USING MONOCULAR DEPTH CUES FOR MODELING STEREOSCOPIC 3D SALIENCY

Iana Iatsun, Mohamed-Chaker Larabi, Christine Fernandez-Maloigne

Departement XLIM-SIC UMR CNRS 7252,
SP2MI, Teleport 2, Bvd Marie et Pierre Curie, BP 30179, 86962, Futuroscope, France

ABSTRACT

Saliency is one of the most important features in human visual perception. It is widely used nowadays for perceptually optimizing image processing algorithms. Several models have been proposed for 2D images and only few attempts can be observed for 3D ones. In this paper, we propose a stereoscopic 3D saliency model relying on 2D saliency features jointly with depth obtained from monocular cues. On the one hand, the use of 2D saliency features is justified psychophysically by the similarity observed between 2D and 3D attention maps. On the other hand, 3D perception is significantly based on monocular cues. The validation of our model using state-of-the-art procedures including Kullback-Leibler divergence (KLD), area under the curve (AUC) and correlation coefficient (CC) in comparison with attention maps showed very good performance.

Index Terms— Saliency, monocular depth cues, stereoscopic 3D, visual attention.

1. INTRODUCTION

With the widespread of three dimensional (3D) imaging in television and movie production, advertisement and gaming, requirement of new algorithms for compression, transmission and display grew significantly. Stereoscopic 3D (S3D) content delivery is more than ever possible and the viewer experience becomes a very hot topic. Therefore, exploring human visual system (HVS) capabilities is undoubtedly an important step towards increasing visual comfort.

The HVS allows to perceive the world in three dimensions and evaluate the distance to objects thanks to not only binocular indices but also monocular ones [1]. On the one hand, binocular cues, *i.e.* stereopsis and vergence, are achieved by using the information coming from left and right eyes. On the other hand, 3D perception relies mostly on monocular cues such as relative size, texture, occlusion, shadow and perspective. They are partly linked to *a priori* knowledge and cognitive information.

Among the prominent characteristics of the HVS, visual attention is playing an important role in exploring and understanding our environment. This process can be either top-down (*i.e.* task-driven) or bottom-up (*i.e.* stimuli-driven). Several works from the literature have focused on mimicking the HVS in order to produce saliency maps that predict areas attracting attention from images. The pioneering work of *Itti and Koch* is based on the construction of conspicuity maps from low level criteria such as intensity, color and orientation [2]. Several other works, mostly belonging to the bottom-up family, have tried to extend and improve the latter model [3, 4, 5].

While these models are efficient and close to human attention, their complexity is often seen as an obstacle for real-time applications. Computational approaches for constructing saliency maps have emerged in the recent years. One can cite the model of *Achanta et al.* [6] based on the fact that salient objects are those which are prominent beyond the local mean of the neighborhood. Another more recent computational method based on interest points detection was proposed by *Nauge et al.* [7]. It showed a correlation between interest points (IP) and gaze points.

Although, the literature is rich in terms of saliency studies, 3D saliency has been rarely tackled. It has been reported that, there is a difference between 3D and 2D visual attention due to depth information. With the aim of studying the effect of depth on saliency and to analyze the difference 2D and 3D attention, several works have been done. For instance, *Jansen et al.* explored the effect of depth on human perception in a free-watching task using 2D and 3D still images [8]. It revealed the time-dependent effect of depth and difference in eye-movement for 2D and 3D. Interestingly, they observed that viewers were fixating earlier on the first-plan objects even in 2D, using monocular cues. In the same vein, *Huynh et al.* performed similar study on 2D and 3D video, and led to the same conclusion as previously [9]. Despite this results, authors highlighted some similarities between 2D and 3D attention maps which pushed new works on the development of computational models of visual attention for 3D content relying on 2D salient features in conjunction with depth cues. In this context, a bottom-up attentional model for 3D video was proposed by *Zhang et al.* by integrating depth, luminance, color, motion and orientation [10]. Unfortunately, this work is lacking in terms of implementation details that help readers to understand the fusion of left and right saliency and in addition no quantitative evaluation was provided. Recently, *Wang et al.* suggest the creation of a separate saliency map for a depth information and its fusion with 2D salient features [11]. Authors reported quantitative results on arbitrarily chosen view, and the main limitation is linked to use of depth contrast only. Several similar work using depth information could be found in the literature [12, 13, 14]. However, this assumes the availability of depth information or the possibility of its computation through disparity.

As stated before, 3D perception is highly relying on monocular cues. One can assume that if it is possible to extract depth from only one view, this information will be of a great importance for prediction of 3D saliency. Following the idea, few attempts of depth extracting from 2D images can be observed in the field. Among them, the method suggested by *Saxena et al.* based on a priori learning process, which is unfortunately lacking in terms of genericity [15]. Monocular depth evaluation based on low level vision features and without any a priori information about scene, was proposed by *Palou et al*. It is based on the detection of T-junction points, that identify occlusions. This approach provides as output segmented areas depending on depth order.

In this paper we propose to exploit monocular depth in addition to 2D salient features in order to develop a 3D saliency model. The idea lies in the fact that there are similarities between 2D and 3D attention behavior. Moreover, depth can be accurately predicted from single 2D view. Therefore, our model performs a fusion be-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tween 2D saliency map obtained using efficient methods from the literature and monocular depth estimated using the aforementioned work described in [16]. Our approach applies on one of the views depending on the notion of master eye that is implemented based on the dominant saliency.

This paper is organized as follows: section 2 introduces the proposed 3D saliency model based on monocular cues. Details of the monocular depth estimation are given in order to explain the major features. Simulation results are presented in section 3 using state-of-the-art measurements to compare between predicted 3D saliency and human gaze. This paper ends with some conclusions and gives opening for future works.

2. PROPOSED 3D SALIENCY MODEL

In the literature, most of the visual attention models for 3D are based on stereo depth or disparity. Generally, this information is accessible by having either two views or, texture and depth representation. We start from the idea that depth information can be obtained from monocular cues, even if the second view or the disparity map are unavailable. So far, we propose a model for 3D saliency prediction by combining 2D salient features and monocular depth. The proposed framework is depicted on figure 1.

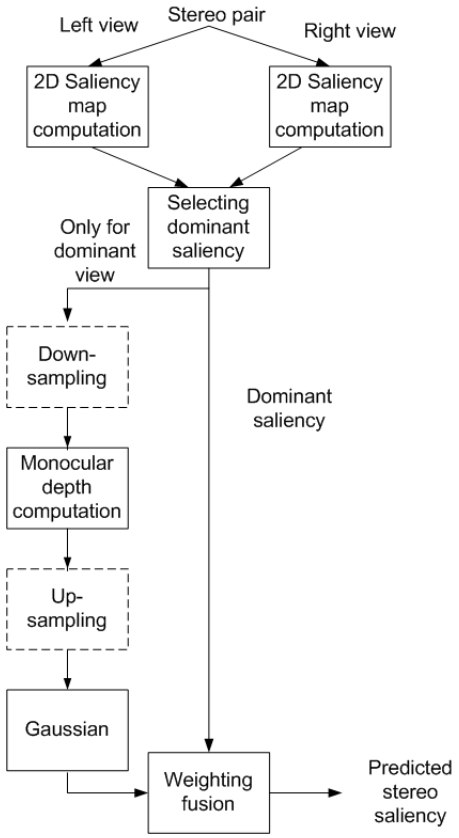

Fig. 1. Flow-chart of the proposed 3D saliency algorithm.

At the first stage of our model, 2D visual features are extracted from both views of the stereo pair. In order to find 2D salient information, we used models that have been validated in the literature namely Itti's model [2] and another based on interest points [7]. We make use of the notion of dominant eye in order to select one of the views for feeding the next stages. The next important step is linked to the extraction of depth from monocular cues using a recent algorithm based on occlusion detection. The latter does not require neither a priori information nor learning process. Procedures of down-sampling and up-sampling, given in dash-line, are optional and allow to reduce complexity. The processing of the mono depth evaluation algorithm creates very strictly define borders, to avoid this effect we convolved final segmented image with a kernel of Gaussian *G*. Finally, a fusion of 2D saliency features and monocular depth is made using an appropriate formulation. Each step of our framework will be described deeply in the following part.

2.1. 2D Saliency features

To perform selection of 2D visual attention features, we applied either Itti's saliency model [2] or Nauge's model based on interest points. This choice is justified by the aim of having two models with different architectures and different behavior. For the still image, Itti's model creates seven low-level feature maps: two for color contrast, one for luminance contrast and four for orientation. Gaussian pyramids are applied with nine levels, beginning from level 0 *i.e.* original size. Center-surround differences are then computed across pyramids levels. Finally, all feature maps are integrated in single saliency map.

The second saliency model is based on the ability of well-known interest points to predict salient features on an image. It allows to use Harris, SIFT or SURF with parameters that make the extraction close to human gaze points [7]. Another advantage lies in its very low complexity.

In order to select one view for the proposed process, we introduced the eye-dominance property. This very important phenomenon reflects the leading role of one eye when gazing on interest object [17]. This idea is already widely explored in image compression [18, 19]. In our model, we consider the dominant saliency map as the one containing more salient information. Therefore, left and right saliency maps are thresholded at increasing values in order to estimate the difference between them. The view corresponding to the map that contains more salient pixels was used to feed the monocular depth estimation algorithm.

2.2. Monocular depth

The process of monocular depth evaluation contains two main phases: building of hierarchical representation of the image and choosing segments to compose final depth order partition. The estimation of low level depth cues are integrated into the construction of binary partition tree (BPT) [20]. At the first step of segmentation, each pixel of image is considered to be a separate region. In this way, computational cost is directly proportional to the image resolution. Each region is characterized by its own three-dimensional histogram (h). The process of BPT building is based on merging neighboring regions. The order in which segments are merged is defined by a similarity measure M (eq.1). This measure is influenced by low-level features of the region such as color M_c, area M_a, shape M_s and depth M_d. Like that, hierarchy order can be established between two given segments S_1 and S_2.

$$M(S_1, S_2) = M_a \left(\alpha M_c + (1-\alpha) M_s\right) M_d \qquad (1)$$

Color similarity M_c (eq.2) between two regions is measured with the help of earth mover's distance (EMD), where h_1 and h_2 are the

three-dimensional histograms of each neighboring region.

$$M_c(S_1, S_2) = EMD(h_1, h_2) \quad (2)$$

The shape parameter M_s (eq.3) measures how much the perimeter, resulting from the fusion, will increase relatively to the biggest merged region.

$$M_s(S_1, S_2) = max\left(0, \frac{min(P_1, P_2) - 2P_{1,2}}{max(P_1, P_2)}\right) \quad (3)$$

The area measure M_a (eq.4) is applied as a weighting function in order to stimulate the merging of small and insignificant regions.

$$M_a(S_1, S_2) = log(1 + min(A_1, A_2)) \quad (4)$$

The influence on the final depth ordering of the image is done by the depth similarity measure M_d. It is based on the evaluation of T-junctions as it is given in eq.5. The region can be considered as T-junction, in case of 3 neighbor regions are found. The geometrical configuration of T-junctions encodes relative depth information of the objects in partial occlusion: the stem of the T belongs to the partially occluded object and the roof to the occluding object. Therefore, the depth similarity measure increases the difference between regions if they belong to different planes, or cause fusion in case of having same depth level. The probability to be occluded is calculated for each region S_1 or S_2, if S_3 is a common neighbor.

$$M_d(S_1, S_2) = \frac{1}{(1 - |p_1 - p_2|)} \quad (5)$$

Monocular depth estimation is founded on two depth cues: T-junctions and convexity. The confidence computation is run during BPT construction to be deployed into region similarity measure. Thus, detected T-junction should have a high contrast, straight branches and strict angle. To obtain this information color, angle, and curvature are investigated at each point of the image. Their values evaluate the confidence to be a real T-junction for a given point. Convexity can be reliable monocular depth cue only on the very long contour, so it is estimated only in the last partition of BPT. Therefore, S_1 can be considered as convex if its boundary is less than S_2.

The final depth ordering is based on pruning BPT, where leaves are represented by regions. They are iteratively merged based on similarity measure *M*. Since depth information from T-junctions and convexity can be contradictory, a conflict resolution step is applied using a probabilistic model. Finally, a depth map is obtained pruning BPT, as described in eq.6.

$$D_{mono} = f_{BPT}(M) \quad (6)$$

However, D_{mono} cannot be used directly because of the sharpness of the obtained segmentation. To cope with this a smoothing using a Gaussian filter *G* is applied as described in eq.7.

$$D^G_{mono} = D_{mono} * G \quad (7)$$

2.3. Fusion process

The fusion step is defined to take advantage of both 2D saliency features and monocular depth. It should take into consideration 3D perception characteristics that have been identified in recent eye-tracking experiments. For instance, it has been demonstrated that first-plan objects are highly salient. Therefore, in case 2D saliency indicates low significance, monocular cues should increase it.

In the proposed model, the saliency map for the dominant view is merged with the monocular depth as given in eq.8. The log-function is used to reduce the dynamic of monocular depth.

$$S = f\left(S_{2D}, D^G_{mono}\right) = S_{2D} \cdot \log_{10}\left(10 + D^G_{mono}\right) \quad (8)$$

3. RESULTS AND DISCUSSIONS

In order to evaluate the proposed 3D saliency model, we used 9 images from Middlebury Stereo Datasets 2005 by choosing view 1 and 5 from the proposed angles [21]. We also used the database described in [11] already containing eye-tracking data. Performance evaluation has been conducted on both databases using state-of-the-art measures.

3.1. Eye-tracking experiment

The experiment was conducted using an eye-tracking device (Tobii TX-120) in a noise-isolated room with diffused lighting. Fifteen healthy naive subjects (9 male, 6 female) participated to this test. Their age was between 19 and 34. All viewers have been screened with RANDOT (Random Dot) stereo test, Ishihara color test and FrACT (Freiburg Visual Acuity Test). Each image (from the dataset described above) was shown to viewers during 3 sec. At the beginning of the test some images were presented during 10 sec to let viewers adapt to 3D watching. Participants were not asked to perform any task, just a free-watching. During the session, positions of eyes and gaze-points were recorded by eye-tracker. It delivers data about gaze point information and eye-movement activity for each eye every 8-10 ms. Afterwards, gaze points and fixation data were extracted and processed. We used Gaussian kernel centered on fixation points and with a radius proportional to the fixation duration in order to construct the attention map.

3.2. Performance assessment

For the sake of validation, we compared objectively the proposed 3D saliency model with generated attention map (ground truth). So far, there is no standard method to perform this comparison, but few measures were proposed in literature [22]. One of the widely-used methods considers the fixation density map and the output of the model as probability distributions and measures the distance between them with help of Kullback-Leibler divergence (*KLD*). In case of presenting the obtained and reference maps as random variables, Correlation Coefficient (*CC*) is used to assess their statistical relation. The ROC curve presents true positive rate versus false positive rate for different threshold values by comparing reference and predicted saliency maps. The area under the ROC curve (*AUC*) shows that the rate of a randomly chosen positive values is higher than a randomly chosen negative one. The evaluation of our model is presented in tables 1, 2 and figure 3 shows stepwise results of the proposed model.

With the aim to compare our results with 3D attention map, we merged fixations resulting from both eyes. As it can be noticed from tables 1 and 2, the use of depth information is important in describing the potential saliency area in 3D. The results from all three similarity measures show significantly higher performance for the proposed model than models based on 2D salient features. Following the results of *KLD*, our model gives values in the range [0,732 , 1,074] much closer to 0 than 2D models, ranging in [1,705, 2,907]. A similar tendency can be seen for the two others performance metrics.

Table 1. Evaluation of proposed saliency model on our dataset. (Perfect prediction case: $KLD \to 0$, $AUC \to 1$, $CC \to \pm 1$).

Model	KLD	AUC	CC
2D Itti + mono depth	1.074	**0.796**	0.592
2D Itti	2.819	0.199	0.221
2D IP + mono depth	**0.815**	0.722	**0.634**
2D IP	2.907	0.447	0.464

Table 2. Evaluation of proposed saliency model on dataset [11]. (Perfect prediction case: $KLD \to 0$, $AUC \to 1$, $CC \to \pm 1$).

Model	KLD	AUC	CC
2D Itti + mono depth	0.775	0.777	0.678
2D Itti + mono depth + downsampling	0.950	0.664	0.563
2D Itti	1.874	0.474	0.341
2D IP + mono depth	**0.732**	**0.808**	**0.687**
2D IP + mono depth + downsampling	0.767	0.726	0.680
2D IP	1.705	0.420	0.607
DSM [11]	0.708	0.656	0.368

Moreover results of our model are highly comparable to those in [11] with even clearly better results for *CC*.

As far as calculation of monocular depth is strongly dependent on image size, this process may have an important complexity for high definition images (HD). Thereby, we tested our model by adding a stage of downsampling before the computation of monocular depth and a stage of up sampling after. The size of image was reduced 4 times, then obtained monocular depth map was restored using bicubic method. It can be seen from table 2, the model with down-sampling procedure decreases slightly performance, *KLD* changes from 0,775 to 0,950 and from 0,732 to 0,767, but nevertheless it shows better results than 2D saliency models (1,705 and 1,874). In such a way, computational cost can be reduced considerably, while saving overall performance.

From the results in tables 1 and 2, it can be noticed also, that models with 2D saliency map based on interest points showed better results than Itti's bottom-up model. Similar conclusion could be made for *CC*. The overall conclusion is that both saliency model presents competitive behavior for 3D saliency map estimation. The

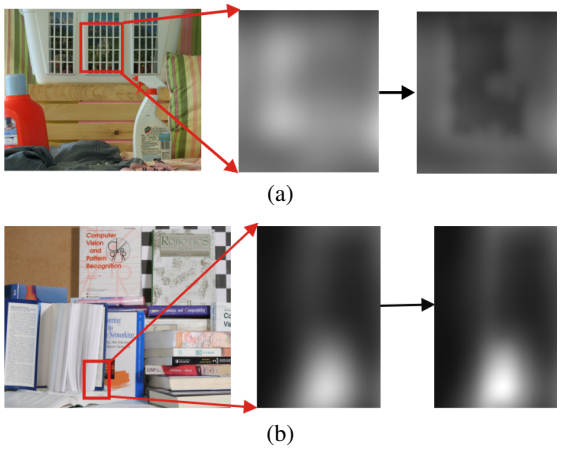

Fig. 2. Effect of using monocular depth map. From left to right: dominant view, zoom on 2D saliency map, zoom on proposed map. *a* - effect of background position, *b* - effect of front position.

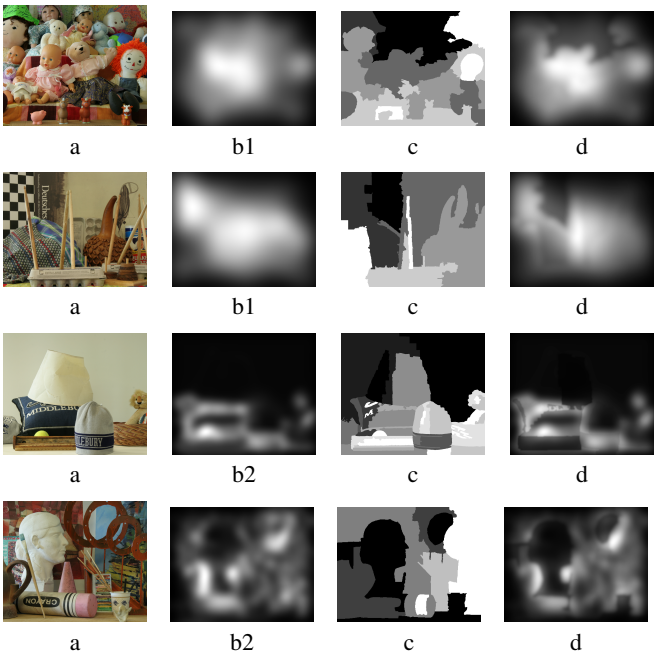

Fig. 3. Step-by-step results; *a* - dominant view of stereo pair; *b* - dominant saliency map: *b1* - based on interest points, *b2* - Itti model; *c* - mono depth map (lighter - closer, darker - further); *d* - the result of the proposed model

one interest points offers the opportunity of using the model in a real-time processing system [23].

From results depicted on figure 2, the effect of depth is quite noticeable. Parts of images that were salient in 2D lose their saliency in the proposed model because of their position in the background, as it can be seen on figure 2a . Other parts that were slightly salient got some extra saliency thanks to their position as foreground (see figure 2b).

4. CONCLUSIONS

In this paper, we proposed model for 3D saliency prediction for still images. The proposed model is based on a weighted fusion of 2D saliency features obtained using state-of-the-art methods on the one hand, and depth obtained from monocular cues, on the other hand. The model predicts visual attention for 3D image from a single view available. From the simulation results, it can be noticed that the results obtained with our model are sound with regards to visual attention obtained by eye-tracking. Therefore, performance metrics such as KLD, CC and AUC demonstrated the good results of our 3D saliency model in comparison with 2D saliency or a recently developed model based on depth. Thanks to additional steps of downsampling and up sampling, this model become computationally efficient. The natural way of extending this work is taking into account temporal phenomena in order to obtain a 3D video saliency model.

5. REFERENCES

[1] K.K. De Valois, *Seeing*, Academic Press, London, UK, 2000.

[2] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 20, no. 11, pp. 1255, 1998.

[3] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau, "A coherent computational approach to model bottom-up visual attention," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 5, pp. 802–817, 2006.

[4] Gert Kootstra, Arco Nederveen, and Bart De Boer, "Paying attention to symmetry," in *Proceedings of the British Machine Vision Conference (BMVC2008)*. The British Machine Vision Association and Society for Pattern Recognition, 2008, pp. 1115–1125.

[5] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[6] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1597–1604.

[7] Michael Nauge, Mohamed-Chaker Larabi, and Christine Fernandez-Maloigne, "A statistical study of the correlation between interest points and gaze points," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 829111–829111.

[8] Lina Jansen, Selim Onat, and Peter König, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *Journal of Vision*, vol. 9, no. 1, 2009.

[9] Quan Huynh-Thu and Luca Schiatti, "Examination of 3d visual attention in stereoscopic video content," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2011, pp. 78650J–78650J.

[10] Yun Zhang, Gangyi Jiang, Mei Yu, and Ken Chen, "Stereoscopic visual attention model for 3d video," in *Proceedings of the 16th international conference on Advances in Multimedia Modeling*. Springer, 2010, pp. 314–324.

[11] Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet, Vincent Ricordel, et al., "A computational model of stereoscopic 3d visual saliency," *IEEE Transactions on Image Processing*, vol. 22, no. 6, 2013.

[12] Nabil Ouerhani and H Hugli, "Computing visual attention from scene depth," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. IEEE, 2000, vol. 1, pp. 375–378.

[13] Christel Chamaret, Sylvain Godeffroy, Patrick Lopez, and Olivier Le Meur, "Adaptive 3d rendering based on region-of-interest," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2010, pp. 75240V–75240V.

[14] Ekaterina Potapova, Michael Zillich, and Markus Vincze, "Learning what matters: combining probabilistic models of 2d and 3d saliency cues," in *Computer Vision Systems*, pp. 132–142. Springer, 2011.

[15] Ashutosh Saxena, Sung H Chung, and Andrew Ng, "Learning depth from single monocular images," in *Advances in Neural Information Processing Systems*, 2005, vol. 18, pp. 1161–1168.

[16] Guillem Palou and Philippe Salembier, "Monocular depth ordering using t-junctions and convexity occlusion cues," *IEEE Transactions on Image Processing*, vol. 22, pp. 1926–1939, 2013.

[17] Alistair P Mapp, Hiroshi Ono, and Raphael Barbeito, "What does the dominant eye dominate? a brief and somewhat contentious review," *Perception & Psychophysics*, vol. 65, no. 2, pp. 310–317, 2003.

[18] Gorkem Saygili, Cihat Goktug Gurler, and A Murat Tekalp, "Quality assessment of asymmetric stereo video coding," in *17th IEEE International Conference on Image Processing (ICIP), 2010*. IEEE, 2010, pp. 4009–4012.

[19] Payman Aflaki, Miska M Hannuksela, J Hakkinen, Paul Lindroos, and Moncef Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," in *17th IEEE International Conference on Image Processing (ICIP), 2010*. IEEE, 2010, pp. 4021–4024.

[20] P Salembier and L Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *Image Processing, IEEE Transactions on*, vol. 9, no. 4, pp. 561–576, 2000.

[21] Heiko Hirschmuller and Daniel Scharstein, "Evaluation of cost functions for stereo matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[22] A Borji and L Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 185–207, 2013.

[23] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417. Springer, 2006.