VISUAL ATTENTION MODELING FOR STEREOSCOPIC VIDEO

Yuming Fang[1], *Chi Zhang*[1], *Jing Li*[2], *Matthieu Perreira Da Silva*[2], *and Patrick Le Callet*[2]

[1]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China
[2]LUNAM Universit, Universit de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, France

ABSTRACT

In this paper, we propose a computational model of visual attention for stereoscopic video. Low-level visual features including color, luminance, texture and depth are used to calculate feature contrast for spatial saliency of stereoscopic video frames. Besides, the proposed model adopts motion features to compute the temporal saliency. Here, we extract the relative planar and depth motion for temporal saliency calculation. The final saliency map is computed by fusing the spatial and temporal saliency together. Experimental results show the promising performance of the proposed method in saliency prediction for stereoscopic video.

Index Terms— Visual attention, saliency detection, stereoscopic video, depth feature, planar motion, depth motion.

1. INTRODUCTION

Visual attention is a significant mechanism in visual perception to rapidly detect the salient information of natural scenes. When observing natural scenes, selective attention would focus on some specific salient information and ignore other unimportant information due to the limited processing resources. Basically, there are two approaches in visual attention: bottom-up and top-down. The bottom-up process is data-driven and task-independent for automatic salient region detection, while the top-down approach is a task-dependent cognitive process related to certain specific task, observers' experiences, *etc*. [1, 2].

Generally, the salient regions extracted by visual attention models can be widely used in many 2D multimedia applications such as visual quality assessment [11], coding [3, 12], segmentation [13, 14], etc. Previously, there have been various computational visual attention models proposed for 2D images/video [4]-[8]. Itti *et al.* proposed an early visual attention model for images by feature contrast from luminance, color, and orientation [4]. Following this work, Harel *et al.* used graph theory to better measure the saliency of image pixels in [5]. Le Meur *et al.* proposed a saliency detection model by the characteristics of the human visual system including contrast sensitivity function, visual masking, and center-surround interactions [9]. Some studies try to extract the saliency information of images from transform domain [6, 7, 8]. Recently, machine learning techniques are widely used for visual attention modeling [25, 26]. More related works can be referred to the study [3].

In the past decade, there have been various emerging stereoscopic multimedia applications with the rapid development of stereoscopic display technologies, including 3D visual quality assessment [15, 16, 17], 3D video coding [18, 19], 3D content generation [20], etc. For these 3D multimedia applications, the integration of visual 3D perception models can be used to improve the performance of these related 3D multimedia processing algorithms. Stereoscopic visual attention, an important stereoscopic visual perception process, can be used to extract the salient regions in stereoscopic visual content for various stereoscopic multimedia applications.

Currently, there are some studies focusing on investigating stereoscopic visual attention modeling. Chamaret *et al.* proposed a saliency detection algorithm for 3D rendering [21]. In that study, the disparity map is used to weight the 2D saliency map to calculate the final saliency map for 3D images [21]. Lang *et al.* constructed an eye tracking database for stereoscopic images and investigated the influence of depth in stereoscopic visual attention modeling [22]. In the study [23], Wang *et al.* designed stereoscopic saliency detection models by combing depth saliency and 2D saliency with the fusion methods of summation and multiplication. An eye tracking database was also built to demonstrate the performance of the stereoscopic saliency detection models [23]. Recently, Fang *et al.* proposed a stereoscopic visual attention model by combining 2D saliency map, depth saliency map and center bias map [24].

Most of the existing stereoscopic visual attention models introduced above are proposed for stereoscopic images. The research on visual attention modeling for stereoscopic video is still limited in the research community. Compared with the visual attention modeling for stereoscopic images, the influence of motion has to be considered in visual attention modeling for stereoscopic video. Basically, there are two types

This work was supported in part by the Natural Science Foundation of China under Grant 61571212 and by the Natural Science Foundation of Jiangxi under Grant 20151BDH80003.

of motion in stereoscopic video: the planar motion and depth motion. In this paper, we propose a novel computational visual attention model for stereoscopic video. Low-level features including color, luminance, texture and depth are extracted to compute the feature contrast for spatial saliency prediction. In addition, planar motion and depth motion are computed for temporal saliency estimation. The final saliency map of stereoscopic video is computed by combining the spatial and temporal saliency maps.

The remaining of this paper is organized as follows. Section 2 introduces the proposed model in detail. In Section 3, we conduct the comparison experiment to demonstrate the performance of the proposed model. The final section concludes the study.

2. PROPOSED METHOD

As described in the previous section, we propose a visual attention model for stereoscopic video by combining the spatial and temporal saliency maps. The proposed model firstly extracts the features of luminance, color, texture and depth from stereoscopic video. Then it computes the feature contrast for these features for spatial saliency calculation. Additionally, the planar and depth motion is extracted from the stereoscopic video for temporal saliency computation. We will introduce the details step by step in the following subsections.

2.1. Feature Extraction

In the proposed method, we extract three types of features: 2D feature, depth feature and motion features. The 2D and depth features are used to calculate the feature contrast, which has been widely used for visual attention modeling in the literature. For the motion feature, we extract the planar and depth motion for temporal saliency calculation.

In the proposed model, we extract the feature contrast from low-level features of color, luminance, texture and depth for the proposed model. It is well known that DCT (Discrete Cosine Transform) is a superior representation of energy compactness and most energy is concentrated on a few low-frequency coefficients. Because of this, DCT has been widely used in various multimedia processing applications. Inspired by a recent study which uses DCT coefficients in saliency estimation [7], DCT coefficients are adopted to compute the feature contrast for stereoscopic video in this study.

The input image is firstly transformed in YCbCr color space and divided into small image patches. For each image patch, the DCT coefficients are adopted to represent the energy. Here, we use the DC coefficient of Y component to represent the luminance feature of each image patch. Assume that Y_{DC} is the DC coefficient of Y component, the luminance feature L can be represented as: $L = Y_{DC}$. The DC coefficients of Cb and Cr components are used to represent the color features of image patches. Assume that Cb_{DC} and Cr_{DC} are the DC coefficients of Cb and Cr components, respectively, The color features of each image patch C_1 and C_2 can be denoted as: $C_1 = Cb_{DC}$ and $C_2 = Cr_{DC}$.

As we know, the Cb and Cr components mainly include color information in images. Thus, we only use AC coefficients in Y component to represent the texture feature for each image patch. In each DCT block, the first several low-frequency coefficients in the left-upper corner include most energy. Thus, similar with the study [24], we only use the first nine AC coefficients in each DCT block with zig-zag scanning to represent the texture feature T: $T = \{Y_{AC1}, Y_{AC2}, ..., Y_{AC9}\}$.

For stereoscopic video, the disparity map shows the parallax of image pixels between the left-view and right-view image pair. It is estimated in unit of pixels for disparity system. Here, we first estimate the disparity map for depth feature extraction. The disparity map P is computed with the variational algorithm by Chambolle and Pock [27]. The depth map P' can be calculated based on the disparity as follows [23]:

$$P' = v/(1 + \frac{d \cdot h}{P \cdot w}) \tag{1}$$

where v denotes the viewing distance; d is the interocular distance; w and h are the width (in cm) and horizontal resolution of the display screen, respectively.

After the depth map is computed, DC coefficients of image patches in the depth map are used as the depth feature in the proposed model. The depth feature D can be represented as: $D = P'_{DC}$.

There are two types of motion in stereoscopic video: the planar motion and depth motion. The motion map of the left view is first estimated by optical flow [28]. The motion map can be represented as the motion vectors in *x* and *y* directions: $M_x(i, j, t)$ and $M_y(i, j, t)$, respectively (where (i, j) denotes the pixel location in the image; t represents the t-th frame.). The depth motion M_d can be computed as follows [29].

$$\begin{aligned} M_d(i_t, j_t, t) = P(i_t + M_x(i_t, j_t, t), j_t + \\ M_y(i_t, j_t, t), t + 1) - P(i_t, j_t, t) \end{aligned} \tag{2}$$

where P is the disparity map.

Since the depth motion would influence the 2D motion estimation in x-direction, the 2D motion map $M_x(i, j)$ is actually a combination of the depth motion and 2D x-direction motion. Thus, the planar x-direction motion can be computed by removing depth motion part from $M_x(x, y)$. The planar motion at x and y directions can be computed as follows [29].

$$M_{px} = M_x - \frac{1}{2} M_d \tag{3}$$

$$M_{py} = M_y \tag{4}$$

2.2. Saliency Estimation

It is well known that the salient regions in images pop out due to its high feature contrast from their surround regions.

The salient regions can be detected by the center-surround differences between image patches. Following the study [7], we compute the saliency value of each image patch by the feature differences between this image patch and all the others in the image. A Gaussian model of spatial distances between image patches is adopted to weight the feature differences for saliency estimation. For each image patch i, its saliency value S_i^m from feature m can be calculated as:

$$S_i^q = \sum_{j \neq i} \frac{1}{\sigma\sqrt{2\pi}} e^{-d_{ij}^2/(2\sigma^2)} R_{ij}^q \tag{5}$$

where $q \in \{L, C_1, C_2, T, D\}$; d_{ij} is the spatial distance between image patches i and j; R_{ij}^q represents the feature difference between image patches i and j; σ is the parameter of Gaussian model, which is used to determine the degree of local and global feature contrast. For the luminance, color and depth features, the feature differences R_{ij}^q between image patches i and j can be computed by the normalized differences between the corresponding DC coefficients. For the texture feature composed of nine AC coefficients, the feature difference R_{ij}^T can be calculated by the L2 norm.

After the feature maps from luminance, color, texture, and depth are computed according to Eq. (5), we calculate the spatial saliency map for stereoscopic video by simply linear combination of these feature maps as follows.

$$S^s = \frac{1}{Q} \sum_q S^q \tag{6}$$

where Q represents the number of features used for spatial saliency prediction.

It is well accepted that object motion is highly correlated with visual attention [30, 31]. In general, an object with strong motion with respect to the background would attract human's attention [30, 31]. The planar motion and depth motion we compute in Eqs. (2) - (4) are absolute local motion. Usually, the object motion we perceive represents the relative motion between the object and background [33, 34]. Thus, we have to calculate the relative planar motion and relative depth motion for temporal saliency extraction of stereoscopic video.

To be aligned with the spatial saliency map computation in Eq.(5), the motion feature maps are estimated as relative planar motion and relative depth motion by motion feature contrast. For each image patch i, the relative motion value from planar/depth motion can be computed as follows:

$$v_i^m = \sum_{j \neq i} \frac{1}{\sigma\sqrt{2\pi}} e^{-d_{ij}^2/(2\sigma^2)} R_{ij}^m \tag{7}$$

where $m \in \{M_d, M_p\}$ (M_d and M_q represent depth motion and planar motion, respectively). R_{ij}^m denotes the depth/planar motion differences between patches i and j. The other parameters are similar with those in Eq. (5). Please note that the depth/planar motion difference R_{ij}^m is computed by motion normalization as follows:

$$R_{ij}^m = \frac{|M_i^m - M_j^m|}{|M_i^m| + |M_j^m| + \mathcal{C}} \tag{8}$$

where $M^m \in \{M_d, M_p\}$; $\mathcal{C}$ is a small constant.

Eq. (7) computes the relative motion in a more localized form within a large neighboring region rather than comparing with the global background motion. After the feature maps of planar and depth motion are computed in Eq. (7), we estimate the temporal saliency by combining these two motion feature maps as follows.

$$S^t = \begin{cases} 0, \max_i(v_i^p) \leq \mathcal{T} \;\&\&\; \max_i(v_i^d) \leq \mathcal{T} \\ v^d, \max_i(v_i^p) \leq \mathcal{T} \;\&\&\; \max_i(v_i^d) > \mathcal{T} \\ v^p, \max_i(v_i^d) \leq \mathcal{T} \;\&\&\; \max_i(v_i^p) > \mathcal{T} \\ \frac{1}{2}(v^d + v^p), otherwise \end{cases} \tag{9}$$

where i denotes the image patch in the video frame; v_p and v_d represent the relative planar motion and relative depth motion computed by Eq. (7), respectively; $\mathcal{T}$ is a threshold value. From Eq. (9), we can see that, the temporal saliency would be zero if both relative planar motion and relative depth motion are small; the temporal saliency would be relative planar motion (relative depth motion) if the relative depth motion (relative planar motion) is small; the temporal saliency would be the linear combination of relative planar motion and relative depth motion if both of them are larger than the threshold.

With the computed spatial and temporal saliency maps, the final saliency map for stereoscopic video can be obtained by combining these two maps. Here, we consider the spatial and temporal saliency as the same important, and thus we use the simple linear combination method to fuse these two types of saliency maps as follows.

$$S = \frac{1}{2}(S^s + S^t) \tag{10}$$

3. EXPERIMENTAL RESULTS

In this section, we conduct the comparison experiment to show the performance of the proposed visual attention model of stereoscopic video. We use a subset of the eye tracking database [35] to conduct the experiment. This database includes video sequences with various content with different levels of 3D effect, aesthetic composition, variations in colour, environment, motion, texture, light, etc. [35]. The stereoscopic video sequences are obtained by the Panasonic AG-3D camera and Intel SATA3 SSDs to record the visual content. These video sequences are recorded with the resolution of 1920×1080. The eye tracking data is obtained from the recording data by the SMI RED (4 firewire) remote eye-tracker working at 60Hz. A set of 40 subjects from

19 to 44 years old were involved in the eye tracking experiment and they were asked not to move during the experiment. The stereoscopic video sequences were displayed on a 26-inch Panasonic BT-3DL2550 LCD screen with the refresh rate of 60Hz and resolution of 1920×1080. The video sequences were viewed by subjects with a pair of passive polarized glasses. The lab environment luminance was adjusted for subjects with an appropriate size for the pupil during the eye tracking experiment. The gaze points recorded by the eye tracker are used to create the fixation density map, which is used as the ground truth of saliency estimation for stereoscopic video.

In this experiment, we evaluate the performance of the proposed model by comparing the saliency maps from the computational visual attention models with the fixation density map from eye tracking data. In general, an effective visual attention model can predict the saliency map similar with the fixation map. We use three common methods to evaluate the performance of the proposed model: Receiver Operating Characteristics (ROC), Linear Correlation Coefficient (CC), and Normalized Scanpath Saliency (NSS).

In the research community, ROC curve is widely adopted for performance evaluation of visual attention models. Through the defined threshold, the saliency map from visual attention model can be divided into salient points and non-salient points. There are target points and background points in the fixation map from eye tracking data. The True Positive Rate (TPR) is computed as the percentage of target points falling into the salient points from the visual attention model, while the False Positive Rate (FPR) is calculated by the percentage of background points falling into the salient points from the visual attention model. The ROC curve of the visual attention model can be obtained as the curve of TPR vs. FPR through defining different thresholds. The area under ROC curve (AUC) provides an overall performance evaluation. A better visual attention model is expected to have a larger AUC value.

CC can be used to measure the degree of linear correlation between the saliency map and fixation map. Here, Pearson's correlation coefficient between two variables is used to compute CC for visual attention models. It is computed by the covariance of the saliency map and fixation map divided by the product of the standard deviations as:

$$CC(s,f) = \frac{cov(s,f)}{\sigma_s \sigma_f} \tag{11}$$

where s and f are saliency map and fixation map, respectively. The CC values are in the range of [0,1] and with larger CC value, the visual attention model can obtain better performance of saliency prediction.

Besides, we also use NSS to evaluate the performance of the proposed model. It is defined by the response value at human fixation locations in the normalized saliency map with zero mean and unit standard deviation as:

$$NSS(s,f) = \frac{1}{\sigma_s}(s(i_f, j_f) - \mu_s) \tag{12}$$

where (i_f, j_f) is the pixel location in the fixation map; μ_s is the mean value of the saliency map; s and f are the saliency map and fixation map, respectively; σ_s is the standard deviation of the saliency map. Usually, with the higher the NSS value, the visual attention model can estimate better saliency results.

In this experiment, we compare the proposed model with several existing studies including Itti-2D [4], Fang-3D [24], Seo-2DV [36]. Specially, Itti-2D is a classical visual attention model for 2D images; Fang-3D is a visual attention model proposed recently for stereoscopic images; Seo-2DV is a saliency detection model for 2D video sequences. The comparison experimental results are shown in Table. 1. From this table, we can see that Fang-3D and Seo-2DV can obtain much better performance than Itti-2D, which mean that the recent 3D saliency detection model [24] and 2D video saliency detection model can predict more accurate saliency information for stereoscopic video than the classic model [4]. Compared with the other existing related saliency detection models, the proposed model can obtain higher values of AUC, CC and NSS. This means that the proposed model can obtain the best performance of saliency prediction for stereoscopic video among the compared models.

In Fig. 1, we also provide some comparison samples from different saliency detection models. From these comparisons, we can find that Itti-2D can only detect the contour information in the images. The Fang-3D model would lost some salient regions. On the contrary, the proposed model can obtain better saliency maps than other existing ones.

4. CONCLUSION

In this paper, we propose a novel visual attention model for stereoscopic video by combining the spatial and temporal saliency. The features of luminance, color, texture and depth are extracted to compute the feature contrast, which is adopted for spatial saliency prediction for stereoscopic video. For temporal saliency estimation, we consider both of the relative planar and depth motion. Experimental results show the promising performance of the proposed model in saliency prediction of stereoscopic video. In the future, we will further investigate the combination methods for different feature maps in the final saliency prediction for stereoscopic video.

5. REFERENCES

[1] A. Treisman and G. Gelade. A feature-integration theory of attention. Cognitive Psychology, 12(1), 97-136, 1980.

Table 1. Comparisons of different saliency detection models

Models	Itti-2D [4]	Fang-3D [24]	Seo-2DV [36]	Proposed Model
AUC	0.6598	0.7294	0.7303	0.7632
CC	0.2302	0.2562	0.2643	0.3102
NSS	0.5886	0.8499	0.8542	0.9155

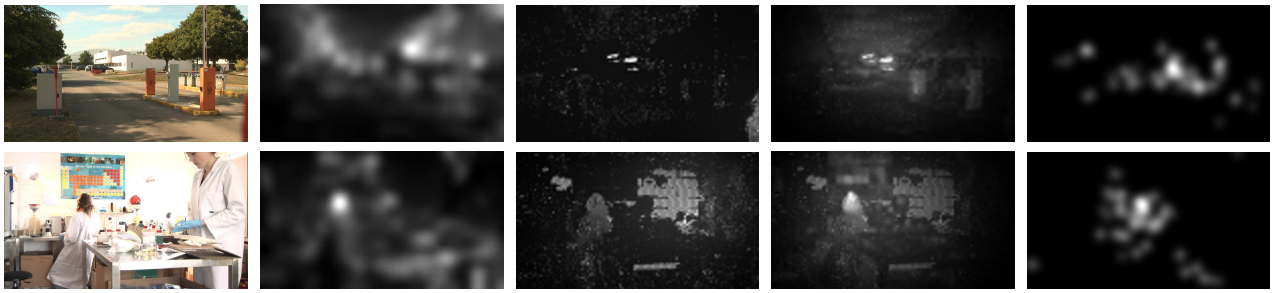


Fig. 1. Comparison of different saliency detection algorithms. First column to the final column: original images; saliency map from Itti-2D, Fang-3D, the proposed model; the ground truth.

[2] Y. Fang, W. Lin, C. T. Lau, and B.-S. Lee, A visual attention model combining top-down and bottom-up mechanisms for salient object detection. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1293-1296, 2011.

[3] A. Borji, and L. Itti, State-of-the-Art in Visual Attention Modeling. IEEE Trans. Pattern Anal. Mach. Intell., 35(1): 185-207, 2013.

[4] L. Itti, C. Koch and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254-1259, 1998.

[5] J. Harel, C. Koch and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 2006.

[6] X. Hou and L. Zhang. Saliency Detection: A spectral residual approach. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[7] Y. Fang, Z. Chen, W. Lin, C.-W. Lin. Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing*, 21(9), 3888-3901, 2012.

[8] J. Li, L.-Y. Duan, X. Chen, T. Huang, and Y. Tian, Finding the Secret of Image Saliency in the Frequency Domain. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(12): 2428-2440, 2015.

[9] O. Le Meur, P. Le Callet, and D. Barba, A coherent computational Approach to model the bottom-up visual attention, *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol.28(5), pp:802-817, 2006.

[10] Z. Liu, W. Zou, and O. Le Meur, Saliency Tree: A Novel Saliency Detection Framework. *IEEE Transactions on Image Processing*, 23(5): 1937-1952, 2014.

[11] W. Lin, and C.-C. Jay Kuo, Perceptual visual quality metrics: A survey. *J. Visual Communication and Image Representation*, 22(4): 297-312, 2011.

[12] C. Guo and L. Zhang. A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1), 185-198, 2010.

[13] Y. Tian, J. Li, S. Yu, and T. Huang, Learning Complementary Saliency Priors for Foreground Object Segmentation in Complex Scenes. *International Journal of Computer Vision*, 111(2): 153-170 (2015)

[14] J. Lei, H. Zhang, L. You, C. Hou, and L. Wang, Evaluation and modeling of depth feature incorporated visual attention for salient object segmentation. *Neurocomputing*, 120: 24-33, 2013.

[15] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, Full-Reference Quality Assessment of Stereoscopic Images by Learning Binocular Receptive Field Properties. *IEEE Transactions on Image Processing*, 24(10), 2971-2983, 2015.

[16] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, Quality Prediction of Asymmetrically Distorted Stereoscopic 3D Images, *IEEE Transactions on Image Processing*, 24(11), 3400-3414, 2015.

[17] Q. Huynh-Thu, M. Barkowsky, and P. Le Callet, The Importance of Visual Attention in Improving the 3DTV

Viewing Experience: Overview and New Perspectives, *IEEE Transactions on Broadcasting*, 57(2), 421-431, 2011.

[18] S. Ma, S. Wang, and W. Gao, Low Complexity Adaptive View Synthesis Optimization in HEVC Based 3D Video Coding, *IEEE Transactions on Multimedia*, 16(1), 266-271, 2014.

[19] Z. Gu, J. Zheng, M. Ling, P. Zhang, Fast segment-wise DC coding for 3D video compression, *IEEE ISCAS*, 2015.

[20] X. Cao, A. C. Bovik, Y. Wang, and Q. Dai, Converting 2D Video to 3D: An Efficient Path to a 3D Experience, *IEEE MultiMedia*, 18(4): 12-17, 2011.

[21] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, Adaptive 3d rendering based on region-of-interest, in *IST/SPIE Electronic Imaging*, 2010.

[22] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: influence of depth cues on visual saliency, *European Conference on Computer Vision*, 2012.

[23] J. Wang, M. P. DaSilva, P. LeCallet, and V. Ricordel, A Computational Model of Stereoscopic 3D Visual Saliency, *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151-2165, 2013.

[24] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, Saliency Detection for Stereoscopic Images. *IEEE Transactions on Image Processing*, 23(6): 2625-2636, 2014.

[25] T. Judd, K. Ehinger, F. Durand, and A. Torralba, Learning to predict where humans look. *International Conference on Computer Vision*, 2009.

[26] J. Li, Y. Tian and T. Huang, Visual Saliency with Statistical Priors, *Intl. J. Comput. Vision*, 107(3), pp.239-253, 2014.

[27] A. Chambolle, and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, Journal of Mathmatical Imaging and Vision, 40(1): 120-145, 2011.

[28] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. IEEE Conf. on Computer Vision and Pattern Recog., 2010.

[29] J. Li, Methods for assessment and prediction of QoE, preference and visual discomfort in multimedia application with focus on S-3DTV, PhD Thesis, 2013.

[30] R. A. Abrams, and S. E. Christ. Motion onset captures attention. Psychological Science, 14, 427-432, 2003.

[31] S. Yantis, and J. Jonides. Abrupt visual onsets and selective attention: Evidence from visual search. Journal of Experimental Psychology: Human Perception and Performance, 10, 601- 621, 1984.

[32] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. Nature Neuroscience, 9, 578 - 585, 2006.

[33] R. J. Snowden. Sensitivity to relative and absolute motion. Perception, 21:563-568, 1992.

[34] D. A. Poggel, H. Strasburger, and M. MacKeben. Cueing attention by relative motion in the periphery of the visual field. Perception, 36, 955-970, 2007

[35] Y. Fang, J. Wang, J. Li, R. Ppion, and P. Le Callet, An eye tracking database for stereoscopic video, QoMEX, 51-52, 2014.

[36] H. J. Seo and P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis., vol. 9, no. 12, p. 15, 2009.