our motion saliency results (bottom) comply with the descriptions that people provided (samples in the second row) for the input images (top). People tended to describe the scene according to the motion information in consecutive frames, as shown in the fourth row. Extracting the dominant objects in each single image, as shown on the third row, might miss the essence of the scene. Conversely, we maintain the key works in the video while all of the vapid information is abandoned.

show that our technique can successfully mark the moving objects. Moreover, by changing the saliency threshold, the disturbance caused by the moving shadow and the shaking background can to a large extent be removed.

In Section 3, we propose a phase spectrum model for consecutive frames. In Section 4, the method for phase spectrum calculation is introduced to obtain saliency maps. Several experimental results comparing our method with others are shown in Section 5, and the conclusion and discussion are given thereafter.

2. Related work

Many visual attention approaches have been proposed for static saliency detection. Generally, these approaches can be divided into two subcategories. One employs low-level image features. Examples include the popular Itti model [18] and its updated editions [19]. Inspired by the attention mechanism of primates, these methods are initialized by searching a few of the fixation points as saliency and salient regions are marked by the numerous points generated through attentional shifts [20]. Input from these methods is comprised of local image features, such as intensity, color, orientation, and texture. As these methods are based on local features, local saliency is excessively emphasized. The disadvantage of local methods can be seen clearly in Fig. 2(b), where the high local contrast includes too many transitions within the background. As a result, most fixation points focus on the background while only a few locate the object of interest. To solve this problem, global features are introduced, such as the center-surround histogram [2], color spatial distribution [2], histogram-based contrast [21], and global-homogeneous information [22]. These global features are either parallely or hierarchically incorporated with the local features. The outperformance using global features for removing background [2], detecting important context [23], and enhancing object saliency [21] has been demonstrated experimentally. Other methods which can prevent too much local saliency operate in the frequency domain. They are commonly called global methods, such as the Spectral Residual (SR) method [10]. The theoretical basis for this category is that variations in the frequency spectrum reflect the saliency in the