2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

RGB-D Saliency Detection under Bayesian Framework

Song-Tao Wang[1,2], Zhen Zhou[1], Han-Bing Qu[2], Bin Li[2]
[1]The Higher Educational Key Laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science & Technology, Harbin, China
[2]Key Laboratory of Pattern Recognition Beijing Academy of Science and Technology, Beijing, China
wangsongtao1983@163.com, zhzh49@126.com, quhanbing@gmail.com, lbn@hit.edu.cn

***Abstract*—In this paper, we propose a saliency detection model for RGB-D images based on the contrasting features of color and depth within a Bayesian framework. The depth feature map is extracted based on superpixel contrast computation with spatial priors. We model the depth saliency map by approximating the density of depth-based contrast features using a Gaussian distribution. Similar to the depth saliency computation, the color saliency map is computed using a Gaussian distribution based on multi-scale contrasts in superpixels by exploiting low-level cues. By assuming that color- and depth-based contrast features are conditionally independent, given the classes, a discriminative mixed-membership naive Bayes (DMNB) model is used to calculate the final saliency map from the depth saliency and color saliency probabilities by applying Bayes' theorem. The Gaussian distribution parameter can be estimated in the DMNB model by using a variational inference-based expectation maximization algorithm. The experimental results on a recent eye tracking database show that the proposed model performs better than other existing models.**

I. Introduction

Saliency detection is considered the problem of identifying the points that attract the visual attention of human beings. Le Callet and Niebur introduced the concepts of overt and covert visual attention and of bottom-up and top-down processing[1]. Visual attention selectively processes important visual information by filtering out less important information and is an important characteristic of the human visual system (HVS) for visual information processing. Visual attention is one of the most important mechanisms that are deployed in the HVS to cope with large amounts of visual information and reduce the complexity of scene analysis. Visual attention models have been successfully applied in many domains, including multimedia delivery, visual retargeting, quality assessment of images and videos, medical imaging, and stereoscopic 3D image applications[1].

Borji and Itti provided an excellent overview of the current state-of-the-art 2D visual attention modeling and included a taxonomy of models (cognitive, Bayesian, decision theoretic, information theoretical, graphical, spectral analysis, pattern classification, and more) [2]. Many saliency measures have emerged that simulate the HVS, which tends to find the most informative regions in 2D scenes[3], [4], [5], [6], [7], [8], [9], [10]. However, most saliency models disregard the fact that the HVS operates in 3D environments and these models can thus investigate only from 2D images. Eye fixation data are captured while looking at 2D scenes, but depth cues provide additional important information about content in the visual field and therefore can also be considered relevant features for saliency detection. Stereoscopic contents carry important additional binocular cues for enhancing human depth perception[11]. Today, with the development of 3D display technologies and devices, there are various emerging applications for 3D multimedia, such as 3D video retargeting[12], 3D video quality assessment[13] and so forth. Overall, the emerging demand for visual attention-based applications for 3D multimedia has increased the need for computational saliency detection models for 3D multimedia content. In contrast to saliency detection for 2D images, the depth factor must be considered when performing saliency detection for 3D images. Therefore, two important challenges when designing 3D saliency models are how to estimate the saliency from depth cues and how to combine the saliency from depth features with those of other 2D low-level features.

In this paper, we propose a new computational saliency detection model for RGB-D images that considers both color- and depth-based contrast features within a Bayesian framework. The main contributions of our approach consist of two aspects: (1) to estimate saliency from depth cues, we propose to model the depth saliency map by approximating the density of depth-based contrast features using a Gaussian distribution and create the depth feature map based on superpixel contrast computation with spatial priors, and (2) by assuming that color-based and depth-based features are conditionally independent given the classes, the discriminative mixed-membership naive Bayes (DMNB) model is used to calculate the final saliency map by applying Bayes' theorem.

II. Related Work

As introduced in the section I, many computational models of visual attention have been proposed for various 2D multimedia processing applications. However, compared with the set of 2D visual attention models, only a few computational models of 3D visual attention have been proposed[14], [15], [16], [17], [18], [19], [20]. These models all contain a stage in which 2D saliency features are extracted and used to compute 2D saliency maps. However, depending on the way in which they use depth information in terms of the development of computational models, these models can be classified into three different categories:

(1) Depth-weighting models−This type of model adopts depth information to weight a 2D saliency map to calculate the final saliency map for 3D images with feature map fusion[15]. The models in this category combine 2D features with a depth

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

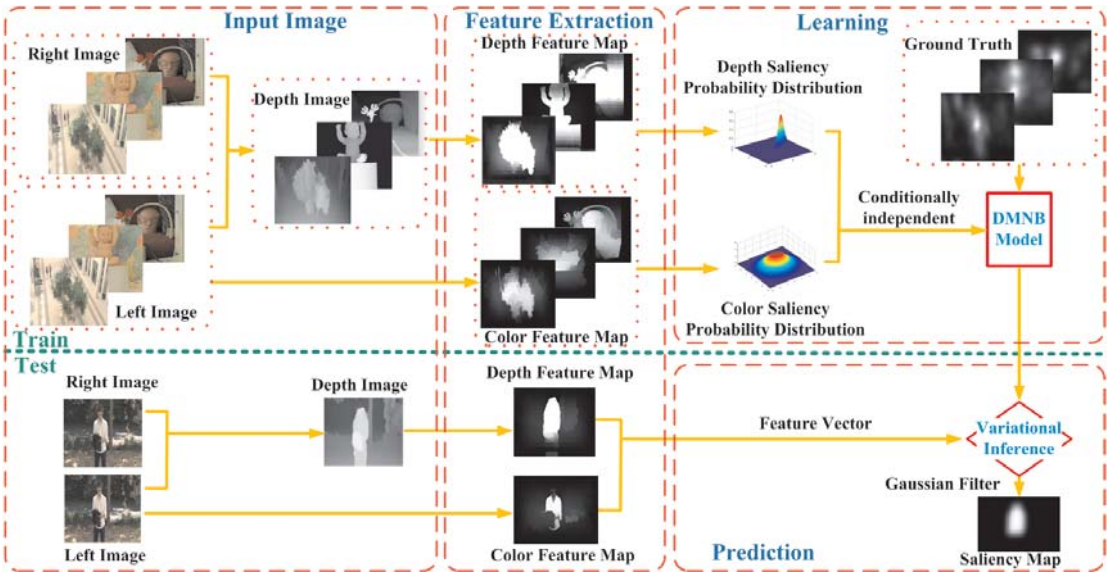


Fig. 1. The flowchart of the proposed model. The framework of our model consists of two stages: the training stage shown in the top part of the figure and the testing stage shown in the bottom part of the figure.

feature to calculate the final saliency map, but they do not include the depth saliency map in their computation processes. Apart from detecting the salient areas by using 2D visual features, these models share a common step in which depth information is used as a weighting factor for the 2D saliency.

(2) Depth-saliency models−This type of model combines a depth saliency map and a traditional 2D saliency map simply to obtain a saliency map for 3D images[16], [17]. The models in this category rely on the existence of "depth saliency maps." Depth features are extracted from the depth map to create additional feature maps, which are then used to generate the depth saliency maps. These depth saliency maps are finally combined with 2D saliency maps using a saliency map pooling strategy to obtain a final 3D saliency map.

(3) Learning-based models−Instead of using a depth saliency map directly, this type of model uses machine learning techniques to build a 3D saliency detection model for stereoscopic images based on extracted 2D features and depth features[18], [19], [20]. Iatsun et al. proposed a visual attention model for 3D video using a machine learning approach. They used artificial neural networks to define adaptive weights for the fusion strategy based on eye tracking data[18]. Desingh et al. investigated the role of depth in saliency detection in the presence of competing saliencies such as appearance, depth-induced blur and centre bias. The computed 3D saliency was combined with 2D saliency models through non-linear regression using a support vector machine (SVM) to improve the saliency maps[19]. Inspired by the recent success of machine learning techniques in building 2D saliency detection models, Fang et al. proposed a learning-based model for stereoscopic images using linear SVM[20].

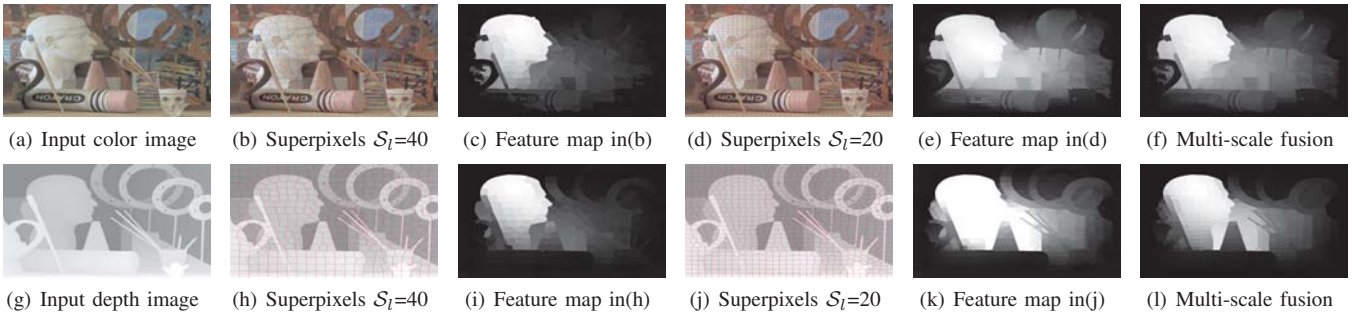
From the above description, the key to a 3D saliency detection model is determining how to integrate the depth cues with traditional 2D low-level features. In this paper, we propose a learning-based stereoscopic saliency detection model with a Bayesian framework that considers both color- and depth-based contrast features. Instead of simply combining a depth map with 2D saliency maps as in previous studies, we propose a computational saliency detection model for RGB-D images based on the DMNB model[21]. Experimental results on a public eye tracking database demonstrate the improved performance of the proposed model over other strategies.

III. The Proposed Approach

In this section, we introduce a method that integrates the color saliency probability with the depth saliency probability computed from Gaussian distributions based on multi-scale superpixel contrast features and yields a prediction of the final 3D saliency map using the DMNB model within a Bayesian framework. First, the input RGB-D images are represented by superpixels using multi-scale segmentation. Then, we compute the color and depth map by the weighted summation and normalization of the color- and depth-based contrast features, respectively, at different scales. Second, the probability distributions of both the color and depth saliency are modeled using the Gaussian distribution based on the color and depth feature maps, respectively. The parameters of the Gaussian distribution can be estimated in the DMNB model using a variational inference-based expectation maximization (EM) algorithm. The general architecture of the proposed framework is presented in Figure 1.

A. Feature extraction using multi-scale superpixels

We introduce a color-based contrast feature and a depth-based contrast feature to capture the contrast information of salient regions with spatial priors based on multi-scale superpixels using simple linear iterative clustering (SLIC)[22], which are generated at various grid interval parameters $\mathcal{S}$. We further impose a spatial prior term on each of the contrast measures holistically, which constrains the pixels that were rendered as salient to be compact as well as centered in the image domain. This spatial prior can also be generalized to consider the spatial distribution of different saliency cues such as the center prior and background prior[23]. We also observe that the background often presents local or global appearance connectivity with each of four image boundaries. These two



(a) Input color image (b) Superpixels $\mathcal{S}_l$=40 (c) Feature map in(b) (d) Superpixels $\mathcal{S}_l$=20 (e) Feature map in(d) (f) Multi-scale fusion

(g) Input depth image (h) Superpixels $\mathcal{S}_l$=40 (i) Feature map in(h) (j) Superpixels $\mathcal{S}_l$=20 (k) Feature map in(j) (l) Multi-scale fusion

Fig. 2. Visual samples of different color and depth feature maps. (a)∼(f) multi-scale superpixel segmentation for color feature maps and (g)∼(l) multi-scale superpixel segmentation for depth feature maps.

features complement each other in detecting 3D saliency cues from different perspectives and, when combined, yield the final 3D saliency value.

Color-based contrast feature An input image is oversegmented at L scales, and the color feature map is formulated as

$$f(p_c) = \frac{1}{L}\sum_{l=1}^{L} C(p_c^l, \Psi_{Bc}^l) e^{-\frac{(x_p^l-\mu_x)^2+(y_p^l-\mu_y)^2}{2\sigma^2}} \quad (1)$$

where p_c^l is a quantified histogram in the CIE Lab color space for each superpixel at any scale l, Ψ_{Bc}^l represents the pseudo-background context that consists of superpixels from the four boundaries of the RGB image, and $C(\cdot,\cdot) = \|\cdot\|_2^2$ is a typical contrast function using the Euclidean metric. In Equation 2, the center prior is formulated as a Gaussian model, where (x_p^l, y_p^l) are the coordinates of the centroid of the superpixel, (μ_x, μ_y) is the image center and σ is the variance. The final pixel-wise color feature map is obtained by assigning the feature value of each superpixel to every pixel belonging to it, where $\mathcal{S}_l$ is the grid interval parameter in [22], as shown in the first row of Figure 2.

Depth-based contrast feature Similar to the construction of the color feature map, we formulate the depth feature map based on multi-scale superpixel contrast in the disparity map that shows the parallax of each pixel between the left- and right-view images:

$$f(p_d) = \frac{1}{L}\sum_{l=1}^{L} C(p_d^l, \Psi_{Bd}^l) e^{-\frac{(x_p^l-\mu_x)^2+(y_p^l-\mu_y)^2}{2\sigma^2}} \quad (2)$$

where p_d^l is the depth value of the centroid calculated as the mean depth value within the superpixel. Visual samples for different depth feature maps are shown in the second row of Figure 2.

B. Bayesian framework for saliency detection

Let the binary random variable $\boldsymbol{z}_s$ denote whether a point belongs to a salient class. Given the observed color-based contrast feature $\boldsymbol{x}_c$ and the depth-based contrast feature $\boldsymbol{x}_d$ of that point, we formulate the saliency detection as a Bayesian inference problem to estimate the posterior probability at each pixel of the image:

$$p(\boldsymbol{z}_s|\boldsymbol{x}_c, \boldsymbol{x}_d) = \frac{p(\boldsymbol{z}_s, \boldsymbol{x}_c, \boldsymbol{x}_d)}{p(\boldsymbol{x}_c, \boldsymbol{x}_d)} \quad (3)$$

where $p(\boldsymbol{z}_s|\boldsymbol{x}_c, \boldsymbol{x}_d)$ is shorthand for the probability of predicting whether a pixel is salient, $p(\boldsymbol{x}_c, \boldsymbol{x}_d)$ is the likelihood of the observed color-based and depth-based contrast features, and $p(\boldsymbol{z}_s, \boldsymbol{x}_c, \boldsymbol{x}_d)$ is the joint probability of the latent class and observed features, defined as $p(\boldsymbol{z}_s, \boldsymbol{x}_c, \boldsymbol{x}_d) = p(\boldsymbol{z}_s)p(\boldsymbol{x}_c, \boldsymbol{x}_d|\boldsymbol{z}_s)$.

In this paper, the class-conditional mutual information (CMI) is used as a measure of dependence between two features $\boldsymbol{x}_c$ and $\boldsymbol{x}_d$, which can be defined as $I(\boldsymbol{x}_c, \boldsymbol{x}_d|\boldsymbol{z}_s) = H(\boldsymbol{x}_c|\boldsymbol{z}_s) + H(\boldsymbol{x}_d|\boldsymbol{z}_s) - H(\boldsymbol{x}_c, \boldsymbol{x}_d|\boldsymbol{z}_s)$, where $H(\boldsymbol{x}_c|\boldsymbol{z}_s)$ is the class-conditional entropy of $\boldsymbol{x}_c$. We employ a CMI threshold τ to discover feature dependencies. For simplicity, we assume that the color-based contrast feature $\boldsymbol{x}_c$ and depth-based contrast feature $\boldsymbol{x}_d$ are conditionally independent given the classes $\boldsymbol{z}_s$, that is, $p(\boldsymbol{x}_c, \boldsymbol{x}_d|\boldsymbol{z}_s) = p(\boldsymbol{x}_c|\boldsymbol{z}_s)p(\boldsymbol{x}_d|\boldsymbol{z}_s)$. This entails the assumption that the distribution of the color-based contrast features does not change with the depth-based contrast feature. Thus, the pixel-wise saliency of the likelihood is given by $p(\boldsymbol{z}_s|\boldsymbol{x}_c, \boldsymbol{x}_d) \propto p(\boldsymbol{z}_s)p(\boldsymbol{x}_c|\boldsymbol{z}_s)p(\boldsymbol{x}_d|\boldsymbol{z}_s)$.

C. DMNB model for saliency estimation

Given the graphical model of DMNB for saliency detection shown in Figure 3, the generative process for $\{x_{1:N}, y\}$ following the DMNB model can be described as follows (Algorithm 1), where $\boldsymbol{x}_{1:N} = (\boldsymbol{x}_c, \boldsymbol{x}_d)$, $\boldsymbol{z}_{1:N} = \boldsymbol{z}_s = (\boldsymbol{z}_c, \boldsymbol{z}_d)$ and $\boldsymbol{y}$ is the label that indicates whether the pixel is salient or not.

Algorithm 1 Generative process for saliency detection following the DMNB model

1: **Input:** α, η.
2: **Choose a component proportion:** $\theta \sim Dir(\theta|\alpha)$.
3: **For each feature:**
choose a component $\boldsymbol{z}_j \sim Mult(\boldsymbol{z}_j|\theta)$;
choose a feature value $\boldsymbol{x}_j \sim p(\boldsymbol{x}_j|\boldsymbol{z}_j, \Omega)$.
4: **Choose the label:** $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{z}_j, \eta)$.

In this work, both the color- and depth-based contrast features are assumed to have been generated from a Gaussian distribution with a mean of $\{\mu_{jk}, [j]_1^N\}$ and a variance of $\{\sigma_{jk}^2, [j]_1^N\}$. The marginal distribution of $(\boldsymbol{x}_{1:N}, \boldsymbol{y})$ is

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{y}|\alpha, \Omega, \eta) = \int p(\theta|\alpha)\Big(\prod_{j=1}^{N}\sum_{\boldsymbol{z}_j} p(\boldsymbol{z}_j|\theta)p(\boldsymbol{x}_j|\boldsymbol{z}_j, \Omega)p(\boldsymbol{y}|\boldsymbol{z}_j, \eta)\Big)d\theta \quad (4)$$

where θ is the prior distribution over K components, $\Omega = \{(\mu_{jk}, \sigma_{jk}^2), [j]_1^N, [k]_1^K\}$, $p(\boldsymbol{x}_j|\boldsymbol{z}_j, \Omega) \triangleq \mathcal{N}(\boldsymbol{x}_j|\mu_{jk}, \sigma_{jk}^2)$. In

two-class classification, $\boldsymbol{y}$ is either 0 or 1 generated from $Bern(\boldsymbol{y}|\eta)$. Because the DMNB model assumes a generative process for both the labels and features, we use both $\mathcal{X} = \{(\boldsymbol{x}_{ij}), [i]_1^{\mathcal{M}}, [j]_1^N\}$ and $\mathcal{Y} = \{\boldsymbol{y}_i, [i]_1^{\mathcal{M}}\}$ as a collection of $\mathcal{M}$ superpixels in trained images from the generative process to estimate the parameters of the DMNB model such that the likelihood of observing $(\mathcal{X}, \mathcal{Y})$ is maximized. Due to the latent variables, the computation of the likelihood in Equation 4 is intractable. In this paper, we use a variational inference method, which alternates between obtaining a tractable lower bound to the true log-likelihood and choosing the model parameters to maximize the lower bound.

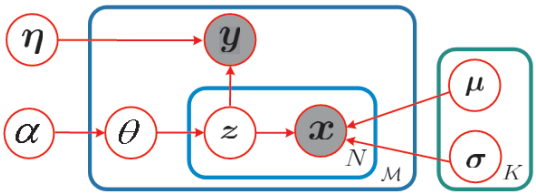


Fig. 3. Graphical models of DMNB for saliency estimation. $\boldsymbol{y}$ and $\boldsymbol{x}$ are the corresponding observed states, and z is the hidden variable.

We use $\mathcal{L}$ to denote the lower bound:

$$\begin{aligned}\mathcal{L} =& \boldsymbol{E}_q[\log p(\theta|\alpha)] + \boldsymbol{E}_q[\log p(\boldsymbol{z}_{1:N}|\theta)] \\ &+ \boldsymbol{E}_q[\log p(\boldsymbol{x}_{1:N}|\boldsymbol{z}_{1:N}, \gamma)] - \boldsymbol{E}_q[\log q(\theta)] \\ &- \boldsymbol{E}_q[\log q(\boldsymbol{z}_{1:N})] + \boldsymbol{E}_q[\log p(\boldsymbol{y}|\boldsymbol{z}_{1:N}, \eta)] \end{aligned} \tag{5}$$

where $\boldsymbol{E}_q[\log p(\boldsymbol{y}|\boldsymbol{z}_{1:N}, \eta)] \geq \sum_{k=1}^K \phi_k(\eta_k \boldsymbol{y} - \frac{e^{\eta_k}}{\xi}) - (\frac{1}{\xi} + \log \xi)$ and $\xi > 0$ is a newly introduced variational parameter. Maximizing the lower-bound function $\mathcal{L}(\gamma_k, \phi_k, \xi; \alpha, \Omega, \eta)$ with respect to the variational parameters yields updated equations for γ_k, ϕ_k and ξ as follows:

$$\phi_k \propto e^{(\Psi(\gamma_k) - \Psi(\sum_{l=1}^K \gamma_l) + \frac{1}{N}(\eta_k \boldsymbol{y}_i - \frac{e^{\eta_k}}{\xi} - \sum_{j=1}^N \frac{(\boldsymbol{x}_{ij} - \mu_{jk})^2}{2\sigma_{jk}^2}))} \tag{6}$$

$$\gamma_k = \alpha + N\phi_k \tag{7}$$

$$\xi = 1 + \sum_{k=1}^K \phi_k e^{\eta_k} \tag{8}$$

The variational parameters $(\gamma^*, \phi^*, \xi^*)$ from the inference step provide the optimal lower bound for the log-likelihood of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and maximizing the aggregate lower bound $\sum_{i=1}^{\mathcal{M}} \mathcal{L}(\gamma^*, \phi^*, \xi^*, \alpha, \Omega, \eta)$ over all of the data with respect to α, Ω and η, respectively, yields the estimated parameters. For η, we have

$$\eta_k = \log(\frac{\sum_{i=1}^{\mathcal{M}} \phi_{ik} \boldsymbol{y}_i}{\sum_{i=1}^{\mathcal{M}} \frac{\phi_{ik}}{\xi_i}}). \tag{9}$$

Based on the variational inference and parameter estimation updates, it is straightforward to construct a variational inference-based EM algorithm to estimate (α, Ω, η). After obtaining the DMNB model parameters from the EM algorithm, we can use η to perform saliency prediction. Given the feature $(\boldsymbol{x}_{1:N})$, we have

$$\boldsymbol{E}[\log p(\boldsymbol{y}|\boldsymbol{x}_{1:N}, \alpha, \Omega, \eta)] = \begin{cases} \eta^T \boldsymbol{E}[\overline{\boldsymbol{z}}] - \boldsymbol{E}[\log(1 + e^{\eta^T \overline{\boldsymbol{z}}})] & \boldsymbol{y} = 1 \\ 0 - \boldsymbol{E}[\log(1 + e^{\eta^T \overline{\boldsymbol{z}}})] & \boldsymbol{y} = 0 \end{cases} \tag{10}$$

where $\overline{\boldsymbol{z}}$ is an average of $\boldsymbol{z}_{1:N}$ over all of the observed features. The computation for $\boldsymbol{E}[\overline{\boldsymbol{z}}]$ is intractable; therefore, we again introduce the distribution $q(\boldsymbol{z}_{1:N}, \theta)$ and calculate $\boldsymbol{E}_q[\overline{\boldsymbol{z}}]$ as an approximation of $\boldsymbol{E}[\overline{\boldsymbol{z}}]$. In particular, $\boldsymbol{E}_q[\overline{\boldsymbol{z}}] = \phi$; therefore, we only need to compare $\eta^T \phi$ with 0.

IV. Experimental Evaluation

In this section, we conduct some experiments to demonstrate the performance of our method. To date, there are no specific and standardized measures for computing the similarity between the fixation density maps and saliency maps created using computational models in 3D situations. Nevertheless, a range of different measures exist that are widely used to perform comparisons of saliency maps for 2D content. We use an evaluation methodology and quantitative evaluation metrics that are similar to those proposed in[15], [16], [20]. The correlation coefficient (CC)[24], the area under the receiver operating characteristic curve (AUC)[25] and normalized scanpath saliency (NSS)[26] are used to evaluate the quantitative performance of the proposed model.

A. Qualitative experiment

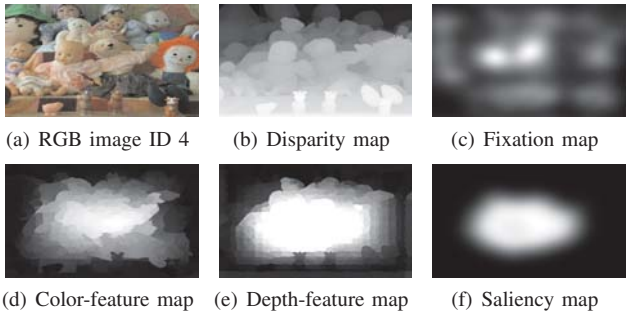
In this experiment, we use the IRC-cyN/IVC 3D Gaze database[27] proposed in [16] to evaluate the performance of the proposed model. As shown in Table I, we compute the CMI for all of the RGB-D images, and we set $\tau = 0.2$, which is a heuristically determined value. In Image 4, the widespread presence of faces and artificial color attracts the viewer's attention to most of the areas in the scene, and the CMI of the color- and depth-based contrast features is affected by the centre bias factor. The saliency map generated based on either the color salient features or depth might predict parts of the salient area, but not the all area, as shown in Figure 4. We are also interested in the contributions of different features in our model. The ROC curves of saliency estimation from different features are shown in Figure 5(a). This may be why color and depth saliency maps show comparable performance, whereas their combination produces a much better result. We divide the databases into two equal subsets and then choose one subset for training and the other for testing. The parameters of the DMNB are determined via 5-fold validation.

TABLE I. CMI of all of the RGB-D images in terms of the IRC-cyN/IVC 3D Gaze database

ID	Image1	Image2	Image3	Image4	Image5	Image6
CMI	0.1816	0.0256	0.0096	0.1916	0.0058	0
ID	Image7	Image8	Image9	Image10	Image11	Image12
CMI	0.0032	0.0977	0.0071	0.1944	0	0.1965
ID	Image13	Image14	Image15	Image16	Image17	Image18
CMI	0.0011	0	0.0735	0.0515	0.1994	0.0818

B. Comparison of 2D models combined with DSM

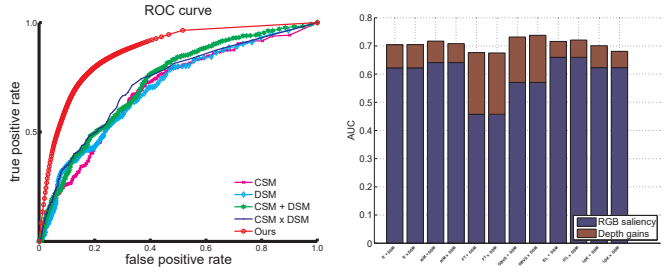
In this experiment, we first compare the performance of existing 2D saliency models before and after fusing the depth saliency map (DSM), which is produced by our proposed depth feature map, as a depth-saliency method [16] in the IRC-cyN/IVC 3D Gaze database. We select six state-of-the-art 2D visual attention models: IT[3], AIM[4], FT[5], GBVS[6],



(a) RGB image ID 4 (b) Disparity map (c) Fixation map

(d) Color-feature map (e) Depth-feature map (f) Saliency map

Fig. 4. Visual samples for different feature maps and saliency map.



(a) (b)

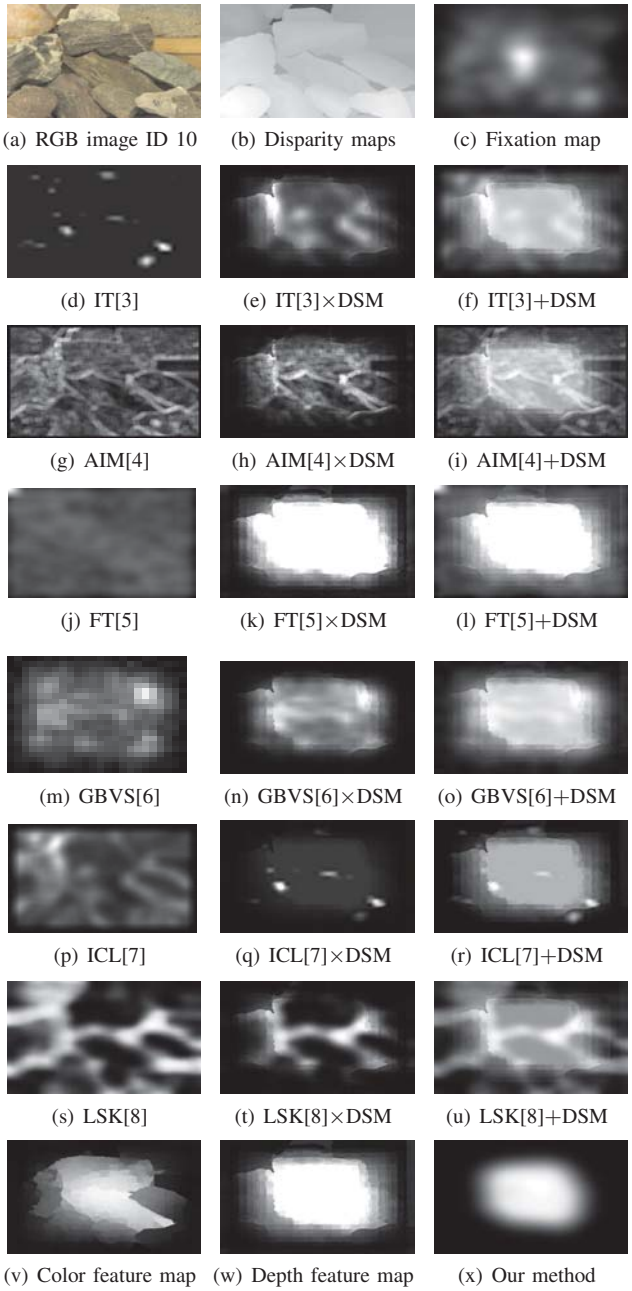
Fig. 5. Experimental results in terms of the IRC-cyN/IVC 3D Gaze database. (a) The ROC curves of different saliency maps. (b) The quantitative comparisons of the performance of depth cues. + indicates a linear combination strategy, and × indicates a weighting method based on multiplication. DSM means depth saliency map, which is produced by our proposed depth feature map. CSM means color saliency map, which is produced by our proposed color feature map.

ICL[7] and LSK[8]. Figure 5(b) presents the experimental results, and we present some visual comparison samples in Figure 6, where + and × denote a linear combination strategy and a weighting method, respectively, based on multiplication as used in [16]. These visual comparison samples illustrate the strong influence of using the DSM on the distribution of visual attention in terms of the viewing of 3D content. Although the simple late fusion strategy achieves improvements, it still suffers from inconsistency in the homogeneous foreground regions, which may be ascribed to treating the appearance and depth correspondence cues in an independent manner.

We calculated the NSS, CC and AUC values of the proposed model on the IRC-cyN/IVC 3D Gaze database as shown in Figure 7. From Figure 7, we can see that the CC and AUC values of the proposed model are larger than those of the other compared models and that the NSS value of the proposed model is lower than those of the compared models. We also provide the ROC curves for several compared models in Figure 8. The ROC curves demonstrate that the proposed stereoscopic saliency detection model performs better than the compared models do.

C. Comparison of 3D models

We compared the proposed model with other existing models described in [15], [16]. In this paper, Wang's model[15] and Fang's model[16] are classified as depth-weighting and depth-saliency models, respectively. Similar to [15], [16], we calculate the AUC value for the proposed model from the database. The quantitative comparison results of the 3D



(a) RGB image ID 10 (b) Disparity maps (c) Fixation map

(d) IT[3] (e) IT[3]×DSM (f) IT[3]+DSM

(g) AIM[4] (h) AIM[4]×DSM (i) AIM[4]+DSM

(j) FT[5] (k) FT[5]×DSM (l) FT[5]+DSM

(m) GBVS[6] (n) GBVS[6]×DSM (o) GBVS[6]+DSM

(p) ICL[7] (q) ICL[7]×DSM (r) ICL[7]+DSM

(s) LSK[8] (t) LSK[8]×DSM (u) LSK[8]+DSM

(v) Color feature map (w) Depth feature map (x) Our method

Fig. 6. Visual comparison of saliency estimations of different methods. (a)∼(c) are original images in the IRC-cyN/IVC 3D Gaze database. (d), (g), (j), (m), (p) and (s) are saliency maps computed using 2D saliency detection approaches. (e)∼(f), (h)∼(i), (k)∼(l), (n)∼(o), (q)∼(r) and (t)∼(u) are saliency maps computed using 2D method confusion with DSM. (v) and (w) are feature maps generated by our feature extraction method.

TABLE II. COMPARISON OF DIFFERENT 3D SALIENCY DETECTION MODELS BASED ON THE IRC-CYN/IVC 3D GAZE DATABASE.

Method	AUC	Method	AUC
IT[3]×DSM in[15]	0.671	IT[3]×DSM in[16]	0.688
IT[3]+DSM in[15]	0.676	IT[3]+DSM in[16]	0.683
AIM[4]×DSM in[15]	0.540	AIM[4]×DSM in[16]	0.671
AIM[4]+DSM in[15]	0.656	AIM[4]+DSM in [16]	0.675
FT[5]×DSM in[15]	0.667	FT[5]×DSM in[16]	0.660
FT[5]+DSM in[15]	0.677	FT[5]+DSM in[16]	0.670
3D framework in [15]	0.740	**Our Method**	**0.773**

saliency detection models are provided in Table II. The AUC values used for the other existing models came from the original papers [15], [16]. From this table, we can see that the AUC value of the proposed model is larger than those of the compared models, which demonstrates that the proposed model can achieve better performance levels than most of the other compared models in terms of saliency. This is mainly because the saliency detection within Bayesian framework enhances the consistency and compactness of salient pathes.

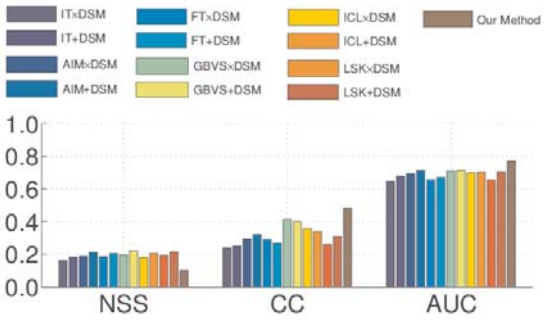


Fig. 7. Comparison results from different 2D saliency detection models combined with DSM in terms of the IRC-cyN/IVC 3D Gaze database.

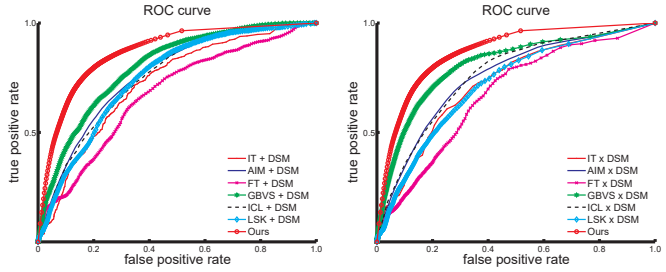


Fig. 8. The ROC curves of different stereoscopic saliency detection models in terms of the IRC-cyN/IVC 3D Gaze database. + indicates a linear combination strategy, and × indicates a weighting method based on multiplication. DSM means depth saliency map, which is produced by our proposed depth feature map.

V. Conclusion

In this study, we proposed a saliency detection model for RGB-D images that considers both color- and depth-based contrast features within a Bayesian framework. The experiments verify that the proposed model's depth-produced saliency can serve as a helpful complement to the existing color-based saliency models. Compared with other competing 3D models, the experimental results based on a recent eye tracking database show that the performance of the proposed saliency detection model is promising. We hope that our work is helpful in stimulating further research in the area of 3D saliency detection.

Acknowledgments

This work was supported in part by the Beijing Academy of Science and Technology Youth Backbone Training Plan (201430, 2015-16) and Innovation Group Plan of Beijing Academy of Science and Technology (IG201506N).

References

[1] P. Le Callet and E. Niebur, *Visual Attention and Applications in Multimedia Technology*, Proceedings of the IEEE Institute of Electrical & Electronics Engineers, 2013, 101(9):2058-2067.

[2] A. Borji and L. Itti, *State-of-the-art in visual attention modeling*. IEEE Transactions on PAMI, 2013, 35(1):185-207.

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

[4] N. D. B. Bruce and J. K. Tsotsos, *Saliency attention and visual search: An information theoretic approach*. Journal of vision, 2009, 9(3):5.1-24.

[5] X. Hou and L. Zhang, *Saliency detection: A spectral residual approach*. IEEE Conference on CVPR, 2007:1-8.

[6] J. Harel, C. Koch and P. Perona, *Graph-based visual saliency*. Advances in NIPS, 2006, 19:545-552.

[7] X. Hou and L. Zhang, *Dynamic visual attention: Searching for coding length increments*. Advances in NIPS, 2008, 21:681-688.

[8] H. J. Seo and P. Milanfar, *Static and space-time visual saliency detection by self-resemblance*. Journal of Vision, 2009, 9(12):15.1-27.

[9] Y. Xie, H. Lu and M.-H. Yan, *Bayesian Saliency via Low and Mid Level Cues*. IEEE Transactions on IP, 2013, 22(5):1689-1698.

[10] G. Li and Y. Yu. *Visual saliency based on multiscale deep features*. IEEE Conference on CVPR, 2015:5455-5463.

[11] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli and S. Yan, *Depth Matters: Influence of Depth Cues on Visual Saliency*. European Conference on Computer Vision-volume Part II, 2012:101-115.

[12] J. Wang, Y. Fang, M. Narwaria, W. Lin and P. L. Callet, *Stereoscopic Image Retargeting based on 3D Saliency Detection*. IEEE Conference on Acoustics, Speech and Signal Processing, 2014:669-673.

[13] H. Kim, S. Lee and A. C. Bovik, *Saliency Prediction on Stereoscopic Videos*. IEEE Transactions on IP, 2014, 23(4):1476-1490.

[14] Y. Zhang, G. Jiang, M. Yu and K. Chen, *Stereoscopic visual attention model for 3D video*. Advances in Multimedia modeling, 2010, 5916(1):314-324.

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

[16] J. Wang, M. P. DaSilva, P. Le Callet and V. Ricordel, *Computational Model of Stereoscopic 3D Visual Saliency*. IEEE Transactions on IP, 2013, 22(6):2151-2165.

[17] N. Ouerhani and H. Hugli, *Computing visual attention from scene depth*. International Conference on PR, 2000,1:375-378.

[18] I. Iatsun, M. C. Larabi and C. Fernandez-Maloigne, *Visual Attention Modeling for 3D Video using Neural Network*. International Conference on 3D Imaging, 2014:1-8.

[19] K. Desingh, K. M. Krishna, D. Rajan and C. V. Jawahar, *Depth really Matters: Improving Visual Salient Region Detection with Depth*. British Machine Vision Conference, 2013:98.1-98.11.

[20] Y. Fang, W. Lin, Z. Fang, J. Lei, P. Le Callet and F. Yuan, *Learning Visual Saliency for Stereoscopic Images*. IEEE International Conference on Multimedia and Expo Workshops, 2014:1-6.

[21] H. Shan, A. Banerjee and N. C. Oza, *Discriminative Mixed-membership Models*. IEEE International Conference on Data Mining, 2009:466-475.

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi and S. Süsstrunk, *Slic superpixels compared to state-of-the-art superpixel methods*. IEEE Transaction on PAMI, 2012, 34(11):2274-2282.

[23] C. Yang, L. Zhang, H. Lu and X. Ruan, *Saliency detection via graph-based manifold ranking*. IEEE Conference on CVPR, 2013:3166-3173.

[24] N. Ouerhani, R. v. Wartburg, H. Hügli and R. Müri, *Empirical validation of the saliency-based model of visual attention*. ELCVIA: electronic letters on computer vision and image analysis, 2004, 3(1):13-24.

[25] T. Fawcett, *ROC graphs: Notes and practical considerations for researchers*. Machine learning, 2004, 31:1-38.

[26] R. Peters, A. Iyer, L. Itti and C. Koch, *Components of Bottom-Up Gaze Allocation in Natural Images*. Vision Research, 2005, 45(18):2397-2416.

[27] J. Wang, M. P. Da Silva and P. Le Callet, *IRC-CyN/IVC 3DGaze database*. http://www.irccyn.ecnantes.fr/spip.php?article1102&lang=en.