UNSUPERVISED SALIENCY DETECTION IN 3D VIDEO BASED ON MULTI-SCALE ARCHITECTURE AND GRAPHICAL MODEL

Anonymous ICME submission

ABSTRACT

In this paper, we propose an unsupervised saliency objects extraction method for 3D video. The proposed framework consists with three main stages: (i) The input video frame is segmented into none-overlapping superpixels by combining both appearance and depth information in the input. The usage of depth can tackle the segmentation of complex regions, *e.g.*, the objects in foreground that have similar appearance with the background. A multi-scale segmentation scheme is also deployed by using different segmentation parameters to extract the varying shapes of the foregrounds in each frame. (ii) The initial saliency score of each segmented superpixel in each scale is calculated via global contrast which is defined by appearance, depth, and motion cues from two consecutive frames. And (iii) the initial saliency scores in each scale are refined by smoothing over a graph built by the spatial neighboring of all the superpixels in the frame. The final result is generated by fusing the saliency maps in all scales. The experiments on two widely-used datasets illustrate that our method excels the state-of-the-art in accuracy, robustness, and reliability.

Index Terms— Saliency detection, Segmentation, Appearance, Motion, Depth

1. INTRODUCTION

Human visual attention is most significant in human visual system, it ensures us to extract the most important object from the very complex scenery. This object is called the salient object and this area in the picture is saliency area.

Center-surround visual attention in [1] used three elementary features, color, luminance and direction, to calculate the contrast between the center and surround areas. By knowing the mechanism of the human visual perception, Itti [1] defined "saliency" as the area which is much different from the surrounding. This definition was broadly accepted. In [2, 3], the center-surround algorithm was upgraded using the KL distance of two feature histograms. Later, instead of the part contrast, global contrast algorithm was designed to segment the saliency area [4]. Considering the feature dispersal effect, a "saliency filter" [5] was designed to calculate both the contrast and dispersal degree of color to determine the saliency area. Apart from the contrast algorithm, graphical model is used in saliency detection starting in 2006. Harel et. [6] modified the method of Itti [1] and proposed a graphical model to represent the similarity between each part within the picture. Then Random Walk (RW) algorithm was adopted to give the grade of each area. In 2013, Yang et. [7] used Manifold Ranking algorithm to distinguish the foreground and background.

In 3D scenery, depth information is naturally incorporated in the saliency detection model. Zhang et. [8] used appearance, depth, motion, illumination and direction to calculate bottom-up saliency in a picture and the closer the object from viewer, the more salient the object. In [9], depth is used to weigh the 2D saliency result. Based on that, Niu [10] give a depth weight curve by presuming that the comfort zone and popping out object are more salient than any other region or object. In [11, 12], instead using depth as a weight bias, they considered it as another feature and calculated the depth-only saliency map via global contrast. Finally, they combined the 2D saliency map with the depth saliency map and obtained 3D saliency result.

Later, with the growing focus on video, motion and depth information is naturally utilized in saliency detection. In 2010, Zhang proposed the 3DV [8] method which fused the depth feature map with appearance and motion saliency map. In 2015, Lino [13] designed a saliency detection method which spatial, motion and depth saliency maps is obtained based on Itti's model, block matching and Difference of Gaussian (DoG) filter, separately These results are fused to yield the final saliency map.

Although there are many proposed methods, their performance is limited in preserving clear edges of the foreground object. The performance is not sufficiently robust for a wide range of image an 3D video.

In order to overcome the aforementioned challenges, we proposed a Multi-scale architecture and a refining method based on Graphical Model to make the detection method general and robust. In our paper, we upgraded simple linear iterative cluster (SLIC) algorithm [14] by adding the depth information as one feature while clustering the pixels into superpixels. As a result, two regions which share a similar appearance but belong to different objects can be separated. Multi-scale architecture is generated by setting the different number of superpixels in each scale and then fusing the final result of each scale together. This architecture enables robustness and generality by combining the merit under each scale and fix-

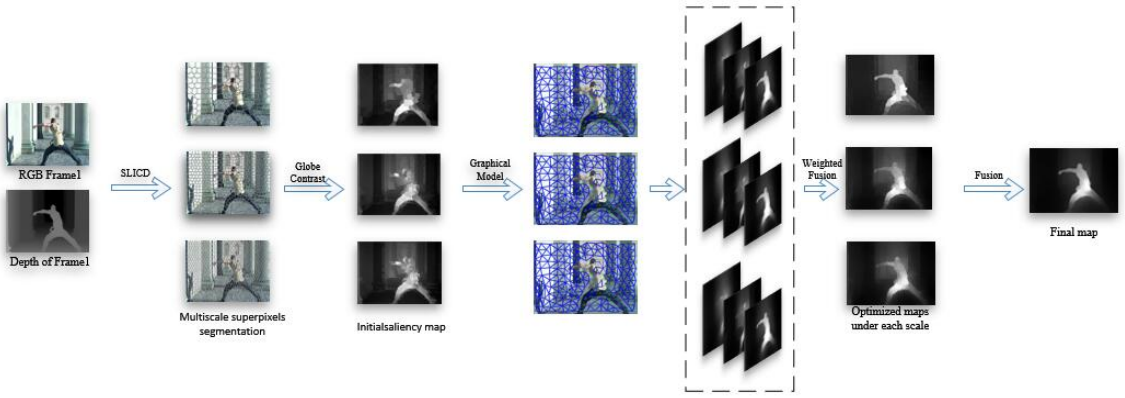


Fig. 1. Flowchart of the proposed method

ing the problem of setting optimized superpixel number due to the continuously changed inputs. After motion feature was calculated by TV-L_1 optical flow [15, 16], the initial saliency map are obtained via global contrast [11] with appearance, depth, and motion. Finally, the refinement based on graphical model [17] is utilized to refine the initial saliency maps.

Overall, the contributions of our method are as follows: (1) The SLIC algorithm is upgraded by adding the depth information and a multi-scale architecture is established. (2) The transition probabilities with appearance, depth and motion are separately utilized to refine the initial saliency maps and than fused three-feature-prior refined saliency maps together. The graphical connections are also drawn based on transition probability function.

2. PROPOSED METHOD

Fig. 1 shows the framework of the proposed method which consists of three parts: (i) Based on upgraded SLIC algorithm, each input image (frame) with appearance and depth is segmented to three images with different superpixel number, as three scales. (ii) The score of each superpixel in each image under each scale is computed using global contrast which is defined by appearance, depth, and motion cues from two consecutive frames. And the initial saliency map in each scale is obtained. (iii) In order to refine the initial saliency map, graphical model is adopted to smooth and regulate the saliency area by appearance, motion and depth prior. Then the final result is obtained by fusing three refined saliency map in all scales. A guided filter is also adopted to process the final result. The proposed method is unsupervised since ground truth is only used for testing. Detailed description of each part is presented in the following subsections.

2.1. Segmentation

In this part, our aim is to decompose the input image into non-overlapping segments and preserve accurate object's shape and edge. While the traditional K-means algorithm in SLIC used appearance, *i.e.*, color, and distance to cluster pixels, we further add depth information with appropriate weight, in the proposed method. We set the depth feature distance as $D_d = |d_j - d_i|/|d_j + d_i|$, where d_j and d_i represent the j^{th} and i^{th} superpixels' depth value, respectively. By adding the weight coefficient of depth feature distance as β, the whole feature distance evaluation coefficient becomes:

$$D' = (1-\beta)D_c + \beta D_d \tag{1}$$

$$D = \sqrt{(\frac{D'}{m})^2 + (\frac{D_S}{S})^2} \tag{2}$$

Here D' is the weighted mean of color and depth feature distance and D_S is the space Euclidean distance calculated by the coordinate of each pixel. m is a parameter related to the largest color distance in one image, with typical range of $[1, 40]$ and S is the expected number of the superpixel. Throughout the experiment, β and m are set to 0.5 and 15, as they give the best segmentation performance.

By computing the color, depth and spatial distance, the algorithm clusters pixels with smallest distance. The algorithm yields excellent segmentation result with 10 iterations.

The number of superpixels is a parameter in image segmentation. Small number of superpixels reduces the algorithm's ability to preserve fine edges of an object, while large number of superpixels misses important corners. Furthermore, the number of optimal superpixels depends on the input images. In order to increase the generality and robustness of the proposed algorithm, a Multi-scale approach is integrated with the SLIC algorithm. The first part of Fig. 1 shows the modified SLIC algorithm's output for a image with appearance and depth information (showing for 200, 600 and 1000 superpixels). The feature value in each superpixel is the mean of all pixel's feature value in the superpixel area.

2.2. Initial saliency

According to visual perceptual research [1], image contrast is the most important factor in visual saliency. In a static image, visual attention centers on the objects which have a high contrast comparing to their surrounding. In 3D videos, the object with hign-contrast appearance, depth and motion would cause the visual attention. In this part, two steps are proposed.

First, we use CIE-Lab color space to represent the appearance of input image, since LAB color correlates better to perceptual of human eyes. Motion is extracted by the TV-L_1 Optical Flow [15, 16] and the two parameters of motion,v_X and v_y, can be obtained by solving the equation $I(x, y, t) = I(x + \mathrm{d}x, y + \mathrm{d}y, t + \mathrm{d}t)$ with the constraint of by $I_x v_x + I_y v_y + I_t = 0$. I is the luminance.

Second, we use global contrast with color, depth, motion and spatial distance to generate the initial saliency maps. Since the images have been segmented by SLIC, feature contrast can be calculated on the superpixel level. We use arithmetic mean of color, motion and depth of all pixels in each superpixel area to represent the feature value of whole superpixel. The spatial distance is calculated among the barycenters of superpixels.

Similar to the distance calculation in SLIC in 2.1, the feature distance of color, denoted as d_C and motion, denoted as d_M can be calculated by Euclidean distance:

$$d_C(R_j, R_i) = \sqrt{(L_{R_j} - L_{R_i})^2 + (a_{R_j} - a_{R_i})^2 + (b_{R_j} - b_{R_i})^2} \quad (3)$$

$$d_M(R_j, R_i) = \sqrt{(v_{x_{R_j}} - v_{x_{R_i}})^2 + (v_{y_{R_j}} - v_{y_{R_i}})^2} \quad (4)$$

where R_i denotes the i^{th} superpixel, and L, a, b are the color values under CIE-Lab.

We set the gray contrast between two superpixels R_j, R_i as the depth feature distance d_D, and D_{R_i} represents the depth value of the R_i^{th} superpixels:

$$d_D(R_j, R_i) = \frac{|D_{R_j} - D_{R_I}|}{|D_{R_j} + D_{R_I}|} \quad (5)$$

After the feature distance calculation, we use weighted mean of these four feature distances, namely color, motion, depth and spatial distance, to represent the saliency score of each superpixel.

According to the visual mechanism of human which the closer the two superpixels are, the more influence they had for each other, we designed the weight coefficient based on spatial distance as follows:

$$\omega(R_j, R_i) = \exp(-\frac{d_S(R_j, R_i)}{\sigma}) \quad (6)$$

d_S represents the spatial distance among the barycenter of the superpixels. When two superpixels are closer, the value of ω is larger. The value of the parameter σ is 0.6.

Then we can obtain the score of each superpixel under each feature:

$$S_F(R_i) = \sum_{R_j \in \Omega} \omega(R_j, R_i) \cdot d_F(R_j, R_i) \quad (7)$$

Where the notation "F" stands for either "C","M" or "D", *i.e.*, color, motion, and depth, correspondingly. $S_F(R_i)$ is the saliency score of R_i^{th} superpixel under "F" feature. By normalizing all superpixels' score, the saliency map under "F" feature is obtained.

In order to get the initial saliency map, three kinds of feature-saliency maps need to be fused. Since the quality of each map is different, a weighted fusion method is conducted.

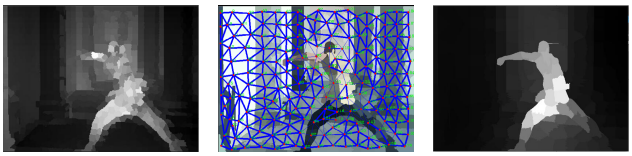
According to perceptual research [1], the more saliency area aggregated, the higher the quality of the saliency map is. Thus we design a way to calculate the aggregation degree:

$$\mu_F = \frac{\sum \sqrt{(y_{R_j} - \overline{p_{y_F}})^2 + (x_{R_j} - \overline{p_{x_F}})^2} \cdot S_F(R_i, R_j)}{\sum S_F(R_i, R_j)} \quad (8)$$

here (x_{R_j}, y_{R_j}) is the coordinate barycenter of each superpixel, $(\overline{p_{x_F}}, \overline{p_{y_F}})$ is the barycenter of the saliency area calculated by saliency scores of all superpixels, and μ_F is the aggregation degree of each feature saliency map.

Based on saliency aggregation degree, the initial saliency maps can be computed as follows:

$$S = \sum_{F \in \{C,M,D\}} 1/\mu_F \times S_F \quad (9)$$



(a) Initial Saliency (b) Graphical Model (c) Optimized Saliency

Fig. 2. Optimization based on Graphical Model

2.3. Saliency refinement based on Graphical Model

Initial saliency map obtained from global contrast has many artifacts at object's edge, as seen in Fig. 2(a). To regularize the saliency map and maintain the object's shape, a refinement method based on graphical model is exploited. A probability transition function [17] can reduce the great saliency value diversity between two near superpixels inside the saliency region and finally make the saliency area smoother and the edge between salient and none-salient area clearer.

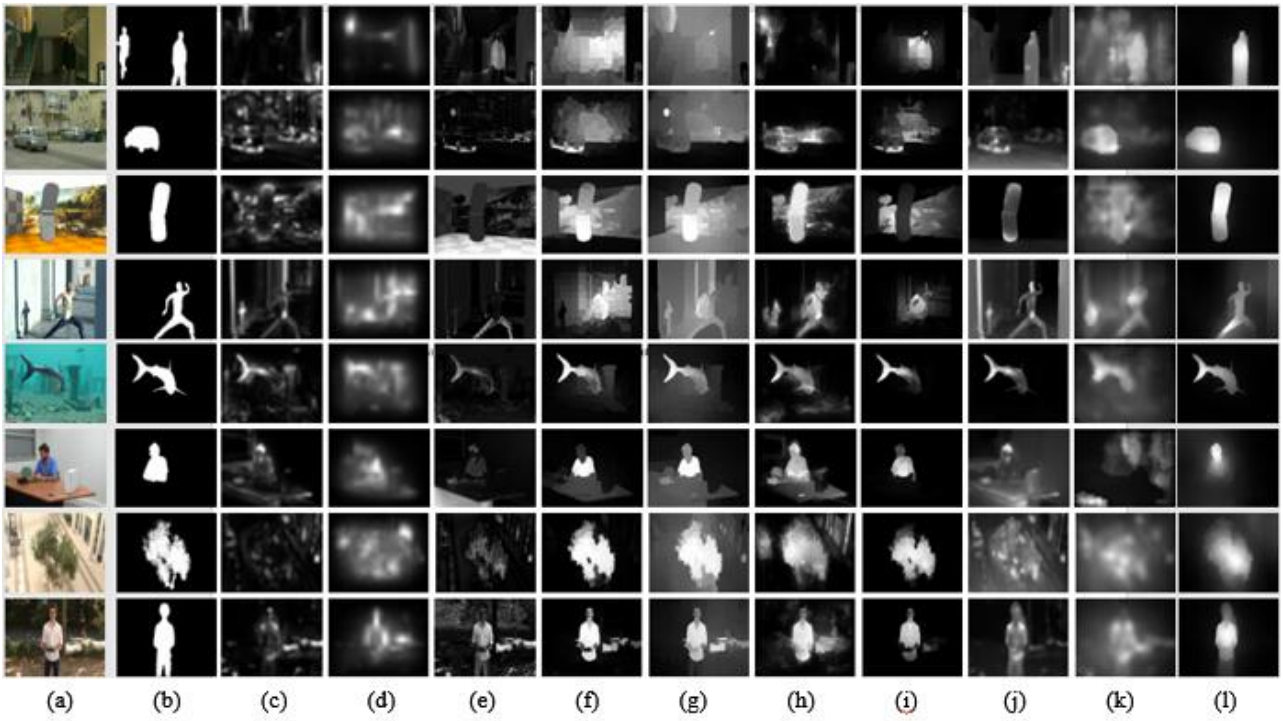


Fig. 3. Example of saliency results of differen methods. (a) Original frame. (b) Ground Truth. (c)-(k) Results of ITTI [1], GBVS [17], FT [18], MR [7], SPL [19], WANG2DV [20], RGBD [12], ZHANG3DV [8], LINO3DV [13]. (l) Our proposal

First, upon initial saliency maps, we separately use motion, color and depth features to calculate the transition probability $P_{R_j,R_i,F}$:

$$P_{R_j,R_i,F} = \frac{W_{R_i,R_j,F}}{\sum_{R_j \in \Omega_{R_i}} W_{R_i,F}} \tag{10}$$

where $F \in \{C, M, D\}$, and $W_{R_i,R_j,F} = \exp(-\frac{d_F(R_j,R_i)}{\sigma})$. We can see that the more different the two superpixels feature is, the smaller the transition probability would be.

Based on the transition probability, we draw the lines among the superpixels' barycenters and use the thickness of the lines to represent the value of transition probability and if the feature contrast of these two superpixels are small, the line would be thick, vice versa (Fig. 2(b)). In Fig. 2(b), the rough outline of the objects could be presented by the thin lines, which reflect a big difference between the objects and the background, also that is the places where the line just cross the objects' edge.

Second, we use the transition probability as the weight to process the initial saliency maps:

$$U_F = \sum_{R_j,R_i \in \Omega} P_{R_j,R_i,F} \cdot S \tag{11}$$

From this equation, we can gain the refined saliency map by each feature prior, for example U_M represents the refined saliency map by motion prior.

Third, the three refined saliency maps by three features prior would be fused by a designed method:

$$U = \sum_{F \in \{C,M,D\}} U_F + \prod_{F \in \{C,M,D\}} U_F \tag{12}$$

The proposed fusion method yields saliency map with well-preserved object's shape and edge, as shown in Fig. 2(c).

Similarly, saliency maps for the other two scales are calculated and fused using a guided filter, yielding the final result.

3. EXPERIMENTS AND EVALUATION

3.1. Datasets and settings

Two public datasets, 3D sequence supplied by 3D-HEVC [21] and NAMA3DS1 [22], are used to evaluate the proposed method. 3D-HEVC provided totally 8 groups of 3D sequences with both appearance and depth, and the resolution is 1280×720 by 'yuv' form. NAMA3DS1 dataset includes ten 3D full HD stereoscopic sequences with 25 frames per seconds and contain the depth images generated by disparity estimation algorithm. Since each dataset contains plenty of depth information and enough sequence to calculate motion, we choose ten groups of sequence in each datasets to test our proposal. The ground truth is made manually by Adobe Photoshop.

3.2. Evaluation Method

In this section, two steps are adapted to valid our experiment. The first step is to vindicate that every part of our proposal is rational and necessary. The second step is to compare our method with other state-of-the-art methods.

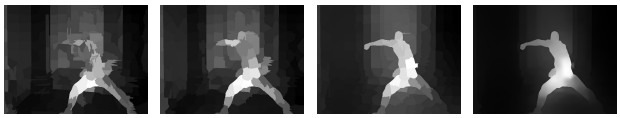


Fig. 4. the images from left to right are:initial saliency map without depth, initial saliency map with depth, refining based on graphical model, result of multi-scale approach

From Fig. 4, by comparing the result of each steps of our approach, we can see the essential of our proposal. As the first two figures shown in Fig. 4, adding depth could help to improve the performance of the initial saliency maps. After refinement, the quality of the saliency maps improved greatly. The multi-scale approach further improves the saliency result in the last subfigure as shown in Fig. 4.

In next experiments, we compare our proposal with other nine the state-of-art salient detection methods: ITTI [1], GBVS [17], FT [18], MR [7], SPL [19], WANG2DV [20], RGBD [12], ZHANG3DV [8] and LINO3DV [13]. In each proposal, the parameters are set as default.

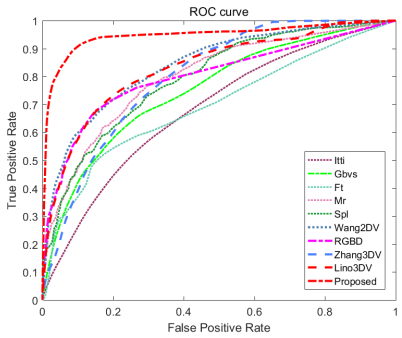


Fig. 5. Evaluation Curves

Results of seven sets are selected (shown in Fig. 3). The first four curves are datasets from 3D-HEVC and the last three are selected from NAMA3DS1. From the results, we can see that our proposal is very good at extracting the saliency object with a well-preserved edge and shape. In the second line, the other proposal failed to extract a whole clear car, while in our method, the car maintains consistent shape and edge and the background is restrained greatly. In complex scenario, the methods in (d), (f) and (i) yield poor results with unrecognizable objects, whereas our method preserves object's shape and edge.

In addition, ROC curves, average Recall, F-measure and accuracy have been measured as well. Fig. 3.2 summarizes the excellent performance of our algorithm over other state-of-the-art algorithms. Table 1 verifies that the proposed algorithm achieves the best performance.

Table 1. Measurement

Models	Precision	Recall	F-measure
ITTI	0.2549	0.2992	0.2464
GBVS	0.3255	0.4143	0.3171
FT	0.4123	0.4023	0.3859
MR	0.4255	0.5811	0.4142
SPL	0.5536	0.3726	0.4322
WANG2DV	0.4197	0.6325	0.4231
RGBD	0.5416	0.5378	0.4661
ZHANG3DV	0.4308	0.4422	0.3931
LINO3DV	0.5507	0.4804	0.4813
Proposed	**0.7132**	**0.7668**	**0.6796**

3.3. Discussion

In our experiments, we also find some limitations of the proposed method. Fig. 6 shows that when there is an another big object with a large contrast with surroundings, our method can not segment the object where the scene also includes a foreground salient object. In this scenario, one could argue that the proposed method yields sensible salient map, as the foreground object could also be salient. This example showcases the robustness of the proposed method.

(a) original image (b) ground truth (c) saliency result

Fig. 6. Limitation illustration

4. CONCLUSION

In this paper, we proposed an improved SLIC segmentation method and a multiscale architecture. Additionally, refining method based on graphical model is adopted to improve the object's shape and edge in the saliency maps. The experiments on 3D-HEVC and NAMA3DS1 dataset verify the excellent performance of our proposal. These datasets cover a wide variety of objects with background. The saliency maps using the proposed method consist of objects with clear shape and edge, demonstrating the generality and robustness of the proposed algorithm.

Future work includes developing an improved multi-scale algorithm with optimal number of superpixels and the corresponding weights.

5. REFERENCES

[1] 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*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[2] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–6.

[3] D. Gao, V Mahadevan, and N Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008.

[4] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, March 2015.

[5] F. Perazzi, P. Krhenbhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 733–740.

[6] Bernhard Schlkopf, John Platt, and Thomas Hofmann, *Graph-Based Visual Saliency*, pp. 545–552, MIT Press, 2007.

[7] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3166–3173.

[8] Yun Zhang, Gangyi Jiang, Mei Yu, and Ken Chen, "Stereoscopic visual attention model for 3d video," in *International Conference on Advances in Multimedia Modeling*, 2010, pp. 314–324.

[9] Christel Chamaret, Sylvain Godeffroy, Patrick Lopez, and Olivier Le Meur, "Adaptive 3d rendering based on region-of-interest," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 7524, 2010.

[10] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 454–461.

[11] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2625–2636, June 2014.

[12] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, "Rgbd salient object detection: A benchmark and algorithms," *National Laboratory of Pattern Recognition*, vol. 8691, pp. 92–109, 2014.

[13] L. Ferreira, L. A. da Silva Cruz, and P. Assuncao, "A method to compute saliency regions in 3d video based on fusion of feature maps," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, June 2015, pp. 1–6.

[14] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[15] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l 1 optical flow," *Pattern Recognition*, pp. 214–223, 2007.

[16] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo, "Tv-l1 optical flow estimation," *Image Processing On Line*, vol. 2013, pp. 137–150, 2013.

[17] Bernhard Scholkopf, John Platt, and Thomas Hofmann, *Graph-Based Visual Saliency*, pp. 545–552, MIT Press, 2007.

[18] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1597–1604.

[19] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 637–640, July 2013.

[20] Wenguan Wang, Jianbing Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3395–3402.

[21] K. Mller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. H. Rhee, G. Tech, M. Winken, and T. Wiegand, "3d high-efficiency video coding for multi-view video and depth data," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3366–3378, Sept 2013.

[22] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutirrez, and N. Garca, "Nama3ds1-cospad1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, July 2012, pp. 109–114.