Spatio-Temporal Visual Saliency Detection Model for High Dynamic Range Content

Lini P Joy1 Meera Panicker P R2 Hari S3
1Student 2,3Assistant Professor
1,2,3Department of Electronic & Communication Engineering
1Mount Zion College of Engineering, Kadammanitta, Pathanamthitta, Kerala, India

Abstract— The spatio-temporal visual system is a computational approach to model the bottom-up visual saliency for HDR input by combining spatial and temporal visual features. The main advantage of this system is that it will reduce the cognitive processing efforts. Computational models of visual attention can be applied to areas such as computer graphics, video coding and quality assessment. The proposed model stands apart from the existing models in the way that it is the only model applicable to HDR videos. Utilizing this method will allow to locally adjust the contrast of HDR images and videos according to the areas of interest provided by the saliency map. This spatio-temporal model provides information about visually important areas. With this information, compression methods for HDR could be more effective by allocating more bit rate resources to visually important areas of each frame and less to the rest of the frame. In this method both spatial and temporal cues are taken into account leading to two saliency maps: the spatial saliency map and the temporal saliency map. A dynamic fusion method is proposed to combine both the spatial and temporal saliency maps. The saliency predictions proposed by this method are evaluated through data collected from eye tracking experiments using an HDR prototype display. Performance evaluations using three quantitative metrics show that the proposed model outperforms the existing state-of-the-art models.

I. INTRODUCTION

Saliency is the state or quality by which something stands out relative to its neighbours. Saliency detection is important in the multimedia, entertainment, astronomy, remote sensing, forensic and medical imaging fields. To focus on this concept effectively it is necessary to know about the visual attention mechanism. Visual attention is an important characteristic in the HVS and the research on visual attention has been reported in 1890. It is a cognitive process of selecting the relevant areas while acquiring the most significant information from the visual scene. Generally, the information captured by the human eyes is much more than that the central nervous system can process. When observers look at a scene, it is impossible for them to recognize all the objects and their relationships in the scene immediately. Thus, the selective attention will allocate processing resources to these salient areas rather than the entire scene. So far, research has focused on two main mechanisms for directing visual attention: top-down and bottom-up. The top-down attention, also called overt attention, is voluntary and task-driven. This mechanism is influenced by cognitive factors such as task, experience, emotions, expectations and knowledge of the observer. It has been studied in various natural environments such as web search and multimedia learning. In contrary to the top-down mechanism, the bottom-up is involuntary, fast, stimulus-driven, and mainly dependent on the intrinsic features of the visual stimuli itself. In the bottom-up process, visual saliency is detected by predicting salient areas through computational models. Since the top-down attention is highly dependent on the task and the observer, most of the existing computational models focus on the bottom-up process.

In the last two decades, many models have been proposed to simulate the bottom-up process. Itti, Koch & Niebur proposed a visual attention model using three feature channels: color, intensity, and orientation for LDR images. This model has become the benchmark for comparing alternative models. Itti, Dhavale & Pighin, extended this model to LDR video content by adding two temporal feature channels: flicker and motion. Apart from the spatio-temporal saliency models proposed, there are also other visual attention models proposed solely for images or for both image and video content. Fang et al. proposed a saliency detection method for compressed videos using MPEG2, H.264, and MPEG4.

In this paper, we address these shortcomings by proposing a new saliency detection method that automatically detects the most visually important areas of HDR images and HDR video frames. The proposed method utilizes the bottom-up structure while bearing the properties of the HVS in mind. Both spatial and temporal cues are taken into account to formulate temporal and spatial saliency maps. Finally, a dynamic fusion method is proposed to combine both the spatial and temporal saliency maps.

II. SYSTEM ARCHITECTURE

Different from LDR content, HDR can describe an expanded color gamut and a wider range of luminance. HDR images and videos store a truthful representation of the depicted scene. To extend the bottom-up framework invented by Itti et al. to HDR content, we propose a new spatio-temporal saliency detection method as depicted in Fig. 1.
Spatio-Temporal Visual Saliency Detection Model for High Dynamic Range Content

CAM
Color perception and luminance perception are tackled in parallel. In HVS model, the perception of colors information is predicted by the Color Appearance Model (CAM), which describes how color information is perceived by HVS under given lighting conditions. For luminance perception, our HVS model takes into account the sensitivity change of the visual perception at different light levels and spatial frequencies using two steps: a) an “amplitude nonlinearity” process and b) the contrast sensitivity function (CSF). In the human eye, cone cells are responsible for color perception. There are three types of cones: L-cones, M-cones, and S-cones (LMS cone space), which are sensitive to long, medium, and short wavelengths. To model how the HVS perceives colors under different lighting conditions, CAMs are developed based on psychophysical studies. Hunt Effect is the basis for color perception and Color Appearance Model. This effect explains that colorfulness increases as luminance increases. Based on the color opponent process theory, the HVS interprets information about color by processing signals from cones in an opposing manner. Take the receptor of red/green as an example, where red creates a positive (or excitatory) response while green creates a negative (or inhibitory) response. Using the psychophysical results on how the responses of cones are combined together, two opponent-color signals are derived.

Amplitude Non-linearity
The luminance range of HDR content covers the full range visible to the HVS. The response of the human eye in this range is neither always linear nor always logarithmic. Inside the HVS model, amplitude nonlinearity accounts for the non-linear response of the HVS to luminance. It is necessary to model the variation of perception at different luminance levels. To this end, through an “amplitude nonlinearity” process, the input luminance is transformed into units of Just Noticeable Difference (JND). This process transforms luminance to JND scaled space, referred as luma, in which adding or subtracting a value of one means just a noticeable change to the human eye. This mapping contains three different functions depending on the intensity of Luminance.

CSF
The sensitivity change at different spatial frequencies is another important property of luminance perception, that needs to be modeled, but was overlooked in the existing models. First, the contrast sensitivity drops at very high and very low spatial frequencies, like a bandpass filter. Moreover, as the level of luminance increases, the peak of CSF shifts to higher spatial frequencies, meaning that certain frequencies become more visible at higher luminance levels. In Fig. 2, the image/frame is filtered in the Fourier domain multiple times, each time using the CSF at a different adaptation luminance level, so that the more visible frequencies are boosted. More specifically, the CSF filters are applied for adaptation luminance (pixel luminance) levels of $L_a = \{0.0001, 0.01, 0.1, 1, 10, 100, 1000\} \text{ cd/m}^2$ as suggested. Then, all filtered images are transformed back to spatial domain. The final filtered image is obtained by interpolation between the two pre-filtered images closest to the adaptation luminance of each given pixel.

Optical Flow
To improve the accuracy of the temporal saliency map, we use an optical-flow-based approach to compute a dense motion vector map between consecutive frames. The magnitude of the motion vector serves as the temporal saliency map. Optical flow is the distribution of the apparent velocities of objects in an image. By estimating the optical flow between video frames, the velocities of objects in the video can be measured.

Dynamic Fusion
A dynamic fusion scheme is used for combining temporal and spatial saliency maps based on some known characteristics of the HVS as follows.
1) Motion is one of the most sensitive cues for HVS.
2) Combined feature targets are more salient than single feature.
3) When watching videos, the human eyes are more sensitive to motion if motion is strong; while the motion is subtle, human attention is attracted more by spatial cues like color, contrast and orientation.

4) In a video sequence, the eye fixation positions are dependent not only on the current frame but also on the frames displayed prior to the current frame, which is known as temporal masking.

III. EYE TRACKING EXPERIMENTS

Eye movements of participants were tracked using the Senso-Motoric Instruments (SMI) iView X RED system. Eighteen individuals (eight males and ten females) participated in the study. All participants had normal or corrected to normal vision, and were screened for normal color vision. Before each participant viewed the stimuli, a calibration was run to ensure accuracy of the eye tracking data. The calibration stage was repeated if the quality of the calibration was not satisfactory. Each participant was asked to ‘free-view’ all images and videos in the stimuli. Each video was presented at its native frame rate and each image was presented for 5 seconds. Before shown each image or video, participants were asked to fixate on a dot presented at one of the four corners of a neutral gray background. Note that by requiring participants to start each trial at one of the corners of the screen, we ensured that they were free to choose where to first begin looking at the material presented on the displays, thereby avoiding any artificial center bias for viewing images and videos. The corner fixation dot was presented for 2 s before each image and video, and the corner location of the dot was randomized from one trial to the next. Fixation density maps (FDM) were obtained from the eye tracking experiment we conducted.

IV. CONCLUSION

This paper addresses these issues by presenting a computational approach to model the bottom-up visual saliency for HDR input by combining spatial and temporal visual features. Utilizing the proposed method in tone-mapping, will allow to locally adjust the contrast of HDR images and videos according to areas of interest provided by the saliency map. The other application of our visual attention model is in designing a quality metric for HDR content. The saliency map predicted by our visual attention model allows us to identify how visible HDR video/image distortions are to the viewer. Distortions appearing in less salient areas are less visible and less annoying compared to the ones appearing in more salient areas. A full reference quality metric that uses our saliency model could be implemented in existing compression standards (e.g., H.264/AVC and HEVC) and potentially improve their compression efficiency. With information about visually important areas, compression methods for HDR could be more effective and efficient by allocating more bit rate resources to visually important areas of each frame and less to the rest of frame. A non-reference quality metric could also be designed based on a similar approach. In this case, the saliency map from our visual attention model could be used to derive a weighting function for the contribution of each pixel to the final quality score. Such a non-reference quality metric is valuable in many applications where a reference image does not exist, such as HDR capturing (i.e., in cameras) or set-top boxes.

ACKNOWLEDGEMENT

We would like to thank everyone who supported us to do this study, especially to Mount Zion College of Engineering and KTU for giving a platform for doing this work.

REFERENCES


