HOSO: Histogram Of Surface Orientation for RGB-D Salient Object Detection

David Feng, Nick Barnes, Shaodi You
Data61, CSIRO; RSE, Australian National University
{david.feng, nick.barnes, shaodi.you}@data61.csiro.au

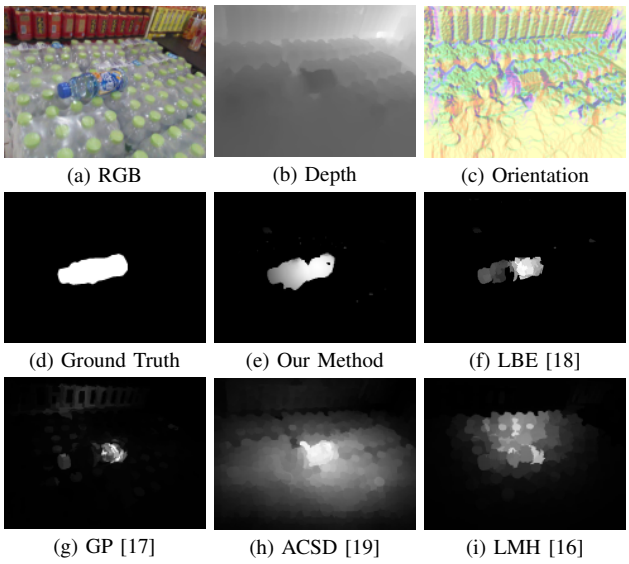
***Abstract*—Salient object detection using RGB-D data is an emerging field in computer vision. Salient regions are often characterized by an unusual surface orientation profile with respect to the surroundings. To capture such profile, we introduce the histogram of surface orientation (HOSO) feature to measure surface orientation distribution contrast for RGB-D saliency. We propose a new unified model that integrates surface orientation distribution contrast with depth and color contrast across multiple scales. This model is implemented in a multi-stage saliency computation approach that performs contrast estimation using a kernel density estimator (KDE), estimates object positions from the low-level saliency map, and finally refines the estimated object positions with a graph cut based approach. Our method is evaluated on two RGB-D salient object detection databases, achieving superior performance to previous state-of-the-art methods.**

I. Introduction

The saliency of a scene component, such as a pixel, patch, or object, refers to how much it stands out with respect to its surroundings. While the majority of saliency methods aim to model and predict human eye fixation points on images [1], [2], recently there has been an increasing number of works on detection and segmentation of salient objects and regions [3], [4]. This is referred to as salient object detection. Salient object detection has many applications, including compression [5], resizing [6], thumbnailing [7], and adaptive image display for small devices [8].

Saliency is typically computed by measuring contrast at a local [1] or global scale [9]. Previous work predominantly operates on RGB input, computing contrast from appearance-based features such as colour, edges, and texture [10], [11]. However, the increasing availability of depth sensing technology has encouraged the exploration of structural features, facilitating improved performance when the foreground and background have similar appearance. Relatively little work has taken advantage of 3D data for saliency computation, and consequently there is scope for better models for the effective representation and integration of structural information. Many RGB-D saliency approaches simply use depth values to modulate RGB saliency maps [12], [13], [14], [15], or measure depth contrast [16], [17] in a similar way to RGB saliency methods. These methods produce false positives when the background is closer than the object or has relatively high depth contrast.

We make the observation that, in terms of depth, saliency consists of not just how close an object is, but that it has an unusual profile of surface orientation with respect to its local region or to other parts of the scene, or has an overall orientation that is unusual. For example the corner between a wall and floor, an obstacle on the ground, or clutter in a tidy space. Surface orientation contrast thus offers a promising structural measure of saliency at multiple scales that operates independently to depth, and can be used to complement depth-based contrast. However, while first order surface properties are commonly used for tasks such as 3D object recognition [20], incorporation of surface orientation for saliency detection has received much less attention [21], [22], [17].



(a) RGB (b) Depth (c) Orientation (d) Ground Truth (e) Our Method (f) LBE [18] (g) GP [17] (h) ACSD [19] (i) LMH [16]

Fig. 1. Saliency output on an image with low foreground depth contrast. Our method measures surface orientation distribution contrast to effectively identify foreground structure. Output is shown for three state-of-the-art methods Global Priors [17], Anisotropic Center Surround Difference [19], and Low Medium High [16].

In this paper, we present a new unified model for salient object detection that integrates surface orientation, depth, and color contrast at multiple scales. Unlike previous approaches, we integrate both orientation and depth contrast in a consistent framework, taking advantage of the complementary information they offer. Surface orientation contrast in existing methods is computed only at a global scale [22] or only with respect to similar regions in the image [21], which can lead to an increased number of false positives and false negatives respectively. Instead, our unified model performs a multi-scale measurement of orientation contrast, based on the intuition that

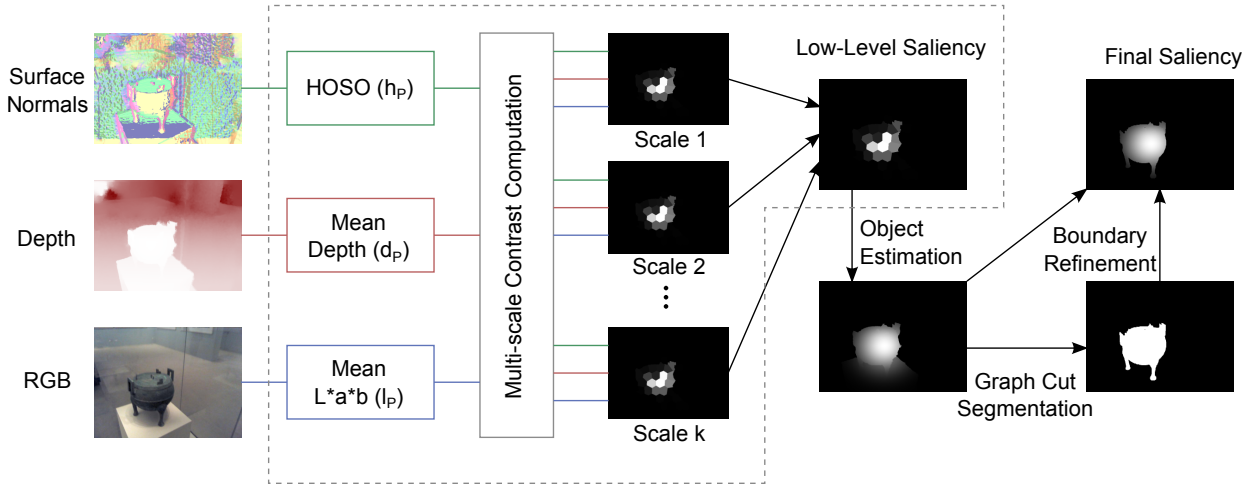


Fig. 2. Overview of the main stages of our method. We measure multi-scale contrast of orientation, depth, and colour to obtain a low-level saliency map. We use the low-level saliency to estimate and object map, and then perform boundary refinement using a graph cut based approach.

salient objects are likely to remain distinct across multiple scales. Unlike purely global formulations of surface orientation, our method captures locally unusual surface orientation profiles that characterize many types of structurally interesting regions, such as wall-floor edge boundaries. Furthermore, while previous work represents regions using mean orientation, we introduce the histogram of surface orientation (HOSO) feature for RGB-D saliency to capture the distribution of surface normals, providing a robust and descriptive characterisation of the underlying region. While histogram based representations of first order image properties are common in feature detection and matching [23], their use and effectiveness is unexplored for RGB-D salient object detection.

Contrast computation in our system is performed using a Gaussian KDE [16]. This allows the integration of different feature types during computation, rather than fusing individually computed feature contrast maps [24], [25], to better exploit the strong complementarities between surface orientation, depth, and color. The incorporation of multiple discriminative features tends to produce a precise but sparse low-level saliency map. We post-process this map using object map estimation and boundary refinement procedures to obtain a uniform saliency response across detected objects.

We evaluate our model on two recently proposed RGB-D datasets for salient object detection, achieving superior performance to state-of-the-art methods on both datasets. Furthermore, we demonstrate the contribution of each feature type and computation stage to the overall performance of our model. Note that we do not provide comparison with recent deep learning based systems, e.g. [26]. Although these systems produce good performance, the focus of this paper is the investigation of low-level structure-based saliency cues. It has been shown that effective saliency cues contribute improved performance in standard deep learning frameworks [27].

The main contributions of this paper are: insight that surface orientation distribution contrast provides valuable cues for determining locally unusual structure that is indicative of salient objects, and a novel feature, HOSO, for capturing these cues; proposal of the first unified multi-scale saliency detection system incorporating surface orientation, colour, and depth contrast; and demonstration of the effectiveness of HOSO and our system through state-of-the-art results on two datasets.

II. Related Work

RGB-D saliency computation is a rapidly growing field, driven by a wide variety of applications including stereoscopic rendering [13], robotic grasping [28], and structure based image retargeting [14].

Early works in depth saliency integrate depth as an additional channel into a classic RGB saliency framework [1]. Ouerhani and Hugli [24] explore which features to incorporate into the framework, selecting depth over depth gradient and curvature. Frintrop et al. [25] apply the framework to depth and intensity to reduce the search space for object detection. These methods fuse individually computed saliency maps from feature type, and do not exploit complementary cues between features during the contrast computation stage.

Based on findings that closer objects are more likely to appear salient in the human visual system [29], a number of existing techniques modulate RGB saliency maps using image depth values. Zhang et al. [12] scales the output of [1] with depth to identify regions of interest in stereoscopic video. Similarly, Chamaret et al. [13] weights an RGB saliency map with depth to identify salient regions for adaptive rendering on a 3D display. In addition to linear depth scaling, reweighting RGB saliency based on a Gaussian distribution over depth has also been explored. Lin et al. [14] use a Gaussian distribution centered on the local maximum of a depth histogram to reweight an RGB saliency map. Tang et al. [15] attenuate saliency using a Gaussian based on the depths of salient regions, filtering object patches for salient object detection.

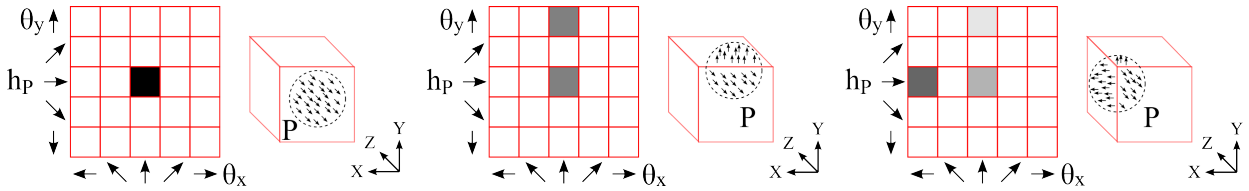


Fig. 3. Illustration of the HOSO feature for three different patches on a cube, with camera direction along the Z axis. Given an image patch P, surface normals within P are parameterized by their 2D orientation and binned into a 5×5 histogram h_P.

Some approaches aim to directly model the influence of depth on human visual attention by learning a non-linear depth prior from eye tracking data. Lang et al. [30] model the joint density between depth distribution and saliency response using a Gaussian Mixture Model learned from 3D eye tracking data, while Wang et al. [31] apply a learned mapping between saliency and Difference of Gaussians response on the depth image. This type of approach does not fully exploit the interaction between RGB and depth feature types. Furthermore, with the exception of [31], these approaches do not consider relative depth, and work best when the range of salient objects is closer than the background, which is a strong assumption.

Many RGB-D saliency techniques compute saliency based on global depth contrast. Niu et al. [32] extend [9] with disparity contrast for salient object detection in stereo image pairs. Fang et al. [33] measure global contrast over depth, colour, luminance, and texture to predict gaze in stereoscopic images. Peng et al. [16] compute saliency using depth and colour at both global and local scales. While the majority of previous work takes absolute depth differences when measuring depth contrast, some methods modulate depth contrast by the relative depth between regions. Cheng et al. [34] use global color and depth contrast for salient object detection, with increased depth contrast from pop-out regions. Ju et al. [19] compute saliency based on the average distance to minimum values encountered along a set of scanlines. This approach is sensitive to noise and the placement of the scan lines, which only provide a partial sample of the neighborhood. Feng et al. [18] propose an enclosure-based formulation of depth saliency. While this method alleviates some of the problems of depth contrast methods, it does not take into account unusual profiles of surface shape when predicting saliency.

Previous approaches are generally unlikely to produce good results when the foreground has low depth contrast to the background. Surface orientation contrast is an alternative structural cue that is useful for identifying salient regions. Potapova et al. [28] show that the surface orientation difference between objects and the supporting surface is an important factor in determining locations that are suitable for robot grasping. Surface orientation is employed as an application-specific cue, whereas we use it as part of a comprehensive model to measure general structural saliency. Ciptadi et al. [21] compute surface orientation contrast between a target region and the K most similar regions using vectorized patches of surface normals. The selection of nearest neighbours in feature space reduces the discriminability of the feature. Desingh et al. [22] represent surface patches with histograms of pairwise angular distances between point-wise normals. Contrast is computed at a global scale and does not capture local distinctness, which is particularly informative in structural analysis. Ren et al. [17] use orientation as a prior, marking surfaces perpendicular to the camera axis as more salient. This method produces false positives for background regions that face the camera, and false negatives for objects not facing the camera or with complex surfaces. Unlike previous work, we examine orientation distribution contrast at multiple scales, facilitating detection of a wider range of salient region types such as structural edges. We directly represent surface orientation distribution using a 2D histogram representation, HOSO, in order to give a rich description of the underlying surface that is robust to noise. Furthermore, previous approaches incorporate either depth contrast or surface orientation contrast to compute saliency. We present a model that exploits contrast with respect to both surface orientation and depth features, taking advantage of the strong complementary relationship between these features.

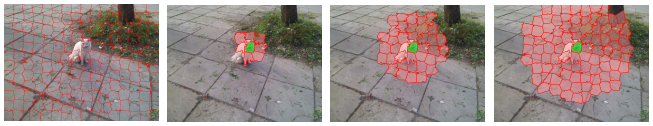
III. HOSO Feature

Our saliency model includes the distribution of surface orientation as a feature, based on the observation that salient objects are more likely to contain orientation structure that contrasts with the surroundings.

We aim to identify structurally salient regions based on their surface orientation profile. In order to perform this task, the representation of patch-level surface orientation must be descriptive as well as robust to noise. First-order surface properties are particularly sensitive to sensor noise, which can impact the performance of a saliency system if used directly [24].

Rather than representing a patch with a single orientation value as in previous work, we use a histogram to capture the distribution of patch normals as the core orientation feature. This provides a more detailed representation of the underlying surface shape, and improves the capacity of the feature for distinguishing locally unusual structure. Furthermore, histograms are more robust to sensor noise than mean values.

The HOSO feature is computed as follows. First, point-wise normals are estimated from the depth image using PCA with an 11×11 support. The large support size was chosen to further alleviate the effect of noise. We parameterize normals by their 2D orientation (θ_x, θ_y) to avoid wrap around issues and facilitate uniform quantization. Normal orientations in a



(a) Segmentation (b) $r = 60$ (c) $r = 140$ (d) $r = 220$

Fig. 4. Example image segmentation and illustration of contexts at multiple scales. The candidate patch P is shown in green. The context sets ϕ_P are shown in red, containing patches within distance r of P.

patch P are binned into a normalized 2D histogram h_P with N bins in each dimension (see Figure 3). The bin mapping function for both dimensions is equal to

$$b(\theta) = \left\lfloor N \cdot \frac{\theta}{\pi} \right\rfloor . \tag{1}$$

Thus, each point in P with orientation (θ_x, θ_y) increments bin $(b(\theta_x), b(\theta_y))$ of h_P. The value of bin (i, j) of h_P is thus given by

$$h_P(i,j) = \frac{1}{\mathrm{card}(P)} \left| \{(\theta_x, \theta_y) \in P \mid b(\theta_x) = i \ \wedge \ b(\theta_y) = j\} \right| . \tag{2}$$

The dissimilarity of the surface orientation distributions of two patches is measured using the Bhattacharyya distance $\mathrm{dist}_B(\cdot,\cdot)$ between their HOSO features:

$$\mathrm{dist}(P, Q) = \mathrm{dist}_B(h_P - h_Q) . \tag{3}$$

1) Patch-level Feature: Given an image patch P, we compute saliency based on the HOSO feature h_P, in addition to the mean depth d_P and mean Lab colour l_P of the patch [16].

IV. Saliency Detection System

The pipeline of our method consists of three major stages, as shown in Figure 2. First, a low-level saliency map is computed from surface orientation, depth, and colour contrast at multiple scales. We use Gaussian Kernel Density Estimation [16] to measure contrast and integrate the different features during the contrast computation stage. This is followed by an object estimation stage, which uniformly highlights foreground regions identified in the low-level saliency map. Each pixel is assigned a probability that it belongs to the foreground, computed using a Gaussian model of the object constructed from the low-level saliency map. In the final step, the boundaries of the estimated object map are refined with a graph cut based approach [35].

A. Low-level Saliency

This section describes our method for computing the low-level saliency map from raw patch level features. We first segment the input image into patches using SLIC [36]. The low-level saliency $S(P)$ of a patch P is formulated as the product of a contrast measurement function across multiple scales, such that:

$$S(P) = \prod_{\phi \in \Phi_P} C(P, \phi) , \tag{4}$$

where $\Phi_P = \{\phi_P^r | r \in \mathbb{R}\}$ denotes the scale space of P, and C measures the contrast between P and its context ϕ_P^r, which consists of all other patches within a radius of r (see Figure 4). That is, $\{\phi_P^r = Q | \|c_P - c_Q\|_2 < r\}$, where c_P and c_Q are patch centroids.

The contrast between a patch P and its context ϕ is measured by estimating the probability $p(P|\phi)$ that P comes from the distribution defined by ϕ in feature space, as in [16]. A low value of p implies that P is unlikely to belong to ϕ, and has a high contrast. The contrast measurement function is thus given by:

$$C(P, \phi) = -\log\left(p(P|\phi)\right) , \tag{5}$$

We use a kernel density estimator to compute p [16]. However, in addition to mean depth and colour, we extend the density estimation to include the HOSO feature, incorporating differences of surface orientation distributions into the density function. If a patch has unusual surface orientation profile compared to its surroundings, such as a ball resting on the ground, then it will have a low estimated probability of being part of the context distribution, and consequently a high saliency score. On the other hand if a patch has an almost identical surface orientation profile to its surroundings, such as a patch on a planar surface, then the estimated probability density function will have a high value at HOSO feature of the point, leading to a low saliency score.

The probability density estimation is thus given by:

$$p(P|\phi) = \frac{1}{\mathrm{card}(\phi)} \sum_{Q \in \phi} K_h(h_P, h_Q) K_d(d_P, d_Q) K_l(l_P, l_Q) , \tag{6}$$

where $K_h(\cdot,\cdot)$, $K_d(\cdot,\cdot)$, and $K_l(\cdot,\cdot)$ are the kernel components corresponding to surface orientation distribution, mean depth, and mean colour respectively.

We define the surface orientation distribution component as a Gaussian kernel with bandwidth $\sigma_{h,P,Q}$:

$$K_h(h_P, h_Q) = \exp\left(-\frac{\mathrm{dist}_B(h_P, h_Q)^2}{2\sigma_{h,P,Q}^2} \right). \tag{7}$$

The estimate is obtained by measuring the Bhattacharyya distance from the HOSO feature h_P of the candidate patch to the density function of the set of patches Q of its context.

As in [34] we observe that objects that are closer than their surroundings are more likely to be salient. We aim to limit the contribution of context patches with lower depth than the candidate patch, since in these cases the candidate patch is more likely to be background. This is achieved by scaling the base KDE bandwidth σ_h, depending on whether the candidate patch is in front of the context patch:

$$\sigma_{h,P,Q} = \begin{cases} \sigma_h & \text{if } d_P > d_Q \\ \alpha \cdot \sigma_h & \text{otherwise.} \end{cases} \tag{8}$$

Setting $\alpha > 1$ increases the bandwidth and reduces the influence of the context patch.

The depth and colour Gaussian kernels $K_d(\cdot,\cdot)$ and $K_l(\cdot,\cdot)$ are defined similarly, using the Euclidean distance between feature values instead of Bhattacharya distance, and with respective bandwidths $\sigma_{d,P,Q}$ and $\sigma_{l,P,Q}$.

The contribution of each feature when computing low-level saliency on RGBD-1000 is shown in Figure 5. Note that incorporating orientation with depth results in a larger improvement than using colour and depth, validating the incorporation of surface orientation as a structural saliency feature. The combination of all three features gives the best performance, indicating that each feature contributes positively to the final result.

1) Priors: The KDE contrast computation process only incorporates relative depth information. Numerous studies report that absolute depth is also a critical component of pre-attentive visual attention, with closer objects more likely to appear salient to the human visual system [29], [37], [30]. Accordingly, scaling saliency by depth is a common refinement step in previous work [32], [34], [19], [15], [22], [13], [12]. We perform absolute depth reweighting, dividing patch saliency with mean patch depth.

The tendency of the human visual system to fixate on objects near the center of an image is well known [38]. A large number of existing saliency methods incorporate a spatial prior to model this effect [39], [40], [41]. Similar to [39], we apply a Gaussian $G(P)$ to reweight patch saliency based on the distance between the centroid of patch P and the image center.

The low-level saliency map with depth and center bias is thus given by:

$$S_b(P) = S(P) \cdot \frac{1}{d_P} \cdot G(P). \tag{9}$$

B. Salient Object Map Estimation

The low-level saliency computation stage tends to produce saliency maps characterized by sparse high-saliency patches. The multiplicative aggregation of complementary and discriminative features can result in a low overall saliency score for a patch if one feature is assigned a low contrast. Thus, only a few highly distinct points produce a high saliency score in the low-level map.

Ensuring a consistently strong saliency response across entire objects is a fundamental objective in salient object detection [4]. We use the low-level saliency map described in Section IV-A to build a Gaussian model of the object based on image position and depth, from which each pixel is assigned a score reflecting the probability that it is part of the salient object. This is implemented in a similar way to the high-level object bias enhancement performed in [16], but with mean and variance computation modified to account for a saliency map with sparse regions of high response.

The probability that a pixel belongs to a salient object is computed based on the estimated location and size of the object, formulated as a Gaussian model H. Let (a_x, a_y, a_z) denote the image position and depth of pixel a. Then the x component of the model is given by

$$H(a_x) = \exp\left[-\left(\frac{a_x - \mu_x}{2\sigma_x}\right)^2 - \left(\frac{a_y - \mu_y}{2\sigma_y}\right)^2 - \left(\frac{a_z - \mu_z}{2\sigma_z}\right)^2 \right] \tag{10}$$

where (μ_x, μ_y, μ_z) is the expected object center, and $(\sigma_x, \sigma_y, \sigma_z)$ is the expected object size. We will now detail the computation of μ_x and σ_x; the y and z components of the model are computed in a similar manner.

Let $S_b(a)$ denote the low-level saliency of a, obtained by propagating patch saliency to member pixels. In order to handle a saliency map with sparse regions of high response, we set the expected object center μ_x along the x dimension as the weighted mean over all pixels:

$$\mu_x = \frac{\sum_{a \in I} S_b(a) \cdot a_x}{\sum_{a \in I} S_b(a)}. \tag{11}$$

The expected object size along the x dimension is based on the unbiased estimate of the weighted sample variance of the image:

$${\sigma_x}^2 = \frac{\sum_{a \in I} \left(S_b(a) \left(a_x - \mu_x\right)^2 \right) \sum_{a \in I} S_b(a)}{\sum_{a \in I} S_b(a) - \sum_{a \in I} S_b(a)^2}. \tag{12}$$

Since low-level saliency may not be high at all the extremities of the object, we scale the variance estimate with a constant factor v_0.

C. Boundary Refinement

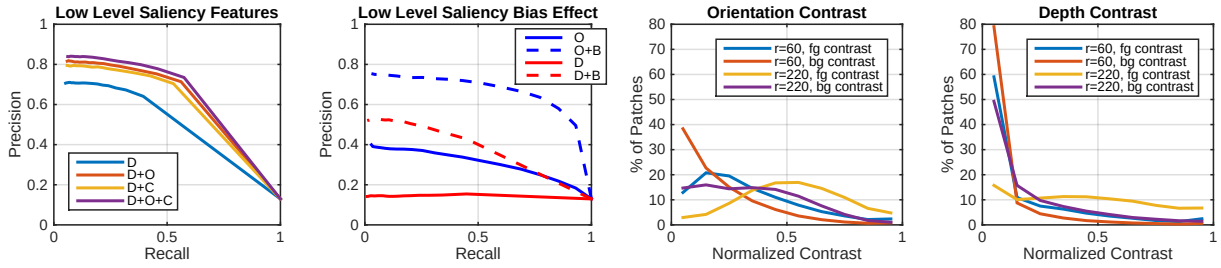
The estimated object map H from the previous stage can contain inaccurate foreground boundaries, particularly when the object occupies a similar depth range to nearby background. Boundary refinement is a common post-processing step employed in existing salient object detection systems (e.g. [34], [16], [35]). We use the graph cut based saliency refinement method described by [35] to obtain object boundaries based on appearance information. The foreground model is initialized with a binary mask obtained by applying a threshold t_0 to H. The output graph cut segmentation mask A is used to prune non-foreground areas from H. The final pixel-wise saliency is thus given by

$$S(a) = A(a) \cdot H(a). \tag{13}$$

V. Experiments

We evaluate our method on two recently proposed datasets for RGB-D salient object detection. The first is RGBD1000 [16], which was introduced to address the lack of a large dataset with depth information for salient object detection. It contains 1000 images featuring diverse scene and object types, with low depth and colour contrast between the foreground and background. We also report the performance of our method on the NJUDS2000 salient object detection dataset [19], containing 2000 disparity images computed from stereo image pairs.

Our method is compared with three state-of-the-art RGB-D salient object detection systems: Low-Medium-High Saliency (LMH) [16] proposed by the authors the RGBD1000 dataset, Anisotropic Center Surround Difference (ACSD) [19], from the authors of the NJUDS2000 dataset, and Global Prior saliency (GP) [17]. We also include comparisons to two top



(a) (b) (c) (d)

Fig. 5. (a) Comparison of low-level saliency results on RGBD1000 when incorporating various patch feature combinations. D=mean depth, O=surface orientation histogram (HOSO), C=mean L*a*b colour. This shows that the effect of surface orientation is large if there is a lack of colour information, for example in a low contrast environment or under low lighting conditions. In the case that colour is incorporated, using HOSO provides additional improvement. (b) The effect of center and depth bias on low-level saliency using our orientation feature (O) and a custom implementation of the low-level depth saliency term of DCS (D). Analysis of contrast for (c) surface orientation and (d) mean depth features at multiple scales on RGBD1000, displaying the percentage of foreground (fg) and background (bg) patches that exhibit the normalized contrast values with respect to a neighbourhood of radius r.

ranking 2D saliency algorithms according to a recent survey [4]: DSR [42], and DRFI [10].

We examine the effect of center and depth bias on low-level orientation contrast saliency compared to the low-level depth saliency from [16]. Note that since [16] is only available as a single executable, we use a custom implementation of the low-level saliency which omits center and depth prior application. We also measure the performance of the low-level and object estimation stages of our framework, and examine the contribution of the different feature types used in our low-level saliency computation method.

A. Contrast Computation Scales

We perform an analysis of structural feature contrast at different scales for foreground identification on the dataset, in order to help inform scale selection for our saliency system.

Figures 5c and 5d show that for a small scale size, foreground patches typically have higher contrast with orientation than depth. In particular, a large number of foreground patches have low local depth contrast, suggesting that depth contrast provides poor discriminability at a local scale, and that orientation contrast is more likely to distinguish foreground regions when the context size is small. However, background regions tend to have greater orientation contrast for larger scales than depth contrast, suggesting that the former is not suited for large context sizes. Based on these observations, we omit depth and orientation when computing contrast with small and large context sizes respectively.

B. Implementation Details

In the experiments, we measure contrast across three scales, $R = \{60, 140, 220\}$. These scales were selected to produce small, medium, and large contexts for each patch. The KDE bandwidths in Equation 6 of the mean depth and Lab colour features were set to $\sigma_d^2 = \sigma_l^2 = 0.025$. For orientation, bandwidths of $\sigma_h^2 = 0.1$ for scale 60 and $\sigma_h^2 = 0.3$ for scale 140 were found to work well.

The expected object size scale v_0 was set to $v_0 = 3$. We found setting $N = 5$ bins for each histogram dimension achieves a good balance between descriptiveness, robustness, and efficiency for HOSO. The threshold used for graph cut initialisation was set to $t_0 = 0.8$. Our unoptimized implementation takes approximately 7 seconds per 640×480 image running on a 2.6GHz i5 processor with 8GB of RAM.

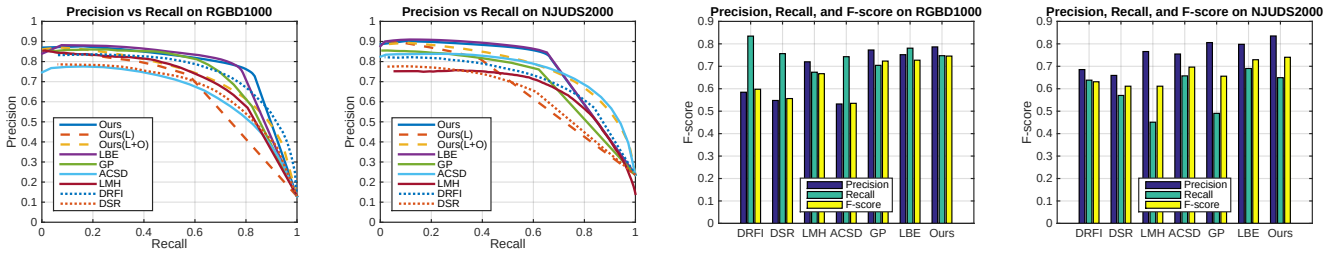
C. Evaluation Metrics

Performance is evaluated through the precision-recall curve and mean F-score, the F_β measure with $\beta = 0.3$ emphasizing precision [3]. The F-score is computed from the saliency output using an adaptive threshold equal to twice the mean of the image [3].

VI. Results

Our method produces a superior F-score compared to all other methods on both datasets, as seen in Figures 6c and 6d. Furthermore, our method achieves a consistently high performance across the two datasets whereas most other methods tend to favour one or the other.

Figure 6a shows that our system achieves higher precision than most other methods at comparable recall rates on RGBD1000. The increased precision is most apparent at just under 0.8 recall. At this point our method is able to identify a larger portion of foreground regions than other methods without affecting precision. Similarly, Figure 6b shows that our method has the highest precision tied with LBE up to just under 0.7 recall. Note that while the precision recall curves of our method are comparable to LBE, our method measures a different type of structural cue which corresponds more closely to salient object shape, as demonstrated by our superior F-scores in Figures 6c and 6d. Figures 6a and 6b also show the contribution of each computation stage in our framework. We see from the figure that applying the object estimation map significantly improves results compared to the low-level saliency map, in particular boosting recall as we expect. The application of boundary refinement subsequently increases the precision of the estimated object map. This pattern of improvement follows the aim of each stage: identification of salient



(a) (b) (c) (d)

Fig. 6. Quantitative comparisons of performance over RGBD1000 and NJUDS2000 datasets. Ours(L) denotes our low level saliency map, and Ours(L+O) denotes our estimated object map.

regions, expansion of candidate regions to cover foreground objects, and boundary refinement to remove background.

We plot the precision-recall for our low-level saliency map using different feature combinations in Figure 5a. As expected, using individual features gives relatively low scores. Combining depth and orientation produces better results than combining depth and colour, which highlights the complementary nature of the two structural features. The relatively high performance of this pairing suggests that orientation may be used as an alternative when colour is not available. The best performance is observed when using all three feature types, demonstrating that each feature offers distinct information that is extracted effectively in our framework. As shown in Figure 5b, the low-level surface orientation saliency of our method outperforms the low-level depth saliency of [16] both with and without the bias terms. This demonstrates that surface orientation contrast is a more reliable indicator of foreground than depth contrast, particularly near image boundaries.

VII. Conclusion

In this paper, we present a unified model for salient object detection that exploits orientation, depth, and colour contrast at multiple scales, using a novel orientation distribution feature HOSO for RGB-D saliency. Low-level saliency computation is performed with a KDE and used to estimate object locations, which are refined with a graph cut based approach. Feature scales are selected based on an analysis of contrast, with orientation suited for small scales and depth applied to larger scales. Experimental results show an improvement in performance compared to the previous state-of-the-art.

References

[1] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. PAMI (1998) 1254–1259

[2] Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR. (2007) 1–8

[3] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR. (2009) 1597–1604

[4] Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. TIP **24** (2015) 5706–5722

[5] Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. TIP **19** (2010) 185–198

[6] Achanta, R., Susstrunk, S.: Saliency detection for content-aware image resizing. In: ICIP. (2009)

[7] Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: ICCV. (2009) 2232–2239

[8] Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. Multimedia systems **9** (2003) 353–364

[9] Cheng, M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.: Global contrast based salient region detection. PAMI **37** (2015) 569–582

[10] Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: CVPR. (2013) 2083–2090

[11] Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: ICCV. (2013) 1529–1536

[12] Zhang, Y., Jiang, G., Yu, M., Chen, K.: Stereoscopic visual attention model for 3d video. In: Advances in Multimedia Modeling. (2010) 314–324

[13] Chamaret, C., Godeffroy, S., Lopez, P., Le Meur, O.: Adaptive 3d rendering based on region-of-interest. In: IS&T/SPIE Electronic Imaging. (2010) 75240V–75240V

[14] Lin, W.Y., Wu, P.C., Chen, B.R.: Image retargeting using depth enhanced saliency. 3DSA (2013) 1–4

[15] Tang, Y., Tong, R., Tang, M., Zhang, Y.: Depth incorporating with color improves salient object detection. The Visual Computer (2015) 1–11

[16] Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: ECCV. (2014)

[17] Ren, J., Gong, X., Yu, L., Zhou, W., Yang, M.Y.: Exploiting global priors for rgb-d saliency detection. In: CVPRW. (2015) 25–32

[18] Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for rgb-d salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2343–2350

[19] Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE (2014) 1115–1119

[20] Tang, S., Wang, X., Lv, X., Han, T.X., Keller, J., He, Z., Skubic, M., Lao, S.: Histogram of oriented normal vectors for object recognition with a depth sensor. In: Asian conference on computer vision, Springer (2012) 525–538

[21] Ciptadi, A., Hermans, T., Rehg, J.M.: An in depth view of saliency. In: BMVC. (2013) 9–13

[22] Desingh, K., K, M.K., Rajan, D., Jawahar, C.: Depth really matters: improving visual salient region detection with depth. In: BMVC. (2013)

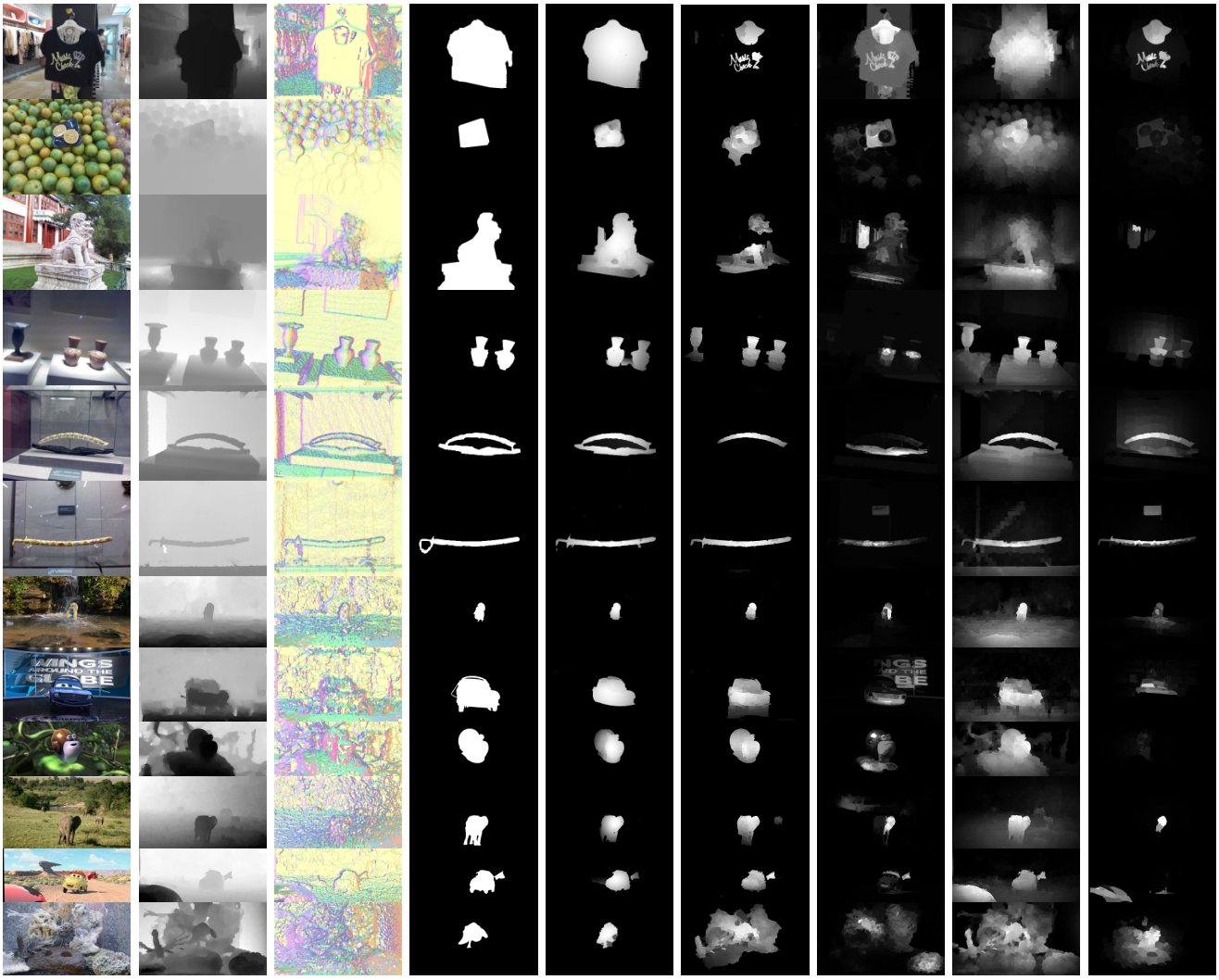
[23] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893

[24] Ouerhani, N., Hügli, H.: Computing visual attention from scene depth. In: CVPR. (2000) 375–378

[25] Frintrop, S., Nüchter, A., Surmann, H.: Visual attention for object recognition in spatial 3d data. In: WAPCV. (2004) 168–182

[26] Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. IEEE Transactions on Image Processing **26** (2017) 2274–2285

[27] Shigematsu, R., Feng, D., You, S., Barnes, N.: Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features. arXiv preprint arXiv:1705.03607 (2017)



(a) RGB (b) Depth (c) Ori (d) G. T. (e) Ours (f) Ours (g) [17] (h) [19] (i) [16]

Fig. 7. Comparison of saliency maps generated by state of the art systems. Our method is shown with GP [17], ACSD [19], and LMH [16]. G. T. denotes Ground Truth and Ori shows surface orientation.

[28] Potapova, E., Zillich, M., Vincze, M.: Learning what matters: combining probabilistic models of 2d and 3d saliency cues. In: Computer Vision Systems. (2011) 132–142

[29] Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience **5** (2004) 495–501

[30] Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: influence of depth cues on visual saliency. In: ECCV. (2012) 101–115

[31] Wang, J., DaSilva, M.P., LeCallet, P., Ricordel, V.: Computational model of stereoscopic 3d visual saliency. TIP **22** (2013) 2151–2165

[32] Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR. (2012) 454–461

[33] Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W.: Saliency detection for stereoscopic images. TIP **23** (2014) 2625–2636

[34] Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: ICIMS. (2014) 23

[35] Mehrani, P., Veksler, O.: Saliency segmentation based on learning and graph cut refinement. In: BMVC. (2010) 1–12

[36] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. PAMI **34** (2012) 2274–2282

[37] Theeuwes, J., Atchley, P., Kramer, A.F.: Attentional control within 3-d space. Journal of Experimental Psychology: Human Perception and Performance **24** (1998) 1476

[38] Tseng, P.H., Carmi, R., Cameron, I.G., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. Journal of Vision **9** (2009) 4

[39] Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: CVPR. (2013) 1139–1146

[40] Shen, X., Wu, Y.: A unified approach to salient object detection via low rank matrix recovery. In: CVPR. (2012)

[41] Duan, L., Wu, C., Miao, J., Qing, L., Fu, Y.: Visual saliency detection by spatially weighted dissimilarity. In: CVPR. (2011)

[42] Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction. In: ICCV. (2013)