DEPTH-AWARE SALIENCY DETECTION USING DISCRIMINATIVE SALIENCY FUSION

Hangke Song[1], *Zhi Liu*[1,*], *Huan Du*[1,2] *and Guangling Sun*[1]

[1]School of Communication and Information Engineering, Shanghai University, Shanghai, China
[2]The Third Research Institute of Ministry of Public Security, Shanghai, China

ABSTRACT

In this paper, we propose a multi-stage depth-aware saliency model for salient region detection. We evaluate saliency on different features at low, mid and high levels, by taking account of primary depth and appearance contrasts, different feature weighted factors and location priors, respectively. Unlike most existing depth-aware saliency models that use a linear or experiential fusion formula to combine saliency maps from different features, we calculate saliency of each feature individually at each level and learn a discriminative saliency fusion (DSF) regressor based on random forest to estimate the saliency measures of regions. Both subjective and objective evaluations on two public datasets designed for depth-aware saliency detection demonstrate that the proposed saliency model consistently outperforms the state-of-the-art saliency models on saliency detection performance.

Index Terms—Depth information, multi-level saliency detection, discriminative saliency fusion, random forest.

1. INTRODUCTION

Saliency detection aims to detect the attractive objects to human viewers in an image. Visual attention is important for the understanding of vision and cognition processes, and two mechanisms of visual attention are usually distinguished: bottom-up and top-down. A great number of saliency models for 2D images have been proposed to effectively exploit bottom-up attention and top-down attention since the seminal work [1], which was mainly used for human fixation prediction. Recently more and more saliency models are proposed for salient object detection with the better performance such as [2]-[5].

With the prevalence of stereo cameras, depth cameras and Kinect sensors, a few tentative researches have shown that the depth information could be powerful in addition to color images for saliency analysis. There are two main sources of depth information: the depth map directly captured for a single image, and the disparity map estimated from stereoscopic images. Depth features extracted from the depth/disparity map can be directly used for measuring saliency, and an integration of depth-induced saliency with the saliency estimated from RGB image is a common paradigm of depth-aware saliency models as summarized in [6]. In [7], the anisotropic center-surround difference on the depth map is utilized to measure saliency. In [8], the depth saliency estimated from point cloud data is integrated with the saliency of RGB image using nonlinear regression. In [9], a saliency model for stereoscopic images exploits the global contrast on disparity map and domain knowledge in stereoscopic photograph to generate saliency map. In [10], both depth weighted color contrast and depth contrast are exploited to measure saliency. In [11], the primitive depth and color contrasts are refined by depth-based object probability and region merging for saliency measurement.

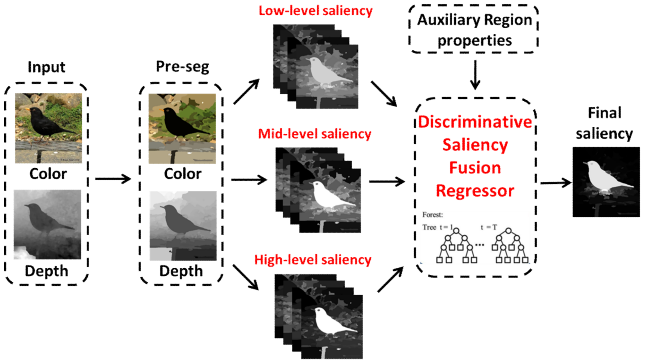

Fig. 1. Illustration of the proposed model.

However, the existing depth-aware saliency models share the following two common limitations. First, they rarely involve top-down information, which can be exploited via machine learning to effectively improve the saliency detection performance. Second, as a critical step, the combination of saliency maps on different features including depth and appearance is usually performed through a simple linear or experiential fusion formula rather than a more discriminative and adaptive fusion. To address the above two limitations, this paper proposes a new depth-aware

This work was supported by National Natural Science Foundation of China under Grants 61171144 and 61471230, and by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

*Corresponding author: Zhi Liu, liuzhisjtu@163.com.

saliency model using discriminative saliency fusion. The framework of the proposed model is presented in Fig. 1. Saliency maps on different features at low, mid and high levels are calculated by taking account of primary depth and appearance contrasts, different feature weighted factors and location priors, respectively. Further, a discriminative saliency fusion (DSF) regressor based on random forest is learned from the training samples to discover the most discriminative saliency maps from the three levels and adaptively integrate them to obtain the saliency measures of regions. Experimental results show that the proposed model achieves the better saliency detection performance on both RGBD images and stereoscopic images.

The rest of this paper is organized as follows. Section 2 describes the proposed saliency model. Experimental results are presented in Section 3, and conclusion is given in Section 4.

2. PROPOSED SALIENCY MODEL

2.1. Low-level saliency

Each RGBD image can be decomposed into a color image and a gray-level depth map. Based on the color image, the gPb-owt-ucm [12] method is used to obtain the primitive segmentation result with approximately 200 regions as illustrated in the second column of Fig. 1. In most RGBD images, salient object regions usually show noticeable feature contrast with background regions. Thus the commonly used low-level center-surround contrast can still work as a fundamental principle of saliency detection. To obtain a group of low-level saliency maps, we take into account multiple low-level regional features including color, depth, texture and geodesic distance. The details of these features are as follows (along with the number of features in the parenthesis): Average color of each channel and histograms in the RGB, HSV and L*a*b* color spaces (12); average depth value and histogram (2); texture features including absolute responses (15×2) and their histograms (1×2) of 15 LM filters [13], HOG [14] histograms (1×2) and LBP [15] histograms (1×2); Geodesic distance [16] (1×2). Note that the texture and geodesic distance features are extracted on both color image and depth map, and thus the numbers of corresponding features are with a multiplication factor of two, "$\times 2$". In total, the number of features for low-level saliency computation is 52.

With the above low-level features, low-level saliency (LS) of each region R_i based on the k^{th} feature contrast is evaluated with respect to the global image and the image border, respectively, as follows:

$$LS^k_{G/B}(R_i) = \sum\nolimits_{R_j \in G/B} w_{i,j} \cdot D^k_{i,j}, \tag{1}$$

where G is the set of all regions in the image, and B is the set of border regions with a distance to the nearest image border less than 20 pixels. $D^k_{i,j}$ is the chi-square distance for features with the form of histogram or the Euclidean distance for features with other forms, on the k^{th} feature between R_i and R_j, i.e., the difference between f^k_i and f^k_j. The weight $w_{i,j}$ takes into account the factors of region size and spatial similarity and is defined as follows:

$$w_{i,j} = \left|R_j\right| \cdot \exp\left(\frac{-\left\|c_i - c_j\right\|}{\alpha \cdot L}\right), \tag{2}$$

where $\left|R_j\right|$ denotes the number of pixels in R_j, L denotes the diagonal length of image, c_i denotes the spatial center position of R_i, and the coefficient α is set to a moderate value, 0.3, to control the influence of spatial distance between regions. Since the low-level saliency is evaluated on a total of 52 features with respect to the global image and image border, respectively, we obtain for each region R_i a 104-dimensional low-level saliency vector $\mathbf{v}^{LS}_i$.

2.2. Mid-level saliency

It is verified that the salient object usually has certain relationship with its depth levels [8]. Besides, the geodesic distance [16] is a simple yet effective feature indicating salient regions directly. Thus the mid-level saliency (MS) is evaluated based on low-level saliency with feature weighting factor from depth or geodesic distance as follows:

$$MS^k_{G/B,DP}(R_i) = \exp(-d_i) \cdot LS^k_{G/B}, \forall f^k \notin \Omega_D, \tag{3}$$

$$MS^k_{G/B,DG}(R_i) = Geo_d(R_i) \cdot LS^k_{G/B}, \forall f^k \notin \Omega_D, \tag{4}$$

$$MS^k_{G/B,CG}(R_i) = Geo_c(R_i) \cdot LS^k_{G/B}, \forall f^k \notin \Omega_C, \tag{5}$$

where $MS^k_{G/B,DP}(R_i)$ is the depth weighted mid-level saliency. d_i is the mean depth value of R_i, and the term $\exp(\cdot)$ indicates that the close regions tend to receive more visual attention. $MS^k_{G/B,DG}(R_i)$ / $MS^k_{G/B,CG}(R_i)$ is the depth/color geodesic distance weighted mid-level saliency, $Geo_d(R_i) / Geo_c(R_i)$ is the depth/color geodesic distance, and Ω_D / Ω_C is the set of features involving depth/color information mentioned in Section 2.1, and such a cross weighting in Eqs. (3)-(5) enables these saliency measures to comprehensively utilize depth and color information. The above three types of weighted saliency measures for each region R_i constitute a 166-dimensional mid-level saliency vector $\mathbf{v}^{MS}_i$.

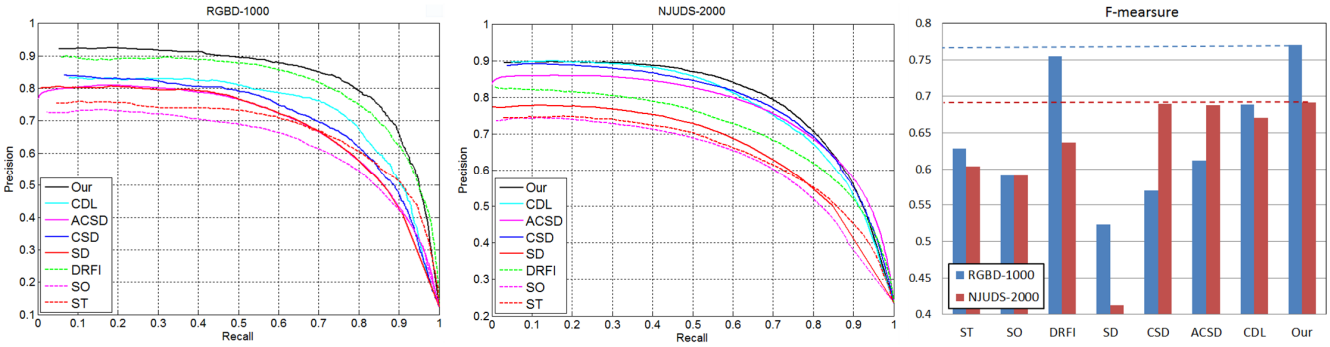

Fig. 2. PR curves (left and middle) and F-measures (right) of different saliency models.

2.3. High-level saliency

In addition to the feature contrast, some high-level location priors are also important in identifying salient regions. In most RGBD images, background regions generally have a higher ratio of connectivity with image borders than salient objects. Based on this observation, the location-based object prior (OP) for each region R_i is defined as follows:

$$OP_i = \left(1 - \left(\frac{NB_i}{NB_{\max}}\right)^{\beta}\right) \cdot \exp\left(\frac{-SDC_i}{L/2}\right), \quad (6)$$

where SDC_i denotes the Euclidean spatial distance from the center position of R_i to the image center position. NB_i is the number of image border pixels contained in R_i, and among all regions touching the image border, $NB_{\max}$ is the maximum number of image border pixels contained in a region. The coefficient β is set to 0.25 for a moderate attenuation effect on location priors of those regions touching image borders.

To obtain the high-level saliency (HS) for each region R_i based on the k^{th} feature contrast, we exploit the k^{th} feature difference between R_i and R_j to assign similar location saliency measures to regions with similar values on the k^{th} feature as follows:

$$HS_i^k = OP_i \cdot \frac{\sum_{j=1, j\neq i}^{n} OP_j \cdot \left(1 - ND_{i,j}^k / ND_{\max}^k\right)}{\sum_{j=1, j\neq i}^{n} \left(1 - ND_{i,j}^k / ND_{\max}^k\right)}, \quad (7)$$

where $ND_{\max}^k$ is the maximum of $D_{i,j}^k$ between all the region pairs. In summary we obtain for each region R_i a 52-dimensional high-level saliency vector $\mathbf{v}_i^{HS}$.

2.4. Discriminative saliency fusion

After calculating regional saliency measures on various features at the three levels as shown in the third column of Fig. 1, we obtain a 322-dimensional saliency vector for each region. Unlike most existing depth-aware saliency models that use some linear or experiential fusion formula to combine saliency measures on different features, we aim to integrate the saliency measures on different features at different levels in a discriminative way by automatically discovering the most discriminative ones.

Besides, we consider some regional properties to better combine saliency measures. For each region R_i, we obtain a 123-dimensional auxiliary property vector $\mathbf{v}_i^{RP}$ including the absolute values and variances of features (111) in Section 2.1, and the geometric features (12) in [5]. As a result, each region R_i is represented by a 445-dimensional saliency vector consisting of multi-level saliency measures and auxiliary regional properties. We integrate them together in a discriminative way, where F is a DSF regressor based on random forest, to obtain the final regional saliency RS_i as shown as the fourth column of the example in Fig.1.

The learning procedure for the DSF regressor F is detailed in the following. First, we randomly select 500 images with binary ground truths of salient objects from the RGBD-1000 dataset [6] as the training images. Multi-scale segmentation is performed on each training image to generate training samples. We use the gPb-owt-ucm method [12] on the color image to generate the real-valued ultrametric contour map (UCM), which indicates for each pixel the likelihood of being a true boundary. Then we set two groups of thresholds to control the maximum number of regions and the area ratio of the smallest region to the largest region, and obtain M-scale segmentation results $\mathbf{S} = \{S_1, S_2, \cdots, S_M\}\ (M = 15)$ with different number of regions. For the segmentation result at the m^{th} scale, $S_m = \{R_1^m, R_2^m, \cdots, R_K^m\}$, we pick out the confident regions as $S_m' = \{R_1^m, R_2^m, \cdots, R_Q^m\}\ (Q \le K)$, in which the number of object/background pixels in each region exceeds 80 percent of the total number of pixels in the region, and set for each

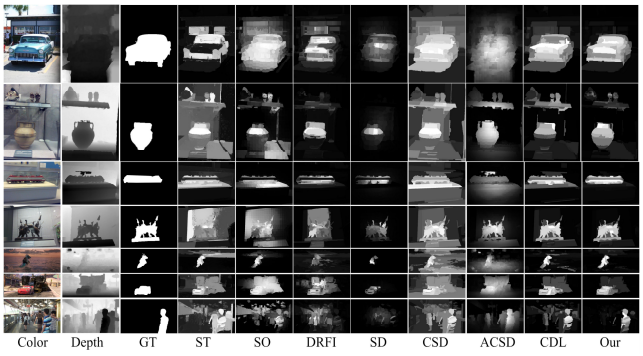

Fig. 3. Saliency maps generated using different models**.**

region its saliency score to 1/0 and obtain the binary-valued vector $A_m = \left\{a_1^m, a_2^m, \cdots, a_Q^m\right\}$.

As aforementioned, each region R_i^m in S_m' can be represented using a 445-dimensional saliency vector $\mathbf{x}_i^m$ consisting of multi-level saliency and regional properties. The DSF regressor F based on random forest is learned from the training data $X_m = \left\{\mathbf{x}_1^m, \mathbf{x}_2^m, \cdots, \mathbf{x}_Q^m\right\}$ and the saliency score vector $A_m = \left\{a_1^m, a_2^m, \cdots, a_Q^m\right\}$. The DSF regressor F can fuse the multi-level saliency values in a discriminative way. For each test image with segmented regions, the DSF regressor F is exploited to estimate the saliency measures of regions and generate the region-level saliency map as shown in the rightmost column of Fig. 1.

3. EXPERIMENTAL RESULTS

3.1. Datasets and experimental settings

We performed experiments on two public datasets designed for depth-aware saliency detection: RGBD-1000 [6] and NJUDS-2000 [7] including 1000 RGBD images and 2000 stereoscopic images, respectively, along with manually labelled ground truths for salient objects. Note that the depth information is captured by Kinect in RGBD-1000 and represented by disparity maps in NJUDS-2000.

Since we used 500 images from the RGBD-1000 dataset to train the random forest regressor in Section 2.4, and thus the remaining 500 images in RGBD-1000 and all the 2000 images in NJUDS-2000 are used as the two test datasets. The test over the two datasets can help to evaluate the adaptability of the proposed model. In order to present a robust evaluation of saliency detection performance, we adopt the commonly used precision-recall (PR) curve, which is plotted by connecting the precision-recall scores at all thresholds. Besides, we use F-measure, which can be interpreted as a weighted average of precision and recall, to quantitatively evaluate the quality of saliency maps. The adaptive thresholding method [OTSU] [17] is performed on each saliency map to obtain the binary mask of salient objects for calculating F-measure, and the weight coefficient in F-measure is set to 1 to weight precision and recall equally.

3.2. Training parameters

Based on some comparison experiments, we set several key training parameters used in Section 2.4 as follows. The maximum region number of the finest segmentation to generate the training samples is set to 300. We totally use 200 trees to construct the random forest for training our DSF regressor. During constructing a decision tree, the number of predictors sampled for splitting at each node is set to 75 in order to balance the efficiency and the effectiveness.

3.3. Results and Discussion

We compared our model with state-of-the-art saliency models including four depth-aware saliency models, i.e., SD [6], ACSD [7], CSD [10] and CDL [11], as well as three high-performing 2D saliency models including SO [3], ST [4] and DRFI [5]. For all saliency models, we used the saliency maps, executables or source codes with default parameter settings provided by the authors. For a fair comparison, we retrained a new model for DRFI [5] using the same training dataset as our model.

Objective comparisons of different models are shown in Fig. 2. It can be seen from Fig. 2 that our model consistently outperforms all the other models on the two datasets in terms of PR curve and F-measure. Saliency maps of several example images are shown in Fig. 3 for a subjective comparison. It can be seen from Fig. 3 that our model can generally better highlight salient objects with well-defined boundaries and suppress background regions effectively, compared to other saliency models. Especially for some complicated images such as the bottom two examples, other models may be distracted by the cluttered background and low contrast between salient object and background regions, while our model also highlights the complete salient objects well.

4. CONCLUSIONS

This paper has proposed a new depth-aware saliency model using discriminative saliency fusion method. Saliency maps of different features at three levels are calculated by taking account of primary depth and appearance contrasts, different feature weighted factors and location priors respectively. Then we learn a random forest regressor to perform the discriminative saliency fusion and generate the final regional saliency map. Both subjective and objective evaluations demonstrate that the proposed model achieves a satisfactory overall saliency detection performance both on RGBD image dataset and stereoscopic image dataset.

5. REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[2] M. M. Cheng, N. J. Mitra, X. Huang, P. Torr and S.M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569-582, Feb. 2015.

[3] W. Zhu, S. Liang, Y. Wei and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2814-2821.

[4] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937-1952, May 2014.

[5] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2083-2090.

[6] H. Peng, B. Li, W. Xiong, W. Hu and R. Ji, "RGBD salient object detection: a benchmark and algorithms," in *Proc. ECCV*, Sep. 2014, pp. 92-109.

[7] R. Ju, Y. Liu, T. Ren, L. Ge and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Processing: Image Communication*, vol. 38, pp. 115-126, Oct. 2015.

[8] K. Desingh, K. M. Krishna, D. Rajan, and C. V. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *Proc. BMVC*, Sep. 2013, pp. 1-11.

[9] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE CVPR*, Jun. 2012, pp. 454-461.

[10] X. Fan, Z. Liu and G. Sun, "Salient region detection for stereoscopic images," in *Proc. IEEE DSP*, Aug. 2014, pp. 454-458.

[11] H. Song, Z. Liu, H. Du, G. Sun and C. Bai, "Saliency detection for RGBD images," in *Proc. ICIMCS*, Aug. 2015, pp. 1-4.

[12] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no.5, pp. 898-916, May 2011.

[13] T. K. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29-44, Feb. 2001.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, Jun. 2005, pp. 886-893.

[15] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognition*, vol. 42, no. 3, pp. 425-436, Mar. 2009.

[16] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic Saliency Using Background Priors," in *Proc. ECCV*, Sep. 2012, pp. 29-42.

[17] N. Otsu. "A threshold selection method from gray-level histograms," *Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62-66, Jan. 1979.