Deep Learning Features Inspired Saliency Detection of 3D Images

Qiudan Zhang[1,2], Xu Wang[1,2]✉, Jianmin Jiang[1,2], Lin Ma[3]

qiudzhang@gmail.com,{wangxu,jianmin.jiang}@szu.edu.cn, forest.linma@gmail.com

1. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
2. Research Institute for Future Media Computing, Shenzhen University, Shenzhen 518060, China
3. Huawei Noah's Ark Lab, Shatin, Hong Kong

ABSTRACT. Saliency detection of 3D images is important for many 3D applications, such as bit allocation in 3D video coding, spatial pooling in stereoscopic image quality assessment and feature extraction in 3D object retrieval. However, traditional saliency detection approaches only target for the 2D images. Meanwhile, the traditional hand-crafted low-level feature extraction process may be not suitable for the 3D images. In this paper, we propose a deep learning feature based 3D visual saliency detection model. The pre-trained CNN model is employed to extract the feature vectors for both color and depth images after multi-level image segmentation. Then, we train a neutral network based classifier to generate the color and depth saliency maps from the feature vectors. Final, the linear fusion method is adopted to obtain the final saliency map for 3D image. Experimental results demonstrate that our proposed model can achieve appealing performance improvement over two public benchmark datasets.

Keywords: Index Terms— 3D image • Visual Saliency • Conventional Neural Network • Deep Learning Feature and Depth Saliency

1 Introduction

With the development of consumer electronic industry, three dimensional (3D) applications become more and more popular in our daily life. Comparing to the traditional 2D viewing experience, 3D applications can offer users with depth perception and immersive viewing experience. However, there are still many open issues that need to be well addressed in 3D processing chain. Saliency detection of stereoscopic images is one of the most fundamental problems in 3D research, which aims to find regions of interest that standout from their neighbors in natural images. It can be applied to optimize the bit allocation in 3D video coding [1], spatial pooling in stereoscopic image quality assessment [2, 3] and compressed domain for image and video [4,5].

adfa, p. 1, 2011.

Existing visual saliency detection methods are mostly related to 2D image. These models estimate saliency from color images via hand-crafted low-level features (such as luminance, color, contrast and texture) [6, 7, 8], which did not exploited the depth cues. Thus, traditional 2D saliency detection models are unable to accurately predict where people look in a 3D scene. To improve the prediction accuracy, some researcher modeled the visual saliency of stereoscopic images by considering the depth information. For example, Fang et al. designed a framework that applied the feature contrast of color, luminance, texture, and depth to estimate saliency of a stereoscopic image [9], in which adopted a traditional hand-crafted method to extract low-level features and depth features in computing the stereoscopic saliency. Qi et al proposed a 3D visual saliency detection model with generated disparity by using low-level features and depth features [10]. For these approaches, the hand-crafted feature extraction stages cannot effectively and accurately extract feature hierarchically from raw pixel [11]. Thus, the performances are limited.

In this paper, we propose a visual saliency detection model for 3D image based on deep learning features. The contributions of our proposed model can be summarized as follows.

- For each stereoscopic image pair, the features map of depth image and color image are extracted by a pre-trained convolution neural network (CNN), respectively.
- The saliency map of a depth image (or color image) is estimated by a fine-tuned deep neural network which is used to infer the saliency value of every image region from the deep learning features maps.
- The final visual saliency map of 3D image is obtained by a linear fusion approach which combines the color saliency map and depth saliency map.

The remaining of this paper is organized as follows. Section II surveys the related work in the literature. Section III describes the proposed computation model in detail. Section IV shows the experiment results. Final, Section V concludes this paper.

2 Background and Motivation

Classical visual saliency model can be classified into two categories, one is bottom-up methods and the other is top-down methods. Bottom-up approaches are low-level features driven which aims to distinguish regions of interest stand out from their background. For example, Itti et al. proposed a model of saliency for rapid scene analysis combined the multi-scale image features into a saliency map [6]. Bruce et al. introduced a saliency measure based on information maximization, in which Shannon's self-information method was applied to the saliency operation [12]. Hou et al. designed a visual saliency detection model based on spectral residual which constructed the saliency map via the log-spectrum of an image [13].

Top-down approaches are specific-task driven to process information of image. For example, Goferman et al. designed an especially context-aware saliency detection model has been proposed to detect the image regions which can present the scene [14]. Yang et al. proposed a visual saliency model based on top–down approach via joint a Conditional Random Field and a discriminative dictionary [15]. Kanan et al. introduced a top-down saliency based on natural image statistics, in which many forms of top-down knowledge was incorporated into saliency detection model [16].

Currently, to further improve the prediction accuracy, combining the bottom-up and top-down approaches together becomes the hot research topic. Itti et al. proposed an overt and covert shifts of visual attention model based on a saliency search mechanism, in which focused on how to combine the information including orientation, intensity, and color information in a specific visual attention task [17]. Cheng et al. put forward a salient object detection algorithm based on a regional contrast. Meanwhile the global contrast and spatial coherence were both applied to detect the salient object [18]. Zhao et al. designed multi-context deep learning framework for salient object detection [19], in which the global context and local context are both considered into saliency detection.

Apart from 2D visual saliency models, a few studies investigated the computation model of 3D visual saliency. Wang et al. added a depth to visual dimension, and utilized the 2D visual features to detect the salient areas in a computation model of 3D visual saliency [20]. The depth saliency map is generated by a Bayesian approach. Fang et al. obtained the feature contrast based saliency map by measuring the spatial distance between image patches [9]. Qi et al. proposed a band-pass filtering based 3D visual saliency detection model [10]. Kim et al. described a saliency prediction model on stereoscopic videos which accounts for diverse low-level features, depth attributes and high-level classifications of scenes [21].

Based on above-mentioned discussions, the performance of computation visual saliency map is significantly influenced by the visual feature representation. Thus, finding the representative visual features is quite important to the 3D visual saliency research. Existing saliency detection model of stereoscopic images are based on handcrafted features [9, 10, 20]. However, these approaches are difficult to achieve high degree of distinction among saliency region and their neighbors. Besides, due to the lack of knowledge on 3D visual perception, how depth information can contribute to the final visual saliency is still not clear.

Deep convolution neural network (CNN) has already been widely applied in hierarchical feature learning and extraction [22], and achieves significantly success on performance improvement of visual saliency models for 2D images [11]. However, existing works mainly focus on the feature learning for color image but not for depth image. Thus, it is necessary to learn the depth feature representative through the CNN tools.

3 Proposed Visual Saliency Model for 3D Images

The framework of proposed deep learning features based visual saliency model contains three basic stages, including deep features extraction, saliency map generation and saliency map fusion as shown in Fig. 1. First, the deep feature vectors for the color and depth image are extracted by employing the CNN model, respectively. Then, the saliency maps for color and depth image are generated from the feature vector through a three-layer neural network. Final, the saliency map of 3D image is fused by the saliency maps of color and depth image.

3.1 Deep Learning Features Extraction

Based the theory of human visual system, the visual attention mechanism contains a hierarchical selection process from the coarsest to finest [23]. Thus, the multi-level image segmentation is employed before feature extraction. Then, the feature extraction is implemented for each region with in the same level. Detailed descriptions are provided as follows.

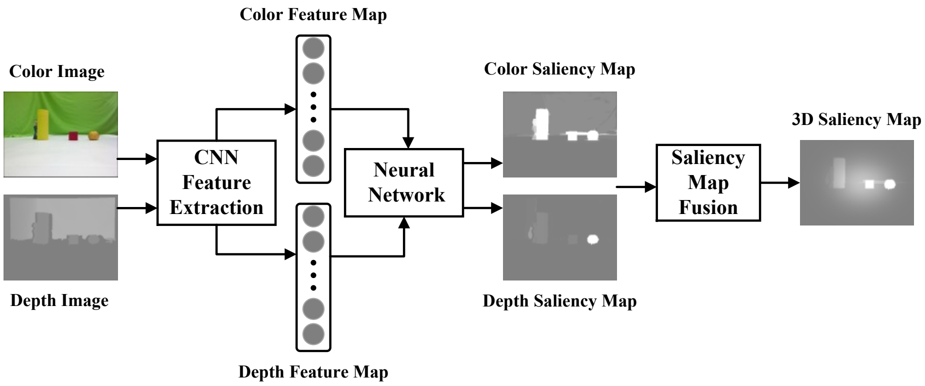

Fig. 1. Framework of proposed visual saliency detection model of 3D image

A. Multi- level Image Segmentation
In this paper, we focus on the depth based 3D image format where each color image is associated with a depth image. For each 3D image, the color image and its associated depth image are decomposed into multi-level non-overlapping regions, respectively. For convenience, we assume the total number of level is *L*. For each level *j*, the non-overlapping region sets of color image I_c and depth image I_d are denoted as $A_c^j=\{R_c^{j,i} \mid \mathrm{i}=1,...,\mathrm{m}_j\}$ and $A_d^j=\{R_d^{j,i} \mid \mathrm{i}=1,...,\mathrm{n}_j\}$, respectively. Detailed description of multi- level image segmentation algorithm can be referred in [23].

B. CNN based Feature Extraction
Due to the lack of depth image capturing technology, the amount of ground truth data public available in the 3D saliency detection domain is not very large. Thus, it is hard to training an accuracy CNN model rely on these available datasets. In this stage, the pre-trained AlexNet model is employed to extract the features for both color and depth images. The AlexNet model is trained over the ImageNet dataset [24], which has five convolutional layers and three full-connected layers.

As we know, the saliency of each local region is not only relying on its own characteristics, but also influenced by the content of its neighborhood and background (the rest part of the region). Thus, for each level j of depth image, we extract the CNN feature vectors of local region $R_d^{j,\mathrm{k}}$, its neighborhood region $H_d^{j,\mathrm{k}}$ and background region $B_d^{j,\mathrm{k}}$. Detailed description of region segmentation and bounding box determination can be referred in [11, 23]. Each region is resized into a 227x227 square and fed into the CNN model. The output for each region is with a 12288-dimensional feature vector denoted as $f_d^{j,\mathrm{k}}$. For color image, the feature vector of region $R_c^{j,\mathrm{k}}$ of color image is denoted as $f_c^{j,\mathrm{s}}$.

3.2 Saliency Map Generation

The output feature vector is just a sparse representation of the local region. To determine whether the current region is salient or not, we need to build a mapping function from the feature vector to the saliency label. In this stage, we trained a neural network (NN) with two fully connected hidden layers. The feature vector is the input and the output is the saliency labels of current region. The NN models are trained for color image and depth image, respectively. The mapping function between the saliency label and the feature vector for color and depth image is denoted as $P_c(f_c^{j,\mathrm{i}})$ and $P_d(f_d^{j,\mathrm{i}})$, respectively. All the pixels belong to the same region share the same saliency labels. Final, the saliency maps of depth image is generated as

$$S_d(\mathrm{x}) = \sum_{j=1,\mathrm{x}\in\mathrm{R}_d^{j,i}}^{L} w_d^j P_d(f_d^{j,\mathrm{i}}) \,. \tag{1}$$

The saliency maps of color image is generated as

$$S_c(\mathrm{x}) = \sum_{j=1,\mathrm{x}\in\mathrm{R}_c^{j,i}}^{L} w_c^j P_c(f_c^{j,\mathrm{i}}) \,. \tag{2}$$

x is the pixel belongs to the region $f_d^{j,\mathrm{i}}$ of depth image and $f_c^{j,\mathrm{i}}$ of color image. w_d^j and w_c^j are the weighting factors of depth image and color image, respectively.

3.3 Saliency Map Fusion and Enhancement

Fusing depth saliency map and color saliency map is important for obtaining a accuracy saliency map. After obtaining the saliency maps S_d and S_c, the saliency map of 3D image is generated by a linear fusion approach. The formula is provided as follows.

$$S = w \cdot S_c + (1-\mathrm{w}) \cdot S_d \,, \tag{3}$$

where w is the parameter to adjust the two components. To further improve the performance, the widely used centre-bias mechanism is also employed to enhance the final saliency map.

4 Experimental Results

4.1 Simulation Setup

To make a fair performance comparison, the computational models including Li's multi-scale features based model [11] (denoted as VSMD), İmamoğlu's wavelet domain based model [25] (denoted as SDLL), Fang's 2D saliency model [9] (denoted as SSDF2D) and 3D saliency model [9] (denoted as SSDF3D) are implemented as benchmark. The VSMD, SDLL and SSDF2D are computation models targeted for 2D images. The SSDF3D model is targeted for 3D images. In our experiment, the number of segmentation level N of our proposed model is set as 15. The parameter w is set as 0.5.

To demonstrate the performance of our proposed 3D visual saliency detection model, we evaluate all the saliency detection models with various datasets. Currently, there are few public available eye-tracking datasets for 3D saliency research. In this paper, we evaluate the saliency detection methods on the two representative datasets, name as NUS3D-Saliency [26] and NCTU-3DFixation [27]. Detailed descriptions of these datasets are provided as follows.

- NUS3D-Saliency dataset (denoted as NUS) has 600 images collected from 80 participants. It involves several 2D and 3D scenes. This dataset provides color stimuli, depth maps, smooth depth maps, 2D and 3D fixation maps.
- NCTU-3DFixation dataset (denoted as NCTU) is consisting of 475 3D images as well as their depth maps. The contents of this dataset are various scenes and mainly came from the existing 3D movies or videos.

4.2 Performance Comparison

To evaluate the performance of the 3D visual saliency detection models, the following three criterions [28] are employed in our experiments for fair comparison.

- Pearson Correlation Coefficient (CC) measures the strength of a linear association between two variables. And it can be used to measure the linear correlation coefficient between the saliency map of an image and the eye-fixation map of the image.
- Earth Mover's Distance (EMD) measures the distance between two probability distributions over a region. Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.
- Similarity score (SIM) measures the similarity of the two distributions. A high similarity score indicates the distributions of two maps are quite similar.

These metrics can evaluate the performance of the proposed model. A good saliency model would have a high CC and SIM score, but a low EMD score. Detailed experimental results are provided in Table I. Our proposed model can achieve better performance than all the 2D saliency models for both the NCTU and NUS datasets as shown in Table 1. For example, the CC, EMD and SIM scores of our proposed 3D model are 0.5225, 2.1547 and 0.4985, respectively. The CC, EMD and SIM scores of VSMD are only 0.3783, 2.8419 and 0.3812. The experimental results show that the performance of saliency prediction model can benefit from fusing color saliency map with depth saliency map.

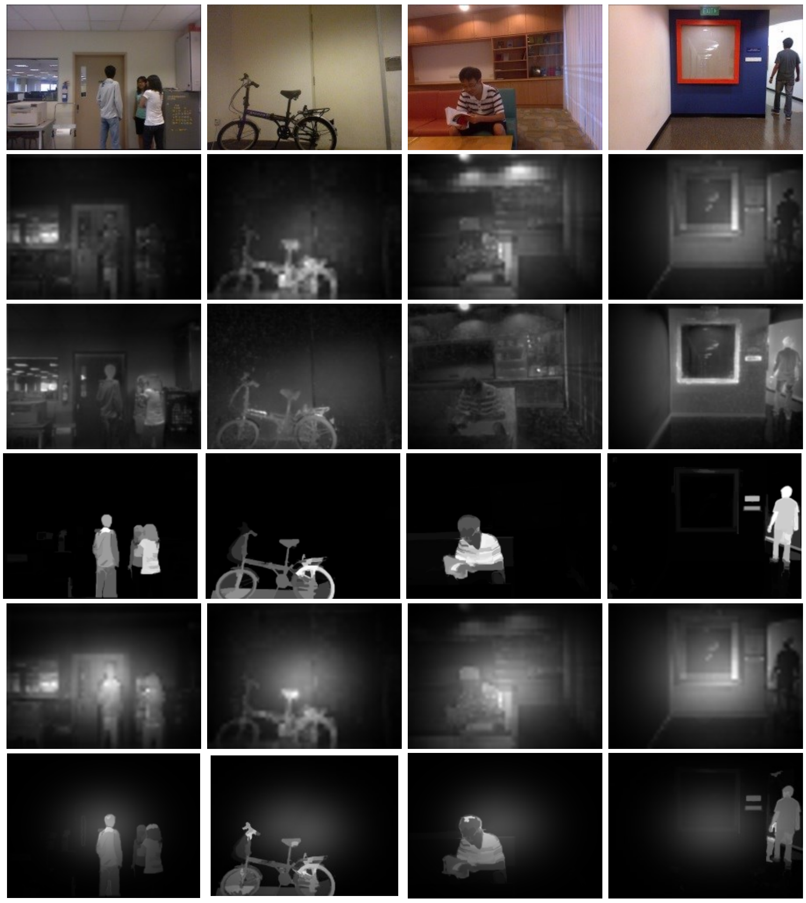

Fig. 2. From left to right are four images from NUS dataset. From the second row to the last row, the order of the model is SSDF2D model, SDLL model, VSMD model, SSDF3D model, and our proposed model.

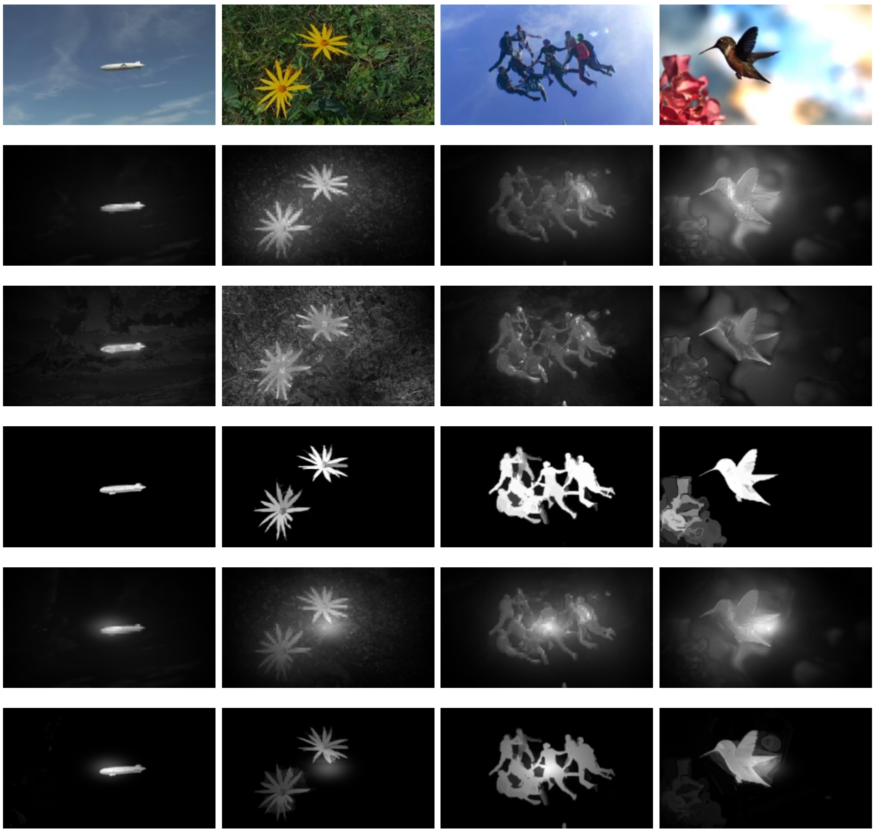

Fig. 3. From left to right are four images from NCTU dataset. From the second row to the last row, the order of the model is SSDF2D model, SDLL model, VSMD model, SSDF3D model, and our proposed model.

Table 1. Performance of the models for the two datasets in terms of CC, EMD and SIM.

Model	NUS			NCTU		
	CC	**EMD**	**SIM**	**CC**	**EMD**	**SIM**
VSMD	0.3783	2.8419	0.3812	0.3121	9.106	0.3198
SDLL	0.3397	3.1359	0.3893	0.2678	11.1806	0.321
SSDF2D	0.4430	2.6513	0.4321	0.4531	9.184	0.3951
SSDF3D	0.5033	2.3965	0.4589	**0.5601**	8.3874	0.4356
Proposed	**0.5225**	**2.1547**	**0.4985**	0.4761	**7.2313**	**0.4445**

The performance comparison between our proposed 3D model and the SSDF3D model on both the NCTU and NUS datasets are also provide in Table 1. It is observed that the CC, EMD, SIM scores of our proposed 3D model is better than those of the SSDF3D model on the NUS dataset. For the NCTU dataset, the CC score of proposed

3D model is less than the SSDF3D model, but the EMD, SIM scores of proposed 3D model are larger than those of SSDF3D model. Experiment results demonstrate that our proposed 3D model have a significantly performance improvement on 3D visual saliency detection. For further illustration, some 3D saliency detection examples among the models are shown in Fig. 2 and Fig. 3, where it can be seen that our proposed model can achieve the best performance.

5 Conclusion

In this paper, we propose a computation model of 3D image visual saliency based on deep learning features. There are three key factors in our approach. First, we extract deep learning features of color and depth image using multi-scale regions by a CNN model. Second, the saliency map from depth image (or color image) is generated based on deep feature vectors and the saliency labels of regions by a NN model which plays the role as a classifier. Final, we adopt a linear fusion method to combine the color and depth saliency map to generate the final 3D image saliency. The centre bias mechanism is also implemented to enhance the saliency map. The proposed model can achieve the remarkable performance on two public available datasets.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grants 61501299 and 61373103, in part by the Guangdong Nature Science Foundation under Grant 2016A030310058, in part by the Shenzhen Emerging Industries of the Strategic Basic Research Project under Grants JCYJ20150525092941043 and JCYJ20160226191842793, in part by the Project 2016049 supported by SZU R/D Fund.

References

1. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on Image Processing 19(1), pp. 185-198 (2010)
2. Ma, L., Lin, W., Deng, C., Ngan, K.N.: Image retargeting quality assessment: A study of subjective scores and objective metrics. IEEE Journal of Selected Topics in Signal Processing 6(6), pp. 626-639 (2012)
3. Ma, L., Li, S., Zhang, F., Ngan, K.N.: Reduced-reference image quality assessment using reorganized DCT-based image representation. IEEE Transactions on Multimedia 13(4), pp. 824-829 (2011)
4. Fang, Y.M., Lin, W.S., Chen, Z.Z., Tsai, C.M., Lin, C.W.: A video saliency detection model in compressed domain. IEEE Transactions on Circuits and Systems for Video Technology 24(1) . pp. 27-38 (2014)
5. Fang, Y.M., Chen, Z.Z., Lin, W.S., Lin, C.W.: Saliency Detection in the Compressed Domain for Adaptive Image Retargeting. IEEE Transactions on Image Processing 21(9), pp. 3888-3901 (2012)

6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), pp. 1254-1259 (1998)
7. Song, X., Zhang, J., Han, Y., Jiang, J.: Semi-supervised feature selection via hierarchical regression for web image classification. Multimedia Systems 22(1), pp.41-49 (2016)
8. Zhang, J., Han, Y., Jiang, J.: Tensor rank selection for multimedia analysis. Journal of Visual Communication and Image Representation 30, pp. 376-392 (2015)
9. Fang, Y., Wang, J., Narwaria, M., Callet, P.L., Lin, W.: Saliency detection for stereoscopic images. IEEE Transactions on Image Processing 23(6), pp. 2625-2636 (2014)
10. Qi, F., Zhao, D., Liu, S., Fan, X.: 3D visual saliency detection model with generated disparity map. Multimedia Tools and Applications. DOI: 10.1007/s11042-015-3229-6 (2016)
11. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5455-5463 (2015)
12. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Advances in neural information processing systems. pp. 155-162 (2005)
13. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8 (2007)
14. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(10), pp. 1915-1926 (2012)
15. Yang, J., Yang, M.H.: Top-down visual saliency via joint CRF and dictionary learning. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2296-2303 (2012)
16. Kanan, C., Tong, M.H., Zhang, L., Cottrell, G.W.: Sun: Top-down saliency using natural statistics. Visual Cognition 17(6-7), pp. 979-1003 (2009)
17. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research 40(10), pp. 1489-1506 (2000)
18. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3), pp. 569-582 (2015)
19. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1265-1274 (2015)
20. Wang, J., DaSilva, M.P., Callet, P.L., Ricordel, V.: Computational model of stereoscopic 3D visual saliency. IEEE Transactions on Image Processing 22(6), pp. 2151-2165 (2013)
21. Kim, H., Lee, S., Bovik, A.C.: Saliency prediction on stereoscopic videos. IEEE Transactions on Image Processing 23(4), pp. 1476-1490 (2014)
22. Song, H.A., Lee, S.Y.: Hierarchical representation using NMF. In: Neural Information Processing. pp. 466-473 Springer (2013)
23. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2083-2090 (2013)
24. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F..: Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255 (2009)
25. İmamoğlu, N., Lin, W., Fang, Y.: A saliency detection model using low-level features based on wavelet transform. IEEE Transactions on Multimedia 15(1), pp. 96-105 (2013)
26. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Influence of depth cues on visual saliency. In: European Conference on Computer Vision (ECCV). pp. 101-115. Springer (2012)

27. Ma, C.Y., Hang, H.M.: Learning-based saliency model with depth information. Journal of Vision 15(6), pp. 19-19 (2015)
28. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. Massachusetts Inst. Technol., MA, USA, Computer Science and Artificial Intelligence Lab (CSAIL), Tech. Rep. MIT-CSAIL-TR-2012-001 (2012)