J. Vis. Commun. Image R. 38 (2016) 367–377

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Perceptual synoptic view of pixel, object and semantic based attributes of video ☆

Sinnu Susan Thomas *, Sumana Gupta, Venkatesh K. Subramanian

Department of Electrical Engineering, Indian Institute of Technology, Kanpur, 208016, India

ARTICLE INFO

Article history:
Received 5 February 2016
Revised 5 March 2016
Accepted 12 March 2016
Available online 16 March 2016

Keywords:
Surveillance
Attention
Perceptual feature
Synoptic view
Object and semantic attributes
Modeling

ABSTRACT

For a scene, what are the object and semantic based attributes, other than the pixel based attributes, and how do they affect our attentional selection are some of the questions we need to address. We studied the effects of various attributes on our attentional perspective. We described a new saliency prediction model that accounts for different pixel-level attributes as color, contrast and intensity; object level attributes such as size, shape of objects and semantic level attributes as motion and speed of objects. We quantified these attributes based on motion contrast, motion energy and motion chromism. With this in view, we examined the problem of information prioritizing and filtering with emphasis on directing this exercise using object and semantic based attributes of the human attention model. We have evaluated proposed approach on different types of videos for their quantitative and qualitative comparison. The promising results create a gateway for synopsis view.

1. Introduction

Nowadays, with the increase of various crime rates, people are worried about their security and safety. The growth of video surveillance is at a rapid pace due to the need to combat increasing security and safety concerns and to monitor activities to detect intrusion, theft and traffic surveillance. According to a new report from Transparency Market Research the video surveillance market, including Video surveillance-as-a-service, is expected to reach $42.81 billion by 2019, growing at a compound annual growth rate (CAGR) of 19.1% from 2013 to 2019. These cameras work round the clock and extensive monitoring of this huge chunk of data is a tiresome and tedious job. The storage of this crucial data is also an expensive task. The entire video may not be of any importance to us, but certain parts may be. To reduce the scanning time of the entire video, synoptic view of the key points of the video seems a better solution for the problem. Moreover, a synoptic view can lead to a minimal storage space.

Each time we open our eyes, our retinas are bombarded with a breathtaking amount of information. In spite of all this, we have a clear understanding of what we see. This requires separating the wheat from the chaff, selecting meaningful information out of the meaningless one, to achieve which the brain must filter its inputs. Attention is a mechanism which prioritizes some aspects of information by focusing on that point of the scene. Visual attention is prompted by low-level image features (Bottom-Up attention) or based on a specific task (Top-Down attention). Top-down visual attention is goal driven, whereas bottom-up visual attention is stimulus driven (e.g., color, contrast and intensity) [10]. These attended regions highlights the prominent objects in an image.

Psychologists are highlighting the facts of human perception in which a selective attention mechanism is employed to filter out irrelevant information and limit processing to information that are relevant to the present task. Still, much is unknown about the underlying cognitive mechanisms of visual attention. A small child when asked about to separate apples from a bunch of bananas will be quick enough to do it. Undoubtedly, without any specific knowledge, categorizing objects based on their overall perceived similarities seems to be reasonable.

Several factors influence one's attention which is classified as external factors and internal factors. External factors are related to features of stimuli that govern attention like change, size, repetition and movement, whereas internal factors are purely psychological factors like motive of an individual, one's interest, mindset and attitudes. Studies reveal that human attention is very sensitive to sudden stimuli in the scene, but usual stimuli are less responsive in the phenomenon of sensory adaptation. The fixation time of the eye for the features which have a low probability of appearing in a scene is more than the features appearing frequently in a scene

☆ This paper has been recommended for acceptance by Qionghai Dai.

* Corresponding author.
E-mail addresses: sinnu@iitk.ac.in (S.S. Thomas), sumana@iitk.ac.in (S. Gupta), venkats@iitk.ac.in (V.K. Subramanian).

http://dx.doi.org/10.1016/j.jvcir.2016.03.015
1047-3203/

[15]. The size of the stimulus draws our attention very easily [33]. Similarly, humans are more sensitive to a moving object than a stationary one [7,9].

1.1. Background and related work

Research work in the area of visual saliency increased after a mathematical model for finding out the salient areas in the image was proposed. Many studies employed different methods [11] to calculate the saliency measure of objects in an image. [10] builds a space based attention model, derived from Features Integration Theory. It combines low-level features together to evaluate how different a given location is from its surrounding. Itti's model provides a complete computation method for bottom-up attentional competition, including center-surround contrast and across-scale integration in terms of primitive feature dimensions. Conventional saliency modeling mainly focus on pixel-level image attributes such as contrast[23], orientation[10], intensity bispectra [14] and color [23] to generate a saliency map. These pixel level attributes are used to find the salient regions from the video. Pixel-level information failed to encode object and semantic level information which is essential in the contribution of saliency.

Few semantic level attributes were used in [16,18] to create a dynamic video skim based on an attention model. Each frame is assigned with an attention value, which indicates its likelihood for the viewer's attention. Ma et al. [18] used only motion and face as attributes to find an attention value which is not enough to categorize the attention. Walther et al. [30] demonstrated object recognition using computational models of bottom-up visual attention. Longfei et al. [16] used only face and some pixel based attributes to describe the attention model. Lopez et al. [17] defined a model for the generation of attention based on some features like color, motion and shape. Dessing et al. [7] stated that few aspects like position, velocity and temporal information streams combine to give catching movements. Shih et al. [26] used context, motion and face as features to create an attention model. [4] proposed that visual attributes like intensity contrast, color contrast, flicker contrast, motion contrast and integrated saliency are most predictive in selection of attention. However these attributes alone are not enough to define an attention model. Hence, when the application is a synopsis of the video, its legitimate to think beyond the pixels.

Fig. 7 poses a question on Itti's model whether this is enough to differentiate a salient object in a video. Each salient map appears to have almost same attributes, whereas the input frame contains salient objects with other attributes. Apart from classical attributes, is there anything else that makes an object salient? This question motivated us to explore this area. This paper helps us to understand various factors that contribute to saliency in a video and to find the key frames from the video.

1.2. Overview of an approach

This work mainly focuses on the stationary camera scenario which is the case in many applications like visual surveillance. In this work, we assume to have a short video clip with us and our job is to find out the key frames using various stimuli other than the conventional stimuli of Itti's model [10]. Most of the existing saliency models like Itti's model focus on pixel-level attributes.

In this paper, we explicitly focus on creating a saliency model based on object and semantics level attributes like motion, size, speed and shape of an object. The need for various attributes is discussed in detail at first. According to Gestalt psychologists, our perception follows the law of proximity, similarity, closure, symmetry, continuity and past experience. The motion stimuli in a video convey information about the position, motion and appearance of an object that is useful for modeling an attention model. Second, a system model is designed with the attributes like motion contrast, motion energy and motion chromism and they are combined to find the key frames. We then combine these key frames to create a synoptic view. To summarize, the main contributions of this paper are as follows:

- Description of the need for various object and semantics level attributes.
- Proposed attention model based on object and semantics level attributes.
- Key frames selection based on the proposed model.

1.3. Organization

The remainder of the paper is organized as follows. Section 2 highlights the need of different object and semantic based attributes for proposed work. In Section 3, model based on the object and semantic level attributes is proposed. Simulation results are presented in Section 4. Section 5 highlights the limitations and future prospects of proposed work. Section 6 discusses the conclusion of this paper.

2. Why other attributes for perception?

Attention is a cognitive process in which the recipient stimuli are an accumulation of various senses on individual consciousness. When we perceive things from a scene, it is not necessary that all stimuli are in focus of one's consciousness- some may be and some may not be. The stimuli that are in less focus are vaguely perceived or attended compared to the stimuli which are in focus and perceived more. Each visual feature contains some information and has the ability to encode data in isolation. Different features can interact with one another in various ways when they are displayed together in a common image [9]. The factors influencing attention are broadly classified as external and internal factors. The paper is confined to the external factors only and internal factor is beyond the scope of this paper.

Object like attributes include external factors as

1. *Size of the stimulus.* Size is an object level attribute which affects our perceptual saliency. Experiments performed by [32] showed that the search time for an object that differed in size took ten seconds longer than an object which differed in color. [21] carried out studies on how consumers get attracted to an advertisement when the size of the product is made large. Studies show that object size plays a significant effect on fixation likelihood, but the influence of surface size on attention remains very strong. [3] studied how object size affects object perception. They measured visual sensitivity by varying the widths of objects with fixed height and found that humans can adapt to variations in wider objects more easily than narrower objects. If more objects are there in a scene, less attention is directed to any object. If the scene contains fewer objects, then the effect of increasing object size on attention is stronger. If all objects have the same size and distance from each other, each object will receive 1/N measures of attention [21]. But, with increasing size, differences in relative attention devoted to central or peripheral objects can be observed. Larger regions are more likely to attract more attention than smaller ones.
2. *Shape of the stimulus.* Kovács and Julesz [13] proved that contour completion and shape is an important phenomena of human gaze. In natural scenes objects appear to be embedded in complex cluttered backgrounds. The detection of an object is difficult when it is occluded in a scene. An important task of the visual system is to segment local elements from their context