# Salient Object Detection in Images by Combining Objectness Clues in the RGBD Space

François Audet [1,a], Mohand Said Allili [1,b], Ana-Maria Cretu [1,c]

[1]Université du Québec en Outaouais, QC, Canada.
[a] francois.audet02@uqo.ca
[b] mohandsaid.allili@uqo.ca
[c] ana-maria.cretu@uqo.ca

**Abstract.** We propose a multi-stage approach for salient object detection in natural images which incorporates color and depth information. In the first stage, color and depth channels are explored separately through objectness-based measures to detect potential regions containing salient objects. This procedure produces a list of bounding boxes which are further filtered and refined using statistical distributions. The retained candidates from both color and depth channels are then combined using a voting system. The final stage consists of combining the extracted candidates from color and depth channels using a voting system that produces a final map narrowing the location of the salient object. Experimental results on real-world images have proved the performance of the proposed method in comparison with the case where only color information is used.

## 1 Introduction

Salient object detection is one of the important problems for object recognition and image understanding [3]. It consists of localizing the most informative objects or regions in images [5, 12]. Saliency detection methods rely either on local or global contrast estimation [3]. Local contrast based methods [11] assume that regions which stand out from their neighborhoods have high saliency values. These methods are more suitable to highlight salient object boundaries instead of entire objects. Global contrast based methods [1, 14] express rarity of regions compared to the overall image in terms of global statistics [5]. They are better at highlighting entire salient regions. However, they are less accurate to detect large-sized objects due to the fact the object statistics dominate the global statistics of the image [1].

Most of existing methods for salient object detection show good performance in general when applied to simple scenarios of images containing single objects situated against uniform and non-cluttered backgrounds [3, 5]. However, when images contain several objects and/or cluttered backgrounds, the performance of these methods drastically decreases [6, 9]. This stems mainly from the assumption that salient objects stand out most of the time from the background, where

color contrast is sufficient for their differentiation from the rest of the image. However, this assumption is no longer valid when images contain multiple objects and/or cluttered backgrounds where objects can hardly be discriminated using color information.

Recently, another trend of methods aiming at detecting directly objects in the image is gaining popularity [4, 8, 15]. These methods apply objectness measures in the color space with the aim of detecting bounding boxes surrounding objects [16]. In a nutshell, objectness quantifies how likely it is for an image window to contain an object of any class (e.g., car, dogs, etc.) as opposed to backgrounds, such as grass and water [2]. Objectness-based methods, as their saliency based counterparts, can locate efficiently objects standing out from their immediate surrounding. However, these methods usually produce a huge number of false positives consisting of parts of objects or details of the background. Therefore, an appropriate filtering is required for these methods for better object detection using more discriminative information, such as object appearance, depth information and a priori knowledge such object location, size and shape.

To exploit color and depth information together, Peng et al. [10] have proposed some techniques for locating salient objects in images in the RGBD space. This approach employs local and global contrast measures from color and depth to discriminate the salient objects from the image background. The approach supposes the salient object is centered in the image and relies on depth maps for locating the boundary of salient objects. However, images often times contain multiple objects in the scene, while depth information can miss some parts of objects or merge them with the background, which may cause awed saliency maps.

In this paper, we propose a method combining color and depth information to extract objectness clues for salient object detection and localisation in natural images. Using the algorithm in [16], we first generate separate object proposals for Lab channels of the image and its smoothed version. In addition, given that highest ranked object proposals do not correlate sometimes with salient objects, a filtering step is applied to prune unlikely candidates based on various spatial characteristics. The depth channel is preprocessed using vertical and horizontal scans for removing false object candidates caused by projection effects and noise. Finally, to combine the bounding boxes obtained from all the channels, we propose a voting system taking into account the image layout, object location and size in the image. Experimental results on a set of real-world images have shown that the proposed method yields better localisation of salient objects compared to recent methods in the literature.

This paper is organized as follows: Section 2 presents our algorithm for regions of interest proposal. Section 3 presents some experimental results validating our approach. We then end the paper with a conclusion.

## 2 The proposed approach

An outline of the algorithm for salient object detection in the RGBD space is shown in Figure 1. The algorithm input consists of the color image $I$ and a depth map of the image $D$. The RGB color image is first transformed into the Lab space which is similar to the human vision perception. Object proposals are then generated from each channel and their combination is performed using a weighted voting and saliency priors. On the other hand, the depth map $D$ is pre-processed to filter projection and noise effects and generate object proposals. Finally, a procedure is proposed to combine color and depth object candidates to form the final bounding box where the salient object is the most likely located. Details of the different steps are presented in the following sections.

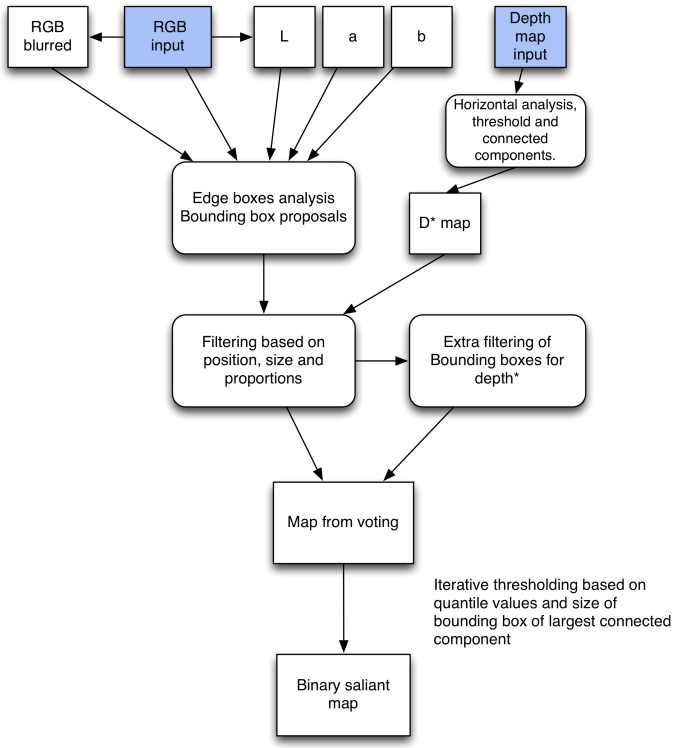



Fig. 1: Outline of the proposed salient object detection algorithm.

### 2.1 Object proposal in the RGB space

As can be seen in Figure 1, from the RGB image we obtain a blurred version using a 2D Gaussian smoothing kernel with standard deviation $\sigma = 5$. The smoothed version generally allows to reduce the number of returned object proposals corresponding to local object details. The RGB image is also decomposed into normalized Lab channels. The two color maps (RGB and its smoothed version) as well as the three grayscale maps (L, a, b) are passed as inputs to the object proposal algorithm [16], which returns for each map an array of bounding boxes ranked by decreasing objectness score.

Figure 2 first row illustrates some cases where the algorithm puts too much emphasis on minute details, for example around the digits of the clock in the first image. The bounding boxes are represented in decreasing order of objectness, with red showing the highest objectness score, followed by green, blue, yellow and finally magenta, representing the lowest objectness score. The ground truth bounding box is shown in white. By performing a filtering step, as described next, we obtain substantial improvements as shown in the second row of the figure. For the filtering step, we gathered statistics about the bounding boxes of the ground truth contained in 848 images. The following features have been computed :

- $\{f_1, f_2\}$: normalized coordinates $(x, y)$ of the centroid of the bounding boxes relative to the width and height of each image.
- $\{f_3, f_4\}$: relative width and height of the bounding boxes.
- $\{f_5\}$: the aspect ratio of the bounding boxes (i.e. width / height).
- $\{f_6, f_7, f_8, f_9\}$: relative distances from each of the 4 corners of the bounding boxes.

Each of these features has been analyzed for all images grouped together, as well as separately for landscape (width/ height $\geq 1$) and portrait images (width/ height $< 1$). The marked distinction observed between the two orientations concerns the aspect ratio (width/ height) of their bounding boxes, which was found to be on average 0.66 and 1.30 for portrait and landscape, respectively, which incidentally is close to the aspect ratio of the whole images (0.75 and 1.33, respectively). Figure 3 shows the histograms of the first five features. Most of these histograms can be approximated by normal distributions. Thus, using the features computed from the ground truth bounding boxes, we obtain two lists $L_p$ (for portrait) and $L_l$ (for landscape) of 9 normal distributions.

The calculated distributions will serve as a prior knowledge to filter future bounding boxes that will returned by the algorithm in the different channels. Given a bounding box candidate $b$, the 9 features are first calculated. Let $\phi_i$ be the normal distribution of feature $f_i$, and $p_i$ the cumulative probability of its value measured on the bounding box $b$. The bounding box is rejected if one of the following conditions is satisfied:

1. $\exists f_i : p_i(1 - p_i) < \delta$.

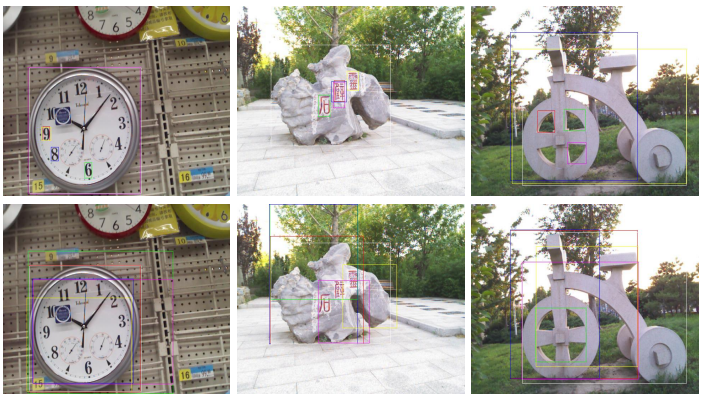


Fig. 2: Effect of filtering on object proposals. The first and second rows represent object proposals before and after filtering, respectively.

2. $\sum_{i=1}^{9} p_i(1-p_i) < \eta$

where $\delta$ and $\eta$ are experimental thresholds. Note that the first condition is equivalent to having an observation which exceeds the statistical *p-value* at risk $\alpha\%$ of the feature $f_i$. A typical value for the risk is $\alpha = 0.02\%$ in which case $\delta = 0.0196$. The second condition aims to enforce global conformity of the bounding box to the prior knowledge. Based on the sum of the p-values for the ground truth (see figure 3), a value of 0.9 for $\eta$ will rejects roughly the 1% least probable candidates.

### 2.2 Object proposal in the depth space

In contrast to color information, depth carries information about the relative distance of objects to the camera. Therefore, objects and converging surfaces to the camera can often have similar depth values and be confounded if depth thrsholding is used. To exploit depth information for more precise salient object detection, we should first perform a transformation to remove non-desirable projection effects and noise caused by errors in depth estimation. Given a depth map $D$ of size $h \times w$, we perform a horizontal scanning using dynamic thresholding to flat out the depth of converging surfaces and noise.

Let $\sigma_l$ by the standard deviation computed for a given line $l$ in $D$ and let $D^*$ be the transformed map. If $\sigma_l < \epsilon$ (i.e., absence of contrast), the line is not processed any further, and its pixels are set to 0 in $D^*$. Otherwise, for each column index $j$ of line $l$, we test if the depth value at the column belongs to an object or the background. For this purpose, we use linear regression to

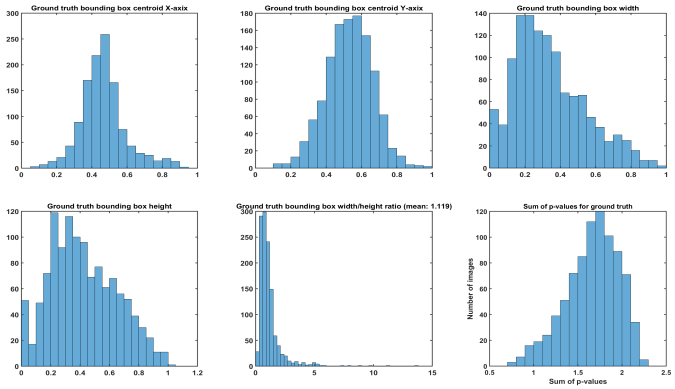


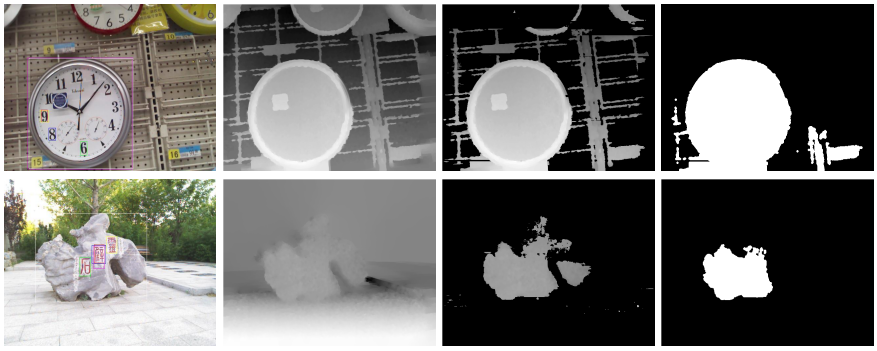

Fig. 3: The histograms of the five first features in the dataset, as well as the p-values. Top row, from left to right, we show histograms of centroid positions x and y and relative width. The second row shows the relative height, aspect ratio and sum of p-values.

approximate the map of the $l$-th line. Let $y(j) = a \times j + D(l, 1)$ be the resulted regression line with slope $a$. Then, if $D(l, j) > y(j)$ and $D(l, j)$ constitutes a local maximum in the line, then $D^*(l, j) \leftarrow D(l, j)$, otherwise $D^*(l, j) \leftarrow 0$. Figure 4 shows some examples of depth filtering, where the salient object has been clearly isolated from the background.

### 2.3 The final object proposal algorithm

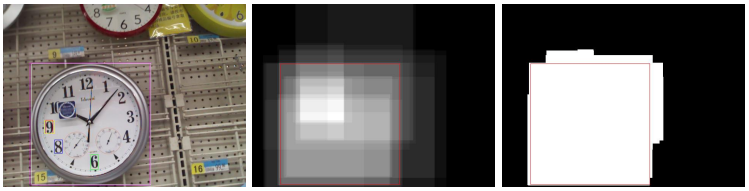
From the filtering process as described above, we retain the five most highly ranked bounding boxes for each channel, and therefore a total of at most 30 bounding boxes (in some instances there is less than 5 bounding boxes that are proposed for $D^*$). Each bounding box votes with 1 for the pixels it contains and the final objectness score is normalized for each pixel. Then, an initial threshold of 25% is used to segment the objectness to obtain a map BC. The size of BC is compared to the ground truth statistics and if its size is above the 95-th percentile, a new map is generated by increasing the threshold by 5%, and the process is repeated until BC area is at a 95-th percentile or below, or that the threshold of 50% has been reached.

The first image of Figure 5 shows a typical map obtained after the votes of each bounding box. The second image shows in white the area thresholded; the ground truth is represented by the red rectangle.



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

Fig. 4: Effect of filtering and thresholding on depth map: (a) and (b) represent the RGB and its depth map, respectively, (c) and (d) represent the transformed depth maps after applying the first and second thresholds, respectively.



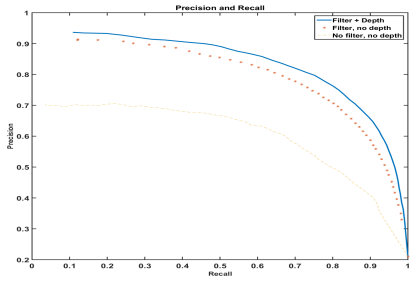
(a) (b) (c)

Fig. 5: Voting and thresholding for the image of the clocks.

## 3 Experimental results

The performance of the proposed model is evaluated quantitatively on the dataset provided by Peng et al. [10]. It is a diversified data set consisting of 1000 images with depth map and ground-truth annotations. It has more than 400 types of common objects under various illumination conditions. Some images contain several salient objects and were not taken into consideration for the tests. Moreover, color and depth areas corresponding to salient objects have wide distributions, and their area varies from 16% and 80% of the image, with a majority under 50%. We evaluated the performance of our method using the *precision*, *recall* and $F_\alpha$ metrics (with $\alpha = 1$). These are calculated by comparing the bounding box of the ground truth in each image with the calculated bounding box obtained by our algorithm.

In Figure 6, we show ROC curve of the obtained object maps using compared methods, as well as best values for precision, recall and $F_\alpha$ measures. We can note that our ROC curve is better than the other methods. We can see also the improvements in the recall and the $F_\alpha$ values that have been obtained. This gives our method the advantage of better localizing all the parts of salient objects. In fact, as shown in the results, using only color information as in [16] yields more precision and less recall, which can happen when objects are partially detected. Therefore, our method is better positioned when it is aimed to detect objects for high-level applications such as object recognition and segmentation.




| Measure | objectness using [16] | Our method |
|---|---|---|
| Precision | 0.7166 | 0.7620 |
| Recall | 0.7899 | 0.8014 |
| F1-measure | 0.7024 | 0.7371 |

Fig. 6: Comparative results of precision, recall and $F_\alpha$.

## 4 Conclusion

We have proposed a multi-stage approach for salient object detection in natural images by combining color and depth information. Color and depth channels

are explored through objectness-based measures to detect potential regions containing salient objects. These regions are then filtered and combined using a voting system to produce a final map narrowing the location of the salient object. Experimental results on real-world images have proved the performance of the proposed method in comparison with the case where only color information is used.

## References


1. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-Tuned Salient Region Detection. *IEEE Conf. on Comp. Vision and Pattern Recognition*, 1597-1604, 2009.
2. B. Alexe, T. Deselaers and V. Ferrari. Measuring the Objectness of Image Windows. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2189-2202, 2012.
3. A. Borji, M-M. Cheng, H. Jiang and J. Li. Salient Object Detection: A Benchmark. *IEEE Trans. on Image Processing*, 24(12):5706-5722, 2015.
4. M-M. Cheng, Z. Zhang, W-Y. Lin and P.H.S. Torr. BING: Binarized Normed Gradients for Objectness Estimation at 300fps. *IEEE Conf. on Computer Vision and Pattern Recognition*, 3286-3293, 2014.
5. M-M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S-M. Hu. Global Contrast Based Salient Region Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(3): 569-582, 2015.
6. I. Filali, M.S. Allili, N. Benblidia. Multi-Scale Salient Object Detection Using Graph Ranking and Global-Local Saliency Refinement. *Signal Processing: Image Communication*, 47: 380-401, 2016.
7. V. Gopalakrishnan, Y. Hu and D. Rajan. Random Walks on Graphs for Salient Object Detection in Images. *IEEE T. on Image Processing*, 19(12):3232-3242, 2010.
8. S. He and R. Lau. Oriented Object Proposals. *IEEE Int'l Conf. on Computer Vision*, 280-288, 2015.
9. Z. Liu, W. Zou and O. Le Meur. Saliency Tree: A Novel Saliency Detection Framework. *IEEE Trans. on Image Pocessing*, 23(5):1937-1952, 2014.
10. H. Peng, B. Li, W. Xiong, W. Hu and R. Ji. RGBD Salient Object Detection: A Benchmark and Algorithms. *Europrean Conf. on Computer Vision*, 92109, 2014.
11. H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *IEEE Conf. on Comp. Vision and Pattern Recognition Workshops*,9(12): 45-52, 2009.
12. J. Wang, H. Jiang, Z. Yuan. M-M. Cheng, X. Hu, N. Zheng. Salient Object Detection: A Discriminative Regional Feature Integration Approach. *IEEE Int'l Journal of Computer Vision*, 1-18, DOI 10.1007/s11263-016-0977-3, 2017.
13. C. Yang, L. Zhang, H. Lu and X. Ruan. Saliency Detection via Graph-Based Manifold Ranking. *IEEE Conf. on Comp. Vision and Pattern Recognition*, 3166-3173, 2013.
14. L. Zhang, Z. Gu and H. Li. SDSP: A Novel Saliency Detection Method by Combining Simple Priors. *IEEE Int'l Conf.on Image Processing*, 171-175, 2013.
15. Z. Zhang and P.H.S. Torr. Object Proposal Generation Using Two-Stage Cascade SVMs. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 38(1): 102-115, 2016.
16. C. Zitnick and P. Dollár. Edge boxes: Locating Object Proposals from Edges. *European Conf. on Computer Vision*, 391-405, 2014.