Saliency Prediction on Stereoscopic Videos

Haksub Kim, Sanghoon Lee, *Senior Member, IEEE*, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—We describe a new 3D saliency prediction model that accounts for diverse low-level luminance, chrominance, motion, and depth attributes of 3D videos as well as high-level classifications of scenes by type. The model also accounts for perceptual factors, such as the nonuniform resolution of the human eye, stereoscopic limits imposed by Panum's fusional area, and the predicted degree of (dis) comfort felt, when viewing the 3D video. The high-level analysis involves classification of each 3D video scene by type with regard to estimated camera motion and the motions of objects in the videos. Decisions regarding the relative saliency of objects or regions are supported by data obtained through a series of eye-tracking experiments. The algorithm developed from the model elements operates by finding and segmenting salient 3D space-time regions in a video, then calculating the saliency strength of each segment using measured attributes of motion, disparity, texture, and the predicted degree of visual discomfort experienced. The saliency energy of both segmented objects and frames are weighted using models of human foveation and Panum's fusional area yielding a single predictor of 3D saliency.**

***Index Terms*—Stereoscopic, scene classification, eye-tracker, saliency strength, saliency energy, human visual system.**

I. Introduction

THE development of 3D display technologies and devices has led to a rapid expansion of the 3D imaging market. This rapid technological growth has been accompanied by greatly increased popularity of 3D entertainment, as exemplified by the many recent impressive 3D-cinema and 3DTV productions. An important ingredient in further improving 3D video processing technologies are efforts to incorporate better models of 3D perception. Among these, saliency detection, or the automated discovery of points of high visual interest, conspicuity, or task relevance, is a particularly challenging problem. Yet it is an exceedingly promising problem, since it has the potential to dramatically affect current approaches to such important applications as object detection and recognition, image/video compression, visual navigation and image/video quality assessment [1].

Despite considerable prior research on visually salient region detection on natural scenes [3]–[13], [17], it remains difficult to precisely define the characteristics of visual attraction or to create models that reliably and automatically detect salient regions. The 3D saliency problem, whereby points of presumed visual interest are found in a reconstructed 3D visual space (e.g., from stereoscopic data) has received much less attention [2], [14]–[16], although such 3D factors as depth and shape certainly affect visual attention. Towards closing this gap of knowledge, we propose a framework for 3D salient region detection in stereoscopic videos that utilizes a bottom-up approach, identifies camera motion, and classifies each dynamic scene accordingly. By analyzing the statistical distribution of expected object motions with regard to camera motion or dynamic focus changes, the detection of salient regions is simplified. We define the notion of 3D "*saliency strength*," which quantifies the degree of likely visual attraction to a region based on measure of visual information content in the region. "*Saliency strength*" is computed over space, time, and disparity, then combined into a 3D space-time "*saliency energy*."

While scene classification is a powerful contextual cue for determining saliency autonomously, most prior efforts in this direction have been focused on scene categorization through pattern recognition. Such techniques attempt, for example, to recognize or classify specific objects or environments (e.g., sky, ground) based on learning feature statistics from a database [18], [19]. However, it is difficult to generalize such an approach to selecting salient regions. We instead describe a saliency detection framework that uses scene classification based on motion information, then extracts visual saliency measurements suitable for each type of scene.

A variety of sophisticated 2D saliency detection algorithms have been devised that analyze spatial image characteristics [3]–[12] such as processed luminance, color and texture features or measured fixations statistics. The principles underlying such 2D saliency detection methods could be extended to 3D saliency detection problems. However to be able to predict 3D video saliency with high accuracy, it will be necessary to account for 3D characteristics such as depth motion, disparity perception and visual discomfort [14]–[17]. For example, the authors of [14] analyzed stereoscopic fixations in 3D as a function of luminance and depth gradients. They found while large luminance and contrast gradients tend to draw fixations, large depth gradients tended to repel the point of gaze in favor of smooth depth regions. It is also important to consider 3D motion information and its effects on human visual attention. The authors of [15] presented a saliency detection approach for stereoscopic video, based on extracting disparity, motion, visual discomfort and object information in space

Manuscript received December 12, 2012; revised September 30, 2013 and December 29, 2013; accepted January 4, 2014. Date of publication January 29, 2014; date of current version February 18, 2014. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) Funded by the Ministry of Education under Grant NRF-2013R1A1A2A10011764. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Damon M. Chandler. *(Corresponding author: S. Lee.)*

H. Kim and S. Lee are with the Center for Information Technology of Yonsei, Yonsei University, Seoul 120-749, Korea (e-mail: khsphillip@yonsei.ac.kr; slee@yonsei.ac.kr).

A. C. Bovik is with the Department of Electrical and Computer Engineering, Laboratory for Image and Video Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

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/TIP.2014.2303640

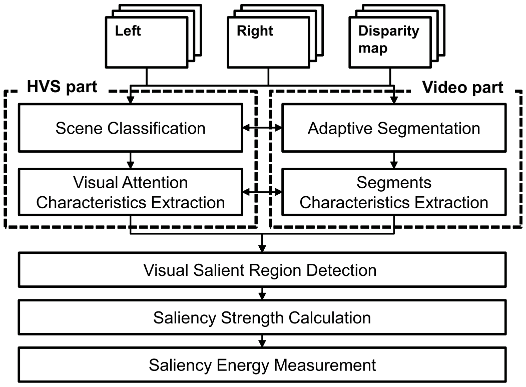

Left
Right
Disparity map
HVS part
Video part
Scene Classification
Adaptive Segmentation
Visual Attention Characteristics Extraction
Segments Characteristics Extraction
Visual Salient Region Detection
Saliency Strength Calculation
Saliency Energy Measurement

Fig. 1. Procedure of the saliency energy measurement from the input sequence and its disparity map.

and time. The authors of [16] propose to detect salient regions based on motion and disparity. Conducting a perception-based analysis of stereoscopic video is complicated by the introduction of such factors as vergence, accommodation, binocular rivalry, visual discomfort, and 3D scene statistics, all of which are related 3D geometry and the perception of 3D space [20]–[23]. In [20] and [21], the authors analyze 3D geometry and its effects on the 3D experience utilizing viewing distance, display size and disparity. The authors of [22] studied depth of focus (DoF) to understand visibility as it is affected by depth perception in stereoscopic video. In [23], a model of Panum's fusional area is used to predict visual discomfort when viewing stereoscopic video. In particular, the level of visual discomfort model is related to aspects of visual attention on stereoscopic video. This further motivates our desire to create 3D saliency models that reflect both 3D geometry and visual discomfort/comfort. Fig. 1 shows the processing framework of our proposed 3D saliency model. A video sequence and its disparity map are analyzed based on camera motion, object motion, and zoom. The saliency model is then uniquely designed for each class of scene. We verify this modeling stage using an eye-tracker [37]. Then, the video is segmented to isolate potential highly salient regions using spatial, temporal, and disparity saliency strength. Finally, we measure the overall saliency energy for each video frame, incorporating models of both the nonuniform resolution of retinal sampling (foveation) and of Panum's fusional area. In Section II, we describe the method of motion-based scene classification and how we extract visual attention as a function of scene type using the results of eye-tracking experiments. Section III explains an adaptive method of detecting salient regions by segmenting each scene type differently based on motion and disparity. Section IV describes the calculation of saliency strength in terms of spatial, disparity and temporal saliency strength factors. Section V outlines how the saliency energy of each stereoscopic video frame is measured using a model of foveation and Panum's fusional area. We conclude the paper in Section VI.

II. Motion-Based Scene Classification

A. Utilization of Motion Information

It is generally quite difficult to recognize visually salient regions in natural 2D or 3D images and videos in a manner that agrees with visual attention or gaze patterns. One feasible strategy to reduce prediction error is to classify each scene, then predict salient regions adaptively. This approach seeks to reduce visual diversity over each candidate region, and hence to increase reliability relative to "general" solutions. Our low-level approach to video scene classification relies on analyzing camera motion and object motion, as shown in Fig. 2.

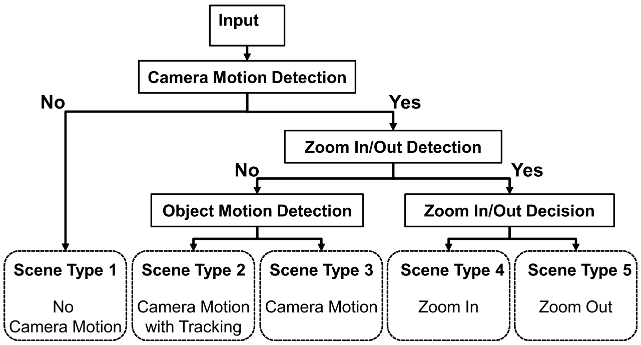
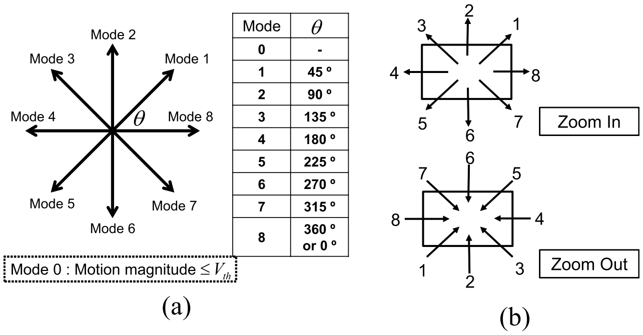


Fig. 2. Block diagram for scene classification.



Mode	θ
0	-
1	45 °
2	90 °
3	135 °
4	180 °
5	225 °
6	270 °
7	315 °
8	360 ° or 0 °

Fig. 3. Depiction of modes of the motion direction. (a) Motion quantization into 9 direction modes. (b) Expected behavior of motion direction in the presence of zoom.

1) Camera Motion Detection: We deploy a camera detection module that utilizes motion information obtained using an efficient optical flow algorithm [40]. The direction of each motion vector is quantized into one of nine modes as shown in Fig. 3(a). The motion at spatial index (u, v) in a frame is denoted $\mathbf{V}^{(u,v)} = (V_x^{(u,v)}, V_y^{(u,v)})$, where $1 \leqq u \leqq w$ and $1 \leqq v \leqq h$ (w and h are the width and height of the frame), and $V_x^{(u,v)}$ and $V_y^{(u,v)}$ are the horizontal and vertical motion components. This may be expressed in polar form $\mathbf{V_p}^{(u,v)} = (V_r^{(u,v)}, V_\theta^{(u,v)})$, where $V_r^{(u,v)}$ is the magnitude of the motion vector and $V_\theta^{(u,v)}$ is its orientation.

We utilize the two components, $V_r^{(u,v)}$ and $V_\theta^{(u,v)}$ in a mode decision process [Fig. 3(a)]. After finding the histogram of the quantized motion directions in each frame, the dominant motion direction is found: $V_{mode}^M = \arg\max_{mode}(\mathrm{hist}(\mathbf{V}_{mode}))$, where hist$(\cdot)$ is the histogram function and $\mathbf{V}_{mode}$ is the set $\{V^{(u,v)}\}$, where $V^{(u,v)}$ is the motion direction at (u, v). Finally, we decide whether camera motion[1] is present from the direction of the dominant

[1] Or other large motion resulting in an extensive motion field having highly coherent directionality.

motion

$$\begin{cases} \text{Scene Type 1} & ; \text{ if } V_{mode}^{M} \le V_{th} \\ \text{Camera motion exists} & ; \text{ otherwise} \end{cases} \tag{1}$$

where V_{th} indicates the threshold below which the velocity is assumed zero.[2]

2) Zoom Detection and Decision: As shown in Fig. 2, we classify scenes containing camera motion as either being dominated by zoom-induced motion, or otherwise. As shown in Fig. 3(a), eight direction modes are used for zoom detection. Fig. 3(b) shows the types of motion direction distributions that are used to classify scenes as "Zoomed" or not. The directions are quantized versions of the actual motion. This distribution is utilized for scene classification by

$$\begin{cases} \text{Scene Type 4 or 5;} & \text{if} \min\left(\dfrac{NumMode(m)}{I_{size} - NumMode(0)}\right) \ge \mathcal{Z}_{th} \\ \text{Scene Type 2 or 3;} & \text{otherwise} \end{cases} \tag{2}$$

where $I_{size} = w \times h$ and $NumMode(m)$ is the number of pixels in mode m in a frame. A scene is classified as "Zoom" if the minimum among $NumMode(m)$ for $1 \le m \le 8$ exceeds a threshold $\mathcal{Z}_{th}$. Moreover, as shown in Fig. 3(b), the polarity of zoom can be detected in a straightforward manner. Thus, scenes containing Zoom In vs. Zoom Out are classified as

$$\begin{cases} \text{Scene Type 4;} \\ \quad \text{if } \dfrac{\sum_{m=1}^{3} NumMode^{up}(m) + \sum_{m=5}^{7} NumMode^{lo}(m)}{\sum_{m=1}^{3} NumMode^{lo}(m) + \sum_{m=5}^{7} NumMode^{up}(m)} \ge 1 \\ \text{Scene Type 5; otherwise} \end{cases} \tag{3}$$

where $NumMode^{up}(m)$ indicates the number of the m^{th} mode in the upper half ($\le h/2$) of a frame and $NumMode^{lo}(m)$ indicates the number of the m^{th} mode in the lower half ($> h/2$) of a frame.

3) Object Motion Detection: Using a simple object motion detection module, scene types 2 and 3 are further classified to reflect object motion in the presence of camera motion. Camera motion occurs when tracking a target object or objects in scene type 2, and without tracking objects in scene type 3. In the case of scene type 2, the motion velocity magnitude is close to zero when the camera fixation tracks a moving object. Thus, the motion magnitude is used to classify scenes types 2 and 3:

$$\begin{cases} \text{Scene Type 2;} \\ \quad \text{if } \dfrac{NumMode(0)}{\sum_{m=1}^{8} NumMode(m) - NumMode(V_{mode}^{M})} \ge 1 \\ \text{Scene Type 3; otherwise.} \end{cases} \tag{4}$$

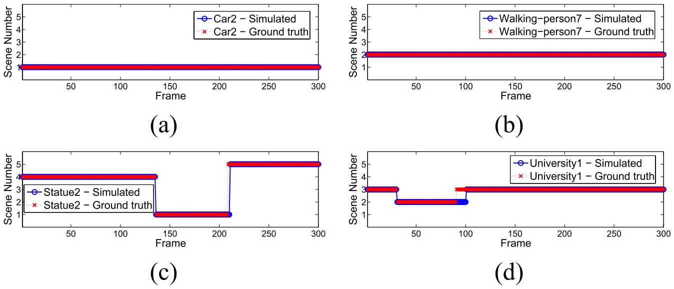


Fig. 4. Scene classification results and ground truth results as a function of frame number using the stereoscopic test sequences in [42]. (a) "Car2." (b) "Walking-person7." (c) "Statue2." (d) "University1".

In (4), the number of mode 0 motion directions (zero motion) is counted in each frame. If this number is larger than those of the other motion direction modes after excluding the dominant motion directions, the scene is classified as scene type 2. To verify the performance of the motion-based scene classification method, we conducted scene classification subjective experiments utilizing twenty human subjects. The subjects watched each test sequence frame-by-frame and classified each into one of the scene types. Ground truth was then taken to be the majority scene classification on each frame. The 3D test sequences were drawn from the IEEE Stereoscopic Imaging Database [42], the EPFL stereoscopic video database [43] and the mobile 3D TV video database [44]. Fig. 4 shows the scene classification results and ground truth examples on the test sequences. For quantitative evaluation, we measure the *Hit-ratio* indicating the ratio of correct classified frames to total frames. Table I shows the performance of the proposed scene classification over all test sequences. It can be seen that the performance exceeded 95% correct for all the sequences. There were a few misclassifications on scene types 2 or 3, likely arising from inaccuracies of the motion estimation or ambiguous motion scenes.

B. Fixation Behavior as a Function of Scene Type

Towards understanding human visual fixation behaviors as a function of scene types 1-5, we conducted an eye-tracking experiment in a dark room involving twenty persons with ages ranging from 20 to 30 years old. We used a "SMART EYE PRO" binocular eye-tracker [37] and a 23" LG polarization stereoscopic display having a resolution 1600x900.

To study the distinguishing characteristics of each scene at points of fixation, we computed angular disparity, motion speed, motion difference at both human fixation locations and at randomly picked "fixation" locations.[3] Fig. 5 shows the results for each scene where the error bars show 95% confidence intervals. Fig. 5(a) shows the normalized mean angular disparity at fixations and at randomly picked locations, where nearer points have larger weights than farther points as defined in (8). It can be seen that the angular disparities at true fixations are generally higher than at random "fixations" on all the five scenes, suggesting that humans tend to fixate on closer rather than farther objects. In Fig. 5(b), the 3D scene

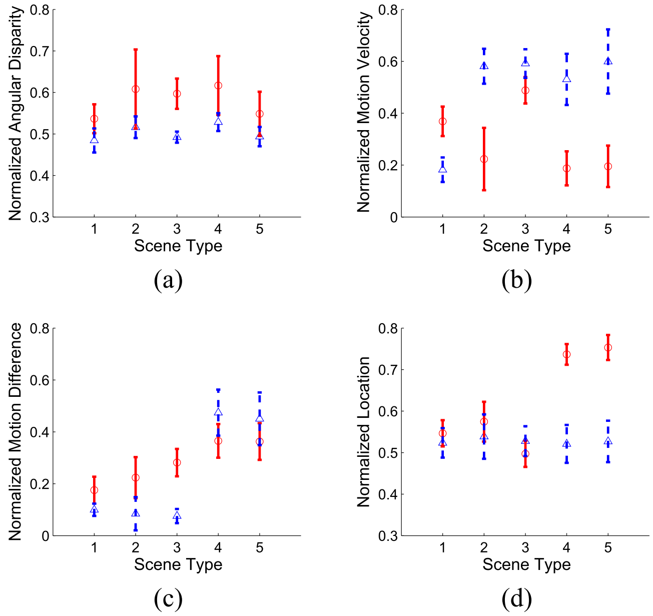
[2] We set $V_{th} = 2$ (pixel/frame).

[3] Up to 300 random fixation locations were generated.

TABLE I

THE PERFORMANCE OF MOTION-BASED SCENE CLASSIFICATION

Seq. Name	Car1	Car2	Walking-person1	Walking-person7	Walking-person8	University1
Scene Type	1	1	2	2	2	2, 3
Hit-ratio (%)	100	100	99.5	100	99.6	96.6
Seq. Name	University2	Statue2	Statue3	Street-lamp1	Crosswalk2	Library3
Scene Type	3	1, 4, 5	1, 4, 5	1, 4, 5	3	1
Hit-ratio (%)	95.5	99.3	99.7	99.5	97.7	100
Seq. Name	Library4	Marathon1	Restaurant1	Sidewalk-lateral1	Bike	Car
Scene Type	3	1	1	3	1	1
Hit-ratio (%)	96.7	100	100	97.3	100	100
Seq. Name	Feet	Hallway	Notebook	Sofa	Street	Balloons
Scene Type	1	1	1	1	1	2, 3
Hit-ratio (%)	100	100	100	100	100	95.7



(a) (b)

(c) (d)

Fig. 5. Plots of the 3D scene characteristics at fixations (red and solid line) and at random locations (blue and dot line) for each scene type using the test sequences [42]–[44]. (a) Normalized mean angular disparity. (b) Normalized mean speed. (c) Normalized mean motion difference. (d) Normalized location between center of frames and the fixation or randomly selected points.

characteristics are plotted against object speed. This suggests that humans tend to fixate on objects moving at higher speeds when there is no camera motion (scene type 1). However, when camera motion exists and objects are moving, humans tend to fixate on and track objects that moving more slowly. For the Zoom In and Out cases (scene types 4 and 5), in our model humans are assumed to tend to fixate on objects that are moving more slowly. In those instances the center of the image is fixated more often, since the motion is near zero there. Fig. 5(c) shows the normalized means of motion differences at fixations and at random locations. From (7), objects exhibiting larger motion differences are assigned larger weights than those having smaller motion differences. Generally, the motion differences at fixations are larger than elsewhere for scene type 3.

To summarize the results of the experiment, we found that most of the subjects directed their attention to foreground objects having large crossed disparities on the stereoscopic video. Moreover, for scenes of type 1, they tended to fixate on moving objects. On the other hand, we found that most subjects fixated on and tracked objects having near zero motion on scenes of type 2. On scenes of type 3, they would largely fixate on objects moving along trajectories not tracked by the camera motion. Finally, for scenes of types 4 and 5, most of the subjects concentrated their attention near the middle of the screen.

III. ADAPTIVE SEGMENTATION FOR DETECTING VISUALLY SALIENT REGIONS

A. Adaptive Segmentation Based on Scene Classification

The conclusions given in Section II-B indicate that human visual attention on 3D stereoscopic videos is, at least in part, directed towards objects based on their motion and disparity. Moreover, depending on the type of scene, the motion and disparity of each object affects visual attraction in different ways. Realizing this, we have designed our model to conduct adaptive segmentation based on the scene classification protocol described in Section II-A that is based on measured motion and disparity. The conclusions given in in Section II-B suggest that motion and disparity are significant saliency factors on scenes of types 1, 2 and 3. Thus, segmentation using motion and disparity is conducted on those scene types, and the motion and disparity segmentation maps are merged. For scenes of types 4 and 5 (zoom scenes), it is difficult to expect good segmentation performance based on motion information. For those types of scenes, we rely only on disparity information to conduct segmentation. The method of segmentation used to obtain motion and disparity based segmentation maps was a combination of k-means clustering [38] and a region growing algorithm [39], as follows. First, the two (left and right) optical flow maps are input to the motion segmentation module, where simple k-means clustering is applied on the x- and y-motions. We set the $k = 15$. The disparity-based segmentation is performed in parallel using the computed disparity map. As a result, two sets associated with motion- and disparity-based segmentation are obtained:

$$\begin{cases} \mathbf{S_m} = \{S_m^i | S_m^i \in \mathbf{S_m}, \quad i = 1, \ldots, n_m\} \\ \mathbf{S_d} = \{S_d^i | S_d^i \in \mathbf{S_d}, \quad i = 1, \ldots, n_d\} \end{cases}$$

where $\mathbf{S_m}$ and $\mathbf{S_d}$ are the motion- and disparity-based segmentation maps, n_m and n_d are the numbers of segments obtained

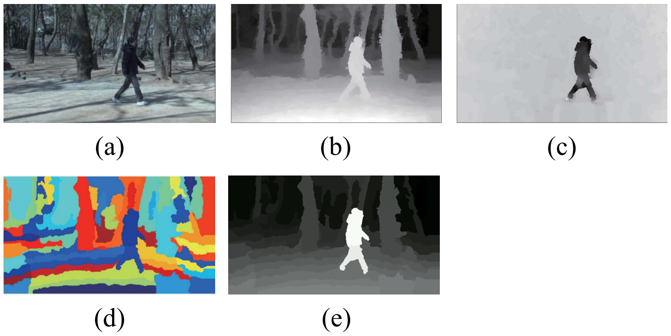


Fig. 6. Example of adaptive segmentation and salient region detection. (a) Original left image of the 109^{th} stereoscopic frame "Walking-person1" (scene type 2) [42]. (b) Disparity map. (c) Motion magnitude map. (d) Adaptive segmentation. (e) Salient region detection.

using motion- and disparity-based segmentation respectively, and i is the segment index. The final adaptive segmentation map, $\mathbf{S}$, is obtained as

$$\begin{cases} \mathbf{S} = \{S^i | S^i \in \mathbf{S_m} \cap \mathbf{S_d}\}, & \text{if Scene Type} = 1, 2 \text{ or } 3 \\ \mathbf{S} = \{S^i | S^i \in \mathbf{S_d}\}, & \text{if Scene Type} = 4 \text{ or } 5 \end{cases} \tag{5}$$

where n_s is the number of segments in $\mathbf{S}$ and $1 \le i \le n_s$. Fig. 6 shows an example of the adaptive segmentation process. Each segment in the segmentation map is represented by a unique value after merging the motion and disparity segmentations as shown in Fig. 6(d).

B. Segment Characteristics for Detecting Salient Regions

We define a few parameters that are used to characterize each segment extracted as in the preceding. First, normalized velocity and velocity difference parameters are defined using the averaged x-, y- and z-motion components of each segment. The x- and y-motion vectors are obtained using the optical flow algorithm in [40], while the z-motion (depth motion) is measured using the difference in disparity between the current and next frames of each segment. The normalized velocity magnitude (speed) parameter of the i^{th} segment, $\mathcal{M}_I^i$, is defined

$$\mathcal{M}_I^i = \widehat{V}^i / \widehat{V}^M \tag{6}$$

where i is the i^{th} segment, $\widehat{V}^i$ is the average speed ($\widehat{V}^i = \sqrt{(V_x^i)^2 + (V_y^i)^2 + (V_z^i)^2}$), V_x^i, V_y^i and V_z^i are the average x-, y- and z-motion components, and $\widehat{V}^M$ is the maximum speed of the segments for ∀ i which is used as a normalized factor.

Next, the velocity difference parameter indicates the difference in motion of each segment relative to a neighboring dominant motion:

$$\mathcal{M}_\Delta^i = \widehat{V}_\Delta^i / \widehat{V}_\Delta^M \tag{7}$$

where $\mathcal{M}_\Delta^i$ is the velocity difference parameter of the i^{th} segment. Using the velocity histogram of each segment, the degree of motion difference is determined. Let $\mathbf{V}_x$ and $\mathbf{V}_y$ be the sets of x- and y-velocities and $\mathbf{V}_z$ be the set of z-velocities for each segment, then $V_x^M = \max(\text{hist}(\mathbf{V}_x))$, $V_y^M = \max(\text{hist}(\mathbf{V}_y))$ and $V_z^M = \max(\text{hist}(\mathbf{V}_z))$. V_x^M, V_y^M and V_z^M are the dominant motions of x-, y- and z-velocities. Then,

$$\widehat{V}_\Delta^i = \sqrt{(V_x^M - V_x^i)^2 + (V_y^M - V_y^i)^2 + (V_z^M - V_z^i)^2}$$

captures the degree of velocity difference between the i^{th} segment and the dominant frame motion. $\widehat{V}_\Delta^M$ is a normalization factor, defined as the maximum difference in the frame for $\forall i$.

TABLE II

SALIENT REGION DETECTION FOR EACH SCENE TYPE

Scene Type 1 : No Camera Motion	
· Parameters : $\mathcal{M}_I^i\ \delta^i$	· $\mathcal{R}^i = \frac{w_I \mathcal{M}_I^i + w_\delta \delta^i}{w_I + w_\delta}$
Scene Type 2 : 2D Camera Motion with Object Tracking	
· Parameters : $\mathcal{M}_I^i$, δ^i	· $\mathcal{R}^i = \frac{w_I (1 - \mathcal{M}_I^i) + w_\delta \delta^i}{w_I + w_\delta}$
Scene Type 3 : 2D Camera Motion without Object Tracking	
· Parameters : $\mathcal{M}_\Delta^i$, δ^i	· $\mathcal{R}^i = \frac{w_\Delta \mathcal{M}_\Delta^i + w_\delta \delta^i}{w_\Delta + w_\delta}$
Scene Type 4 or 5 : Zoom (In or Out) Camera Motion	
· Parameters : $\mathcal{L}^i$, δ^i	· $\mathcal{R}^i = \frac{w_L \mathcal{L}^i + w_\delta \delta^i}{w_L + w_\delta}$

In a stereoscopic video, disparity is also an important saliency factor. Disparity is obtained from left and right frame using the depth estimation software in [41]. Define the disparity parameter

$$\delta^i = 1 - \frac{\mathcal{D}^i - \mathcal{D}_N^M}{\mathcal{D}_F^M - \mathcal{D}_N^M} \tag{8}$$

where $\mathcal{D}^i$ is the average disparity of the i^{th} segment, and $\mathcal{D}_N^M$ ($\mathcal{D}_F^M$) is the nearest (farthest) disparity of all segments in the frame ($\mathcal{D}_N^M \le \mathcal{D}^i \le \mathcal{D}_F^M$). This normalization causes nearer objects to have larger weights than farther objects.

Last, a segment location parameter $\mathcal{L}^i$ is obtained, which indicates the distance between the i^{th} segment and the central point of the frame:

$$\mathcal{L}^i = 1 - \frac{\sum_{j=1}^{n_s^i} \sqrt{(x_j^i - x_c)^2 + (y_j^i - y_c)^2}}{n_s^i \cdot \mathcal{L}^M} \tag{9}$$

where n_s^i is the number of pixels in the i^{th} segment, (x_j^i, y_j^i) is the j^{th} pixel of the i^{th} segment ($1 \le j \le n_s^i$), (x_c, y_c) is the center of the display and $\mathcal{L}^M$ is the maximum distance of each segment to the central point (x_c, y_c) for $\forall i$, which is used as a normalization factor ($\mathcal{L}^M = \sqrt{(x_c - w)^2 + (y_c - h)^2}$). As the distance between the i^{th} segment and the center of the screen is increased, $\mathcal{L}^i$ decreases. If the value of $\mathcal{L}^i$ is large, then the i^{th} segment is located close to the center of the screen. In order to better connect the characteristics derived from the eye tracking study with the scene classification/segmentation model, Table II tabulates the saliency parameters for each segment.

In Table II, $\mathcal{R}^i$ is the degree of saliency for the i^{th} segment expressed as a function of the motion, disparity and location parameters. A set of weights on speed (w_I), velocity difference (w_Δ), disparity (w_δ) and location (w_L) are employed. Those regions in each frame that are deemed salient form a spatial set $\mathbf{R} = \{\mathcal{R}^i | \mathcal{R}^i > \mathcal{R}_{th},\ i = 1, \ldots, n_s\}$ where $\mathcal{R}_{th}$ is a salient threshold and n_s is the number of segments in a frame. Finally, the overall saliency set is $\mathbf{R} = \{\mathcal{R}^k,\ \ k = 1, \ldots, n_r\}$, where $n_r = n_s - \text{num}(\mathcal{R}^i \leq \mathcal{R}_{th})$ for $1 \leq i \leq n_s$, which is the number of salient regions. Fig. 6 shows an example of this process, where brightness indicates the degree of saliency. In Fig. 6(e), the final map is shown.

IV. Saliency Strength

It has been well known that the density of photoreceptors in the retina is densest at the fovea and decreases exponentially to the retinal periphery. The size of the neuronal receptive fields in the retinal output and in the cortical map of the retina increases towards the visual periphery [29], [30]. Thus, the area perceived by the human eyes is a function of the viewing angle w.r.t the fovea. Thus, as people watch natural scenes, it is natural for them to move their fixations to capture information across the scene, and to fixate at more informative regions (relative to the task at hand) or objects in the image [29]. In order to capture such behavior, we define the notion of saliency strength, which quantifies the degree to which eyes are drawn to salient regions. The saliency strength measures the amount of certain types of information contained in salient regions of stereoscopic video. However, when watching stereoscopic video, visual discomfort can occur due to the presence of large disparities that are sustained along the temporal axis, which adversely affects visual attention. To capture the effects of visual discomfort in the saliency strength, we include certain visual factors related to stereoscopic visual discomfort [22].

A. Spatial Saliency Strength

The strength of spatial saliency is based on luminance, color, size and compactness. Our measure of spatial saliency strength is computed on a fused representation of the left and right frames obtained using the cyclopean image model described in [24].

1) Luminance: Several studies have been conducted on visual attention as a function of the luminance distribution of a stereoscopic image [3]–[14]. In [14], the authors found that the luminance contrast and luminance gradient of stereoscopically fixated patches are generally higher than in randomly selected patches. They defined the fixation-to-random luminance contrast and gradient ratios, and found that the ratios are generally larger than 1. We use this attentional characteristic as part of luminance saliency strength factors, using the luminance contrast and luminance gradient maps. Define luminance contrast and luminance gradient factors

$$\mathcal{C}_l^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} \mathcal{C}_l(x_n, y_n)}{\mathcal{R}_{size}^k}, \quad \mathcal{G}_l^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} \mathcal{G}_l(x_n, y_n)}{\mathcal{R}_{size}^k} \quad (10)$$

where $\mathcal{C}_l^k$ ($\mathcal{G}_l^k$) is the luminance contrast (gradient) factor of the k^{th} salient region in the original left frame as shown in Fig. 7(b) and (c). $\mathcal{R}_{size}^k$ is the number of pixels in the k^{th} salient region and (x_n, y_n) is the n^{th} pixel of the k^{th} salient region ($(x_n, y_n) \in \mathcal{R}^k$, $n = 1, \ldots, \mathcal{R}_{size}^k$). Using these elements, the luminance contrast saliency strength is expressed as

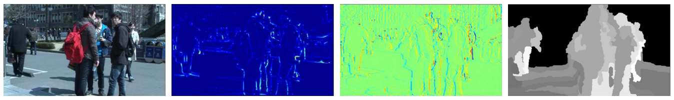


Fig. 7. Luminance saliency strength. (a) Original left image of the 109^{th} stereoscopic frame "University2" [42]. (b) Luminance contrast map (window size 10 pixels). (c) Luminance gradient map. (d) Luminance saliency strength map.

$$\begin{cases} \mathcal{W}_{lc}^k &= 1, \qquad \text{if } \frac{\mathcal{C}_l^k}{\widehat{\mathcal{C}_l}} > 1 \\ \mathcal{W}_{lc}^k &= \frac{\mathcal{C}_l^k}{\widehat{\mathcal{C}_l}}, \qquad \text{otherwise} \end{cases} \quad (11)$$

where $\mathcal{W}_{lc}^k$ is the luminance contrast saliency strength of the k^{th} salient region, $\mathbf{C_l}$ is luminance contrast map, and $\widehat{\mathcal{C}_l} = \text{mean}(\mathbf{C_l})$. This saliency strength measure takes its highest value when the luminance contrast factors of the k^{th} salient region are bigger than the average luminance contrast. It decreases when the salient region contains lower luminance contrast values than $\widehat{\mathcal{C}_l}$. Similarly, the luminance gradient saliency strength is expressed as

$$\begin{cases} \mathcal{W}_{lg}^k &= 1, \qquad \text{if } \frac{\mathcal{G}_l^k}{\widehat{\mathcal{G}_l}} > 1 \\ \mathcal{W}_{lg}^k &= \frac{\mathcal{G}_l^k}{\widehat{\mathcal{G}_l}}, \qquad \text{otherwise} \end{cases} \quad (12)$$

where $\mathcal{W}_{lg}^k$ is the luminance gradient saliency strength of the k^{th} salient region, $\mathbf{G_l}$ is the luminance contrast map, and $\widehat{\mathcal{G}_l} = \text{mean}(\mathbf{G_l})$.

Using (11) and (12), the luminance saliency strength is

$$\mathcal{W}_l^k = \omega_{lc}\mathcal{W}_{lc}^k + \omega_{lg}\mathcal{W}_{lg}^k \quad (13)$$

where $\mathcal{W}_l^k$ is the luminance saliency strength of the k^{th} salient region. There is correlation between luminance contrast and luminance gradient values, we combine these two values using weighted sum. ω_{lc} and ω_{lg} are the weights of luminance gradient and luminance contrast. It increases (decreases) when the salient region contains higher (lower) luminance contrast and gradient values. Fig. 7(d) shows the luminance saliency strength where again, brightness indicates the importance of saliency.

2) Color: Numerous studies on saliency detection have been conducted on the effects on visual attention of color information [3], [4]. The authors of [3] considered the role of color contrast and found it to be highly correlated with visual fixation locations. Accordingly, we use color contrast measurements to calculate the saliency strength. We also use measurements of the color gradient to capture the local rate of color changes. These are computed after converting the images into the perceptually uniform CIELab color space [46]. Define the color contrast and gradient factors respectively as

$$\mathcal{C}_c^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} \mathcal{C}_c^a(x_n, y_n) + \mathcal{C}_c^b(x_n, y_n)}{2\mathcal{R}_{size}^k} \quad (14)$$

$$\mathcal{G}_c^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} \mathcal{G}_c^a(x_n, y_n) + \mathcal{G}_c^b(x_n, y_n)}{2\mathcal{R}_{size}^k} \quad (15)$$

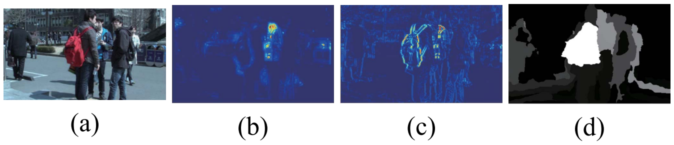

(a) (b) (c) (d)

Fig. 8. Color saliency strength. (a) Original left image of the 109^{th} stereoscopic frame "University2" [42]. (b) Color contrast map (window size 10 pixels). (c) Color gradient map. (d) Color saliency strength map.

on the k^{th} salient region in the cyclopean frame, depicted in Fig. 8(b) and (c). $\mathcal{C}_c^a$ and $\mathcal{C}_c^b$ are the color contrast maps of the color maps in Fig. 8(a) and (b). $\mathcal{G}_c^a$ and $\mathcal{G}_c^b$ are the color gradient maps of the color maps in Fig. 8(a) and (c). Using these elements, the color contrast saliency strength is represented as

$$\begin{cases} \mathcal{W}_{cc}^k &= 1, \quad \text{if } \frac{\mathcal{C}_c^k}{\widehat{\mathcal{C}}_c} > 1 \\ \mathcal{W}_{cc}^k &= \frac{\mathcal{C}_c^k}{\widehat{\mathcal{C}}_c}, \quad \text{otherwise} \end{cases} \tag{16}$$

where $\mathcal{W}_{cc}^k$ is the color contrast saliency strength of the k^{th} salient region. Let $\mathbf{C_c}$ be the color contrast map, and take $\widehat{\mathcal{C}}_c = \text{mean}(\mathbf{C_c})$. Similarly, the color gradient saliency strength is expressed

$$\begin{cases} \mathcal{W}_{cg}^k &= 1, \quad \text{if } \frac{\mathcal{G}_c^k}{\widehat{\mathcal{G}}_c} > 1 \\ \mathcal{W}_{cg}^k &= \frac{\mathcal{G}_c^k}{\widehat{\mathcal{G}}_c}, \quad \text{otherwise} \end{cases} \tag{17}$$

where $\mathcal{W}_{cg}^k$ is the color contrast saliency strength of the k^{th} salient region. Let $\mathbf{G_c}$ be the color gradient map, and $\widehat{\mathcal{G}}_c = \text{mean}(\mathbf{G_c})$. Using (16) and (17), the color saliency strength of the k^{th} salient region is

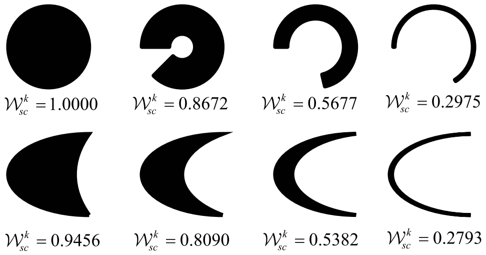
$$\mathcal{W}_c^k = w_{cc}\mathcal{W}_{cc}^k + w_{cg}\mathcal{W}_{cg}^k. \tag{18}$$

Naturally, there is correlation between the color contrast and the color gradient. Hence, we combine the two elements by a weighted sum where w_{cc} and w_{cg} weight the color contrast and the color gradient, respectively. The color saliency strength increases (decreases) when the salient region contains higher (lower) color contrast and gradient values. Fig. 8(d) shows the color saliency strength map where brightness indicates the importance of saliency.

3) Size and Compactness: Many studies have addressed the relationship between visual attention and the size and shapes of objects [31], [32]. For example, the authors of [31] relate object size to human recognition and perception. They define the relative object size (ROS) as the ratio of the number of pixels of an object to the number of pixels in the image, and define a simple threshold on suitable object size (ROS > 5%). Generally, humans perceive, recognize and more frequently fixate larger (and often nearer) objects. Using this observation, we define an object size saliency strength factor in terms of the ROS of each salient region:

$$\begin{cases} \mathcal{W}_{sc1}^k &= 1, \quad \text{if } ROS^k \geq ROS_{th} \\ \mathcal{W}_{sc1}^k &= \dfrac{ROS^k}{ROS_{th}}, \quad \text{otherwise} \end{cases} \tag{19}$$

where $ROS^k = \frac{R_{size}^k}{w \times h} \times 100$ and $ROS_{th} = 5$.



$\mathcal{W}_{sc}^k = 1.0000$
$\mathcal{W}_{sc}^k = 0.8672$
$\mathcal{W}_{sc}^k = 0.5677$
$\mathcal{W}_{sc}^k = 0.2975$
$\mathcal{W}_{sc}^k = 0.9456$
$\mathcal{W}_{sc}^k = 0.8090$
$\mathcal{W}_{sc}^k = 0.5382$
$\mathcal{W}_{sc}^k = 0.2793$

Fig. 9. Spatial compactness saliency strength of various regions.

The authors of [32] studied the question of how object size affects object perception. They varied the widths of objects with height fixed, then measured visual sensitivity. They found that human can adapt to variations in wider objects more easily than narrower objects. Inspired by this result, define the compactness of a salient region as the mean distance between the central point and other points inside the salient region:

$$\mathcal{W}_{sc2}^k = \frac{\eta_{den}^k}{\text{mean}(\sqrt{(x_c^k - x_n)^2 + (y_c^k - y_n)^2})}, \tag{20}$$

$$x_c^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} x_n}{\mathcal{R}_{size}^k}, \quad y_c^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} y_n}{\mathcal{R}_{size}^k} \tag{21}$$

where $(x_c^k,\ y_c^k)$ is the center of $\mathcal{R}^k$. Since a circle has the highest possible compactness amongst objects of a given size, the compactness $\eta_{den}^k = \text{mean}(\sqrt{\mathbf{x}^2 + \mathbf{y}^2})$ is used to normalize the saliency strength of other objects of the same size, where $\mathbf{x} = \{1, \ldots, \text{round}(\sqrt{\mathcal{R}_{size}^k/\pi})\}$ and $\mathbf{y} = \{1, \ldots, \text{round}(\sqrt{\mathcal{R}_{size}^k/\pi})\}$. Fig. 9 shows the region densities of various shapes.

Finally, we express the overall saliency strength factor incorporating size and compactness saliency using (19) and (20):

$$\mathcal{W}_{sc}^k = \mathcal{W}_{sc1}^k \cdot \mathcal{W}_{sc2}^k. \tag{22}$$

Thus, large and dense objects have high saliency strength, while small and low compactness objects have low saliency strength.

Using (13), (18) and (22), the overall spatial saliency strength is expressed

$$\mathcal{W}_S^k = (w_l\mathcal{W}_l^k + w_c\mathcal{W}_c^k) \cdot \mathcal{W}_{sc}^k \tag{23}$$

where w_l (w_c) weights the luminance (color) saliency strength. There is correlation between luminance and color saliency strengthes, so that we combine the two elements by using the weighted sum. In contrast, we multiply the size and compactness saliency strength to the weighted sum of luminance and color saliency strengthes because it is difficult to find correlation between the two factors.

B. Disparity Saliency Strength

1) Depth Discontinuities: Depth discontinuities are an important factor in defining disparity saliency strength. In particular, the authors of [14] defined the relationship between 3D visual fixations, disparity contrast and disparity

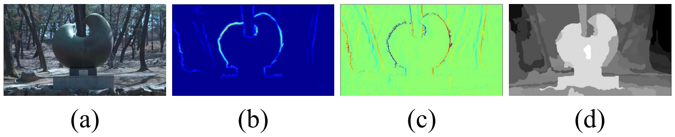


Fig. 10. Depth discontinuity saliency strength. (a) Original left image of the 250th stereoscopic frame of “Statue2” [42]. (b) Disparity contrast map (window size 10 pixels). (c) Disparity gradient map. (d) Depth discontinuity saliency strength map.

gradient. They found that fixated disparity contrasts and disparity gradients are generally *lower* than disparity contrasts and disparity gradients at random location. They found that the ratios of disparity contrasts and gradients of true fixations relative to random fixations are generally smaller than 1. They conclude that humans tend to fixate away from large disparity gradients and contrasts, preferring instead smooth depth regions. Thus, define disparity contrast and disparity gradient factors

$$\mathcal{C}_d^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} \mathcal{C}_d(x_n, y_n)}{\mathcal{R}_{size}^k}, \quad \mathcal{G}_d^k = \frac{\sum_{n=1}^{\mathcal{R}_{size}^k} \mathcal{G}_d(x_n,y_n)}{\mathcal{R}_{size}^k} \qquad (24)$$

where $\mathcal{C}_d^k$ ($\mathcal{G}_d^k$) is the disparity contrast (gradient) factor of the k^{th} salient region. $\mathcal{C}_d$ ($\mathcal{G}_d$) is the disparity contrast (gradient) map of the original left frame as shown in Fig. 10(b) and (c). These then define the disparity contrast saliency strength factor

$$\begin{cases} \mathcal{W}_{dc}^k = 1, & \text{if } \frac{\mathcal{C}_d^k}{\widehat{\mathcal{C}}_d} < 1 \\ \mathcal{W}_{dc}^k = \frac{\widehat{\mathcal{C}}_d}{\mathcal{C}_d^k}, & \text{otherwise} \end{cases} \qquad (25)$$

where $\mathcal{W}_{dc}^k$ is the disparity contrast saliency strength of the k^{th} salient region, $\mathbf{C_d}$ is the disparity contrast map, and $\widehat{\mathcal{C}}_d = \text{mean}(\mathbf{C_d})$. This strength is highest when the disparity contrast factors of the k^{th} salient region are smaller than the average disparity contrast. It decreases when the salient region has a higher disparity contrast than $\widehat{\mathcal{C}}_l$. Similarly, the disparity gradient saliency strength is

$$\begin{cases} \mathcal{W}_{dg}^k = 1, & \text{if } \frac{\mathcal{G}_d^k}{\widehat{\mathcal{G}}_d} < 1 \\ \mathcal{W}_{dg}^k = \frac{\widehat{\mathcal{G}}_d}{\mathcal{G}_d^k}, & \text{otherwise} \end{cases} \qquad (26)$$

where $\mathcal{W}_{dg}^k$ is the disparity gradient saliency strength of the k^{th} salient region, $\mathbf{G_d}$ is the disparity gradient map, and $\widehat{\mathcal{G}}_d = \text{mean}(\mathbf{G_d})$. Using (25) and (26), the depth discontinuity saliency strength is

$$\mathcal{W}_{dd}^k = w_{dc}\mathcal{W}_{dc}^k + w_{dg}\mathcal{W}_{dg}^k \qquad (27)$$

where $\mathcal{W}_{dd}^k$ is the depth discontinuity saliency strength of the k^{th} salient region. There is correlation between disparity contrast and disparity gradient, thus we combine two elements by weighted sum. w_{dc} and w_{dg} are the weights of the disparity contrast and disparity gradient elements. This decreases (increases) when the salient region contains higher (lower) disparity contrast and gradient values. Fig. 10(d) depicts an example of depth discontinuity saliency strength, where brightness indicates the degree of saliency.

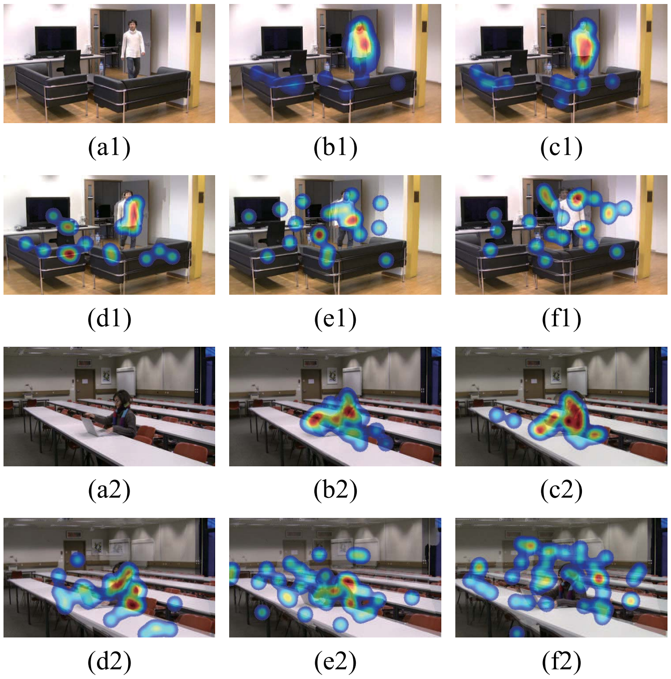


Fig. 11. Fixation tracing using an eye tracker [37] for EPFL stereoscopic video database [43]. (a1) Original left video of the 62th frame “sofa.” (a2) Original left video of the 20th frame “notebook.” (b1)-(b2) Eye-tracker results (disparity level=1). (c1)-(c2) Eye-tracker results (disparity level=2). (d1)-(d2) Eye-tracker results (disparity level=3). (e1)-(e2) Eye-tracker results (disparity level=4). (f1)-(f2) Eye-tracker results (disparity level=5).

2) Visual Discomfort: The disparity of an object and its size affect whether fixating on it causes visual discomfort. If the disparity does not lie within a zone of comfort, the viewer will feel increased visual discomfort which can subsequently affect visual attention. Previous studies have focused on visual discomfort associated with disparity [21]–[23]. The EPFL database [43] was constructed with five different disparity levels for analyzing visual discomfort in terms of disparity. They found that when disparity levels are increased, visual discomfort also tends to increase. Thus, we conducted an eye-tracking experiment using this database using the experimental environment described in Section II-B. Fig. 11 shows the experimental results obtained using the eye-tracker. At low disparity levels 1 and 2, fixations tended to land on salient objects. By contrast, at larger disparity levels are 3, 4 and 5, the fixation distribution became diffused as compared to the lower disparity levels.

To capture this observation in the calculation of saliency strength, we employ the zone of comfort that defines boundaries on disparity outside of which visual comfort is lost. The zone of comfort is typically defined to be where the angular disparity is in the range of (−1°, 1°) [22]. The average angular disparity is used here to define the visual discomfort saliency strength. The relationship between measured pixel disparity and angular disparity is laid out in Fig. 12.

Fig. 12(a) shows the stereoscopic geometry of a frame relative to the size of the stereoscopic display and the distance (in meters) from the human eye. In Fig. 12(b), the stereoscopic geometry is shown from above. Table III shows the parameters

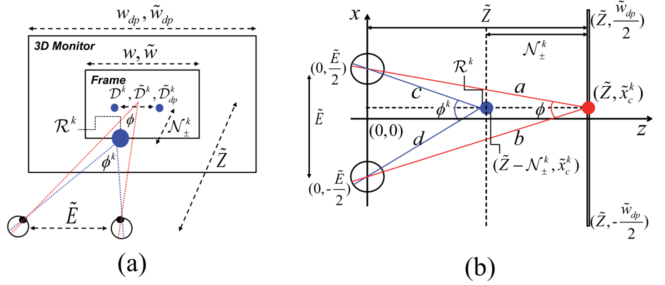

$w_{dp}, \tilde{w}_{dp}$
3D Monitor
$w, \tilde{w}$
Frame
$\mathcal{D}^k, \tilde{\mathcal{D}}^k, \tilde{\mathcal{D}}^k_{dp}$
$\mathcal{R}^k$
ϕ
$\mathcal{N}^k_{\pm}$
ϕ^k
$\tilde{Z}$
$\tilde{E}$
(a)
x
$\tilde{Z}$
$(\tilde{Z}, \frac{\tilde{w}_{dp}}{2})$
$\mathcal{N}^k_{\pm}$
$(0, \frac{\tilde{E}}{2})$
$\mathcal{R}^k$
a
c
$(\tilde{Z}, \tilde{x}^k_c)$
ϕ^k
ϕ
$\tilde{E}$
$(0,0)$
z
d
b
$(\tilde{Z} - \mathcal{N}^k_{\pm}, \tilde{x}^k_c)$
$(0, -\frac{\tilde{E}}{2})$
$(\tilde{Z}, -\frac{\tilde{w}_{dp}}{2})$
(b)

Fig. 12. Stereoscopic geometry (a) 3D space (front view). (b) 2D space (top-down view).

TABLE III

PARAMETER DESCRIPTION OF STEREOSCOPIC GEOMETRY

• Domain : Frame
· w ($\tilde{w}$) : Frame width in pixels (meters)
· $\mathcal{D}^k$ ($\tilde{\mathcal{D}}^k$) : Disparity of the k^{th} salient region in pixels (meters)
• Domain : Display
· w_{dp} ($\tilde{w}_{dp}$) : Display width in pixels (meters)
· $\tilde{\mathcal{D}}^k_{dp}$: Disparity on display of the k^{th} salient region in meters
· $\tilde{x}^k_c$: Horizontal location of the k^{th} salient region on display in meters
• Domain : 3D space
· $\mathcal{N}^k_{\pm}$: Real depth on the 3D space in meters of the k^{th} salient region
· $\tilde{Z}$: Viewing distance in meters
· ϕ : Focus angle on display in degrees
· ϕ^k : Focus angle on the k^{th} salient region with $\mathcal{N}^k_{\pm}$ in degrees
· $\overline{\mathcal{D}}^k$: Angular disparity of the k^{th} salient region in degrees (ϕ - ϕ^k)
• Domain : Eyes
· $\tilde{E}$: average pupil distance in meters (0.063m)

of these models, which can be used to estimate the relative size of the display and the frame size associated with the ratio

$$\rho = \frac{w_{dp}}{w} = \frac{\tilde{w}_{dp}}{\tilde{w}} = \frac{\tilde{\mathcal{D}}^k_{dp}}{\tilde{\mathcal{D}}^k} \tag{28}$$

where ρ is the ratio of the frame and display sizes. By using (28), the disparity of the k^{th} salient region in pixels ($\mathcal{D}^k$) can be converted to disparity on the display ($\tilde{\mathcal{D}}^k_{dp}$)

$$\tilde{\mathcal{D}}^k_{dp} = \rho \cdot \tilde{\mathcal{D}}^k = \rho \cdot \frac{\tilde{w}}{w} \cdot \mathcal{D}^k = \frac{\tilde{w}_{dp}}{w} \cdot \mathcal{D}^k \tag{29}$$

where $\tilde{\mathcal{D}}^k = \frac{\tilde{w}}{w} \cdot \mathcal{D}^k$. Similarly, the exact horizontal location (in meters) of the k^{th} salient region on the display ($\tilde{x}^k_c$) can be written as

$$\tilde{x}^k_c = \rho \cdot \frac{\tilde{w}}{w} \cdot x^k_c = \frac{\tilde{w}_{dp}}{w} \cdot x^k_c \tag{30}$$

where x^k_c defined in (21) is the horizontal location of the k^{th} salient region on the frame (in pixels). Finally, the real depth of the k^{th} salient region in 3D space is

$$\mathcal{N}^k_- = \frac{\tilde{Z} \cdot \tilde{w}_{dp} \cdot \mathcal{D}^k}{w \cdot \tilde{E} + \tilde{w}_{dp} \cdot \mathcal{D}^k}, \quad \mathcal{N}^k_+ = \frac{\tilde{Z} \cdot \tilde{w}_{dp} \cdot \mathcal{D}^k}{w \cdot \tilde{E} - \tilde{w}_{dp} \cdot \mathcal{D}^k} \tag{31}$$

where $\mathcal{N}^k_-$ is (-) depth and $\mathcal{N}^k_+$ is (+) depth. Using (31), the angular disparity of the k^{th} salient region ($\overline{\mathcal{D}}^k$) is obtained in units of pixel disparity via

$$\begin{aligned} \overline{\mathcal{D}}^k &= \phi - \phi^k, \\ &= \arccos(\frac{a^2 + b^2 - \tilde{E}^2}{2ab}) - \arccos(\frac{c^2 + d^2 - \tilde{E}^2}{2cd}). \end{aligned} \tag{32}$$

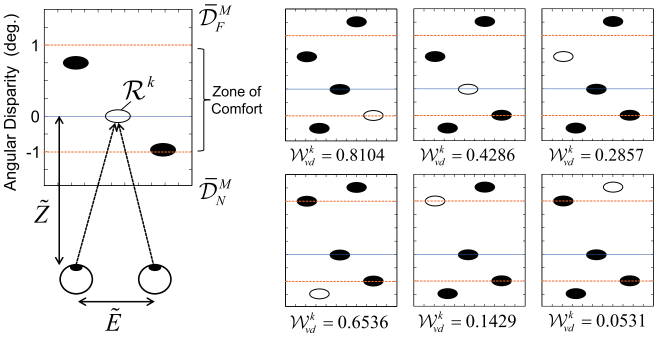

Angular Disparity (deg.)
1
0
-1
$\overline{\mathcal{D}}^M_F$
Zone of Comfort
$\mathcal{R}^k$
$\overline{\mathcal{D}}^M_N$
$\tilde{Z}$
$\tilde{E}$
$\mathcal{W}^k_{vd} = 0.8104$
$\mathcal{W}^k_{vd} = 0.4286$
$\mathcal{W}^k_{vd} = 0.2857$
$\mathcal{W}^k_{vd} = 0.6536$
$\mathcal{W}^k_{vd} = 0.1429$
$\mathcal{W}^k_{vd} = 0.0531$

Fig. 13. Visual discomfort saliency strength for various disparity cases (top-down view).

If it has (-) depth, then $a = (\tilde{Z}^2 + (\tilde{x}^k_c - \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$, $b = (\tilde{Z}^2 + (\tilde{x}^k_c + \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$, $c = ((\tilde{Z} - \mathcal{N}^k_-)^2 + (\tilde{x}^k_c - \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$ and $d = ((\tilde{Z} - \mathcal{N}^k_-)^2 + (\tilde{x}^k_c + \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$. Otherwise, $a = (\tilde{Z}^2 + (\tilde{x}^k_c - \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$, $b = (\tilde{Z}^2 + (\tilde{x}^k_c + \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$, $c = ((\tilde{Z} + \mathcal{N}^k_+)^2 + (\tilde{x}^k_c - \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$ and $d = ((\tilde{Z} + \mathcal{N}^k_+)^2 + (\tilde{x}^k_c + \frac{\tilde{E}}{2})^2)^{\frac{1}{2}}$. Using the calculated angular disparity of the k^{th} salient region, the visual discomfort saliency strength is defined as

$$\mathcal{W}^k_{vd} = \begin{cases} -\frac{1}{\overline{\mathcal{D}}^k} \cdot \eta_d, & \text{if } \overline{\mathcal{D}}^k \leq -1^\circ \\ \eta_d, & \text{if } -1^\circ < \overline{\mathcal{D}}^k < 1^\circ \\ \frac{1}{\overline{\mathcal{D}}^k} \cdot \eta_d, & \text{if } \overline{\mathcal{D}}^k \geq 1^\circ \end{cases} \tag{33}$$

where $\mathcal{W}^k_{vd}$ is the visual discomfort saliency strength expressed as a function of the angular disparity of the k^{th} salient region, and

$$\eta_d = 1 - \frac{\overline{\mathcal{D}}^k - \overline{\mathcal{D}}^M_N}{\overline{\mathcal{D}}^M_F - \overline{\mathcal{D}}^M_N}$$

as an index ($0 \leq \eta_d \leq 1$) on $\overline{\mathcal{D}}^M_N$ ($\overline{\mathcal{D}}^M_F$), which is the nearest (farthest) angular disparity of all the salient regions in the frame, where $\overline{\mathcal{D}}^M_N \leq \overline{\mathcal{D}}^k \leq \overline{\mathcal{D}}^M_F$. $\mathcal{W}^k_{vd}$ takes its maximum value when $\overline{\mathcal{D}}^k$ is in the zone of comfort. When $\overline{\mathcal{D}}^k$ is out of the zone of comfort, then $\mathcal{W}^k_{vd}$ decreases in proportion to the angular disparity. Fig. 13 shows the visual discomfort saliency strength of a salient region for the various disparity cases. The white circular object is the k^{th} salient region and the other black objects represent neighboring regions. The visual discomfort saliency strength takes the largest value when it is in the zone of comfort and has a nearer angular disparity, as in the top left of Fig. 13. In addition, when the salient region is out of the zone of comfort with a nearer angular disparity (bottom left), the saliency strength is reduced. When the salient region is out of the zone of comfort with a greater angular disparity, the strength is greatly reduced as shown at bottom right.

Finally, we express overall saliency strength factor incorporating depth discontinuity and visual discomfort saliency strength using (27) and (33):

$$\mathcal{W}^k_D = \mathcal{W}^k_{dd} \cdot \mathcal{W}^k_{vd} \tag{34}$$

where $\mathcal{W}^k_D$ is the disparity saliency strength of the k^{th} salient region. We multiply the depth discontinuity factor of saliency strength by the visual discomfort factor of saliency strength.

C. Temporal Saliency Strength

There has been a significant amount of research on the topic of eye-movements as a function of object motion [25]–[28]. Human eye movements related to visual attention can be classified into four main types [28]: 1) *Vestibular* eye movements hold the image of the world on the retina steady as the head moves; 2) *Smooth pursuit* which, broadly holds the image of a fixated moving target on the fovea; 3) *Saccade* which brings images of objects of interest onto the fovea and 4) *Vergence* which moves the eyes so that the images of a single object are placed or held simultaneously on both foveas. Amongst these, we shall focus on using smooth pursuit [26]. As mentioned, smooth pursuit eye movement occurs when the eyes are tracking a moving object and is affected by object speed. The human eye is able to track objects angular velocities of up to about 80 deg/sec although it is often reported that the maximum velocity of smooth pursuit eye movement is 20-30 deg/sec when observers are tracking the moving objects perfectly [25]. Following this latter guideline, define temporal saliency strength in terms of object motion speed, using the geometric parameters in Table III and ρ in (28), by converting the units of motion speed from pixel/sec to meter/sec at the display.

$$\widetilde{V}^k_{dp} = \rho \cdot \widetilde{V}^k = \rho \cdot \frac{\widetilde{w}}{w} \cdot V^k = \frac{\widetilde{w}_{dp}}{w} \cdot V^k \qquad (35)$$

where V^k_x, V^k_y and V^k_z are the average spatial and depth motions (speeds) in pixels of the k^{th} salient region, V^k is the average speed (pixel/sec) of the k^{th} salient region ($V^k = \sqrt{(V^k_x)^2 + (V^k_y)^2 + (V^k_z)^2} \times f_r$), f_r is the frame rate of the stereoscopic video, $\widetilde{V}^k$ is the salient region's speed in meter/sec and $\widetilde{V}^k_{dp}$ is the projected speed in meter/sec on the display. Next, transform the speed of the k^{th} salient region to deg/sec as

$$\dot{V}^k = \arctan\left(\frac{\widetilde{V}^k_{dp}}{\widetilde{Z}}\right). \qquad (36)$$

Finally, we define the temporal saliency strength factor

$$\mathcal{W}^k_T = \begin{cases} 1, & \text{if } 0 \leq \dot{V}^k \leq V^m_{th} \\ \frac{V^m_{th}}{\dot{V}^k}, & \text{if } V^m_{th} < \dot{V}^k \leq V^M_{th} \\ 0, & \text{if } \dot{V}^k > V^M_{th} \end{cases} \qquad (37)$$

where $V^m_{th} = 20$ deg/sec and $V^M_{th} = 80$ deg/sec.

Table IV summarizes the each saliency strength factor.

D. Saliency Strength Performance Analysis

To evaluate the performance of our model, we computed the area under the receiver operating characteristics (ROC) curve, i.e., the area under the curve (AUC) score [11]. In [47], it was described that the center-biased human attention characteristics reduce the accuracy of AUC scores. They found that the starting fixation point for all observers is biased at the central fixation. Thus, we excluded 20 fixations from the first for ameliorate the center-bias of visual attention, and calculated the AUC scores.

TABLE IV
SUMMARIZATION OF SALIENCY STRENGTH FACTORS

• Spatial saliency strength factor ($\mathcal{W}^k_S$)
· $\mathcal{W}^k_l$: Luminance saliency strength in terms of luminance contrast and gradient · $\mathcal{W}^k_c$: Color saliency strength in terms of color contrast and gradient · $\mathcal{W}^k_{sc}$: Size and compactness saliency strength
• Disparity saliency strength factor ($\mathcal{W}^k_D$)
· $\mathcal{W}^k_{dd}$: Depth discontinuity saliency strength in terms of disparity contrast and gradient · $\mathcal{W}^k_{vd}$: Visual discomfort saliency strength in terms of angular disparity
• Temporal saliency strength factor ($\mathcal{W}^k_T$)
· $\mathcal{W}^k_T$: Temporal saliency strength in terms of human eye-movement ability

1) The Sensitivity of the Each Saliency Strength Factor: We calculated the sensitivity of each of these features: spatial saliency strength, disparity saliency strength and temporal saliency strength. Table V shows the AUC score of each saliency strength factor. It can be seen that each sequence reveals different feature sensitivities depending on the sequence characteristics and the scene type. Spatial saliency strength is a highly sensitive predictor for most of the test sequences as it embodies the established salient factors luminance, color, size and compactness. Disparity saliency strength delivers high sensitivity on test sequences exhibiting large disparity variations or large crossed disparities such as "Marathon1," "Sidewalk-lateral1" and "Hallway." On scene types 4 or 5 (Zoom In or Out), the temporal saliency strength is highly predictive as on "Statue2," "Statue3" and "Street-lamp1."

2) The Final Saliency Strength Calculation: We used three different fusion methods: *Weighted sum*, *Max* and *Multiplication* to calculate the final saliency strength map.

∘ *Weighted sum*: the weighted sum of the three saliency strength maps

$$\mathcal{R}^k_S = (w_s \mathcal{W}^k_S + w_D \mathcal{W}^k_D + w_T \mathcal{W}^k_T) \cdot \mathcal{R}^k. \qquad (38)$$

∘ *Max* : the maximum value of the three saliency strength maps

$$\mathcal{R}^k_S = \text{Max}(\mathcal{W}^k_S, \mathcal{W}^k_D, \mathcal{W}^k_T) \cdot \mathcal{R}^k. \qquad (39)$$

∘ *Multiplication* : multiplicative fusion of the three saliency strength maps

$$\mathcal{R}^k_S = (\mathcal{W}^k_S \cdot \mathcal{W}^k_D \cdot \mathcal{W}^k_T) \cdot \mathcal{R}^k \qquad (40)$$

where $\mathcal{W}^k_S$, $\mathcal{W}^k_D$ and $\mathcal{W}^k_T$ are the spatial, disparity and temporal saliency strengths of the k^{th} segment, and w_S, w_D and w_T are the spatial, disparity and temporal saliency strength weights in the weighted sum of the three saliency strength maps. The overall performance of the weighted sum method is generally higher than that attained via the other two fusion methods (*Weighted sum* = 0.7652; *Max* = 0.7079; *Multiplication* = 0.7291).

3) The Performance Analysis: Fig. 14 shows saliency strength maps using our model with the weighted sum method and the results of various other saliency detection methods. Fig. 14(h) show the saliency strength weight maps obtained using our model and indicating the spatial, temporal and disparity saliency strength factors. Regions having low saliency

TABLE V
THE SENSITIVITY OF THE EACH SALIENCY STRENGTH

Sequence name	Scene type	Spatial saliency strength			Disparity saliency strength		Temporal saliency strength
		$\mathcal{W}_l^k$	$\mathcal{W}_c^k$	$\mathcal{W}_{sc}^k$	$\mathcal{W}_{dd}^k$	$\mathcal{W}_{vd}^k$	$\mathcal{W}_T^k$
Car1	1	0.3984	0.7092	**0.7604**	0.2858	0.3623	0.3211
Car2	1	0.3813	0.4483	**0.8274**	0.1287	0.3583	0.3363
Walking-person1	2	0.3501	0.2976	0.3948	0.3643	**0.4319**	0.2895
Walking-person7	2	0.3178	0.3730	0.4311	0.4363	**0.4478**	0.3070
Walking-person8	2	0.3630	0.4501	**0.6025**	0.3213	0.3475	0.3162
University1	2, 3	**0.6148**	0.5817	0.4462	0.3824	0.2237	0.3384
University2	2, 3	0.4429	**0.7450**	0.4697	0.1575	0.2924	0.4054
Statue2	1, 4, 5	0.2601	0.2340	0.3790	0.1308	0.1361	**0.6580**
Statue3	1, 4, 5	0.2785	0.4526	0.2236	0.6523	0.1820	**0.6968**
Street-lamp1	1, 4, 5	0.2426	0.3285	0.0851	0.1336	0.0809	**0.6246**
Crosswalk2	1	0.4269	0.4426	**0.4754**	0.3321	0.3051	0.2960
Library3	1	0.3087	**0.5677**	0.2423	0.2065	0.3290	0.3302
Library4	3	0.5758	**0.7681**	0.2818	0.2957	0.1971	0.2105
Marathon1	1	0.4113	0.5370	0.4048	**0.6085**	0.3974	0.2633
Restaurant1	1	0.4128	**0.7547**	0.0803	0.1641	0.1647	0.1360
Sidewalk-lateral1	3	0.5053	0.4623	0.3568	**0.5142**	0.1630	0.1547
Bike	1	0.4326	0.4862	**0.7090**	0.4563	0.4947	0.3240
Car	1	0.4558	0.4467	**0.7065**	0.4712	0.3961	0.3328
Feet	1	0.3056	**0.6914**	0.3224	0.4950	0.5845	0.3883
Hallway	1	0.3897	0.3925	0.5939	0.7214	**0.7782**	0.7401
Notebook	1	0.3146	**0.6319**	0.3420	0.3311	0.2191	0.3215
Sofa	1	**0.6642**	0.3671	0.3423	0.3539	0.3137	0.3298
Street	1	0.4465	0.3029	0.7227	0.7090	**0.7765**	0.6003
Balloons	1, 2, 3	0.6694	0.2620	0.6891	**0.7769**	0.4320	0.6915
Average		0.4154	**0.4889**	0.4537	0.3929	0.3506	0.3922

strength, such as road, ground or sky, are excluded from the saliency strength weight maps. Moreover, objects exhibiting high luminance gradients or contrast have relatively large weights. Rapid motions in the 3D videos, such as the movement of the person in the rear in Fig. 14(h)-(2) and the cars in Fig. 14(h)-(3) lead to low values in the saliency strength maps. Fig. 14(h)-(5) and (h)-(6) show that, in similar environments and scenes, objects that are wide and dense with few or no depth discontinuities tend to have higher saliency strength. Moreover, in videos containing Zoom In or Out, motion at the periphery of the image tends to be faster than at the center of the image. Thus, the peripheries of the videos in Fig. 14(h)-(5) and (h)-(6) have low saliency strength. Fig. 14(i) show results using the eye-tracker in the same environment as shown in Section II-B. The eye-tracking results are similar to the saliency strength maps.

We compare our model with algorithms proposed by Bruce and Tsotsos [5], Zhai and Shah [6], Itti and Baldi [7], Marat *et al.* [8], Zhang *et al.* [9], and Seo and Milanafar [10]. Fig. 14(b)–(g) show the saliency results of these saliency models. It can be seen that our model relies heavily on identifying salient object regions. Our model achieves high correlations relative to results obtained using an eye-tracker. Furthermore, much of the success of our saliency strength model arises from adaptation to scene type. In particular, it significantly outperforms other models (Fig. 14(2)–(6) and (10)) when camera motions occur. Table VI shows the AUC score of each algorithm. Our proposed saliency model outperforms other methods in most cases. In particular, it achieves much higher accuracy when the test sequences include camera motion or Zoom In/Out.

V. SALIENCY ENERGY MEASUREMENT

The overall saliency energy is obtained by applying saliency weights to the saliency strength computed on each frame over time. Here we utilize two important additional visual factors when a human watches stereoscopic video. One is foveation [33], [34], which describes the space-variant resolution of photoreceptors in the retina. The other is Panum's fusional area [23] which measures the depth resolution and range of the eyes. Using these two factors, we obtain visual weights on each region and calculate the overall saliency energy as a weighted sum.

A. Stereoscopic Saliency Weight

1) Foveation Saliency Weight: Fig. 15(a) depicts the foveal region of the eye. Since foveation is a process of nonuniform sampling that allows preferential acquisition of visual information at the human retina, we define the foveation saliency weights as follows [33], [34]. The model of foveation we use is

$$f_c(x) = \min[\ \frac{e_2 \ln(\frac{1}{CT_0})}{\alpha[c_2 + \arctan(\frac{d(x)}{w\gamma})]}, \frac{\pi w \gamma}{360}\ ] \tag{41}$$

where γ indicates the distance between the human eyes and the fixation point (center of the fixated region), $d(x)$ is the pixel distance between the foveal region and other neighboring locations (x_n, y_n), $\gamma = \frac{d(x)}{w}$ and $\arctan(\frac{d(x)}{w\gamma})$ is the eccentricity e. Other parameters in the foveation model have been estimated in [36] yielding $e_2 = 2.3$, $\alpha = 0.106$ and $CT_0 = \frac{1}{64}$. Using this foveation model, the foveation weight is defined

$$\mathcal{W}_{2D}(x_n, y_n, x_f, y_f) = \frac{f_c(x_n, y_n)}{f_c(x_f, y_f)} \tag{42}$$

where $f_c(x_f, y_f)$ is the spatial cutoff frequency at the foveation point, with cutoff frequency f_c at (x_f, y_f). Fig. 16(b) shows examples of foveation visual weights.

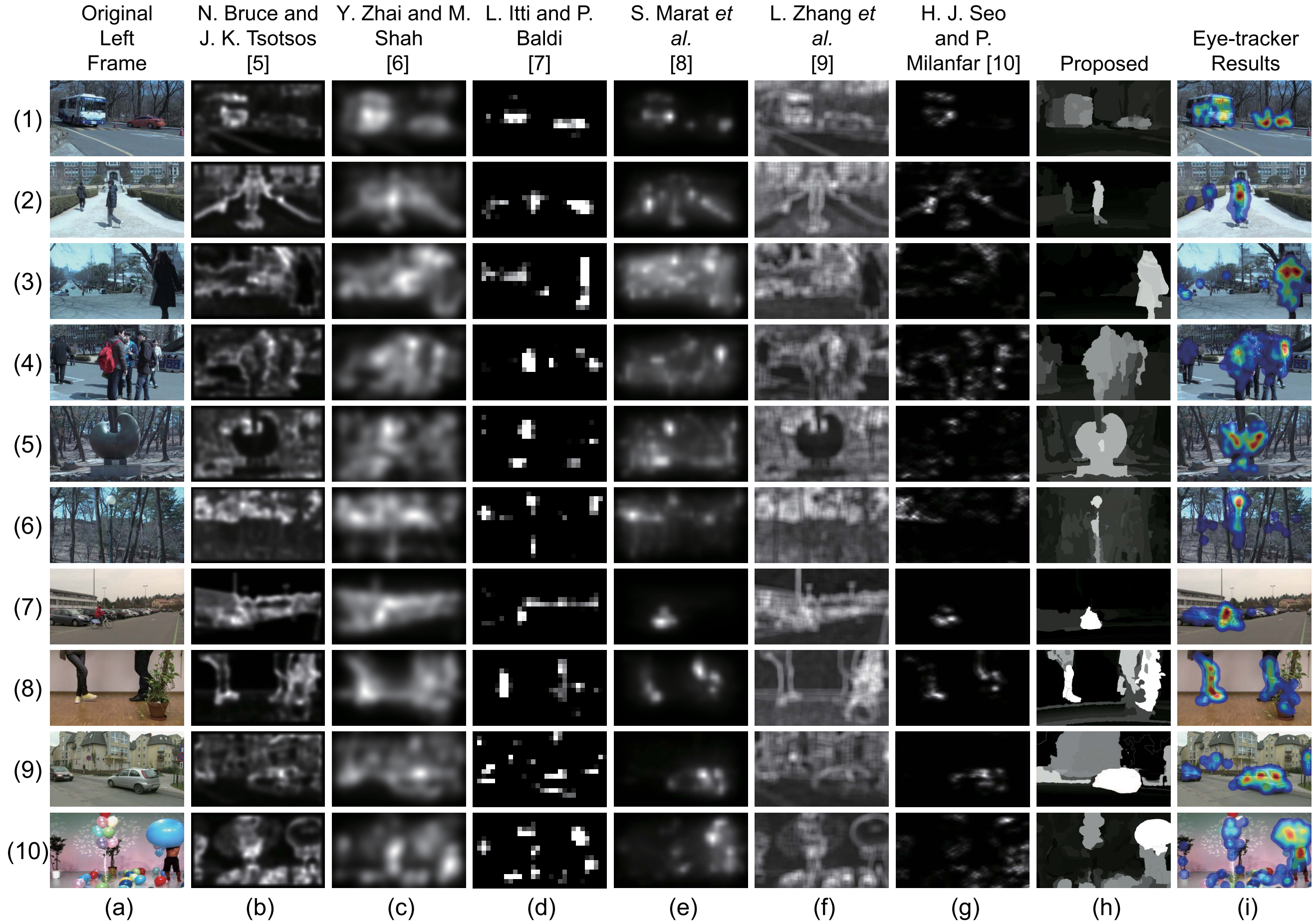


Fig. 14. The results of saliency detection of the various methods. (a) Original left images: (1) 130^{th} frame, "Car2" (scene type 1), (2) 215^{th} frame "Walking-person7" (scene type 2), (3) 48^{th} frame "University1" (scene type 3), (4) 109^{th} frame "University2" (scene type 3), (5) 111^{th} frame "Statue2" (scene type 4), (6) 86^{th} frame "Street-lamp1" (scene type 5), (7) 60^{th} frame "Bike" (scene type 1), (8) 124^{th} frame "Feet" (scene type 1), (9) 8^{th} frame "Street" (scene type 1) and (10) 62^{th} frame "Balloons" (scene type 3). (b)-(g) The saliency detection results of the various methods. (h) The proposed saliency strength maps. (i) Eye-tracker simulation results.

2) 3D Saliency Weight - Panum's Fusional Area: Panum's fusional area is defined as the area on the retina of one eye over which a point-sized image can range, while still being able to provide a single image with a specific point of stimulus on the retina of the other eye. Therefore, the region in visual space over which we perceive "single vision" is Panum's fusional area, i.e., where stereoscopic fusion is clearly performed. Any objects to the front and back of this area are not fused completely, a phenomenon known as diplopia (double vision) [23]. Fig. 15(b) shows Panum's fusional area. The horopter is the locus of 3D points having the same angular disparity. We define the 3D saliency factor relative to Panum's fusional area by

$$\begin{cases} \mathcal{W}_{3D}(\phi^k,\phi^s)=1, & \text{if } \ 0\le\Delta\phi\le\beta \\ \mathcal{W}_{3D}(\phi^k,\phi^s)=\exp(-\frac{\Delta\phi-\beta}{\psi}), & \text{if } \Delta\phi\ge\beta \end{cases} \tag{43}$$

where ϕ^k is the viewing angle of the k^{th} salient region when the depth is $\mathcal{N}_-^k$ or $\mathcal{N}_+^k$ from the display to the fixation region, ϕ^s is the viewing angle of another neighboring region when the depth is $\mathcal{N}_-^s$ or $\mathcal{N}_+^s$, and $\Delta\phi$ is the difference between ϕ^k and ϕ^s ($\Delta\phi=|\phi^k-\phi^s|$). The parameter β is a threshold needed to define Panum's fusional area ($\beta=0$ in general). $\psi\approx 0.62°$ is a fixed coefficient that has been estimated in physiological experiments [23]. Fig. 16(c) shows the weighting computed used (43) as a function of depth w.r.t. the fixation region. Regions lying within Panum's fusional area exhibit higher saliency weight while those lying outside have much lower weight. Using these foveation and 3D saliency weights, the saliency energy of a stereoscopic video is obtained as shown in Fig. 16(d).

B. Stereoscopic Video Saliency Energy

Using the foveation and 3D saliency weights, the overall saliency energy, $\mathcal{S}_E$ of a stereoscopic video is defined as

$$\mathcal{S}_E=\frac{1}{n_r}\sum_{s=1}^{n_r} \times\left(\frac{\sum_{k=1}^{n_r}\sum_{n=1}^{\mathcal{R}^k_{size}}\mathcal{W}_{2D}(x_n^k,y_n^k,x_c^s,y_c^s)\cdot\mathcal{W}_{3D}(\phi^k,\phi^s)\cdot\mathcal{R}^k_S}{\sum_{k=1}^{n_r}\sum_{n=1}^{\mathcal{R}^k_{size}}\mathcal{W}_{2D}(x_n^k,y_n^k,x_c^s,y_c^s)\cdot\mathcal{W}_{3D}(\phi^k,\phi^s)}\right) \tag{44}$$

TABLE VI
THE PERFORMANCE EVALUATION OF THE VARIOUS SALIENCY METHODS

Sequence name	Scene type	Bruce & Tsotsos [5]	Zhai & Shah [6]	Itti & Baldi [7]	Marat *et al.* [8]	Zhang *et al.* [9]	Seo & Milanafar [10]	Proposed
Car1	1	0.3183	0.4382	0.3562	0.4428	0.6853	0.7400	**0.7958**
Car2	1	0.3375	0.5389	0.4332	0.4543	0.7546	0.7128	**0.8032**
Walking-person1	2	0.2129	0.2245	0.4178	0.4202	0.6471	0.5518	**0.7535**
Walking-person7	2	0.4143	0.1481	0.4078	0.3742	0.7555	0.7465	**0.8177**
Walking-person8	2	0.1366	0.0983	0.4287	0.3310	0.6753	0.6820	**0.7932**
University1	2, 3	0.1257	0.1173	0.4244	0.3723	0.7802	0.7671	**0.7897**
University2	2, 3	0.3152	0.3680	0.4232	0.3388	0.7168	**0.7787**	0.7610
Statue2	1, 4, 5	0.1019	0.1413	0.4403	0.1137	0.4144	0.4327	**0.7812**
Statue3	1, 4, 5	0.1236	0.1532	0.4011	0.1983	0.3922	0.4366	**0.7959**
Street-lamp1	1, 4, 5	0.0281	0.0822	0.4168	0.1317	0.6296	0.6225	**0.7731**
Crosswalk2	1	0.1135	0.2275	0.4189	0.4414	0.7087	0.7303	**0.7877**
Library3	1	0.2569	0.2464	0.4249	0.4008	0.6839	0.6962	**0.7535**
Library4	3	0.3480	0.3566	0.3970	0.5371	0.7732	0.6306	**0.7768**
Marathon1	1	0.2469	0.3383	0.3443	0.4403	0.7285	0.7791	**0.7947**
Restaurant1	1	0.1670	0.1947	0.4417	0.4872	0.6903	0.6996	**0.7769**
Sidewalk-lateral1	3	0.4413	0.3307	0.4202	0.5436	0.7350	**0.7568**	0.7518
Bike	1	0.4041	0.3290	0.4468	0.4172	0.7521	0.7573	**0.7640**
Car	1	0.1467	0.3061	0.4348	0.4283	0.6955	0.7150	**0.7154**
Feet	1	0.2159	0.3298	0.4064	0.5407	0.2478	0.3641	**0.6447**
Hallway	1	0.2430	0.4549	0.4044	0.5144	0.5722	0.5702	**0.6912**
Notebook	1	0.2176	0.2970	0.4467	0.4731	0.5636	0.5346	**0.7440**
Sofa	1	0.2843	0.1855	0.4340	0.4855	0.7997	0.6347	**0.7483**
Street	1	0.2499	0.1905	0.3969	0.4222	0.6893	0.7272	**0.7919**
Balloons	1, 2, 3	0.3142	0.2021	0.3958	0.4178	0.5763	0.5490	**0.7581**
Average		0.2402	0.2625	0.4151	0.4053	0.6528	0.6507	**0.7652**

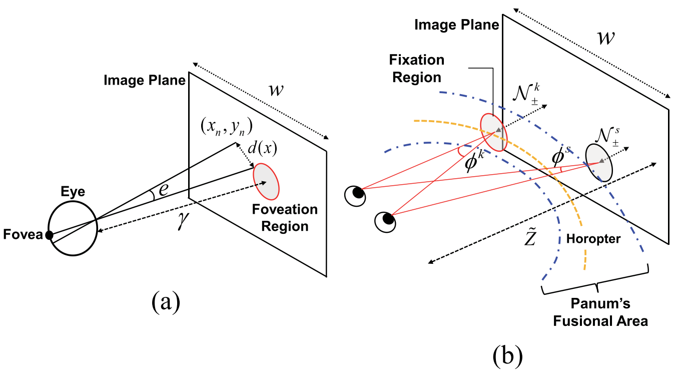


Fig. 15. Stereoscopic saliency energy weight model. (a) Foveation. (b) Panum's fusional area.

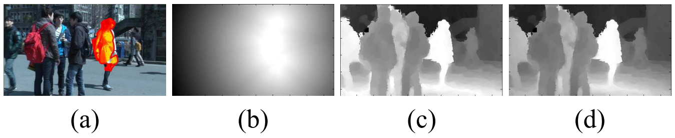


Fig. 16. Saliency weight. (a) The 139th left frame "University2" (scene type 3) [42] for a given fixation region (10th segment). (b) foveation saliency weight. (c) 3D saliency weight. (d) Saliency energy weight using the foveation and 3D weights.

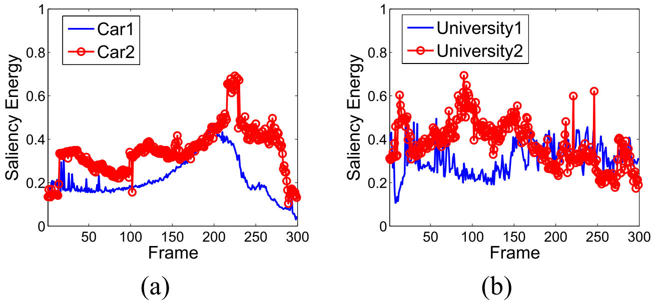


Fig. 17. Saliency energy graph (a) "Car1" (scene type 1) and "Car2" (scene type 1) and (b) "University1" (scene types 1, 2 and 3)" and "University2" (scene types 1, 2 and 3).

where n_r is the number of salient regions after segmentation in a frame, (x_n^k, y_n^k) is a pixel in the k^{th} salient region, (x_c^s, y_c^s) is the location information of the s^{th} salient region using (21), and $\phi^k(\phi^s)$ is the average viewing angle of the k^{th} (s^{th}) salient region. If the fixation is in the s^{th} salient region, the saliency energy weight of each region is calculated w.r.t. that of the s^{th} salient region. We carry out this process with reference to all salient regions sequentially under the assumption that the fixation is on each salient region, and obtain the final saliency energy by taking the weighted sum.

Fig. 17 plots saliency energy measurements against frame number for two video sequences. Since "Car2" has more objects than "Car1," it has higher saliency energy than "Car1" overall, as shown in Fig. 17(a). The saliency energies of the two sequences in Fig. 17(a) increases up to around the 220th frame (when the bus arrives), then rapidly decreases afterwards, when the bus disappears from the scene. In Fig. 17(b), the patterns of occurrence of salient regions are irregular in the sequences "University1" and "University2" due to random object motions in addition to camera motion (scene type 3). However, the saliency energy of "University2" is higher than that of "University1" since more of the objects in it have a high saliency strength. Fig. 18 demonstrates frames having the maximum (minimum) saliency energy (Fig. 17(a), (c), (e) and (g) ((b), (d), (f) and (h))), where more (less) objects having high saliency strength are distributed. It may be observed that high saliency energy occurs on frames containing more regions with high saliency strength.

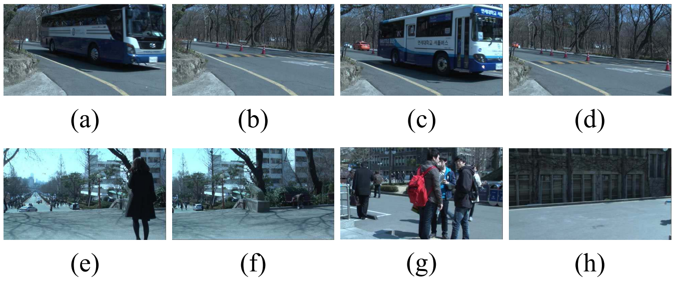


Fig. 18. Frames with maximum and minimum saliency energy. (a) Maximum saliency energy frame; 197^{th} frame "Car1" ($\mathcal{S}_E = 0.4691$). (b) Minimum saliency energy frame; 298^{th} frame "Car1" ($\mathcal{S}_E = 0.0300$). (c) Maximum saliency energy frame; 225^{th} frame "Car2" ($\mathcal{S}_E = 0.6919$). (d) Minimum saliency energy frame; 289^{th} frame "Car2" ($\mathcal{S}_E = 0.1040$). (e) Maximum saliency energy frame; 57^{th} frame "University1" ($\mathcal{S}_E = 0.4951$). (f) Minimum saliency energy frame; 8^{th} frame "University1" ($\mathcal{S}_E = 0.1064$). (g) Maximum saliency energy frame; 90^{th} frame "University2" ($\mathcal{S}_E = 0.6941$). (h) Minimum saliency energy frame; 296^{th} frame "University2" ($\mathcal{S}_E = 0.1737$).

VI. Conclusion

When analyzing the saliency of stereoscopic video, it is important to include relevant perceptual mechanisms in particular since human visual responses in 3D are different than in 2D, owing to, for example, accommodation and vergence processes. Here we proposed a new and detailed methodology to quantify the degree of saliency of objects and regions in 4D space-time by defining novel saliency strength and saliency energy measures on stereoscopic videos, using well-known perceptual and attentional principles. The simulation results indicate that the new saliency strength and energy models enable detection of visually important regions and visually important frames on stereoscopic videos. Models such as these can be used in many applications such as 3D video compression and 3D quality assessment.

References

[1] J. Lee and T. Ebrahimi, "Perceptual video compression: A survey," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 684–697, Oct. 2012.

[2] Q. Huynh-Thu, M. Barkowsky, and P. L. Callet, "The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 421–431, Jun. 2011.

[3] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neuro Sci.*, vol. 2, no. 3, pp. 194–203, 2001.

[4] L. Itti and C. Koch, "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.

[5] N. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, pp. 1–24, 2009.

[6] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[7] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.

[8] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guerin-Dugue, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, 2009.

[9] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.

[10] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, 2009.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[12] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 564–573, Apr. 2008.

[13] P. Reinagel and A. M. Zador, "The effect of gaze on natural scene statistics," in *Proc. NICW*, Snowbird, UT, USA, 1997, pp. 1–15.

[14] Y. Liu, L. K. Cormack, and A. C. Bovik, "Dichotomy between luminance and disparity features at binocular fixations," *J. Vis.*, vol. 10, no. 12 pp. 1–17, 2010.

[15] N. Bruce and J. K. Tsotsos, "An attentional framework for stereo vision," in *Proc. Can. Conf. Comput. Robot Vis.*, 2005, pp. 88–95.

[16] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3D video," *Advances in Multimedia Modeling* (ser. Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, pp. 314–324, 2010.

[17] Y. Ma, X. Hua, L. Lu, and H. 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. 2005.

[18] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[19] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," in *Proc. IEEE CVPR Workshop Semantic Learn. Appl. Multimedia*, Jun. 2006, pp. 100–105.

[20] A. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," *Proc. SPIE*, vol. 1915, pp. 36–48, Jan. 1993.

[21] D. Kim and K. Sohn, "Visual fatigue prediction for stereoscopic image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 231–236, Feb. 2011.

[22] M. Lambooij, W. I. Jsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *J. Imag. Sci. Technol.*, vol. 53, no. 3, pp. 1–4, 2009.

[23] T. Qhshima, H. Yamamoto, and H. Tamura, "Gaze-directed adaptive rendering for interacting with virtual space," in *Proc. VRAIS*, 1996, pp. 103–110.

[24] M. J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3379–3391, Sep. 2013.

[25] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," *Proc. SPIE*, vol. 3299, pp. 180–191, Jul. 1998.

[26] P. E. Hallett, *Chapter 10 in Handbook of Perception and Human Performance*. New York, NY. USA: Wiley, 1986.

[27] L. Festinger, H. A. Sedgwick, and J. D. Holtzman, "Visual perception during smooth pursuit eye movements," *J. Vis. Res.*, vol. 16, no. 12, pp. 1377–1386, 1976.

[28] R. J. Leigh and D. S. Zee, *The Neurology of Eye Movements*. Oxford, U.K.: Oxford Univ. Press, 1999.

[29] I. A. Rybak, V. I. Gusakova, A. V. Golovan, L. N. Podladchikova, and N. A. Shevtsova, "A model of attention-guided visual perception and recogniton," *J. Vis. Res.*, vol. 38, nos. 15–16, pp. 2387–2400, 1998.

[30] J. R. Wilson and S. M. Sherman, "Receptive-field characteristics of neurons in cat striate cortex: Changes with visual field eccentricity," *J. Neurophysiol.*, vol. 29, no. 3, pp. 512–533, 1976.

[31] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "On the usefulness of attention for object recognition," in *Proc. 8th Eur. Conf. Comput. Vis.*, 2004, pp. 96–103.

[32] K. I. Beverley and D. Regan, "Visual perception of changing size: The effect of object size," *J. Vis. Res.*, vol. 19, no. 10, pp. 1093–1104, 1979.

[33] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[34] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 149–162, Feb. 2003.

[35] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[36] W. S. Geisler and J. S. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," *Proc. SPIE*, vol. 3299, pp. 294–305, 1998.

[37] (2013). *Smart Eye Pro-Remote 3D Eye Tracking for Research* [Online]. Available: http://www.smarteye.se

[38] Z. Chi and H. Yan, "Image segmentation using fuzzy rules derived from K-means clusters," *J. Electron. Imaging*, vol. 4, no. 2, pp. 199–206, 1995.

[39] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing and bayes/MDL for multiband image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 9, pp. 884–900, Sep. 1996.

[40] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun 2010, pp. 2432–2439.

[41] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," in *Proc. IEL*, Apr. 2008, pp. 1–8.

[42] (2012). *Stereoscopic (3D Imaging) Database* [Online]. Available: http://grouper.ieee.org/groups/3dhf/or ftp://165.132.126.47/

[43] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Proc. SPIE*, vol. 7526, pp. 11–15, Jan. 2010.

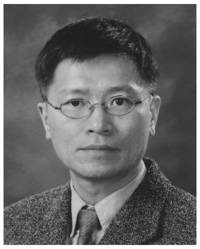
[44] A. Smolic, G. Tech, and H. Brust, "Report on generation of stereo video data base," *Mobile3DTV Tech. Rep.*, vol. 2, no. 1, pp. 1–15, 2010.

[45] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp, "Camera motion-based analysis of user generated video," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 28–41, Jan. 2010.

[46] K. T. Mullen, "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings," *J. Physiol.*, vol. 359, no. 1, pp. 381–400, 1985.

[47] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, 2005.

Haksub Kim received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2008 and 2011, respectively. He is currently pursuing the Ph.D. degree since 2011. His research interests include 2D/3D image and video processing based on human visual system, human visual attention, quality assessment of 2D/3D image and video, cross-layer optimization, and wireless multimedia communications.



Sanghoon Lee (M'05–SM'12) received the B.S. degree in electrical engineering from Yonsei University in 1989 and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology in 1991. From 1991 to 1996, he was with Korea Telecom. He received the Ph.D. degree in electrical engineering from the University of Texas at Austin in 2000. From 1999 to 2002, he was with Lucent Technologies on 3G wireless and multimedia networks. In 2003, he joined the faculty of the Department of Electrical and Electronics Engineering, Yonsei University, Seoul, Korea, where he is a Full Professor. He has been an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2010 and an Editor of the *Journal of Communications and Networks* since 2009, and the Chair of the IEEE P3333.1 Quality Assessment Working Group since 2011. He has been serves as the Technical Committee of the IEEE IVMSP since 2014, the General Chair of the 2013 IEEE IVMSP Workshop, and a Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has received the 2012 Special Service Award from the IEEE Broadcast Technology Society and the 2013 Special Service Award from the IEEE Signal Processing Society. His research interests include image/video quality assessments, medical image processing, cloud computing, wireless multimedia communications, and wireless networks.

Alan Conrad Bovik is the Curry/Cullen Trust Endowed Chair Professor with The University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering. He is a Faculty Member with the Department of Electrical and Computer Engineering and the Center for Perceptual Systems, Institute for Neuroscience. His research interests include image and video processing, computational vision, and visual perception. He has published more than 650 technical articles and holds two U.S. patents. His several books include the recent companion volumes *The Essential Guides to Image and Video Processing* (Academic Press, 2009).

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including: the Best Paper Award in 2009; the Education Award in 2007; the Technical Achievement Award in 2005, and the Meritorious Service Award in 1998. He was named recipient of the Honorary Member Award of the Society for Imaging Science and Technology in 2013, the SPIE Technology Achievement Award in 2012, and was the IS&T/SPIE Imaging Scientist of the Year for 2011. He received the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin, the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana in 2008, the IEEE Third Millennium Medal in 2000, and two journal paper awards from the International Pattern Recognition Society in 1988 and 1993. He is a fellow of the Optical Society of America, the Society of Photo-Optical and Instrumentation Engineers, and the American Institute of Medical and Biomedical Engineering. He has been involved in numerous professional society activities, including: Board of Governors for the IEEE Signal Processing Society from 1996 to 1998; co-founder and Editor-in-Chief for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1996 to 2002; Editorial Board for THE PROCEEDINGS OF THE IEEE from 1998 to 2004; Series Editor for *Image, Video, and Multimedia Processing* (Morgan and Claypool Publishing Company, 2003); and Founding General Chairman for the First IEEE International Conference on Image Processing, Austin, TX, USA, in 1994.

Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.