Depth Enhanced Saliency Detection Method

Yupeng Cheng[1], Huazhu Fu[2], Xingxing Wei[1], Jiangjian Xiao[4], Xiaochun Cao[1,3]✉

[1] School of Computer Science and Technology, Tianjin University, China
[2] School of Computer Engineering, Nanyang Technological University
[3] State Key Laboratory of Information Security, Chinese Academy of Sciences
[4] Ningbo Industrial Technology Research Institute, Chinese Academy of Sciences
chengyupeng2008@hotmail.com, caoxiaochun@iie.ac.cn

ABSTRACT

Human vision system understands the environment from 3D perception. However, most existing saliency detection algorithms detect the salient foreground based on 2D image information. In this paper, we propose a saliency detection method using the additional depth information. In our method, saliency cues are provided to follow the laws of the visually salient stimuli in both color and depth spaces. Simultaneously, the 'center bias' is also extended to 'spatial' bias to represent the nature advantage in 3D image. In addition, We build a dataset to test our method and the experiments demonstrate that the depth information is useful for extracting the salient object from the complex scenes.

Categories and Subject Descriptors

I.4.6 [**Image Processing And Computer Vision**]: Segmentation—*Region growing, partitioning*; I.4.9 [**Image Processing And Computer Vision**]: application

General Terms

Theory

Keywords

depth map, saliency detection, RGB-D image

1. INTRODUCTION

Visual saliency detection is employed to extract the conspicuous or meaningful objects, which stand out from their surrounding and catch a viewer's attention in the image [5, 6, 2]. Most existing methods focus on 2D image information, such as color, texture, and edge [3, 8, 12]. They estimate saliency of regions and capture eye fixation data in 2D space, which might result in inaccurate saliency detection in some images. However, human visual system understands the environment from 3D perception. The depth information provides not only the additional space attention rule, but also the discriminative power against the complex background.

Depth information has been demonstrated useful on many applications [11, 7, 10].Nonetheless, its researches on saliency detection are not yet refined. Lang et al. [6] propose a model to calculate probability density of saliency using depth. However, one main difference is that their result is visual attention rather than salient object for their lack of the concept of segmentation. Niu et al. [9] offer a method by computing stereo saliency based on the global disparity contrast. Although they have many similarities, disparity map differs from depth map in span. This kind of difference leads to the loss of depth-of-field information, which influences the weights of depth [6]. Moreover, both of their works [9, 6] focus only on the depth (or stereo) rather than the entire 3D saliency. Thus, in this paper, we introduce a Depth Enhanced Saliency detection method (DES), which combines the color and depth information (i.e. RGB-D image). Our work has the following contributions: Firstly, the saliency map produces an object level segmentation combining color and depth information, which is more proximate to the real human visual system and has more robustness to complex background. Secondly, extending by 'center bias', we propose a 'spatial bias' which is testified powerful in 3D saliency detection. At last, to demonstrate the effectiveness of our algorithm, we build a RGB-D saliency dataset and compare our DES and other state-of-the-art methods.

2. OUR METHOD

In our DES, we start by pixel clustering, and then measure each cluster's salient value using three saliency cues: color contrast, depth contrast and spatial bias. The final saliency map is obtained by combing these saliency cues.

Inputting a RGB-D image, we group the pixels of the image into K clusters based on color, depth and spatial information. In our method, we are not constrained to specific choice of the clustering methods, and herein K-means is used. Since each cluster is a connected region, we use 'region' instead of 'cluster' in the rest of the letter for facilitating the description. The regions are denoted by $\mathcal{R} = \{\mathbf{r}_k\}_{k=1}^{K}$, and each region $\mathbf{r}_k$ is represented by its cluster center in a 6-dimensional vector: RGB color (3D), 3D spatial information (image coordinate and depth).

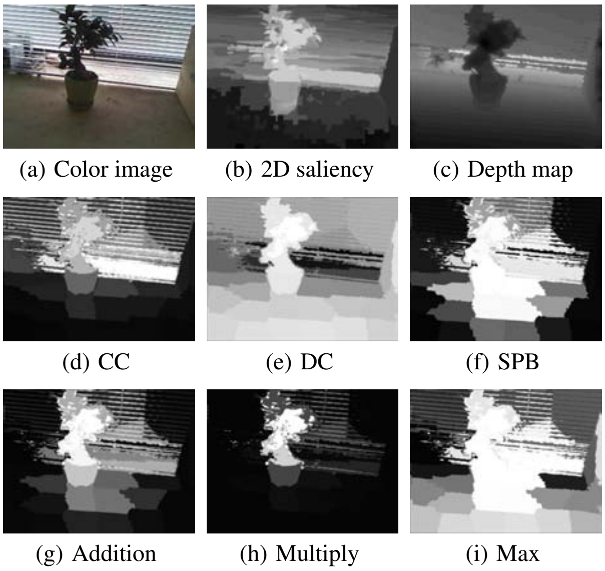
2.1 Feature contrast

Contrast cue is widely used in measuring saliency since the contrast operator simulates the human visual receptive fields [3, 8]. This rule is also valid for saliency detection in RGB-D images. We define the feature contrast cue of region $\mathbf{r}_k$ as:

$$W_f(\mathbf{r}_k) = \sum_{i=1, i\neq k}^{K} \frac{n_i}{N} \omega(\mathbf{r}_k, \mathbf{r}_i) F(\mathbf{r}_k, \mathbf{r}_i), \qquad (1)$$

where n_i is the number of pixels in the region $\mathbf{r}_i$, N denotes the pixel number of the entire image, $F(\mathbf{r}_k, \mathbf{r}_i)$ is the distance between

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICIMCS '14, July 10-12, Xiamen, Fujian, China.
Copyright 2014 ACM 978-1-4503-2810-4/14/07 ...$15.00.


(a) Color image (b) 2D saliency (c) Depth map

(d) CC (e) DC (f) SPB

(g) Addition (h) Multiply (i) Max

Figure 1: Examples of saliency detection. (a) Input RGB image. (b) Most saliency detection methods depend on the significance of an object in a specific field. When the object contrast does not stand out, the detected result may be unsatisfactory. (c) Depth map. (d) Color contrast cue highlights the most salient object in color space. (e) Depth contrast cue extracts the object popping out of its surroundings. (f) Spatial bias cue depresses the far-range background in the edge. Final result: Fusion by using multiplication (g), addition (h), and max (i) operations.

regions $\mathbf{r}_k$ and $\mathbf{r}_i$ in feature space, and $\omega(\mathbf{r}_k, \mathbf{r}_i)$ acts as a spatial weighting term:

$$\omega(\mathbf{r}_k, \mathbf{r}_i) = \exp(-d(\mathbf{r}_k, \mathbf{r}_i)/\sigma^2), \tag{2}$$

where $d(\mathbf{r}_k, \mathbf{r}_i)$ represents the spatial distance between regions $\mathbf{r}_i$ and $\mathbf{r}_k$, and σ^2 controls the strength of spatial weighting. In our method, color and depth features are considered separately:

2.1.1 Color contrast

Inspired by [3], the color of an object contrasts strongly with the background typically draw more attention. We compute the color contrast cue (CC) of each region $\mathbf{r}_k$ and rename $W_f(\mathbf{r}_k)$ in Eq. (1) as $W_{cc}(\mathbf{r}_k)$. Moreover, term $F(\mathbf{r}_k, \mathbf{r}_i)$ here is the Euclidean distance $F_{cc}(\mathbf{r}_k, \mathbf{r}_i)$ in color space:

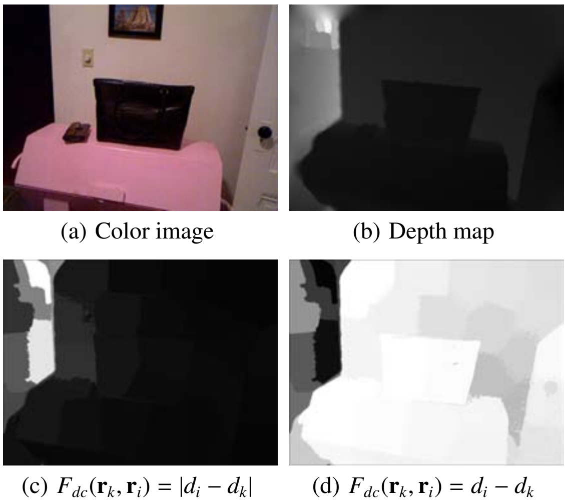
$$F_{cc}(\mathbf{r}_k, \mathbf{r}_i) = ||\mathbf{c}_k - \mathbf{c}_i||_2, \tag{3}$$

where $\mathbf{c}_i$ is the color center of region $\mathbf{r}_i$. To distinguish with the next depth contrast, here we call σ^2 in $\omega(\mathbf{r}_k, \mathbf{r}_i)$ as σ^2_{cc} and assign $\sigma^2_{cc} = 0.4$ with pixel coordinates normalized to [0,1] as suggested in [3]. Color contrast cue is useful to a wide range. However, the validity of color contrast cue often degrades by the complex background, e.g. the window-shades in Fig. 1(d).

2.1.2 Depth contrast

Similar to the color contrast, the unique position in depth space also makes object obvious [9]. To integrate this cue into our method, $W_f(\mathbf{r}_k)$ in Eq. (1) is defined as depth contrast (DC) $W_{dc}(\mathbf{r}_k)$ to all other regions. Therefore, we explain $F(\mathbf{r}_k, \mathbf{r}_i)$ as the depth difference $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$ between regions $\mathbf{r}_k$ and $\mathbf{r}_i$:

$$F_{dc}(\mathbf{r}_k, \mathbf{r}_i) = d_i - d_k, \tag{4}$$


(a) Color image (b) Depth map

(c) $F_{dc}(\mathbf{r}_k, \mathbf{r}_i) = |d_i - d_k|$ (d) $F_{dc}(\mathbf{r}_k, \mathbf{r}_i) = d_i - d_k$

Figure 2: Comparison between the different definition of $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$. (a-b) RGB-D input. (c) The depth contrast cue when $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$ is the absolute difference (also the l_2 normal distance) between $\mathbf{r}_k$ and $\mathbf{r}_i$. (d) The depth contrast cue when $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$ changes into the simple difference value between $\mathbf{r}_k$ and $\mathbf{r}_i$.

where d_i represents the depth center of region $\mathbf{r}_i$. As CC does, σ^2 in Eq. (2) is called as σ^2_{dc} which is discussed in the following experiments.

Note that, $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$ is explained as how region r_k stands out relative to the other region $\mathbf{r}_i$. It is a signed value, and $F_{dc}(\mathbf{r}_k, \mathbf{r}_i) = -F_{dc}(\mathbf{r}_i, \mathbf{r}_k)$. In other words, far-range makes close-range more salient, while the close object reduces significance of the far one. As shown in Fig. 2(c), when $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$ defines as l_2 norm, the shadow on the left side is wrongly highlighted. Considering the natural advantage in depth space, absolute value can only detect the relative depth which would generate wrongly detection in high depth-of-field scene. Therefore, we define $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$ as $(d_i - d_k)$, which appreciates a lower (closer) depth region $\mathbf{r}_k$ to avoid this kind of error. Fig. 2(d) illustrates the differences between two definitions of $F_{dc}(\mathbf{r}_k, \mathbf{r}_i)$. Unlike (c), the shadow on the left side is depressed, and the bag is reasonable highlighted.

2.2 Spatial bias

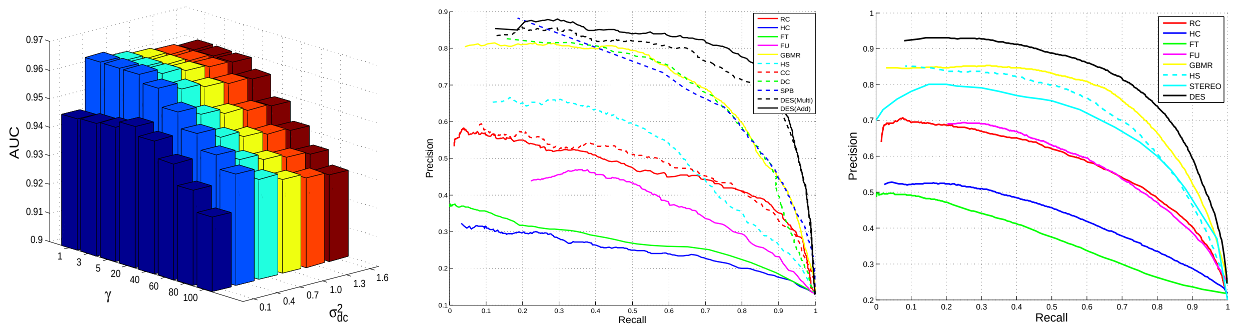
In human 2D visual system, there is a tendency of human subjects to preferentially look near the image center, which called 'center bias'. Similarly, in 3D scene, we extended 2D 'center bias' into 3D 'spatial bias' (SPB) and formulated it as:

$$W_{spb}(\mathbf{r}_k) = \frac{1}{n_k}\mathcal{N}(||p_k - o||_2 \mid 0, \sigma^2)\mathcal{D}(d_k), \tag{5}$$

where n_k, p_k and d_k denote the pixel number, image coordinates and depth value of region $\mathbf{r}_k$ respectively. The first term reflects center bias, where $\mathcal{N}$ is Gaussian kernel which plays a penalty term of Euclidean distance between p_k and the center o of image. The variance σ^2 is the normalized radius of images. This term assigns a maximum saliency value to the center. The second term represents depth bias and we define $\mathcal{D}(\cdot)$ as:

$$\mathcal{D}(d_k) = (\max_{1\leq i\leq K}\{d_i\} - d_k)^{\gamma/DOF}, \tag{6}$$

where DOF, denoted depth-of-field, represents the distance between the nearest region and the farthest one . Its definition is $\max_{1\leq i\leq K}\{d_i\} - \min_{1\leq j\leq K}\{d_j\}$. For one image, DOF is fixed. γ is a fixed



(a) (b) (c)

Figure 3: (a) Performance with respect to parameter settings. Each bar represents the AUC of our algorithm with a pair of parameters (σ^2_{dc} and γ). (b) Precision-recall curves of different saliency detection methods on our benchmark dataset. (c) Precision-recall curves of various saliency detection methods on STEREO [9]'s dataset.

parameter. This term decreases as the region goes far and *DOF* controls its weight in Eq. 5.

2.3 Final saliency map

So far, we have three cues to detect the salient object on RGB-D images. CC cue extracts the object whose color appears less frequently in the picture. However, this cue loses its power when foreground object does not show visible difference with its surroundings in color space. In that case, DC cue, which measures saliency in depth, provides a significant supplement. Moreover, with the addition of prior SPB, our result DES provides a satisfactory result:

$$W_{DES}(\mathbf{r}_k) = W_{cc}(\mathbf{r}_k) \odot W_{dc}(\mathbf{r}_k) \odot W_{spb}(\mathbf{r}_k), \qquad (7)$$

where $\odot$ is an integration scheme(e.g. $+$, $*$, or max). Fig. 1(g-i) illustrates various effects of the three operators. Fig. 1(i) 'Max' is the first to be excluded for it retaining most wrong information from previous steps. 'Addition' and 'Multiply' are both work well. So we plot their Precision-Recall curve (Fig. 3(b)) and select 'Addition' as our final integration scheme.

3. EXPERIMENTS

In our experiments, we provide a new RGB-D saliency detection dataset. We take 135 RGB-D indoor images by Kinect with the resolution 640×480. Then, three users are asked to mark the salient object of each image. We employ the overlapping areas of the manually labelled object as the ground truth. Some examples are shown in Fig. 4(c). Before the comparison with other algorithms, we test a set of parameters, i.e. σ^2_{dc} in DC (Eq. 2) and γ in SPB (Eq. 6), to find how they influence the final performance. Fig. 3(a) shows the result. The AUC reaches a relatively stable value with $\sigma^2_{dc} = 0.4$ and $\gamma \in [1, 5]$. We employ values $\sigma^2_{dc} = 0.4$ and $\gamma = 5$ in the following experiments.

In the experiments, we estimate our the final saliency map (DES) (Eq. (7)), and three additional saliency maps by using the independent cue, i.e. the color contrast cue (CC), the depth contrast cue (DC), and the spatial bias cue (SPB). We also compare our method with six state-of-the-art algorithms, RC [3], HC [3], FT [1], FU [4], HS [13], and GBMR [14]. Fig. 4 shows the results in our RGB-D dataset, where the first two columns are the color image and the depth map. The third column shows the ground truth, which labels the salient foreground. Fig. 4(d-i) show the results of other methods. And all of them detect saliency from only color image. Therefore, they are not particularly effective in some cases. In the first image, for example, the foot of a wall top right can not be distinguished from the basketball by RC. HS missed part of the foreground. The HC, FT methods even tend to select the wall rather than the basketball. FU and GBMS perform better, but they provide undesirable results of the second and third images which are representations of simple but confusing background and complex background respectively. Fig. 4(j-l) show the process on our three cues. Fig. 4(m) is our DES result that combines (j-l). The second image in Fig. 4(j) also shows CC, as a color-depend cue, can not discriminate between the tube and the wall around it. As expected, without additional depth information, the salient object can not be identified. Fig. 4(k) shows that depth contrast cue distinguishes the basketball and tube from the confusing background successfully. And the sofa in the third image, which is highlighted by most methods, is also depressed for its smooth change in depth. Spatial bias cue provides a reliably prior which supply the above cues in some case. For instance, the fan in the second image, which is not well depressed by both color and depth cues, is wiped out by SPB. The final combination results (Fig. 4(l)) indicate DES is indeed capable of filtering out the RGB-D salient foreground, which stands out in any field: color or depth.

To evaluate the performance quantitatively, we plot the precision-recall curves for all the methods (RC [3], HC [3], FT [1], FU [4], HS [13], and GBMR [14], and the proposed DES). We threshold the saliency map by each method with a gradually increasing value t (0-255) and get a binary segmentation of salient region. We intersect this binary map with the ground truth mask to compute the precision-recall curve. As shown in Fig. 3(b), FU, FT and HC have lower performances comparing with other color based methods. The main reason is that they are pixel based algorithms while the ground truth is in object level. Region based method GBMR is a typical representative of 2D algorithm and has a relatively good result. Our DES method has a improvement comparing with 2D methods. To better understand how each cue contributes, we also provide the precision-recall curves of CC, DC, and SPB cues. CC curve in Fig. 3(b) is similar to the performance of RC, which is however lower than DC and SPB curves. Without surprise, DES integrates the advantage of all cues, and excels in the comparison.

To compare with STEREO [9], we employ its dataset, and use the STEREO's result reported in [9]. As the STEREO's dataset does not have depth map, we treat the disparity map as depth map. The left image is regarded as color input in our test. Fig. 3(c) shows precision-recall curves of our method, STEREO and the 2D saliency methods mentioned above. STEREO performs better than most existing competitors, but fail to HS [13] and GBMR [14]. Our

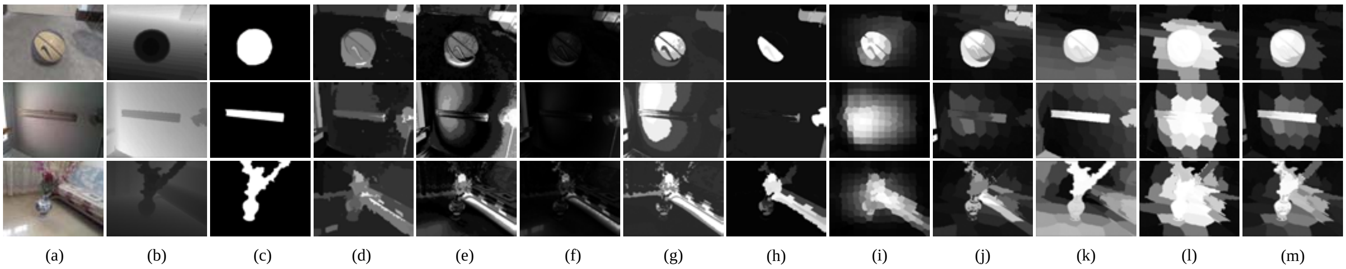
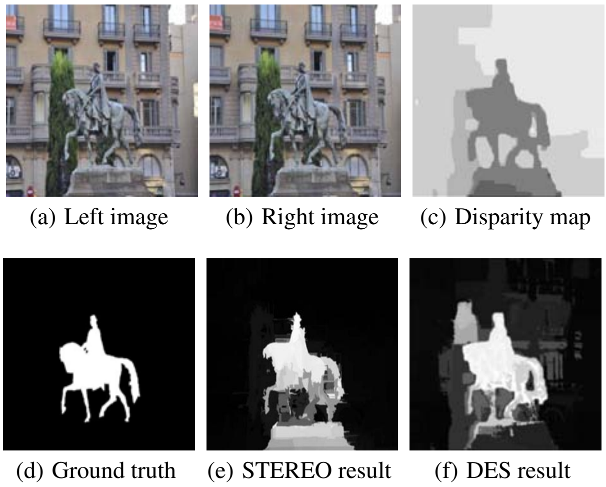


Figure 4: Visual results of saliency detection on our RGB-D dataset. (a) Color image. (b) Depth map. (c) Ground truth. Other saliency maps by RC [3] (d), HC [3] (e), FT [1] (f), FU [4] (g), HS [13] (h), and GBMR [14] (i). (j) Saliency map with color contrast cue (CC). (k) Saliency map with depth contrast cue (DC). (l) Saliency map with spatial bias (SPB). (m) Our final saliency result.

method (DES) is still the outperformer and improves 10% and 25% in precision than STEREO when recall is 0.2 and 0.9 respectively. Fig. 5 illustrates the difference between DES and STEREO. Fig. 5(e) shows the result of STEREO, where STEREO segments the base and sculpture together due to their similar values in disparity map. And in our DES result (f) the base is well depressed with the help of 'color contrast' (CC) and 'spatial bias' (SPB). This example further reflects that color and depth information can collaborate to complete the task of detecting salient foregrounds.


(a) Left image (b) Right image (c) Disparity map

(d) Ground truth (e) STEREO result (f) DES result

Figure 5: Comparison between DES and STEREO [9]. (a-b) are left and right images and (c) is disparity map which is calculated from (a) and (b). Left image and disparity map are regarded as color image and depth map in DES method. (d) is the ground truth coming from STEREO [9]'s dataset. (e) and (f) are the results of DES method and STEREO respectively.

4. CONCLUSIONS

In this paper, we have proposed a simple and effective algorithm to detect the saliency from the RGB-D images. Based on color and depth attention rules, this approach calculated three cues. The experimental results demonstrated that the additional depth information is a useful complement to existing visual saliency analysis.

5. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 61332012, 61272264, 61273276), National High-tech R&D Program of China (2013AA01A601), 100 Talents Programme of The Chinese Academy of Sciences, and the Opening Project of State Key Laboratory of Digital Publishing Technology.

6. REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[2] Z. Chen, J. Yuan, and Y. Tan. Hybrid saliency detection for images. *Signal Processing Letters*, 20(1):95–98, 2013.

[3] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[4] H. Fu, X. Cao, and Z. Tu. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 22(10):3766–3778, 2013.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998.

[6] C. Lang, T. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. *ECCV*, pages 101–115, 2012.

[7] S. Liu, Y. Wang, L. Yuan, J. Bu, P. Tan, and J. Sun. Video stabilization with a depth camera. In *CVPR*, pages 89–95, 2012.

[8] T. Liu, Z. Yuan, J. Sun, N. Z. J. Wang, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *TPAMI*, 33(2):353–367, 2011.

[9] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012.

[10] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304, 2011.

[11] H. Simon and B. Richard. Kinecting the dots: Particle based scene flow from depth sensors. In *ICCV*, pages 2290–2295, 2011.

[12] J. Yan, M. Zhu, H. Liu, and Y. Liu. Visual saliency detection via sparsity pursuit. *Signal Processing Letters*, 17(8):739–742, 2010.

[13] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[14] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.