

Human Visual System-Based Saliency Detection for High Dynamic Range Content

Yuanyuan Dong, *Student Member, IEEE*, Mahsa T. Pourazad, and Panos Nasiopoulos, *Senior Member, IEEE*

Abstract—The human visual system (HVS) attempts to select salient areas to reduce cognitive processing efforts. Computational models of visual attention try to predict the most relevant and important areas of videos or images viewed by the human eye. Such models, in turn, can be applied to areas such as computer graphics, video coding, and quality assessment. Although several models have been proposed, only one of them is applicable to high dynamic range (HDR) image content, and no work has been done for HDR videos. Moreover, the main shortcoming of the existing models is that they cannot simulate the characteristics of HVS under the wide luminous range found in HDR content. 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. An analysis of eye movement data affirms the effectiveness of the proposed model. Comparisons employing three well-known quantitative metrics show that the proposed model substantially improves predictions of visual attention for HDR content.

Index Terms—Eye tracking, high dynamic range (HDR), saliency map, visual attention.

I. INTRODUCTION

HUMAN EYES are able to handle intensities in the range of $10^5 : 1$ in real time and up to $10^{14} : 1$ with long-time adaptation. However, conventional cameras and displays only support a dynamic range of approximately $10^2 : 1$ to $10^3 : 1$, known as Low Dynamic Range (LDR). High Dynamic Range (HDR) technology is overcoming this limitation by recording information that covers a wider luminance range and an expanded color gamut. In the last few years, HDR imaging has been gaining widespread acceptance and the related technologies are maturing rapidly. Cameras have been developed with the capability to capture HDR images and videos by fusing multiple pictures at different exposures. In terms of displaying HDR

content, prototype designs are paving the way for an emerging market.

With the truthful representation of the real world, more details and information about the scenes, and a life-like visual experience, HDR imaging is affecting not only specialized fields like entertainment and photography, but also every other application involving digital imaging. High-end cinematographic cameras are already utilizing HDR techniques to provide content with better quality. Video game developers and graphics card vendors are incorporating HDR into video game engines to deliver more believable virtual worlds. In medical imaging, HDR displays can show better contrast than existing medical displays, thus improving the accuracy of diagnosing diseases. In addition to the above, HDR techniques can find applications in computer vision, surveillance and scientific visualization.

Understanding how humans perceive HDR, in other words coming up with visual attention model that best represents such content, is essential to many different stages of the HDR imaging pipeline, such as HDR image and video capturing, compression, content resizing, and displaying. Visual attention research investigates the properties of HVS and simulates the cognitive selection, geared towards predicting salient regions or objects. Existing visual environments contain far more information than the human visual system (HVS) is able to process. Human cognition reacts to such over stimuli by selectively attending to some objects while ignoring other regions of the visual environment. This cognitive selection enables the effective allocation of limited visual processing resources, reducing mental efforts in object detection and recognition.

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 [1]. This mechanism is influenced by cognitive factors such as task, experience, emotions, expectations and knowledge of the observer [1], [2]. It has been studied in various natural environments such as web search [3] and multimedia learning [4]. 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 [2], [5]. 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 [6]. This model has

Manuscript received March 23, 2015; revised September 25, 2015; accepted January 14, 2016. Date of publication January 27, 2016; date of current version March 15, 2016. This work was supported in part by the NSERC under Grant STPGP 447339-13, and in part by the ICICS/TELUS People & Planet Friendly Home Initiative at UBC. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Qi Tian.

Y. Dong was with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: yuand@ece.ubc.ca).

M. T. Pourazad is with TELUS Communications Inc. and also with the Institute for Computing, Information and Cognitive Systems (ICICS), University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: pourazad@icics.ubc.ca).

P. Nasiopoulos is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and also with the Institute for Computing, Information and Cognitive Systems (ICICS), University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: panos@icics.ubc.ca).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2016.2522639

become the benchmark for comparing alternative models. Itti, Dhavale & Pighin [7], extended this model to LDR video content by adding two temporal feature channels: flicker and motion. Apart from the spatio-temporal saliency models proposed in [6] and [7], there are also other visual attention models proposed solely for images or for both image and video content. Fang *et al.* [8] proposed a saliency detection method for compressed videos using MPEG2, H.264, and MPEG4. The saliency map is constructed from four types of features including luminance, color, texture, and motion, which are extracted from the discrete cosine transform coefficients and motion vectors in video bitstream. A new video saliency detection algorithm (called ROCT), which is based on feature learning, was recently proposed in [9]. Using the support-vector machine (SVM) learning algorithm, spatial and temporal features are classified and combined to provide saliency maps. According to [10], at least 65 computational models of visual attention were proposed over the last two decades. However, to the best of our knowledge, the only model that has been proposed for HDR image content is the Contrast Features (CF) model described in [11]. When applied to HDR content, the state-of-the-art models such as Itti *et al.*'s model and the CF model have some shortcomings [6], [7], [11]. Itti *et al.*'s model neglects the HVS properties related to the wide luminance range and rich color gamut associated with HDR content, while the CF model does not address the color perception of HDR images and it is not applicable to video content.

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, leading to two saliency maps: the spatial saliency map and the temporal saliency map. To obtain the spatial saliency map, we use the HVS model to decompose feature channels from an HDR input and then follow the procedure of the classical bottom-up method. To compute the temporal saliency map, an optical flow based method is used to estimate motion. Finally, a dynamic fusion method is proposed to combine both the spatial and temporal saliency maps.

The proposed spatio-temporal model is used to predict the saliency of natural HDR images and videos. The predictions 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.

The rest of this paper is organized as follows: an overview of related work on saliency detection is given in Section II. Our proposed method is outlined in Section III, followed by a description of the eye tracking experiment conducted using a prototype HDR display in Section IV. In Section V, the performance of the proposed model is evaluated using an HDR image database and an HDR video dataset. Finally, future work and potential applications are discussed and conclusions are drawn in Section VI.

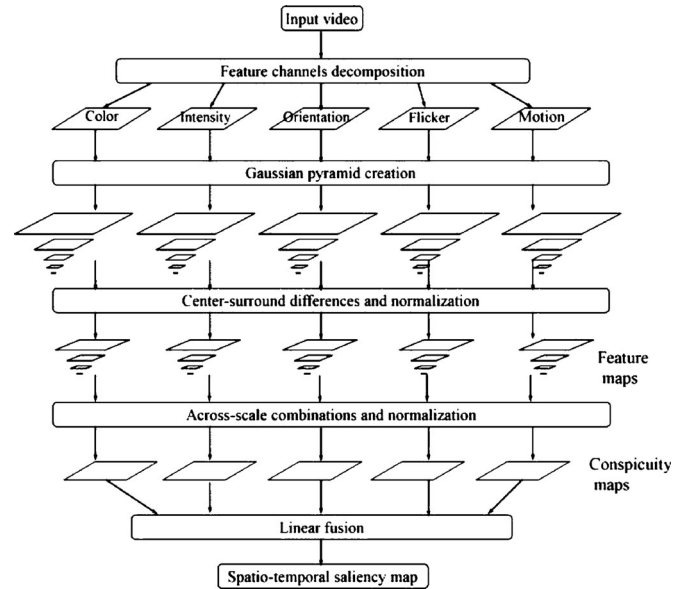


Fig. 1. Architecture of the model proposed by Itti *et al.*

II. RELATED WORK

This section provides a brief overview of the Itti *et al.*'s visual attention model [7], the most widely used LDR approach, and the CF visual attention model [11], which is an extension of the Itti *et al.*'s model for HDR image applications. These two state-of-the-art models are selected as the benchmarks that our proposed model is compared against (Section V).

A. Itti *et al.*'s Model

Itti *et al.*'s model described in [7] is an implementation of the bottom-up framework proposed by Koch and Ullman [12]. Fig. 1 depicts the structure of Itti *et al.*'s model for video content. First of all, the visual input is decomposed into several parallel channels (color, intensity, orientation, flicker and motion) and a Gaussian pyramid is formed in each channel by subsampling (scale 0 to 8, with scale 0 corresponding to the resolution of the original image). Then, early visual features are extracted to form a set of topographic feature maps in each channel by center-surround operations, which imitate the visual receptive fields of the human eye. The center-surround operation is implemented as across-scale subtraction between a center scale c ($c = \{2, 3, 4\}$) and a surrounding scale s ($s = \{c + 3, c + 4\}$) in the Gaussian pyramid. The across-scale subtraction is obtained by interpolation to the center scale and point-by-point subtraction [6]. Afterwards, all feature maps are combined into a "conspicuity map" in each channel. Conspicuity maps from all channels are averaged to form one master saliency map, which topographically represents the local saliency. When the input is an image, only the color, intensity and orientation channels are used [7]. Although this model has shown good results with various LDR images and videos, it has the following drawbacks when applied to HDR content.

- 1) The input signal is assumed to be a perceptually uniform and linear RGB signal. Although this is generally true for

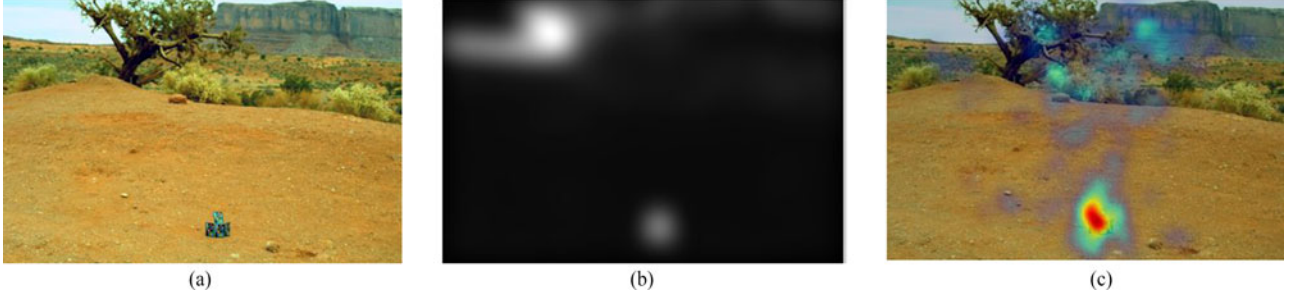


Fig. 2. Itti *et al.*'s model detects the brightest area in an HDR image as the most salient area. (a) HDR image. (b) Saliency map using Itti *et al.*'s method. (c) Human fixation map from eye tracking experiment on top of original image (red areas represent the most attended areas).

LDR images, HDR images and videos, however, don't have perceptual uniformity. For this reason, the color, intensity and orientation feature maps are generally not able to capture all the corresponding salient regions.

- 2) Itti's *et al.* model handles the intensity linearly; however, the HVS perceives the light intensity in a non-linear manner [13]. Thus, if there is a very bright area in an HDR image, the proposed model detects that area as the most salient area, but fails to detect other salient regions. For instance, in the case of having a visually important moving object in the scene with a very bright area, Itti's *et al.* model may detect the bright area as the most salient region. This is evident in Fig. 2, where the saliency map using Itti's *et al.* method does not match the human fixation map obtained from an eye tracking experiment. Even though this method carefully and thoroughly models the neurobiological process of human visual attention (Gabor filter models the property of primary visual cortex in the HVS, and the center-surround operations are inspired by the ganglion cells in the visual receptive fields of the HVS), it does not take into account the characteristic of HVS in perceiving light intensity.
- 3) The fusion step simply involves averaging across the spatial and temporal channels. However, studies have revealed that motion is the strongest attractor of attention [14], [15].
- 4) Parameters such as the viewing distance and the pixel density of the display have been overlooked.

B. Contrast Features Model

In [11] Bremond, Petit & Tarel proposed an algorithm, denoted CF, for computing saliency maps of HDR images by defining new visual features within the framework of Itti *et al.* [6]. The main difference between the proposed algorithm and the framework of Itti *et al.* [6] is using intensity normalized values for the intensity and orientation channels. The rationale is that HVS is sensitive to contrast rather than absolute intensities. More specifically, in Itti *et al.*'s model, the intensity feature map between the center scale c and the surround scale s is defined as: $I(c, s) = |I(c) \ominus I(s)|$, where " \ominus " stands for the across-scale subtraction between two maps. However, in the CF model, the

intensity contrast is used instead of the absolute intensity

$$I'(c, s) = |I(c) \ominus I(s)| / I(s). \quad (1)$$

In Itti *et al.*'s model, the orientation pyramid is computed by convolving the levels of the intensity pyramid with Gabor filters at different angles, $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. The orientation feature maps, obtained through center-surround operation, are calculated as: $O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|$. In the CF model, a new definition of orientation feature maps is used to detect the "borders of borders" so this feature is homogeneous to contrast [11]. The new orientation feature map between the center scale c and surround scale s is calculated as

$$O'(c, s, \theta) = O(c, \theta) / I(s) \quad (2)$$

where $O(c, \theta)$ is the orientation pyramid obtained by convolving the intensity with Gabor filters and $I(s)$ is the intensity of the s^{th} level in the pyramid.

In the CF model, the color channel remains the same as in Itti *et al.*'s model in [6], because the color feature is already normalized by the intensity at every pixel.

The predictions obtained by CF were compared with eye tracking experiment results, and it was shown that a better fit exists compared to the predictions obtained by Itti *et al.*'s model. However, the following limitations exist in the CF model:

- 1) the eye tracking experiment was conducted with physical scenes and light sources in front of subjects, instead of using an HDR display;
- 2) only one scene was studied, so the robustness of the model is not fully validated; and
- 3) some mechanisms of visual adaptations were taken into consideration, but there are other critical properties like intensity and color perception under wide luminance range, which are not modeled in CF.

III. THE PROPOSED MODEL

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. 3. Compared to Itti *et al.*'s bottom-up model, three new steps/modules are introduced in our proposed

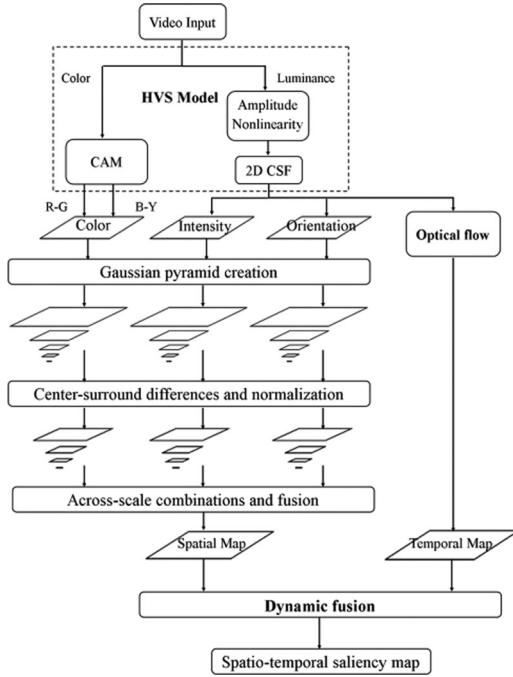


Fig. 3. Architecture of the proposed model.

model to address the limitations of the existing method. First, before feature channels are extracted we add a module, denoted HVS model in Fig. 3, to model the human perception on HDR pixel values. Our proposed HVS model mainly addresses two issues: 1) HDR color perception, and 2) luminance perception under the wider HDR luminance range. Second, we use a method based on optical flow to obtain the temporal saliency map so the motion information is correctly extracted under various illumination conditions. Third, the average fusion scheme is replaced by a dynamic fusion scheme, which takes into account the known characteristics of the HVS regarding perceiving the temporal and spatial cues. The following subsections describe these modules in more details.

A. HVS Model

The HVS model used in the proposed method simulates the human perception of HDR pixel values. Color perception and luminance perception are tackled in parallel (see Fig. 3). In our 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). The following subsections elaborate on our HVS model.

1) *Color Appearance Model*: In Itti *et al.*’s approach, the two color opponent signals are obtained as linear combinations of R, G, and B. However, studies have shown that the HVS perceives colors differently under different lighting conditions. For example, colorfulness increases with higher luminance levels,

which is known as the Hunt effect [16]. 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. Among the existing widely used CAMs, the one proposed by Kim *et al.* [17] is designed and optimized based on the psychophysical experiments with high luminance range settings (up to 16860 cd/m²), while others such as CIECAM97s [18] and CIECAM02 [19] are designed based on the psychophysical experiments on the LUTCHI data set [20] at low luminance ranges (mainly below 690 cd/m²).

Thus, to ensure that color perception under a wide luminance range is correctly modeled, we use the proposed CAM by [17] in our study. As suggested in [17] to emulate color perception, first the tristimulus XYZ values are transformed into LMS cone space using the Hunt-Pointer-Estevéz (HPE) transform described in [21]. Then, the cones’ absolute responses are modeled by

$$L' = \frac{L^{n_c}}{L^{n_c} + L_a^{n_c}}, M' = \frac{M^{n_c}}{M^{n_c} + L_a^{n_c}}, S' = \frac{S^{n_c}}{S^{n_c} + L_a^{n_c}} \quad (3)$$

where L_a is the absolute level of adaptation luminance measured in cd/m², and the exponent n_c is equal to 0.57, which is derived empirically in [17]. Here, we assume that eyes adapt to a single pixel and the adaptation luminance is equal to the luminance of each pixel. Ideally, adaptation luminance should be obtained through an adaptation map, which would consider the fact that eyes adapt to a small area in the visual field [17].

Based on the color opponent process theory [22], 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 [23], two opponent-color signals are derived as

$$\begin{aligned} a &= \frac{1}{11} (11L' - 12M' + S') \text{ red-green channel;} \\ b &= \frac{1}{9} (L' + M' - 2S') \text{ yellow-blue channel.} \end{aligned} \quad (4)$$

2) *Amplitude Nonlinearity*: 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 of Fig. 3, 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), as suggested in [24]. This process transforms luminance to JND scaled space, referred as luma [24], in which adding or subtracting a value of one means just a noticeable change to the human eye (Fig. 4). This mapping contains three different functions depending on the intensity of

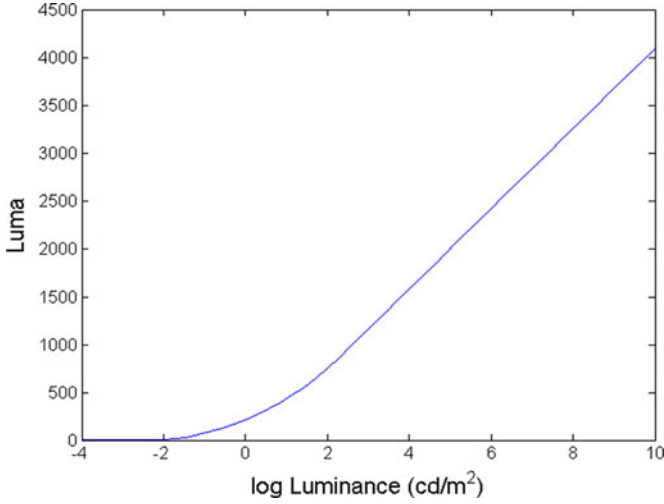


Fig. 4. Luminance to luma mapping.

luminance (L_a) as follows:

$$\text{luma}(L_a) = \begin{cases} 769.18 \cdot L_a, & L_a < L_1 \\ 449.12 \cdot L_a^{0.17} - 232.25, & L_1 \leq L_a < L_2 \\ 181.7 \cdot \ln(L_a) - 90.16, & L_a \geq L_2 \end{cases}$$

$$L_1 = 0.061843 \text{ cd/m}^2$$

$$L_2 = 164.1 \text{ cd/m}^2. \quad (5)$$

As it is observed, at very low luminance levels, below L_1 , linear mapping is used. At high luminance levels, above L_2 , a logarithmic mapping is used, which corresponds to Weber's law (which states the size of change needed to be perceived varies against background luminance levels to produce a constant proportion JND). A power function segment exists between the two luminance thresholds L_1 and L_2 . For details about the derivation of the above coefficients and the thresholds L_1 and L_2 , please see [24].

3) *Contrast Sensitivity Function*: Amplitude nonlinearity compensates for the nonlinearity of luminance perception over the entire luminance range. Yet, 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.

The CSFs shown in Fig. 5 depict the visual sensitivity as a function of spatial frequency. Formulas for CSF can be found in [13]. There are a few things to notice in Fig. 5. 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 our proposed model to account for the sensitivity-change of the visual perception at different spatial frequencies, we use the 2D CSF proposed in [13] to filter the luma image and boost the spatial frequencies more sensitive to the HVS. While in the existing models, the effect of the display resolution, display size, and viewing distance have been overlooked, our study

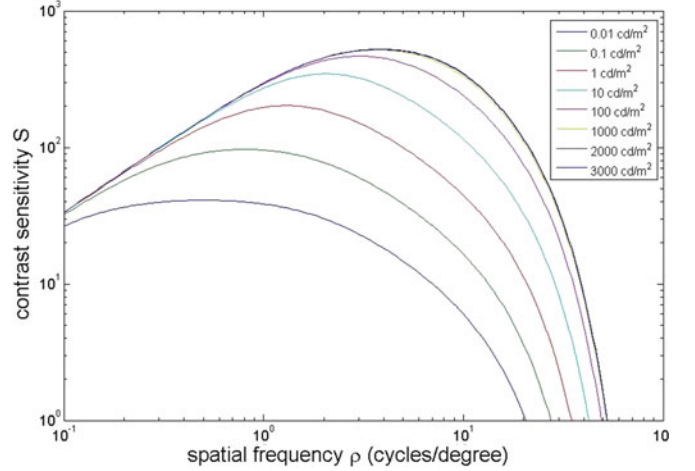


Fig. 5. CSF depends on luminance.

takes into account these parameters when calculating the spatial frequency in cycles-per-degree. Equation (6) is derived based on the parameters of experiment settings (see Fig. 10), including vertical resolution (number of pixels), display size, and viewing distance. A screen with 800 pixels resolution viewed at 1m has roughly the same cycles-per-degree as a screen with 1600 pixels resolution viewed at 2 m

$$\text{pixel per degree} = \frac{\text{vertical resolution}}{\frac{180}{\pi} \tan^{-1} \left(\frac{0.5 \times \text{display height}}{\text{viewing distance}} \right) \times 2} \quad (6)$$

$$\text{cycles per degree} = \frac{\text{pixel per degree}}{\text{vertical resolution}}. \quad (7)$$

Ideally, in order to apply the CSF filter to an HDR image with a wide luminance range, we should use a separate CSF filter for every pixel, since each pixel has a different luminance level, and every luminance level corresponds to a different CSF with the peak at a certain frequency (as shown in Fig. 5). This is computationally not feasible, given that an HD frame has more than 2 million pixels. To address this issue, we adopt a more effective approach, the multi-CSF method described in [25]. In our proposed approach, after the “amplitude nonlinearity” process, as shown in Fig. 6, 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 by [25]. Note that the adaptation luminance levels beyond 1000 cd/m^2 are excluded, because as shown in Fig. 5 the shape of CSF remains constant for the adaptation luminance larger than 1000 cd/m^2 [25]. 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. For example, if the original luminance of a pixel is 70 cd/m^2 , its value in the final filtered image is calculated by linearly interpolating the filtered images with $L_a = 10 \text{ cd/m}^2$ and $L_a = 100 \text{ cd/m}^2$.

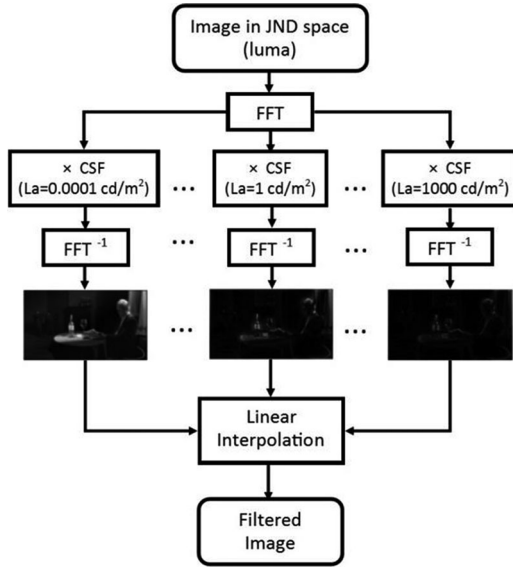


Fig. 6. To apply CSF on image using the multi-CSF method as in [22]. $L_a = \{0.0001, 0.01, 0.1, 1, 10, 100, 1000\}$ cd/m².

B. Optical Flow

In Itti *et al.*'s model, the temporal saliency map is derived from two features, flicker and motion. Flicker is expressed as the difference between the intensity of video frames at different time instances and motion is defined as the spatially-shifted differences between Gabor pyramids from the current and previous frames [7]. Our experiments with HDR videos showed that neither of these two features could accurately represent the motion information in videos. 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. Horn and Schunck first proposed dense optical flow in [26]. Their method is based on two assumptions: 1) the brightness consistency of objects, which states that the apparent brightness of objects in a scene remains constant; and 2) neighboring pixels move similarly, in other words, the optical flow field is smooth. Since optical flow relies on the assumption of intensity consistency between frames, videos with varying lighting, such as the sky, tend to cause more artifacts. To make sure that the difference of exposure over time and the variation of lighting in sequences don't cause artifacts in the motion vector, we adopt the residual based optical flow presented in [27], [28]. In this approach, residuals are used for optical flow computation rather than the original frames; the residual in this case is the difference between an intensity image and its smoothed version, i.e., the high frequencies of the image. Because lighting changes in a sequence mostly affect the low frequencies, this approach improves the accuracy of motion vectors provided by the optical flow.

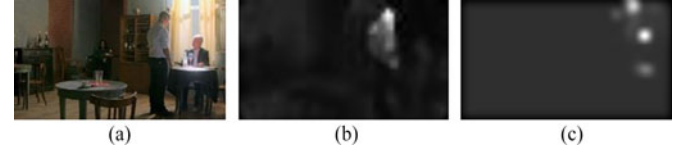


Fig. 7. Temporal saliency map. (a) Example frame from sequence *bistro02*. (b) The temporal saliency map using the proposed approach. (c) The temporal saliency map from Itti *et al.*'s method.

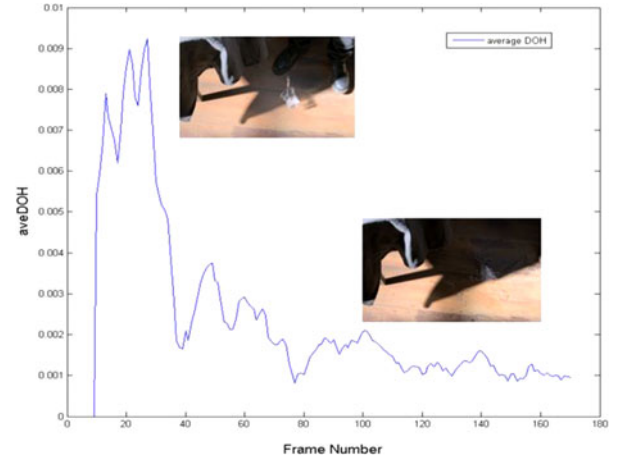


Fig. 8. Average DOH of sequence *bistro 03*.

In the proposed model, the Horn-Schunck method is used and the residuals are generated using a Gaussian filter. For each frame, a Gaussian filter is convolved with the intensity frame generated from the HVS model (see Fig. 3). The residual is calculated by taking the difference between the original image/frame and the Gaussian filtered image/frame. Fig. 7 shows a frame from sequence *bistro02* and the temporal saliency map. The moving object in this frame is the walking man, but Itti *et al.*'s method fails to detect this object.

C. Dynamic Fusion

After the temporal and the spatial maps are obtained, they have to be combined to form a unique spatial-temporal map. The best way to fuse these maps, which are generated from different visual attention cues, is still an open issue. Since the biological and neurological mechanisms for feature fusion in the HVS are not fully understood yet, we propose a dynamic fusion scheme 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. It is the most dominant attractor of attention among the commonly studied features [14].
- 2) Combined feature targets are more salient than single-feature as concluded in the study by Nothdurft [29].
- 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 [15].
- 4) In a video sequence, the eye fixation positions are dependent not only on the current frame but also on the frames

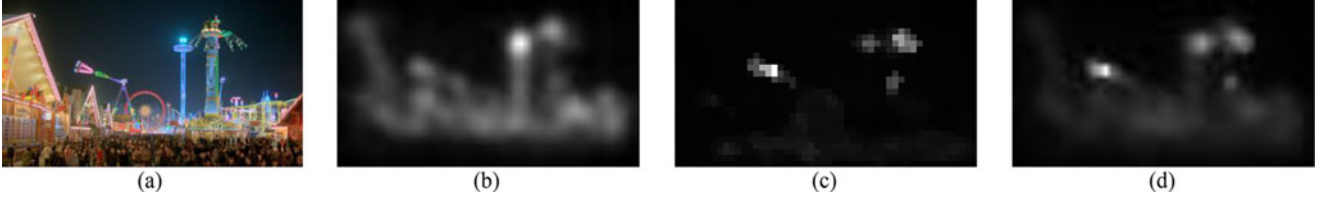


Fig. 9. Example of spatio-temporal saliency map. (a) Frame from sequence *park*. (b) Spatial saliency map. (c) Temporal saliency map. (d) Fused spatio-temporal saliency map.

displayed prior to the current frame, which is known as temporal masking.

In order to address the fact that objects salient both spatially and temporally tend to stand out more than objects that are important only spatially or temporally, our proposed fusion method includes the product of the spatial and temporal maps. Since the relative importance of spatial cues and temporal cues varies based on the intensity of motion in the video, the proposed fusion method contains weighted spatial and temporal maps, rather than the average of them which is used in Itti *et al.*'s model. The formula of the spatio-temporal saliency map is given by

$$\begin{aligned}
 S &= a \cdot S(s) + b \cdot S(t) + c \cdot S(s) S(t) \\
 a &= \max \left\{ 0, 0.5 \left(1 - \frac{\text{ave DOH}}{\varepsilon} \right) \right\} \\
 b &= \min \left\{ 1, 0.5 \left(1 + \frac{\text{ave DOH}}{\varepsilon} \right) \right\} \\
 c &= 1
 \end{aligned} \quad (8)$$

where $S(s)$ and $S(t)$ are the normalized spatial and temporal saliency maps of a frame, a , b , and c are the parameters to control the relative strength of the spatial saliency map, the temporal saliency maps and the product of two maps, respectively. Considering that the luminance histogram is reported to be a very efficient indicator of motion intensity [30], we utilized it in defining the controlling parameters in (8). More specifically, the average Difference of Histograms of consecutive frames (*ave DOH*) in (8) serves as an indicator for motion intensity and ε is an empirical threshold for motion intensity. When the intensity of movement in video is high, i.e., $\text{ave DOH} > \varepsilon$, the temporal saliency map is dominant and the spatial map has little contribution to the spatio-temporal saliency map. The average Difference of Histograms (*ave DOH*) of consecutive frames is defined by

$$\begin{aligned}
 \text{ave DOH}(n) &= \frac{1}{\text{no. of pixels} \times N} \sum_{m=1}^n \sum_{l=1}^q |\text{hist}(L_n) - \text{hist}(L_{n-1})| \\
 m &= \begin{cases} 1, & n \leq N \\ n - N + 1, & n > N' \end{cases}
 \end{aligned} \quad (9)$$

where hist is the histogram, L_n and L_{n-1} are the luminance of frame n and frame $n - 1$ respectively, q is the number of levels used in the histograms, no. of pixels is the number of pixels in a frame, and N is the group size of frames to compensate for the temporal masking effect. Considering the temporal masking

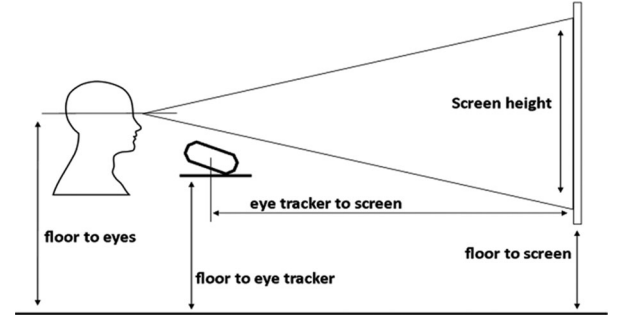


Fig. 10. Setup of eye tracker [40].

effect and the fact that eye fixations last more than 300 msec on average, the group size can be determined based on the video frame rate. For example, in the case of 30 fps video, the group size is set to 10, which corresponds to 333 ms. The average *DOH* in the group of frames divided by the total number of pixels is used as an indicator of the average motion level. This average *DOH* value determines the relative weights of the temporal and spatial saliency maps. Fig. 8 shows the average *DOH* for the sequence *bistro 03*. The peaks before frame 40 match the action of the glass falling and the low average *DOH* in the later part of the video indicates that there is less motion, thus both observations correlate well with the actual intensity of motion. Fig. 9 shows an example of a fused spatio-temporal saliency map using our dynamic fusion scheme.

IV. EYE TRACKING EXPERIMENTS

A. Eye Tracking System

Eye movements of participants were tracked using the SensoMotoric Instruments (SMI) iView X RED system. The eye tracker was mounted on a tripod and participants were seated in a chair that allowed their eye height to be adjusted to meet the set up requirements of the SMI system. This setup is shown in Fig. 10. The sampling frequency of the SMI is 250 Hz and the resolution accuracy is $0.4 \pm 0.03^\circ$.

B. Participants

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. All subjects were naïve to the purpose of the experiment. Before each task, each participant's eye height and position was adjusted so that their eyes could be tracked accurately.

TABLE I
VIDEO SEQUENCES USED IN EYE TRACKING STUDY






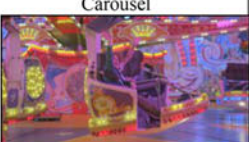



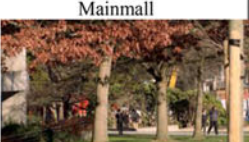
Clip		Source	
Balloon		200 frames 30 fps 1920 * 1080	Technicolor
Market		400 frames 50 fps 1920 * 1080	Technicolor
Bistro 01		151 frames 30 fps 1920 * 1080	Froehlich <i>et al.</i>
Bistro 02		300 frames 25 fps 1920 * 1080	Froehlich <i>et al.</i>
Bistro 03		170 frames 30 fps 1920 * 1080	Froehlich <i>et al.</i>
Carousel		339 frames 30 fps 1920 * 1080	Froehlich <i>et al.</i>
Park		439 frames 30 fps 1920 * 1080	Froehlich <i>et al.</i>
Fishing		371 frames 30 fps 1920 * 1080	Froehlich <i>et al.</i>
Playground		222 frames 30 fps 2048* 1080	DML-HDR
Mainmall		241 frames 30 fps 2048* 1080	DML-HDR



Fig. 11. HDR display system.

C. Stimuli

The stimuli of the eye tracking experiment included twenty-three HDR images and ten HDR video clips. Twenty-three HDR images in non-compressed Radiance RGBE format representing a wide variety of content were selected for the study. Ten HDR videos (Table I) from three different sources in total were used in this study: Technicolor [31], Froehlich *et al.* [32] and DML-HDR.¹ Content was chosen to span a wide variety of scenes that included day and night lighting conditions, different ranges of motion (i.e., minimal motion to fast moving objects) and a wide range of color spectrums. We limited the social context of the scenes, and those scenes that did include social components (i.e., people or human interaction) were kept neutral in nature, in order to reduce the inter-subject bias. The HDR video eye tracking dataset generated from these ten videos, called “DML-iTrack-HDR”, is publicly available.² For more details on “DML-iTrack-HDR” dataset see [33].

D. HDR Display

Tests were performed on a Dolby prototype HDR display system (Fig. 11). The measured output peak luminance of the system is 2700 cd/m². To display HDR content, HDR images and videos are processed to generate two calibrated streams sent to the projector and LDR respectively, based on the procedure described in [34]. The stream sent to the projector consists of a monochrome intensity signal, while the other stream sent to the LCD screen contains the color components.

E. Procedure

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

¹[Online]. Available: <http://dml.ece.ubc.ca/data/DML-HDR>

²[Online]. Available: <http://dml.ece.ubc.ca/data/DML-iTrack-HDR>



Fig. 12. Fixation density maps of an HDR video. (a) Frames from sequence *playground*. (b) Human fixation density maps from eye tracking study.

presented on the displays, thereby avoiding any artificial center bias for viewing images and videos [2]. 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.

F. Human Fixation Density Maps

Fixation density maps (FDM) were obtained from the eye tracking experiment we conducted. Fig. 12 depicts a few frames from one of the video clips, *playground*. These maps represent subjects' region of interest (RoI) and serve as ground truth for assessing the performance of our proposed model.

For a given image, all fixation data from different subjects are combined together to provide a spatial distribution of human fixation. All fixation hits are filtered by a 2D Gaussian filter whose standard deviation is equal to 1° of visual angle (which corresponds to $\sigma = 20$ pixels in our setup). Gaussian filtering is applied to compensate for the eye-tracking errors and visual sensitivity reduction due to the distance from the fovea [35]. For video clips, spatial distribution of human fixation for every frame is computed for every subject. If the duration of a given fixation is longer than the length of one frame, this fixation hit appears in more than one frame. Fixations from all subjects are combined together and filtered by a 2D Gaussian filter. The same Gaussian filter is used for both image stimuli and video stimuli.

V. PERFORMANCE EVALUATION

To evaluate the performance of our proposed method we applied our proposed scheme to two data sets, an HDR image dataset with 23 images, and an HDR video dataset with 10 videos (details of videos can be found in Table I), and compared the results with those of Itti *et al.*'s [7] and CF's models [11]. In our implementation, ε in (8) is empirically chosen as the 0.3% of the total number of pixels in a frame, and to compute the average difference of histograms of consecutive frames in (9), q is set to 30, and the group size of the frames (N) is set to 10 frames (as the frame rate of the test videos is equal to 30 fps). To have a fair comparison, we extended the Matlab code of Itti *et al.*'s model available online [39] to work on HDR images in RGBE format. In order to quantitatively evaluate our model's performance against the existing benchmark models, we chose to use three different statistical metrics. The reason for using

multiple metrics in this quantitative assessment is to ensure that the resulting conclusions are robust and independent of the metrics. A detailed description of these metrics and our experiment results are provided in the following subsections.

A. Quantitative Assessment Statistical Metrics

1) *Linear Correlation Coefficient (CC)*: The first metric we used is the linear correlation coefficient, and measures the strength of linear relationship between two data sets

$$CC(p, g) = \frac{\text{cov}(p, g)}{\sigma_p \sigma_g} \quad (10)$$

where p is the prediction map derived from the computational model, g is the FDM derived from eye tracking, $\text{cov}(p, g)$ is the covariance value between the p and g , and σ_p and σ_g are the standard deviations of p and g , respectively. The range of CC is between -1 and 1 . When CC is equal to $+1$, there is a perfect positive linear relationship while -1 indicates a perfect negative linear relationship between the two variables.

2) *Kullback-Leibler Divergence (KL)*: The second metric is the Kullback-Leibler divergence (KL), which measures the overall dissimilarity between two probability density functions as follows:

$$KL(p|g) = \sum_x p(x) \log \left(\frac{p(x)}{g(x)} \right) \quad (11)$$

where $p(x)$ and $g(x)$ are probability density functions deduced from the prediction map and the ground truth map. This can be done by dividing each location in the map by the sum of all the pixel values. The KL value varies from zero to infinity. When the KL-divergence value is zero, it means that the two probability density functions are exactly the same. However, KL-divergence is not a distance and it is not symmetric, which means that $KL(p|g) \neq KL(g|p)$. In some published results, $KL(p|g)$ is used. In our comparison, similar to [36], the symmetric KL-divergence is reported, which is the average of $KL(p|g)$ and $KL(g|p)$.

3) *Receiver Operating Characteristic (ROC) Analysis*: The Receiver Operating Characteristic (ROC) Analysis is probably the most widely used qualitative metric for assessing saliency models [35] and is the third metric we used in our evaluations. In this metric, the inputs include human fixations and the prediction saliency map (generated using the saliency model). All human

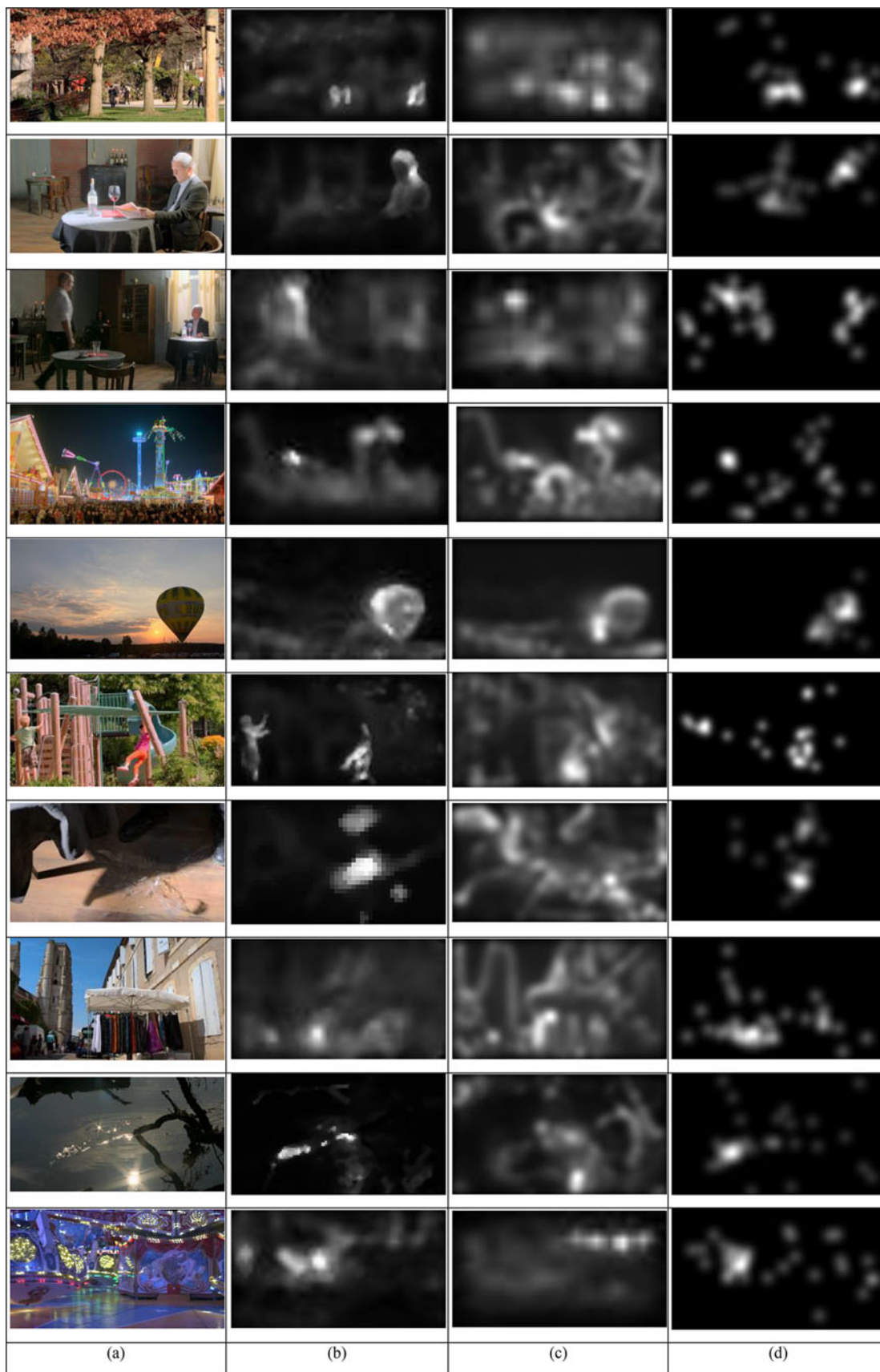


Fig. 13. Ideal ROC and ROCs of proposed, Itti's, and CF models for the HDR image dataset. (a) Distribution of ROC using the split-data technique, providing the theoretical upper bound of ROC. **Ideal ROC** = 0.783. (b) Distribution of proposed model performance. **Mean ROC** = 0.711. (c) Distribution of CF model performance. **Mean ROC** = 0.682. (d) Distribution of Itti *et al*'s model performance. **Mean ROC** = 0.641.

TABLE II
AVERAGE RESULTS OF 23 HDR IMAGES

Model	ROC		CC		KL	
	Ave.	ST dev.	Ave.	ST dev.	Ave.	ST dev.
Proposed	0.71	0.07	0.46	0.12	1.68	0.56
CF	0.68	0.07	0.41	0.16	1.85	0.74
Itti-CIO	0.64	0.08	0.37	0.18	1.86	0.72
Gaussian	0.54	0.03	0.28	0.14	2.90	0.61

fixations are considered as a positive set. Different threshold values are used on the continuous prediction saliency map to convert it into binary maps. For each threshold, fixations are laid on top of the binary map to determine the number of true positives and false positives. The true positive rate vs. false positive rate is plotted as an ROC curve for each image or video frame. The area underneath the ROC curve is often used as a numerical score to measure how well the prediction is aligned with the ground truth. An area of 1 means the prediction is perfect.

Given the inter-subject variability in the eye-tracking data, the natural dispersion of fixations among different subjects looking at the same image [35], no saliency algorithm can perform better (on average) than the ROC dictated by inter-subject variability. We calculate an ideal ROC as the theoretical upper bound to fully evaluate the predictive power of computational models. The ideal ROC is obtained by a split-data technique, performing ROC analysis to measure how well the fixations of half of the subjects can be predicted by the other half of the subjects. All subjects are repeatedly split into two sub-groups with the same number of subjects in a random manner (100 times) and ROC analysis is performed. The reported theoretical upper bound is the value averaged over 100 random samples. Similar techniques are used in previous publications such as [37] and [38].

B. Experiment Results and Discussion

To evaluate the performance of our proposed method, we compared it against two benchmark models, the Itti *et al.*'s model [7] (the most widely used model for LDR content), and the CF model [11], which is designed for HDR images. Since the majority of fixations tend to be near the image center, often referred to as center bias, we also reported the performance of a simple Gaussian filter (sigma = 10, window size 50×50) for comparison. The following subsections elaborate on our experiment results and the comparison analysis based on the statistical metrics explained in Section 5.1.

1) *Experiment 1—Using the HDR Image Dataset:* Table II provides the average results of *CC*, *ROC*, and *KL* on 23 HDR images. The results of the simple center-bias Gaussian filter are also reported to provide a reasonable lower bound for evaluations. As it is observed, even though the CF method improves the prediction accuracy compared to Itti *et al.*'s method, our proposed method outperforms both CF and Itti *et al.*'s methods according

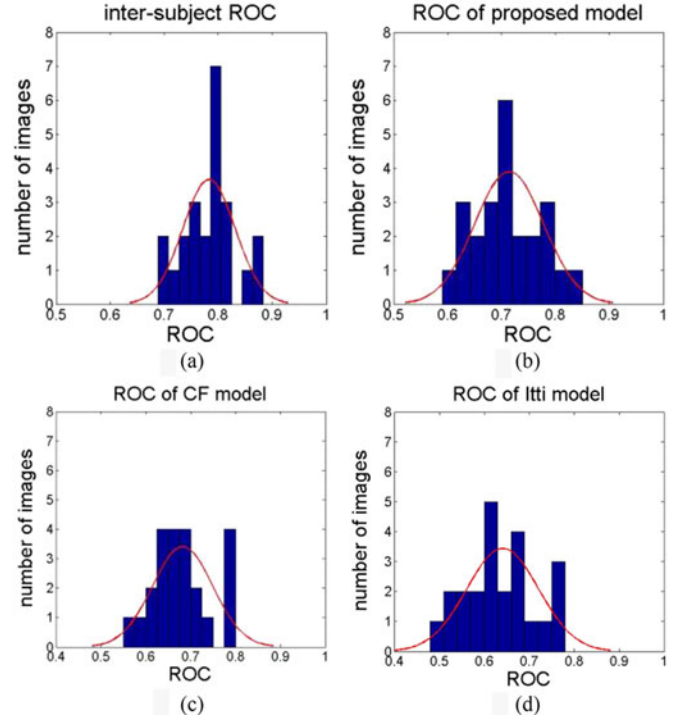


Fig. 14. (a) Example frames of HDR videos. (b) The spatio-temporal saliency map from the proposed model. (c) The spatio-temporal saliency map from Itti *et al.*'s model. (d) The fixation density map (FDM) from eye tracking study.

to all three metrics. Fig. 13 shows the histogram/distribution of an ideal ROC and that of the ROCs obtained by our proposed model, the CF model and Itti *et al.*'s model for the HDR image dataset. The proposed model performs at approximately 91% of the ideal ROC (idealROC = 0.783 and mean ROC of our proposed model = 0.711), the theoretical limit determined by inter-subject variability. Moreover, the lower standard deviation values in Table II across images achieved by our proposed method imply that this method is more robust in predicting visually important areas of HDR images compared to the two benchmark models.

2) *Experiment 2—Using the HDR Video Dataset:* Fig. 14 illustrates the saliency maps of a few of videos obtained using our proposed model in comparison with those generated by Itti *et al.*'s model and the ground truth fixation maps. As it is observed, our proposed model generates more realistic saliency maps compared to the Itti *et al.*'s model. Table III shows the average CC, ROC and KL for each HDR video sequence. Since the CF method does not support video input, it is excluded from this comparison. The average gain of the proposed spatio-temporal method is about 0.05, 0.12 and -0.53 for ROC, CC and KL, respectively (a KL value closer to 0 means a better prediction, thus the gain of KL is a negative value). As Table III shows, the standard deviation values using the proposed model are reduced according to all three metrics, suggesting that our approach leads to a more robust model than Itti *et al.*'s. A few more observations can be made from Table III : 1) the lowest scores of both methods belong to the sequence carousel, a very busy and cluttered scene with multiple moving objects and

TABLE III
RESULTS FOR 10 HDR VIDEOS

	ROC			CC			KL		
	Itti	Proposed	Gain (Proposed - Itti)	Itti	Proposed	Gain (Proposed - Itti)	Itti	Proposed	Gain (Proposed - Itti)
bistro01	0.59	0.7	0.11	0.08	0.19	0.11	5.32	4.73	-0.59
fishing	0.67	0.73	0.06	0.17	0.37	0.2	5.24	4.24	-0.99
park	0.74	0.76	0.02	0.35	0.39	0.04	4.42	3.80	-0.62
mainmall	0.73	0.74	0.01	0.27	0.42	0.15	4.97	4.79	-0.18
market	0.65	0.73	0.08	0.17	0.3	0.13	5.08	4.69	-0.39
bistro03	0.66	0.73	0.07	0.14	0.36	0.22	5.53	4.74	-0.79
balloon	0.80	0.81	0.01	0.43	0.49	0.06	4.65	4.39	-0.27
carousel	0.55	0.66	0.11	0.05	0.19	0.14	5.64	5.11	-0.54
playground	0.73	0.74	0.01	0.28	0.43	0.15	5.26	4.70	-0.56
bistro02	0.77	0.77	0.00	0.35	0.35	0	4.93	4.50	-0.44
Average	0.69	0.74	0.05	0.23	0.35	0.12	5.10	4.57	-0.53
ST dev.	0.08	0.04	-0.04	0.13	0.10	-0.03	0.38	0.36	-0.02

A KL closer to 0 means a better prediction, thus the gain of KL is a negative value.

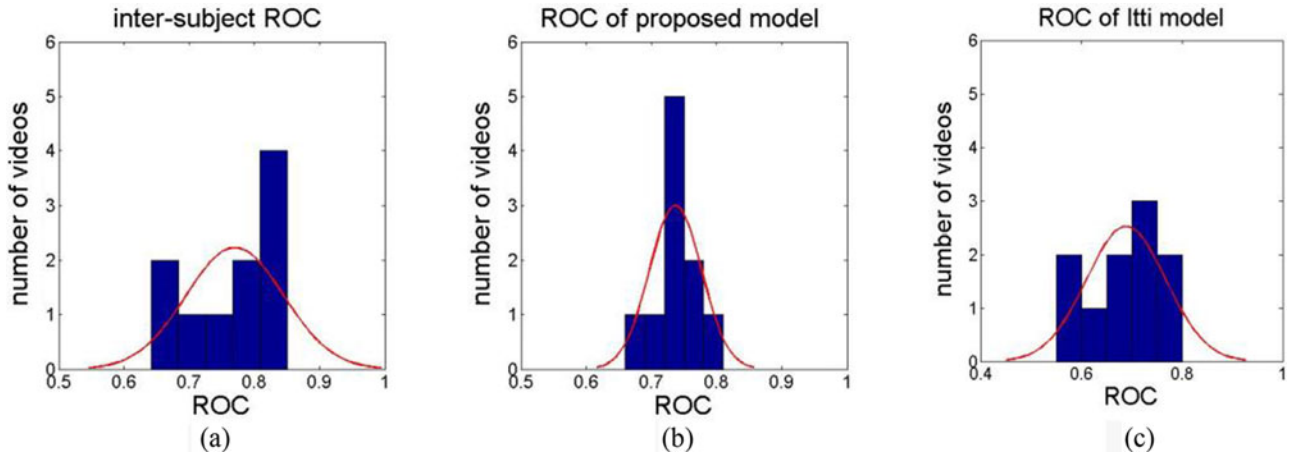


Fig. 15. Ideal ROC and ROC of proposed model for the HDR video dataset. (a) Distribution of ROC using the split-data technique, providing the theoretical upper bound of ROC. Ideal ROC = 0.770. (b) Distribution of proposed model performance. Mean ROC = 0.737. (c) Distribution of Itti *et al.*'s model performance. Mean ROC = 0.689.

blinking lights. A possible explanation might be that when the visual stimuli are complicated and busy, there are more top-down factors, such as experience, emotions, expectations and knowledge, driving the attention, which are not yet taken into consideration in both models. 2) our method has achieved the largest performance gain on the *bistro 01* and *carousel* sequences according to ROC, and on *fishing* and *bistro 03* according to CC and KL. What these sequences have in common is that salient objects are not always in the brightest regions. This suggests that the improvement is due to the use of the HVS model which accounts visual sensitivity under a broad luminance range. 3) it is worth noting that there are some discrepancies among the three metrics, for example, the highest score using Itti *et al.*'s model is produced on the *balloon* sequence according to CC and ROC, but KL shows that Itti *et al.*'s model performs the best on the *park* sequence. This suggests that these metrics are not always unanimous in deciding the degree of similarity between predictions and eye tracking data. Therefore, the performance should be assessed by a combination of metrics to ensure a fair comparison.

TABLE IV
AVERAGE RESULTS OF 10 HDR VIDEOS

Model	ROC		CC		KL	
	Ave.	ST dev.	Ave.	ST dev.	Ave.	ST dev.
Proposed ST with dynamic fusion	0.74	0.05	0.34	0.10	4.57	0.36
Proposed S	0.72	0.05	0.27	0.11	4.82	0.45
Proposed T	0.68	0.08	0.29	0.12	4.69	0.51
Proposed ST with average fusion	0.68	0.06	0.29	0.10	4.68	0.44
Itti-ST	0.69	0.08	0.23	0.13	5.10	0.38
Itti-S	0.68	0.08	0.21	0.12	5.08	0.45
Itti-T	0.66	0.07	0.20	0.13	5.29	0.20
Gaussian	0.56	0.05	0.15	0.11	6.38	0.35

ST stands for spatio-temporal, S stands for spatial and T stands for temporal. Proposed ST with average fusion uses the average instead of dynamic fusion.

Fig. 15 shows the ideal ROC and the ROC obtained by our proposed model and Itti *et al.*'s model for the HDR video dataset. As it is observed, the predictive efficiency of the proposed model is 95% of the ideal ROC.

Table IV shows the performance of the spatial saliency map and the temporal saliency map for both models. The best performance stems from incorporating both spatial and temporal information for both models. The results of the simple center-bias Gaussian filter are also reported in Table IV to provide a reasonable lower bound for evaluations. Moreover, to illustrate the merit of the proposed dynamic fusion method, the results of average fusion are also included in Table IV. As it is observed, the dynamic fusion method outperforms other methods. We should also note that all three metrics show that the proposed spatial and temporal saliency maps outperform Itti *et al.*'s method.

VI. CONCLUSION AND FUTURE WORK

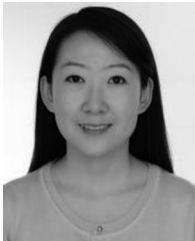
In this paper, we proposed a spatio-temporal model simulating the bottom-up visual attention for HDR images and videos. Considering the HVS's properties related to HDR's characteristics of wide luminance ranges and wide color gamut, the proposed model addresses limitations of existing state-of-the-art models. The spatio-temporal saliency map is obtained by combining low level visual features such as color, orientation, intensity and motion. A dynamic fusion method is used to control the relative weights of the spatial map and temporal map based on the motion intensity of HDR video. Quantitative evaluations have been conducted with data collected from an eye tracking study. The proposed method shows better performance on both image and video datasets.

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.

REFERENCES

- [1] Y. Fang *et al.*, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia.*, vol. 14, no. 1, pp. 187–198, Feb. 2012.
- [2] A. Borji, D. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.
- [3] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky, "Eye tracking in web search tasks: design implications," in *Proc. Symp. ETRA 2002: Eye Tracking Res. Appl. Symp.*, 2002, pp. 51–58.
- [4] T. van Gog, and K. Scheiter, "Eye tracking as a tool to study and enhance multimedia learning," *Learn. Instruction*, vol. 20, no. 2, pp. 95–99, Apr. 2010.
- [5] Y. F. Ma, X. S. Hua, L. Lu, and H. J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia.*, vol. 7, no. 5, pp. 907–919, Oct. 2012.
- [6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
- [7] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE*, 2004, vol. 5200, pp. 64–78.
- [8] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.
- [9] S.-H. Lee, J.-H. Kim, K. P. Choi, J.-Y. Sim, and C.-S. Kim, "Video saliency detection based on spatiotemporal feature learning," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 1120–1124.
- [10] A. Borji, and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
- [11] R. Brémond, J. Petit, and J.-P. Tarel, "Saliency maps of high dynamic range images," in *Proc. 11th Eur. Conf. Trends Topics Comput. Vis.*, 2012, pp. 118–130 2012.
- [12] C. Koch, and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Matters of Intelligence*, Berlin, Germany: Springer-Verlag, 1987, pp. 115–141.
- [13] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Proc. SPIE*, 1992, vol. 1616, pp. 2–15.
- [14] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vis. Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.
- [15] Y. Zhai, and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.
- [16] R. W. G. Hunt, *The Reproduction of Color* 6th, Hoboken, NJ, USA: Wiley, 2004.
- [17] M. H. Kim, T. Weyrich, and J. Kautz, "Modeling human color perception under extended luminance levels," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 27, 2009.
- [18] CIE Publication 131, "The CIE 1997 interim colour appearance model (simple version), CIECAM97s," *Color Res. Appl.*, vol. 23, no. 6, pp. 431, Dec. 1998.
- [19] N. Moroney *et al.*, "The CIECAM02 color appearance model," in *Proc. IS&T 10th Color Imaging Conf.*, 2002, pp. 23–27.
- [20] M. R. Luo *et al.*, "Quantifying color appearance. Part I. LUTCHI color appearance data," *Color Res. Appl.*, vol. 16, no. 3, pp. 166–180, 1991.
- [21] O. E. Uscanga, *On the fundamental data-base of normal and dichromatic color vision*, Ph.D. dissertation, Dept. of Math. and Natural Sci., Univ. of Amsterdam, Amsterdam, The Netherlands, 1979.
- [22] G. E. Müller, "Ueber die farbeempfindungen," *Psychophysische Untersuchungen*, Leipzig, Germany: Barth, 1930.
- [23] J. J. Vos, and P. L. Walraven, "On the derivation of the foveal receptor primaries," *Vision Research*, vol. 11, no. 8, pp. 799–818, 1971.
- [24] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Lossy compression of high dynamic range images and video," in *Proc. SPIE*, 2006, vol. 6057, pp. 311–320.
- [25] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting visible differences in high dynamic range images: Model and its calibration," in *Proc. SPIE*, 2005, vol. 5666, pp. 204–214.
- [26] B. Horn, and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.
- [27] T. Vaudrey, A. Wedel, C.-Y. Chen, and R. Klette, "Improving optical flow using residual and Sobel edge images," in *Proc. Arts Technol.: 1st Int. Conf.*, 2010, pp. 215–222.
- [28] C. Tomasi, and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th IEEE Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.
- [29] H.-C. Nothdurft, "Saliency from feature contrast: Additivity across dimensions," *Vision Research*, vol. 40, no. 10, pp. 1183–1201, 2000.
- [30] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, 1993.
- [31] *Description of HDR Sequences*, ISO/IEC JTC1/SC29/WG11 MPEG2014/m31957, Oct. 2014.

- [32] J. Froehlich *et al.*, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in *Proc. IS&T/SPIE Electron. Imaging*, 2014 Art. ID 90230X.
- [33] Y. Dong, E. Nasiopoulos, M. Pourazad, and P. Nasiopoulos, "High dynamic range video eye tracking dataset," in *Proc. 2nd Int. Conf. Electron., Signal Process. Commun.*, 2014, pp. 56–59.
- [34] H. Seetzen *et al.*, "High dynamic range display systems," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 760–768, 2004.
- [35] O. Le Meur, and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behavior Res. Methods*, vol. 45, no. 1, pp. 251–266, 2013.
- [36] Z. Ren, S. Gao, D. Rajan, L.-T. Chia, and Y. Huang, "Spatiotemporal saliency detection via sparse representation," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2012, pp. 158–163.
- [37] B. J. Stankiewicz, N. J. Anderson, and R. J. Moore, "Using performance efficiency for testing and optimization of visual attention models," in *Proc. SPIE*, 2010, vol. 7867, pp. 78670Y.
- [38] Q. Zhao, and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vision*, vol. 11, no. 3, pp. 1–15, 2011.
- [39] J. Harel, "A Saliency implementation in MATLAB," [Online]. Available: <http://www.vision.caltech.edu/harel/share/gbvs.php>. Accessed on Jan. 15, 2014.
- [40] "Experiment Center 2 Manual," SensoMotoric Instruments, Tetlow, Germany, 2010, Version 2.4.



Yuanyuan Dong (S'13) received the B.Sc. degree in telecommunication engineering from Tianjin University, Tianjin, China, 2011, and the M.A.Sc. degree in electrical and computer engineering from the University of British Columbia, Vancouver, BC, Canada, in 2014.

Her research interests include stereoscopic and high dynamic range image/video.



Mahsa T. Pourazad received the B.A.Sc. degree in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 2000, the M.A.Sc. degree in electrical and computer engineering from the University of Manitoba, Winnipeg, MB, Canada, in 2004, and the Ph.D. degree in electrical and computer engineering from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2010.

She is currently a Research Scientist with TELUS Communications Inc., Vancouver, BC, Canada, and a System Consultant with the Institute for Computing, Information, and Cognitive Systems, UBC. Her research interests include 3-D video processing, 3-D quality of experience, and multiview video compression.

Ms. Pourazad is a Member of the Standard Council of Canada (SCC) and the Moving Picture Experts Group (MPEG).



Panos Nasiopoulos (S'91–M'91–SM'12) received the B.Sc. degree in physics from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1980, and the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical and computer engineering from the University of British Columbia (UBC), Vancouver, BC, Canada, in 1985, 1988, and 1994, respectively.

He was President of Daikin Comtec US (founder of DVD) and Executive Vice President of Sonic Solutions. He is currently the Director of the Institute for Computing, Information and Cognitive Systems, UBC. He is also a Professor with the Department of Electrical and Computer Engineering, UBC, and Director of the Master of Software Systems Program at UBC, Vancouver, BC, Canada. He was the inaugural holder of the Dolby Professorship in Digital Multimedia, UBC.

Prof. Nasiopoulos is a Registered Professional Engineer in British Columbia. He has been an active Member of the SCC and the Association for Computing Machinery.