

Visual Saliency Detection for RGB-D Images with Generative Model

Song-Tao Wang^{1,2}, Zhen Zhou^{1(✉)}, Han-Bing Qu², and Bin Li²

¹ The Higher Educational Laboratory for Measuring & Control Technology
and Instrumentations of Heilongjiang Province,
Harbin University of Science and Technology, Harbin, China
zhzh49@126.com

² Key Laboratory of Pattern Recognition,
Beijing Academy of Science and Technology, Beijing, China

Abstract. In this paper, we propose a saliency detection model for RGB-D images based on the contrasting features of colour and depth with a generative mixture model. 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 colour saliency map is computed using a Gaussian distribution based on multi-scale contrasts in superpixels by exploiting low-level cues. By assuming that colour- 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 colour 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.

1 Introduction

Saliency detection is 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 the concepts of bottom-up and top-down processing [11]. 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 3D image applications [11].

Borji and Itti provided an excellent overview of the current state-of-the-art 2D visual attention modelling and included a taxonomy of models (cognitive,

Bayesian, decision theoretic, information theoretical, graphical, spectral analysis, pattern classification, and more) [3]. Many saliency measures have emerged that simulate the HVS, which tends to find the most informative regions in 2D scenes [4, 13, 18]. 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. The stereoscopic content carries important additional binocular cues for enhancing human depth perception [5, 10]. Today, with the development of 3D display technologies and devices, there are various emerging applications for 3D multimedia, such as 3D video retargeting [16], 3D video quality assessment [9, 19] 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 RGB-D 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 colour- and depth-based contrast features with a generative mixture model. The main contributions of our approach consist of two aspects: (1) to estimate saliency from depth cues, we propose the creation of depth feature maps based on superpixel contrast computation with spatial priors and model the depth saliency map by approximating the density of depth-based contrast features using a Gaussian distribution, and (2) by assuming that colour-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.

2 Related Work

As introduced in the Sect. 1, 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 [6–8, 12, 14, 17, 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 RGB-D images with feature map fusion [6, 17]. Fang et al. proposed a novel 3D saliency detection framework based on colour, luminance, texture and depth

contrast features, which designed a new fusion method to combine the feature maps to obtain the final saliency map for RGB-D images [6]. In [17], colour contrast features and depth contrast features are calculated to construct an effective multi-feature fusion to generate saliency maps, and multi-scale enhancement is performed on the saliency map to further improve the detection precision focused on the 3D salient object detection. The models in this category combine 2D features with a depth feature to calculate the final saliency map, but they do not include the depth saliency map in their computation processes.

- (2) Depth-saliency models—This type of model combines depth saliency maps and traditional 2D saliency maps simply to obtain saliency maps for RGB-D images [8, 12, 14]. Ren et al. presented a two-stage 3D salient object detection framework, which first integrates the contrast region with the background, depth and orientation priors to achieve a saliency map and then reconstructs the saliency map globally [14]. Peng et al. proved a simple fusion framework that combines existing RGB-produced saliency with new depth-induced saliency: the former one is estimated from existing RGB models while the latter one is based on the multi-contextual contrast model [12]. Furthermore, Ju et al. proposed a novel saliency method that worked on depth images based on anisotropic centre-surround difference [8]. 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 (DSM). 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 RGB-D images based on extracted 2D features and depth features [7, 20]. Inspired by the recent success of machine learning techniques in building 2D saliency detection models, Fang et al. proposed a learning-based model for RGB-D images using linear SVM [7]. Zhu et al. proposed a learning-based approach for extracting saliency from RGB-D images, in which discriminative features can be automatically selected by learning several decision trees based on the ground truth, and those features are further utilized to search the saliency regions via the predictions of the trees [20].

From the above description, the key to 3D saliency detection models is determining how to integrate the depth cues with traditional 2D low-level features. In this paper, we propose a learning-based 3D saliency detection model with a generative mixture model that considers both colour- 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 [15]. Experimental results from a public eye tracking database demonstrate the improved performance of the proposed model over other strategies.

3 The Proposed Approach

In this section, we introduce a method that integrates the colour 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 colour and depth map using the weighted summation and normalization of the colour- and depth-based contrast features, respectively, at different scales. Second, the probability distributions of both the colour and depth saliency are modelled using the Gaussian distribution based on the colour 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 Fig. 1.

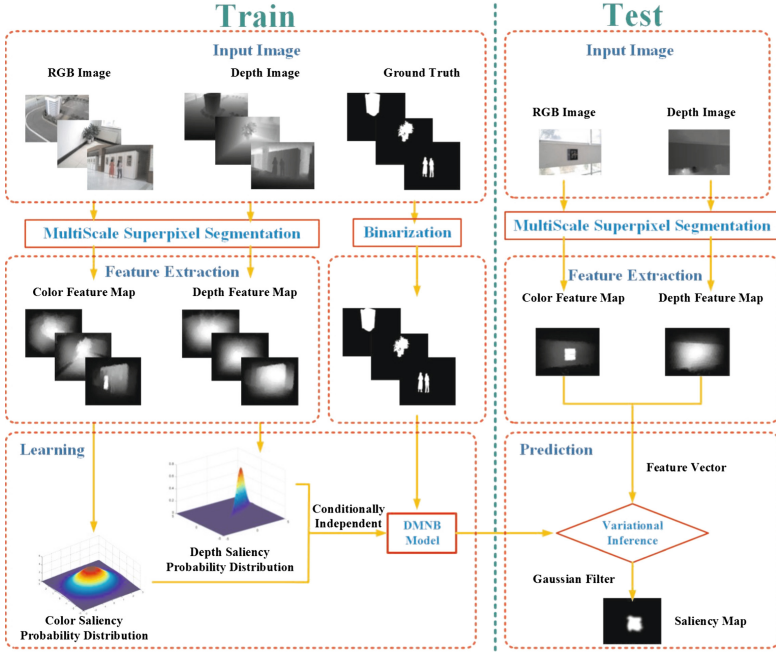


Fig. 1. The flowchart of the proposed model. The framework of our model consists of two stages: the training stage shown in the left part of the figure and the testing stage shown in the right part of the figure. In this work, we perform experiments based on the NLPR dataset in [12].

3.1 Feature Extraction Using Multi-scale Superpixels

We introduce a colour-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, which are generated at various grid interval parameters $\mathcal{S}$, similar to simple linear iterative clustering (SLIC) [1]. 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 centred in the image domain. This spatial prior can also be generalized to consider the spatial distribution of different saliency cues such as the centre prior and background prior [18]. We also observe that the background often presents local or global appearance connectivity with each of four image boundaries. These two features complement each other in detecting 3D saliency cues from different perspectives and, when combined, yield the final 3D saliency value.

RGB-D Images Multi-scale Superpixel Segmentation. For an RGB-D image pair, superpixels are segmented according to both colour and depth cues. We notice that when applying the SLIC algorithm directly to the RGB image and depth map, the segmentation result is unsatisfactory due to the lack of a mutual context relationship. We redefine the distance measurement incorporating depth as shown in Eq. 1:

$$D_s = \sqrt{d_{lab}^2 + \omega_d d_d^2 + \frac{m}{S} d_{xy}^2} \quad (1)$$

where $d_d = \sqrt{(d_j - d_i)^2}$ denotes the depth distance weighted by ω_d between pixel i and j in the depth map, d_{lab} and d_{xy} are the original distance measurements of colour and spatiality normalized with $\frac{m}{S}$ in [1], and D_s is the final distance between two pixels in the RGB-D image pair.

We obtain more accurate segmentation results as shown in Fig. 2 by considering the colour and depth cues simultaneously. The boundary between the foreground and the background is segmented more accurately.

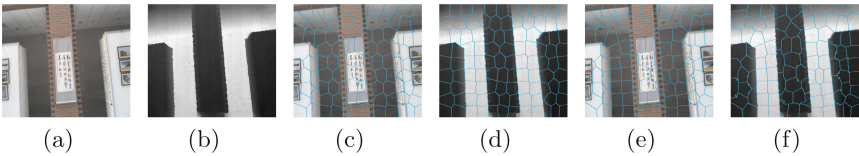


Fig. 2. Visual samples for superpixel segmentation of RGB-D images with $\mathcal{S} = 40$. (a) RGB image, (b) Depth image, (c) Colour-based segmentation, (d) Depth-based segmentation, (e) Colour- and depth-based segmentation result on colour image and (f) Colour- and depth-based segmentation result on depth image. (Color figure online)

Colour-Based Contrast Feature. An input image is oversegmented at L scales, and the colour feature map is formulated as

$$f(p_c^l) = \omega_c^l S C_{GMR}^l \quad (2)$$

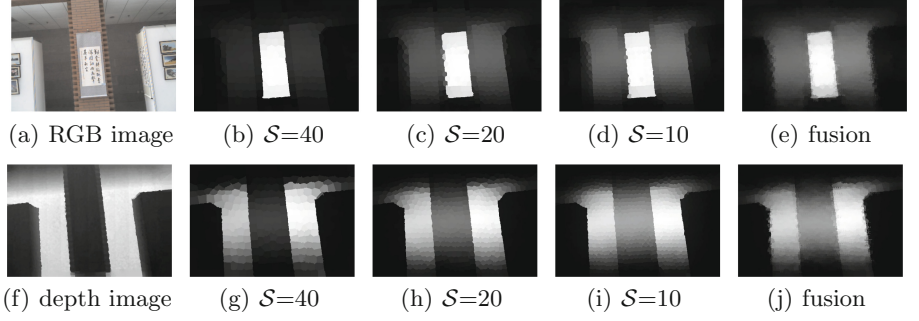


Fig. 3. Visual samples of different colour and depth feature maps. Row 1: colour feature maps of the NLPD dataset. Row 2: depth feature maps of the NLPD dataset.

where p_c^l is a quantified histogram in the CIE Lab colour space for each superpixel at any scale l , and SC_{GMR}^l is the colour saliency map generated by graph-based manifold ranking only with background cues similar to [18], in which the RGB image is represented as a single-layer graph with superpixels as nodes at any l scale. In contrast to [18], the definition of the background priors is inspired by the observation that the patches from the corners of images are more likely to be background and contain considerable scene information that helps distinguish salient objects. With multi-scale fusion, the colour feature map is constructed by weighted summation of $f(p_c^l)$, where the weights are determined by $\sum_{l=1}^L \omega_c^l = 1$. The final pixel-wise colour feature map is obtained by assigning the feature value of each superpixel to every pixel belonging to it, as shown in the first row of Fig. 3.

Depth-Based Contrast Feature. Similar to the construction of the colour feature maps, we formulate the depth feature maps based on multi-scale superpixels in the depth maps:

$$f(p_d^l) = \omega_d^l SD_{GMR}^l \quad (3)$$

where p_d^l is the depth value of the centroid calculated as the mean depth value within the superpixel and SD_{GMR}^l is the depth saliency map generated via graph-based manifold ranking only with background cues. In this work, the weight of the affinity matrix between two nodes in a depth map at any l scales is defined by

$$\omega_{ij}^l = e^{-\frac{(\overline{d_j^l} - \overline{d_i^l})^2}{\sigma^2}} \quad (4)$$

where $\overline{d_j^l}$ and $\overline{d_i^l}$ denote the mean of the superpixel i and superpixel j corresponding to two nodes, respectively, and σ is a constant that controls the strength of the weight in [18]. With multi-scale fusion, the depth feature map is constructed by weighted summation of $f(p_d^l)$, where the weights are determined by $\sum_{l=1}^L \omega_d^l = 1$. Visual samples for different depth feature maps are shown in the second row of Fig. 3.

3.2 Bayesian Framework for Saliency Detection

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

$$p(\mathbf{z}_s|\mathbf{x}_c, \mathbf{x}_d) = \frac{p(\mathbf{z}_s, \mathbf{x}_c, \mathbf{x}_d)}{p(\mathbf{x}_c, \mathbf{x}_d)} \quad (5)$$

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

In this paper, the class-conditional mutual information (CMI) is used as a measure of dependence between two features $\mathbf{x}_c$ and $\mathbf{x}_d$, which can be defined as $I(\mathbf{x}_c, \mathbf{x}_d|\mathbf{z}_s) = H(\mathbf{x}_c|\mathbf{z}_s) + H(\mathbf{x}_d|\mathbf{z}_s) - H(\mathbf{x}_c, \mathbf{x}_d|\mathbf{z}_s)$, where $H(\mathbf{x}_c|\mathbf{z}_s)$ is the class-conditional entropy of $\mathbf{x}_c$. We employ a CMI threshold τ to discover feature dependencies, as shown in Fig. 4. For CMI between the colour-based contrast feature and depth-based contrast feature less than τ , we assume that $\mathbf{x}_c$ and $\mathbf{x}_d$ are conditionally independent given the classes $\mathbf{z}_s$, that is, $p(\mathbf{x}_c, \mathbf{x}_d|\mathbf{z}_s) = p(\mathbf{x}_c|\mathbf{z}_s)p(\mathbf{x}_d|\mathbf{z}_s)$. This entails the assumption that the distribution of the colour-based contrast features does not change with the depth-based contrast features. Thus, the pixel-wise saliency of the likelihood is given by $p(\mathbf{z}_s|\mathbf{x}_c, \mathbf{x}_d) \propto p(\mathbf{z}_s)p(\mathbf{x}_c|\mathbf{z}_s)p(\mathbf{x}_d|\mathbf{z}_s)$.

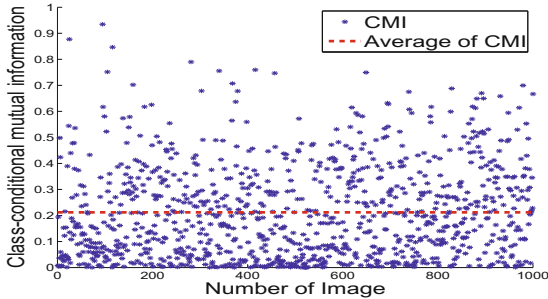


Fig. 4. Visual results for class-conditional mutual information between colour-based contrast features and depth-based contrast features on two RGB-D image datasets.

3.3 Generative Model for Saliency Estimation

Given the graphical model of DMNB for saliency detection shown in Fig. 5, the generative process for $\{x_{1:N}, y\}$ following the DMNB model can be described as follows (Algorithm 1), where $Dir()$ is shorthand for a Dirichlet distribution,

Algorithm 1. Generative process for saliency detection following the DMNB model

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

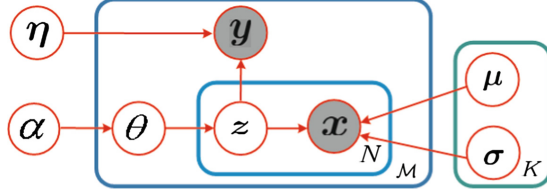


Fig. 5. Graphical models of DMNB for saliency estimation. $\mathbf{y}$ and $\mathbf{x}$ are the corresponding observed states, and z is the hidden variable.

$\text{Mult}()$ is shorthand for a Multinomial distribution, $\mathbf{x}_{1:N} = (\mathbf{x}_c, \mathbf{x}_d)$, $\mathbf{z}_{1:N} = \mathbf{z}_s = (\mathbf{z}_c, \mathbf{z}_d)$, N is the number of features, and $\mathbf{y}$ is the label that indicates whether the pixel is salient or not.

In this work, both the colour- 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 $(\mathbf{x}_{1:N}, \mathbf{y})$ is

$$p(\mathbf{x}_{1:N}, \mathbf{y}|\alpha, \Omega, \eta) = \int p(\theta|\alpha) \left(\prod_{j=1}^N \sum_{z_j} p(z_j|\theta) p(\mathbf{x}_j|z_j, \Omega) p(\mathbf{y}|z_j, \eta) \right) d\theta \quad (6)$$

where θ is the prior distribution over K components, $\Omega = \{(\mu_{jk}, \sigma_{jk}^2), [j]_1^N, [k]_1^K\}$ are the parameters for the distributions of N features respectively, $p(\mathbf{x}_j|z_j, \Omega) \triangleq \mathcal{N}(\mathbf{x}_j|\mu_{jk}, \sigma_{jk}^2)$. In two-class classification, $\mathbf{y}$ is either 0 or 1 generated from $\text{Bern}(\mathbf{y}|\eta)$. Because the DMNB model assumes a generative process for both the labels and features, we use both $\mathcal{X} = \{(\mathbf{x}_{ij}, [i]_1^M, [j]_1^N)\}$ and $\mathcal{Y} = \{(\mathbf{y}_i, [i]_1^M)\}$ 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. In practice, we may find a proper K using the Dirichlet process mixture model (DPMM) [2]. The DPMM thus provides a nonparametric prior for the parameters of a mixture model that allows the number of mixture components to grow as the training set grows, as shown in Fig. 6.

Due to the latent variables, the computation of the likelihood in Eq. 6 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. By a direct application

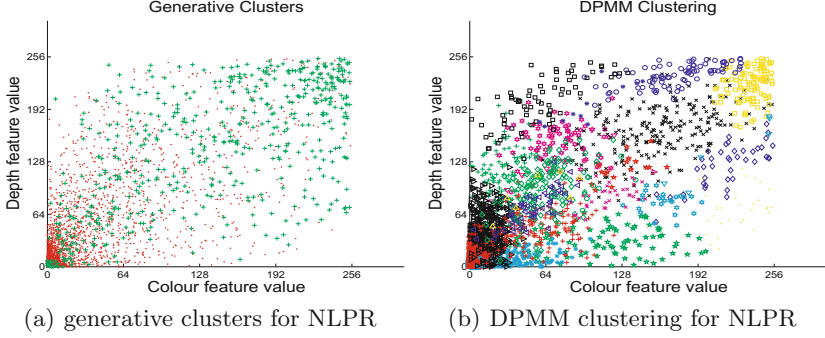


Fig. 6. Visual result for the number of components K in the DMNB model: generative clusters vs DPMM clustering. We find $K = 28$ using DPMM on the NLPR dataset.

of Jensen’s inequality [15], the lower bound to $\log p(\mathbf{y}, \mathbf{x}_{1:N} | \alpha, \Omega, \eta)$ is given by

$$\log p(\mathbf{y}, \mathbf{x}_{1:N} | \alpha, \Omega, \eta) \geq \mathbf{E}_q(\log p(\mathbf{y}, \mathbf{x}_{1:N}, \mathbf{z}_{1:N} | \alpha, \Omega, \eta)) + \mathbf{H}(q(\mathbf{z}_{1:N}, \theta | \gamma, \phi)) \quad (7)$$

Noticing that $\mathbf{x}_{1:N}$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}_{1:N}$, we use a variational distribution:

$$q(\mathbf{z}_{1:N}, \theta | \gamma, \phi) = q(\theta | \gamma) \prod_{j=1}^N q(\mathbf{z}_j | \phi) \quad (8)$$

where $q(\theta, \gamma)$ is a K -dimensional Dirichlet distribution for θ , $q(\mathbf{z}_j | \phi)$ is Discrete distribution for $\mathbf{z}_j$. We use $\mathcal{L}$ to denote the lower bound:

$$\begin{aligned} \mathcal{L} = & \mathbf{E}_q[\log p(\theta | \alpha)] + \mathbf{E}_q[\log p(\mathbf{z}_{1:N} | \theta)] + \mathbf{E}_q[\log p(\mathbf{x}_{1:N} | \mathbf{z}_{1:N}, \gamma)] \\ & - \mathbf{E}_q[\log q(\theta)] - \mathbf{E}_q[\log q(\mathbf{z}_{1:N})] + \mathbf{E}_q[\log p(\mathbf{y} | \mathbf{z}_{1:N}, \eta)] \end{aligned} \quad (9)$$

where $\mathbf{E}_q[\log p(\mathbf{y} | \mathbf{z}_{1:N}, \eta)] \geq \sum_{k=1}^K \phi_k (\eta_k \mathbf{y} - \frac{e^{\eta_k}}{\xi}) - (\frac{1}{\xi} + \log \xi)$ 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:

$$\phi_k \propto e^{(\Psi(\gamma_k) - \Psi(\sum_{l=1}^K \gamma_l) + \frac{1}{N}(\eta_k \mathbf{y}_i - \frac{e^{\eta_k}}{\xi} - \sum_{j=1}^N \frac{(\mathbf{x}_{ij} - \mu_{jk})^2}{2\sigma_{jk}^2}))} \quad (10)$$

$$\gamma_k = \alpha + N\phi_k \quad (11)$$

$$\xi = 1 + \sum_{k=1}^K \phi_k e^{\eta_k} \quad (12)$$

Variational parameters $(\gamma^*, \phi^*, \xi^*)$ from the inference step gives the optimal lower bound to the log-likelihood of $(\mathbf{x}_i, \mathbf{y}_i)$, and maximizing the aggregate lower bound $\sum_{i=1}^{\mathcal{M}} \mathcal{L}(\gamma^*, \phi^*, \xi^*, \alpha, \Omega, \eta)$ over all data points with respect to α , Ω and

Algorithm 2. Variational EM algorithm for DMNB

-
- 1: **repeat**
 - 2: **E-step:** Given $(\alpha^{m-1}, \Omega^{m-1}, \eta^{m-1})$, for each feature value and label, find the optimal variational parameters
 $(\gamma_i^m, \phi_i^m, \xi_i^m) = \arg \max \mathcal{L}(\gamma_i, \phi_i, \xi_i; \alpha^{m-1}, \Omega^{m-1}, \eta^{m-1})$.
Then, $\mathcal{L}(\gamma_i^m, \phi_i^m, \xi_i^m; \alpha, \Omega, \eta)$ gives a lower bound to $\log p(\mathbf{y}_i, \mathbf{x}_{1:N} | \alpha, \Omega, \eta)$.
 - 3: **M-step:** Improved estimate of the model parameters (α, Ω, η) are obtained by maximizing the aggregate lower bound:
 $(\alpha^m, \Omega^m, \eta^m) = \arg \max_{(\alpha, \Omega, \eta)} \sum_{i=1}^N \mathcal{L}(\gamma_i^m, \phi_i^m, \xi_i^m; \alpha, \Omega, \eta)$.
 - 4: **until** $\sum \mathcal{L}(\gamma_i^m, \phi_i^m, \xi_i^m; \alpha^m, \Omega^m, \eta^m) - \sum \mathcal{L}(\gamma_i^{m+1}, \phi_i^{m+1}, \xi_i^{m+1}; \alpha^{m+1}, \Omega^{m+1}, \eta^{m+1}) \leq \text{threshold}$
-

η , respectively, yields the estimated parameters. As for μ , σ and η , we have $\mu_{jk} = \frac{\sum_{i=1}^M \phi_{ik} \mathbf{x}_{ij}}{\sum_{i=1}^M \phi_{ik}}$, $\sigma_{jk} = \frac{\sum_{i=1}^M \phi_{ik} (\mathbf{x}_{ij} - \mu_{jk})^2}{\sum_{i=1}^M \phi_{ik}}$, $\eta_k = \log(\frac{\sum_{i=1}^M \phi_{ik} \mathbf{y}_i}{\sum_{i=1}^M \frac{\phi_{ik}}{\xi_i}})$.

Based on the variational inference and parameter estimation updates, it is straightforward to construct a variant EM algorithm to estimate (α, Ω, η) . Starting with an initial guess $(\alpha^0, \Omega^0, \eta^0)$, the variational EM algorithm alternates between two steps, as follows (Algorithm 2).

After obtaining the DMNB model parameters from the EM algorithm, we can use η to perform saliency prediction. Given the feature $\mathbf{x}_{1:N}$, we have

$$\mathbf{E}[\log p(\mathbf{y} | \mathbf{x}_{1:N}, \alpha, \Omega, \eta)] = \begin{cases} \eta^T \mathbf{E}[\bar{\mathbf{z}}] - \mathbf{E}[\log(1 + e^{\eta^T \bar{\mathbf{z}}})] & \mathbf{y} = 1 \\ 0 - \mathbf{E}[\log(1 + e^{\eta^T \bar{\mathbf{z}}})] & \mathbf{y} = 0 \end{cases} \quad (13)$$

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

4 Experimental Evaluation

4.1 Experimental Setup

Dataset. In this section, we conduct some experiments to demonstrate the performance of our method. We use NLPR dataset¹ to evaluate the performance of the proposed model. The NLPR dataset includes 1000 images of diverse scenes in real 3D environments, where the ground-truth was obtained by requiring five participants to select regions where objects are presented, i.e., the salient regions were marked by hand.

Evaluation Metrics. We introduce two types of measures to evaluate algorithm performance on the benchmark. The first one is the gold standard: F-measure [13]. The second is the receiver operating characteristic (ROC) curve and the

¹ <http://sites.google.com/site/rgbdsaliency>.

area under the ROC curve (AUC). A continuous saliency map can be converted into a binary mask using a threshold, resulting in a pair of precision and recall values when the binary mask is compared against the ground truth. A ROC curve is then obtained by varying the threshold from 0 to 1.

Parameter Setting. To evaluate the quality of the proposed approach, we divided the datasets into two subsets according to their CMI values, and we held out 10% of the data for testing purpose and trained on the remaining 90% whose CMI values are less than CMI threshold τ . As shown in Fig. 4, we compute the CMI for all of the RGB-D images, and the parameter τ is set to 0.35, which is a heuristically determined value. We set the $m = 20$ and $\omega_d = 1.0$ in Eq. 1. We set the $L = 3$, $\omega_c^l = 0.2, 0.3, 0.5$, $\omega_d^l = 0.3, 0.3, 0.4$ and $\sigma^2 = 0.1$ in Eqs. 2, 3 and 4 respectively. We initialize the model parameters using all data points and their labels in the training set in Algorithm 1. In particular, we use the mean and standard deviation of the data points in each class to initialize Ω and $\frac{D_c}{D}$ to initialize α_i , where D_c is the number of data points in class c and D is the total number of data points. For the η in the DMNB model, we run a cross validation by holding out 10% of the training data as the validation set and use the parameters generating the best results on the validation set. We find the initial number of components K using the DPMM.

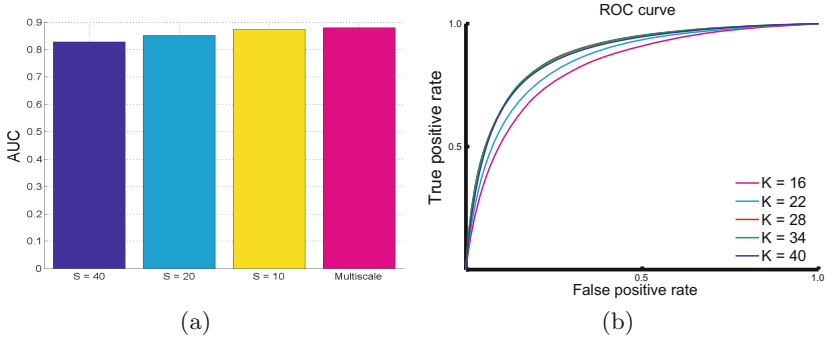


Fig. 7. (a) The effects of the number of scales S on the NLPR dataset. A single scale produces inferior results. (b) The ROC curves for different K components in the DMNB model in terms of the NLPR dataset. The K found using DPMM was adjusted over a wide range to compare the performance. The ROC curves show that changing the parameter K has only a slight effect on the performance.

The Effect of the Parameters. In particular, we performed the experiments while varying S from Eq. 1 and K from Algorithm 1. Figure 7(a) shows typical results when varying S from Eq. 1, which illustrates the AUC obtained from the different numbers of superpixels. If only one scale is used, the results are inferior. This justifies our multi-scale approach.

The parameter K in Algorithm 1 is set according to the training set based on DPMM, as shown in Fig. 6. Figure 7(b) shows the ROC curve from changing

the number of components K in Algorithm 1. Finally, for all the experiments described below, the parameter K was fixed at 28 - no user fine-tuning was done.

4.2 Qualitative Experiment

During the experiments, we compare our algorithm with five state-of-the-art saliency detection methods, among which three are developed for RGB-D images and two for traditional 2D image analysis. One RGB-D method performs saliency detection at Low-level, Mid-level, and High-level stages and is therefore referred to as LMH [12]. One RGB-D method is based on anisotropic centre-surround difference and is therefore denoted ACS [8]. The other RGB-D method exploits global priors, which include the background, depth, and orientation priors to achieve a saliency map and is therefore denoted GP [14]. The two 2D methods are Hemami’s frequency-tuned method [13], which is denoted FT, and the approach from the graph-based manifold ranking [18], which is denoted GMR. For the two 2D saliency approaches, we also add and multiple their results with the DSM produced by our proposed depth feature map; these results are denoted FT+DSM, FT×DSM, GMR+DSM and GMR×DSM. All of the results are produced using the public codes that are offered by the authors of the previously mentioned literature reports.

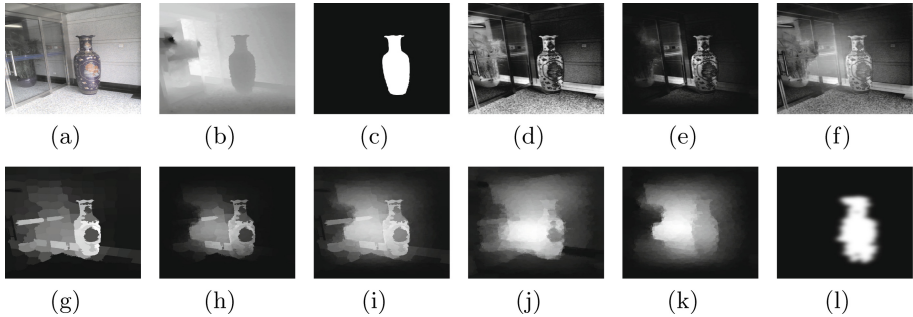


Fig. 8. Visual comparison of the saliency estimations of the different 2D methods with DSM. (a) RGB image, (b) depth image, (c) ground truth, (d) FT, (e) FT×DSM, (f) FT+DSM, (g) GMR, (h) GMR×DSM, (i) GMR+DSM, (j) CSM, (k) DSM, (l) Ours. + indicates a linear combination strategy, and × indicates a weighting method based on multiplication. DSM means depth saliency map, which is produced by our proposed depth feature map. CSM means colour saliency map, which is produced by our proposed colour feature map.

Figure 8 compares our results with FT, FT+DSM, FT×DSM, GMR, GMR+DSM and GMR×DSM. FT detects many uninteresting background pixels as salient because it does not consider any global features. The experiments show that both FT+DSM and FT×DSM are highly improved when incorporated with

the DSM. GMR fails to detect many pixels on the prominent objects because it does not define the pseudo-background accurately. Although the simple late fusion strategy achieves improvements, it still suffers from inconsistency in the homogeneous foreground regions and lacks precision around object boundaries, which may be ascribed to treating the appearance and depth correspondence cues in an independent manner. Our approach consistently detects the pixels on the dominant objects within a Bayesian framework with higher accuracy to resolve the issue.

The comparison of the ACSD, LMH and GP RGB-D approaches is presented in Fig. 9. ACSD works on depth images on the assumption that salient objects tend to stand out from the surrounding background, which takes relative depth into consideration. ACSD generates unsatisfying results without colour cues. LMH uses a simple fusion framework that takes advantage of both depth and appearance cues from the low-, mid-, and high-levels. In [12], the background is nicely excluded; however, many pixels on the salient object are not detected as salient. Ren et al. proposed two priors, which are the normalized depth prior and the global-context surface orientation prior [14]. Because their approach uses the two priors, it has problems when such priors are invalid. We can see that the proposed method can accurately locate the salient objects, and produce nearly equal saliency values for the pixels within the target objects.

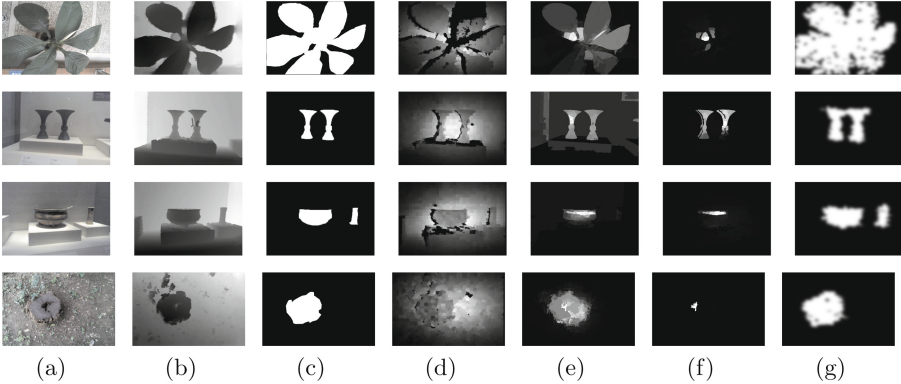


Fig. 9. Visual comparison of the saliency estimations of different 3D methods based on the NLPR dataset. (a) RGB image, (b) Depth image, (c) Ground truth, (d) ACSD, (e) GP, (f) LMH, (g) Ours.

4.3 Quantitative Evaluation

Comparison of the 2D Models Combined with DSM. In this experiment, we first compare the performances of existing 2D saliency models before and after DSM fusing. Figure 10 presents the experimental results, where + and $\times$ denote a linear combination strategy and a weighting method, respectively. From Fig. 10(a), we can see the strong influence of using the DSM on the distribution

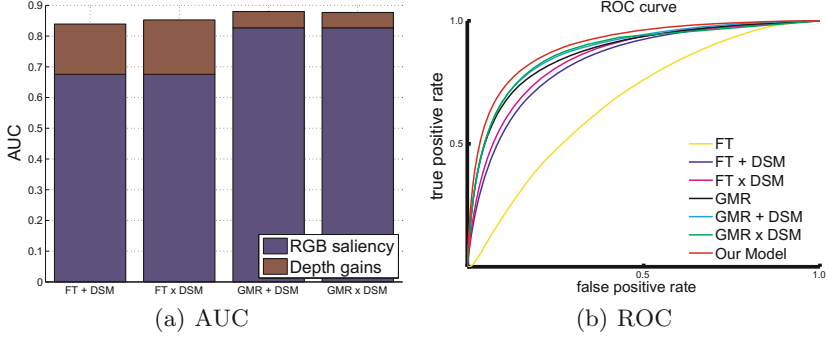


Fig. 10. The quantitative comparisons of the performance of depth cues. + means a linear combination strategy and $\times$ means a weighting method based on multiplication.

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, as shown in Fig. 8. We also provide the ROC curves for several compared methods in Fig. 10(b). The ROC curves demonstrate that the proposed 3D saliency detection model performs better than the compared methods do.

Comparison of 3D Models. In this paper, the GP model, LMH model and ACS D model are classified as depth-saliency models. Figure 11 shows the quantitative comparisons among these method on the constructed RGBD datasets in terms of ROC curves and F-measures. Interestingly, the LMH method, which uses Bayesian fusion to fuse depth and RGB saliency by simple multiplication, has lower performance compared to the GP method, which uses the Markov Random Field model as a fusion strategy, as shown in Fig. 11(a). However, LMH

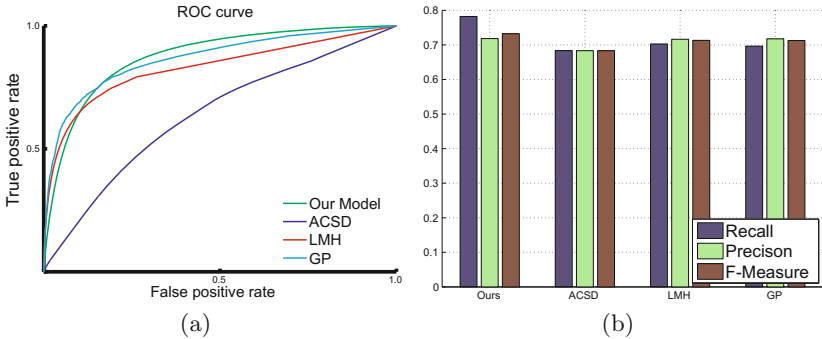


Fig. 11. (a) The ROC curves of different 3D saliency detection models in terms of the NLPR dataset. (b) The F-measures of different 3D saliency detection models when used on the NLPR dataset.

and GP achieve better performances than ACSD by using fusion strategies. The proposed RGBD method is superior to the baselines in terms of all the evaluation metrics. Although the ROC curves are very similar, Fig. 11(b) shows that the proposed method improves the recall and F-measure when compared to LMH and GP. This is mainly because the feature extraction using multi-scale superpixels enhances the consistency and compactness of salient patches.

5 Conclusion

In this study, we proposed a saliency detection model for RGB-D images that considers both colour- and depth-based contrast features with a generative mixture model. The experiments verify that the proposed model's depth-produced saliency can serve as a helpful complement to the existing colour-based saliency models. Compared with other competing 3D models, the experimental results based on a recent eye tracking databases 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.

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

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slc superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2282 (2012)
2. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1**, 121–143 (2006)
3. Borji, A., Itti, L.: State-of-the-art in visual attention modelling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 185–207 (2013)
4. Borji, A., Cheng, M., Jiang, H., Li, J.: Salient object detection: a benchmark. *IEEE Trans. Image Process.* **24**, 5706–5722 (2015)
5. Desingh, K., Madhava, K.K., Rajan, D., Jawahar, C.V.: Depth really matters: improving visual salient region detection with depth. In: *BMVC* (2013)
6. Fang, Y., Wang, J., Narwaria, M., Le Callet, L., Lin, W.: Saliency detection for stereoscopic images. *IEEE Trans. Image Process.* **23**, 2625–2636 (2014)
7. Fang, Y., Lin, W., Fang, Z., Lei, J., Le Callet, P., Yuan, F.: Learning visual saliency for stereoscopic images. In: *ICME* (2014)
8. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic centre-surround difference. In: *ICIP* (2014)
9. Kim, H., Lee, S., Bovik, C.A.: Saliency prediction on stereoscopic videos. *IEEE Trans. Image Process.* **23**, 1476–1490 (2014)
10. Lang, C., Ngugen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: influence of depth cues on visual saliency. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012. LNCS*, vol. 7573, pp. 101–115. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-33709-3_8](https://doi.org/10.1007/978-3-642-33709-3_8)

11. Le Callet, P., Niebur, E.: Visual attention and applications in multimedia technology. *Proc. IEEE* **101**, 2058–2067 (2013)
12. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: a benchmark and algorithms. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014*. LNCS, vol. 8691, pp. 92–109. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-10578-9_7](https://doi.org/10.1007/978-3-319-10578-9_7)
13. Radhakrishna, A., Sheila, H., Francisco, E., Sabine, S.: Frequency-tuned salient region detection. In: *CVPR* (2009)
14. Ren, J., Gong, X., Yu, L., Zhou, W.: Exploiting global priors for RGB-D saliency detection. In: *CVPRW* (2015)
15. Shan, H., Banerjee, A., Oza, N.C.: Discriminative mixed-membership models. In: *ICDM* (2009)
16. Wang, J., Fang, Y., Narwaria, M., Lin, W., Le Callet, P.: Stereoscopic image retargeting based on 3D saliency detection. In: *ICASSP* (2014)
17. Wu, P., Duan, L., Kong, L.: RGB-D salient object detection via feature fusion and multi-scale enhancement. In: Zha, H., Chen, X., Wang, L., Miao, Q. (eds.) *CCCV 2015*. CCIS, vol. 547, pp. 359–368. Springer, Heidelberg (2015). doi:[10.1007/978-3-662-48570-5_35](https://doi.org/10.1007/978-3-662-48570-5_35)
18. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph based manifold ranking. In: *CVPR* (2013)
19. Zhang, Y., Jiang, G., Yu, M., Chen, K.: Stereoscopic visual attention model for 3D video. In: Boll, S., Tian, Q., Zhang, L., Zhang, Z., Chen, Y.-P.P. (eds.) *MMM 2010*. LNCS, vol. 5916, pp. 314–324. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-11301-7_33](https://doi.org/10.1007/978-3-642-11301-7_33)
20. Zhu, L., Cao, Z., Fang, Z., Xiao, Y., Wu, J., Deng, H., Liu, J.: Selective features for RGB-D saliency. In: *CAC* (2015)

Computer Vision – ACCV 2016

13th Asian Conference on Computer Vision, Taipei,
Taiwan, November 20–24, 2016, Revised Selected
Papers, Part V

Lai, S.-H.; Lepetit, V.; Nishino, K.; Sato, Y. (Eds.)

2017, XIII, 434 p. 170 illus., Softcover

ISBN: 978-3-319-54192-1