7-2015

Salient Object Detection via Augmented Hypotheses

Tam Nguyen
University of Dayton, tnguyen1@udayton.edu

Jose Sepulveda
National University of Singapore

Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Salient Object Detection via Augmented Hypotheses

Tam V. Nguyen, Jose Sepulveda
Department for Technology, Innovation and Enterprise
Singapore Polytechnic
{nguyen_van_tam, sepulveda_jose}@sp.edu.sg

Abstract

In this paper, we propose using *augmented hypotheses* which consider objectness, foreground and compactness for salient object detection. Our algorithm consists of four basic steps. First, our method generates the objectness map via objectness hypotheses. Based on the objectness map, we estimate the foreground margin and compute the corresponding foreground map which prefers the foreground objects. From the objectness map and the foreground map, the compactness map is formed to favor the compact objects. We then derive a saliency measure that produces a pixel-accurate saliency map which uniformly covers the objects of interest and consistently separates fore- and background. We finally evaluate the proposed framework on two challenging datasets, MSRA-1000 and iCoSeg. Our extensive experimental results show that our method outperforms state-of-the-art approaches.

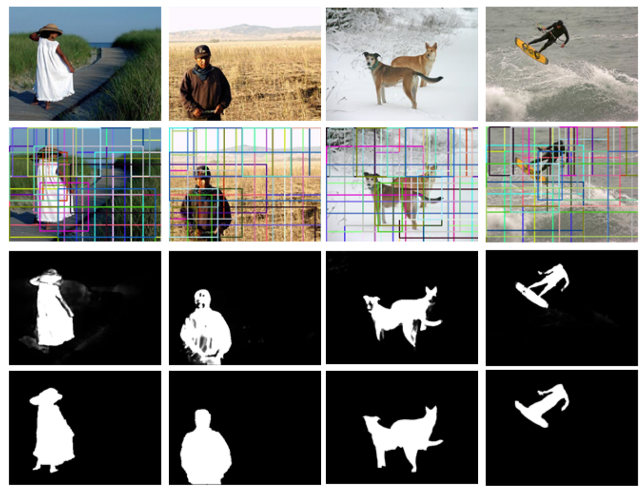


Figure 1: From top to bottom: original images, the objectness hypotheses, results of our saliency computation, and ground truth labeling. For a better viewing, only 40 object hypotheses are displayed in each image.

1 Introduction

The ultimate goal of salient object detection is to search for salient objects which draw human attention on the image. The research has shown that computational models simulating low-level stimuli-driven attention [Koch and Ullman, 1985; Itti *et al.*, 1998] are quite successful and represent useful tools in many practical scenarios, including image resizing [Achanta *et al.*, 2009], attention retargeting [Nguyen *et al.*, 2013a], dynamic captioning [Nguyen *et al.*, 2013b], image classification [Chen *et al.*, 2012] and action recognition [Nguyen *et al.*, 2015]. The existing methods can be classified into biologically-inspired and computationally-oriented approaches. On the one hand, works belonging to the first class [Itti *et al.*, 1998; Cheng *et al.*, 2011] are generally based on the model proposed by Koch and Ullman [Koch and Ullman, 1985], in which the low-level stage processes features such as color, orientation of edges, or direction of movement. One example of this model is the work by Itti et al. [Itti *et al.*, 1998], which use a Difference of Gaussians approach to evaluate those features. However, the resulting saliency maps are generally blurry, and often overemphasize small, purely local features, which renders this approach less useful for applications such as segmentation, detection, etc [Cheng *et al.*, 2011].

On the other hand, computational methods relate to typical applications in computer vision and graphics. For example, frequency space methods [Hou and Zhang, 2007] determine saliency based on spectral residual of the Fourier transform of an image. The resulting saliency maps exhibit undesirable blurriness and tend to highlight object boundaries rather than its entire area. Since human vision is sensitive to color, different approaches use local or global analysis of color contrast. Local methods estimate the saliency of a particular image region based on immediate image neighborhoods, e.g., based on dissimilarities at the pixel-level [Ma and Zhang, 2003] or histogram analysis [Cheng *et al.*, 2011]. While such approaches are able to produce less blurry saliency maps, they are agnostic of global relations and structures, and they may also be more sensitive to high frequency content like image edges and noise. In a global manner, [Achanta *et al.*, 2009] achieves globally consistent results by computing color dissimilarities to the mean image color. Murray et al. [Mur-

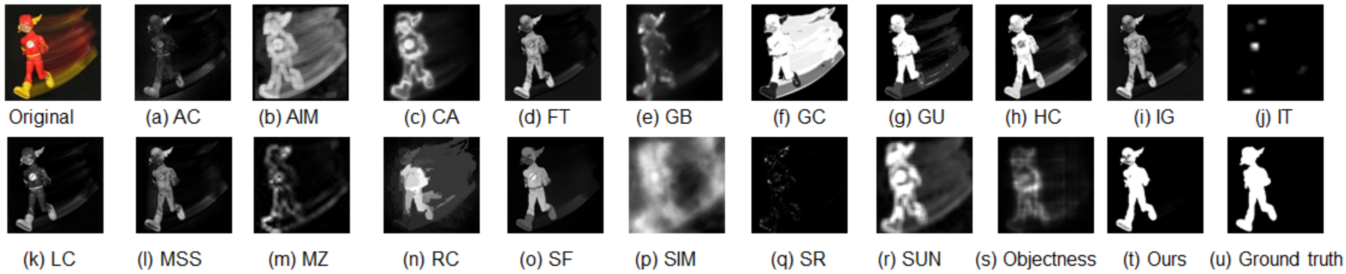


Figure 2: Saliency maps computed by our proposed AH method (t) and state-of-the-art methods (a-r), salient region detection (AC [Achanta *et al.*, 2008]), attention based on information maximization (AIM [Bruce and Tsotsos, 2005]), context-aware (CA [Goferman *et al.*, 2010]), frequency-tuned (FT [Achanta *et al.*, 2009]), graph based saliency (GB [Harel *et al.*, 2006]), global components (GC [Cheng *et al.*, 2013]), global uniqueness (GU [Cheng *et al.*, 2013]), global contrast saliency (HC and RC [Cheng *et al.*, 2011]), spatial temporal cues (LC [Zhai and Shah, 2006]), visual attention measurement (IT [Itti *et al.*, 1998]), maximum symmetric surround (MSS [Achanta and Süsstrunk, 2010]), fuzzy growing (MZ [Ma and Zhang, 2003]), saliency filters (SF [Perazzi *et al.*, 2012]), induction model (SIM [Murray *et al.*, 2011]), spectral residual (SR [Hou and Zhang, 2007]), saliency using natural statistics (SUN [Zhang *et al.*, 2008]), and the objectness map (s). Our result (t) focuses on the main salient object as shown in ground truth (u).

ray *et al.*, 2011] introduced an efficient model of color appearance, which contains a principled selection of parameters as well as an innate spatial pooling mechanism. There also exist different patch-based methods which estimate dissimilarity between image patches [Goferman *et al.*, 2010; Perazzi *et al.*, 2012]. While these algorithms are more consistent in terms of global image structures, they suffer from the involved combinatorial complexity, hence they are applicable only to relatively low resolution images, or they need to operate in spaces of reduced image dimensionality [Bruce and Tsotsos, 2005], resulting in loss of salient details.

Despite many recent improvements, the difficult question is still whether "the salient object is a real object". That question bridges the problem of salient object detection into the traditional object detection research. In the latter object detection problem, the efficient sliding window object detection while keeping the computational cost feasible is very important. Therefore, there exist numerous objectness hypothesis generation methods proposing a small number (e.g. 1,000) of category-independent hypotheses, that are expected to cover all objects in an image [Lampert *et al.*, 2008; Alexe *et al.*, 2012; Uijlings *et al.*, 2013; Cheng *et al.*, 2014]. Objectness hypothesis is usually represented as a value which reflects how likely an image window covers an object of any category. Lampert et al. [Lampert *et al.*, 2008] introduced a branch-and-bound scheme for detection. However, it can only be used to speed up classifiers that users can provide a good bound on highest score. Alexe et al. [Alexe *et al.*, 2012] proposed a cue integration approach to get better prediction performance more efficiently. Uijlings et al. [Uijlings *et al.*, 2013] proposed a selective search approach to get higher prediction performance. However, these methods are time-consuming, taking 3 seconds for one image. Recently, Cheng et al. [Cheng *et al.*, 2014] presented a simple and fast objectness measure by using binarized normed gradients features which compute the objectness of each image window at any scale and aspect ratio only requires a few bit operations. This method can be run 1,000+ times faster than popular alternatives.

In this work, we investigate applying objectness to the problem of salient object detection. We utilize the object hypotheses from the objectness hypothesis generation augmented with foreground and compactness constraint in order to produce a fast and high quality salient object detector. The exemplary object hypotheses and our saliency prediction are shown in the second and the third row of Figure 1, respectively. As we demonstrate in our experimental evaluation, each of our individual measures already performs close to or even better than some existing approaches, and our combined method currently achieves the best ranking results on two public datasets provided by [Achanta *et al.*, 2009; Batra *et al.*, 2010]. Figure 2 shows the comparison of our saliency map to other baselines in literature. The main contributions of this work can be summarized as follows.

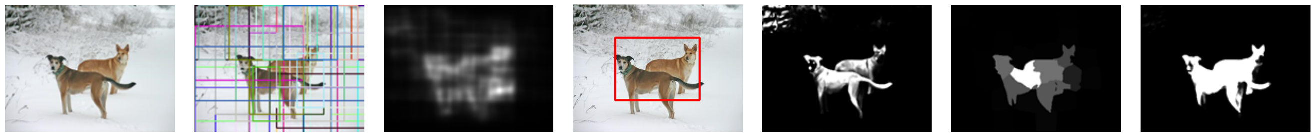
- We conduct the comprehensive study on how the objectness hypotheses affect the salient object detection.
- We propose the foreground map and compactness map, derived from the objectness map, which can cover both global and local information of the saliency object.
- Unlike other works in the literature, we evaluate our proposed method on two challenging datasets in order to know the impact of our work in different settings.

2 Methodology

In this section, we describe the details of our augmented hypotheses (AH), and we show how the objectness measures as well as the saliency assignment can be efficiently computed. Figure 3 illustrates the overview of our processing steps.

2.1 Objectness Map

In this work, we extract object hypotheses from the input image to form the objectness map. We assume that the salient objects attract more object hypotheses than other parts in the



(a) Original image (b) Hypotheses (c) Objectness (d) Margin (e) Foreground (f) Compactness (g) Saliency map

Figure 3: Illustration of the main phases of our algorithm. The object hypotheses are generated from the input image. The objectness map is later formed by accumulating all hypotheses. The foreground map is then created from the difference between the pixel's color and the background color obtained following the estimated margins. We then oversegment the image into superpixels and compute the compactness map based on the spatial distribution of superpixels. Finally, a saliency value is assigned to each pixel.

image. As aforementioned, the objectness hypothesis generators propose a small number n_p (e.g. 1,000) of category-independent hypotheses, that are expected to cover all objects in an image. Each hypothesis P_i has coordinate (l_i, t_i, r_i, b_i), where l_i, t_i are the coordinate of the top left point, whereas r_i, b_i are the coordinate of the bottom right point. Here, we formulate each hypothesis $P_i \in \mathbb{R}^{H \times W}$, where H and W are the height and the width of the input image $\boldsymbol{I}$, respectively. The value of each element $P_i(x, y)$ is defined as:

$$P_i(x,y) = \begin{cases} 1 & \text{if } t_i \leq x \leq b_i \text{ and } l_i \leq y \leq r_i \\ 0 & otherwise \end{cases} . \quad (1)$$

The objectness map is constructed by accumulating all object hypotheses:

$$OB(x,y) = \sum_{i=1}^{n_p} P_i(x,y). \quad (2)$$

The objectness map is later rescaled into the range [0..1]. We observe that the objectness map discourages the object parts locating close to the image boundary. Thus we extend the original image by embedding an image border with the size is 10% of the original image's size. The addition image border is filled with the mean color of the original image. We perform the hypothesis extraction and compute the objectness map similar to the aforementioned steps. The final objectness map is cropped to the size of the original image. Figure 4 demonstrates the effect of our image extension and the shrinkage of the objectness map.

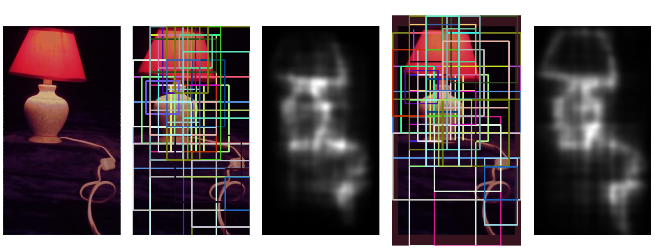


Figure 4: From left to right: the original image, the object hypotheses and the corresponding objectness map, the extended object hypotheses and the corresponding objectness map.

2.2 Foreground Map

The salient object tends to be distinctive from its surrounding context. Thus, we aim to model the background which can facilitate the object localization. In particular, the foreground map is computed by finding the difference between the color of the original image and the background image. In order to model the background, we first localize the salient object by the margin shown as the red rectangle in Fig 3d. To this end, we compute the accumulate objectness level by four directions n_r, namely, top, bottom, left, and right. For each direction, the accumulated objectness level is bounded by a threshold θ. To boost this process, we utilize the integral image [Viola and Jones, 2001] computed from the objectness map. Finally, there are n_r, 4 in this work, corresponding rectangles surrounding the salient object. Each bounding rectangle r_i is represented by its mean color μ_{r_i}. The foreground value computed for each pixel (x, y) is computed as follows,

$$FG(x,y) = \prod_{i=1}^{n_r} \|\boldsymbol{I}(x,y) - \boldsymbol{\mu}_{r_i}\|, \quad (3)$$

where $\boldsymbol{I}(x, y)$ is the color vector of the pixel (x, y).

2.3 Compactness Map

The foreground map prefers the color of the salient object of the foreground. Unfortunately, it also favors the similar color appearing in the background. We observe that though the colors belonging to the background will be distributed over the entire image exhibiting a high spatial variance, the foreground objects are generally more compact [Perazzi *et al.*, 2012]. Therefore, we compute the compactness map in order to remove the noise from the background. First, we compute the centroid of interest $(x_c, y_c) = (\frac{\sum_{(x,y)} x \times OF(x,y)}{\sum_{(x,y)} OF(x,y)}, \frac{\sum_{(x,y)} y \times OF(x,y)}{\sum_{(x,y)} OF(x,y)})$, where the objectness-foreground value $OF(x, y) = OB(x, y) \times FG(x, y)$. Intuitively, the pixel close to the centroid of interest tends to be more salient, whereas the farther pixels tend to be less salient. In addition, the saliency value of a certain pixel reduces if the path between the centroid and that pixel contains many low saliency values. The naive method is to compute the path from the centroid of interest to other pixels. However, it is time-consuming to perform this task in the pixel-level. Therefore, we transform it to superpixel-level. The image is oversegmented into superpixels, and the OF value of a superpixel

Algorithm 1 Superpixel compactness computation

$\boldsymbol{l} = \{v_c\}$.
$\boldsymbol{c} = \mathbf{0} \in \mathbb{R}^{n_{sp}}$.
$\boldsymbol{t} = \emptyset$
while $l \neq \emptyset$ **do**
    for each vertex $\boldsymbol{v}_i$ in $\boldsymbol{l}$ **do**
        for each edge $(\boldsymbol{v}_i, \boldsymbol{v}_j)$ **do**
            if $c(v_j) < \sqrt{c(v_i) \times OF(v_j)}$ **then**
                $c(v_j) \leftarrow \sqrt{c(v_i) \times OF(v_j)}$
                $\boldsymbol{t} \leftarrow \boldsymbol{t} \bigcup \boldsymbol{v}_j$;
            end if
        end for
    end for
    $\boldsymbol{l} \leftarrow \boldsymbol{t}$
    $\boldsymbol{t} = \emptyset$
end while
return compactness values $\boldsymbol{c}$ of superpixels.

is computed as the average OF values of all containing pixels. The over-segmented image can be formulated as a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, where $\mathbb{V}$ is the list of vertices (superpixels) and $\mathbb{E}$ is the list of edges connecting the neighboring superpixels.

The procedure to compute the compactness values of superpixels is summarized in Algorithm 1. Denote $\boldsymbol{v}_c$ as the superpixel containing the centroid of interest. The algorithm transfers the OF value from the $\boldsymbol{v}_c$ to all other superpixels. The procedure performs a sequence of relaxation steps, namely assigning the compactness value $\boldsymbol{c}(\boldsymbol{v}_j)$ of superpixel $\boldsymbol{v}_j$ by the square root of its neighboring superpixel's compactness value and its own OF value. Our algorithm only relaxes edges from vertices $\boldsymbol{v}_j$ for which $\boldsymbol{c}(\boldsymbol{v}_j)$ has recently changed, since other vertices cannot lead to correct relaxations. Additionally, the algorithm may be terminated early when no recent changes exist. Finally, the compactness value CN is computed as:

$$CN(x, y) = \boldsymbol{c}(sp(x, y)), \tag{4}$$

where $sp(x, y)$ returns the index of the superpixel containing pixel (x, y).

2.4 Saliency Assignment

We normalize the objectness map OB, foreground map FG, and compactness map CN to the range [0..1]. We assume that all measures are independent, and hence we combine these terms as follows to compute a saliency value S for each pixel:

$$S(x, y) = OB(x, y) \times FG(x, y) \times CN(x, y). \tag{5}$$

The resulting pixel-level saliency map may have an arbitrary scale. In the final step, we rescale the saliency values within [0..1] and to contain at least 10% saliency pixels.

2.5 Implementation Settings

We apply the state-of-the-art objectness detection technique, i.e., binarized normed gradients (BING) [Cheng *et al.*, 2014], to produce a set of candidate object windows. Our selection of BING is two-fold. First, BING extractor has a weak training from the simple feature, e.g., binarized normed gradients. Therefore, it is useful comparing to bottom-up edge extractor. Second, the BING extractor is able to run 10 times faster than real-time, i.e., 300 frames per second (fps). BING hypothesis generator is trained with VOC2007 dataset [Everingham *et al.*, 2010] same as in [Cheng *et al.*, 2014]. In order to compute the foreground map, θ is set as 0.1 and we convert the color channels from RGB to Lab color space as suggested in [Achanta *et al.*, 2009; Perazzi *et al.*, 2012]. Regarding the image over-segmentation, we use SLIC [Achanta *et al.*, 2012] for the superpixel segmentation. We set the number of superpixels as 100 as a trade-off between the fine over-segmentation and the processing time.

3 Evaluation

3.1 Datasets and Evaluation Metrics

We evaluate and compare the performances of our algorithm against previous baseline algorithms on two representative benchmark datasets: the **MSRA 1000** salient object dataset [Achanta *et al.*, 2009] and the Interactive cosegmentation Dataset (**iCoSeg**) [Batra *et al.*, 2010]. The MSRA-1000 dataset contains 1,000 images with the pixel-wise ground truth provided by [Achanta *et al.*, 2009]. Note that each image in this dataset contains a salient object. Meanwhile, the iCoSeg contains 643 images with single or multiple objects in a single image.

The first evaluation compares the precision and recall rates. High recall can be achieved at the expense of reducing the precision and vice-versa so it is important to evaluate both measures together. In the first setting, we compare binary masks for every threshold in the range [0..255]. In the second setting, we use the image dependent adaptive threshold proposed by [Achanta *et al.*, 2009], defined as twice the mean saliency of the image:

$$T_a = \frac{2}{W \times H} \sum_{(x,y)} S(x, y). \tag{6}$$

In addition to precision and recall we compute their weighted harmonic mean measure or $F - measure$, which is defined as:

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{7}$$

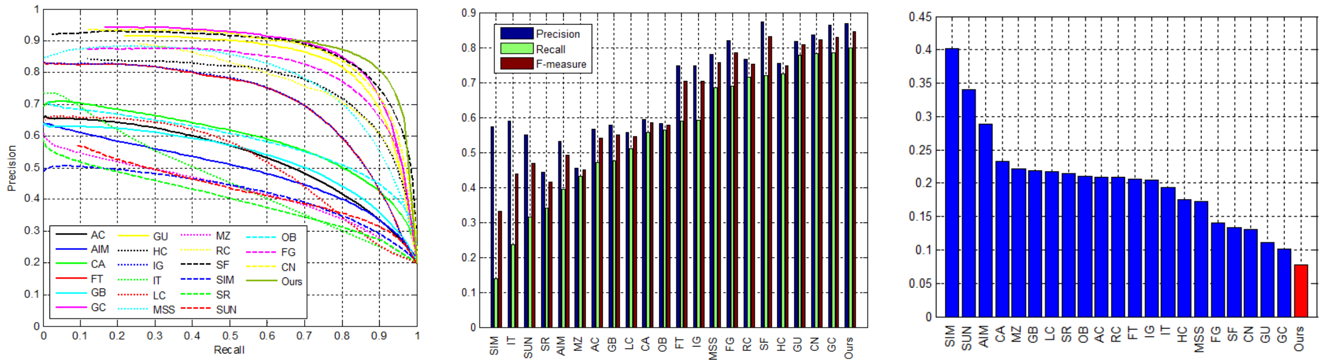
As in previous methods [Achanta *et al.*, 2009; Cheng *et al.*, 2013; Perazzi *et al.*, 2012], we use β^2 = 0.3.

For the second evaluation, we follow Perazzi et al. [Perazzi *et al.*, 2012] to evaluate the mean absolute error (MAE) between a continuous saliency map S and the binary ground truth G for all image pixels (x, y), defined as:

$$MAE = \frac{1}{W \times H} \sum_{(x,y)} |S(x, y) - G(x, y)|. \tag{8}$$

3.2 Performance on MSRA1000 dataset

Following [Achanta *et al.*, 2009; Perazzi *et al.*, 2012; Cheng *et al.*, 2013], we first evaluate our methods using a precision/recall curve which is shown in Figure 5. Our work



(a) Fixed threshold

(b) Adaptive threshold

(c) Mean absolute error

Figure 5: Statistical comparison with 18 saliency detection methods using all the 1000 images from MSRA-1000 dataset [Achanta *et al.*, 2009] with pixel accuracy saliency region annotation: (a) the average precision recall curve by segmenting saliency maps using fixed thresholds, (b) the average precision recall by adaptive thresholding (using the same method as in FT [Achanta *et al.*, 2009], SF [Perazzi *et al.*, 2012], GC [Cheng *et al.*, 2013], etc.), (c) the mean absolute error of the different saliency methods to ground truth mask. Please check Figure 2 for the references to the publications in which the baseline methods are presented.

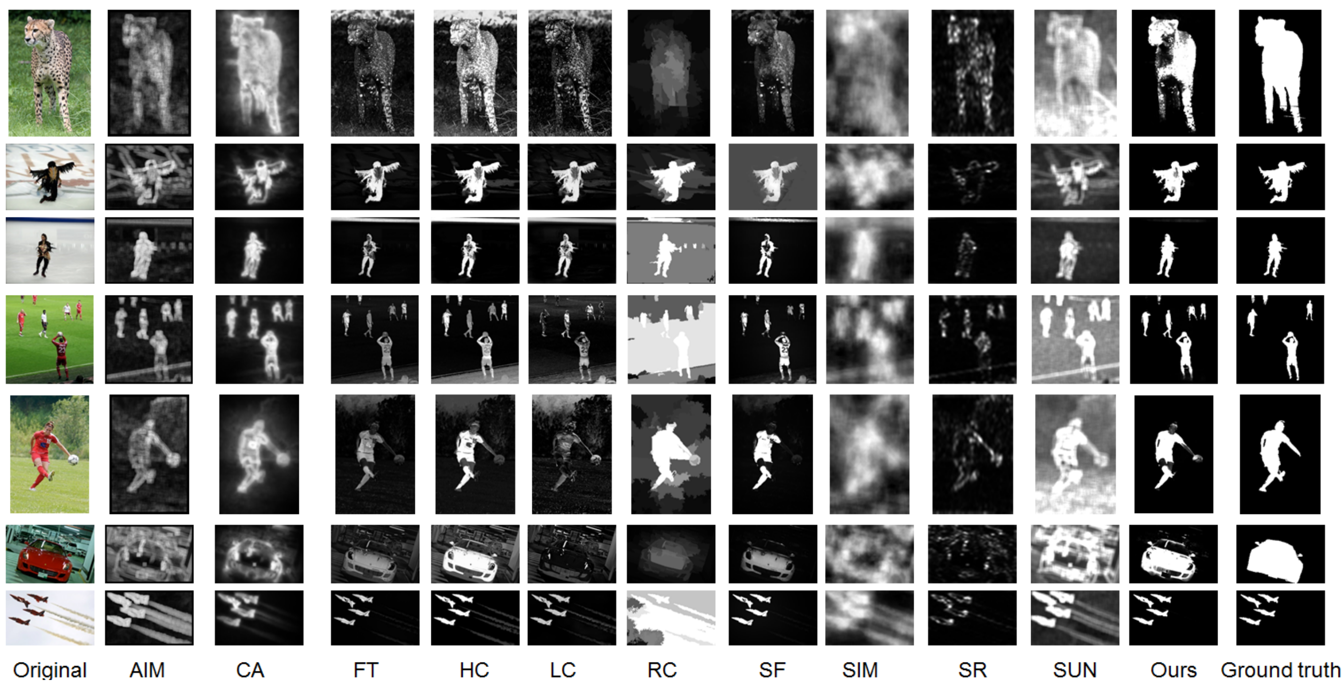
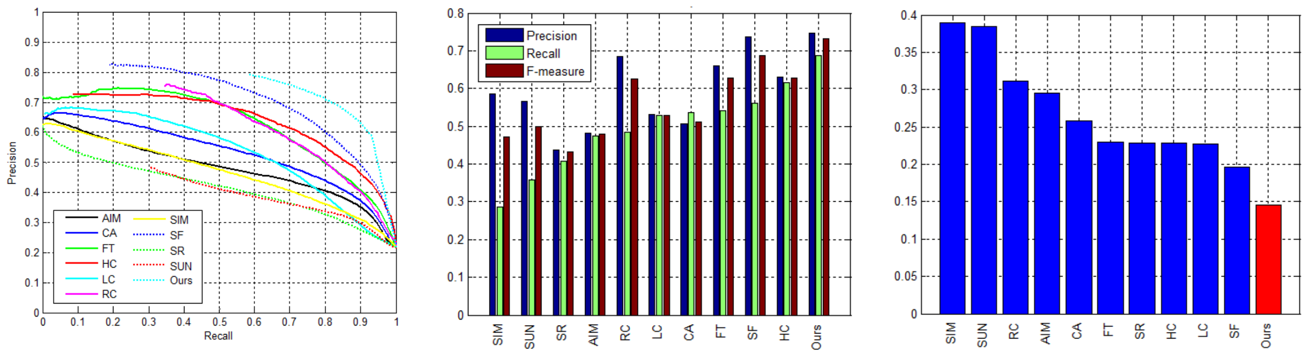


Figure 6: Visual comparison of saliency maps on iCoSeg dataset. We compare our method (AH) to other 10 alternative methods. Our results are close to ground truth and focus on the main salient objects.

reaches the highest precision/recall rate over all baselines. As a result, our method also obtains the best performance in terms of F-measure. We also evaluate the individual components in our system, namely, objectness map (OB), foreground map (FG), and compactness map (CN). They generally achieve the acceptable performance which is comparable to other baselines. The performance of the objectness map itself is outperformed by our proposed augmented hypotheses. In this work, our novelty is that we adopt and augment the conventional hypotheses by adding two key features: foregroundness and compactness to detect salient objects. When fusing them together, our unified system achieves the state-of-the-art performance in every single evaluation metric.

As discussed in the SF [Perazzi *et al.*, 2012] and GC [Cheng *et al.*, 2013], neither the precision nor recall measure considers the true negative counts. These measures favor methods which successfully assign saliency to salient pixels but fail to detect non-salient regions over methods that suc-



(a) Fixed threshold (b) Adaptive threshold (c) Mean absolute error

Figure 7: Statistical comparison with 10 saliency detection methods using all the 643 images from iCoSeg benchmark [Batra *et al.*, 2010] with pixel accuracy saliency region annotation: (a) the average precision recall curve by segmenting saliency maps using fixed thresholds, (b) the average precision recall by adaptive thresholding (using the same method as in FT [Achanta *et al.*, 2009], GC [Cheng *et al.*, 2013], etc.), (c) the mean absolute error of the different saliency methods to ground truth mask.

cessfully do the opposite. Instead, they suggested that MAE is a better metric than precision recall analysis for this problem. As shown in Figure 5c, our work outperforms the state-of-the-art performance [Cheng *et al.*, 2013] by 24%. One may argue that a simple boosting of saliency values similar as in [Perazzi *et al.*, 2012] results would improve it. However, a boosting of saliency values could easily result in the boosting of low saliency values related to background that we also aim to avoid.

3.3 Performance on iCoSeg dataset

The iCoSeg dataset is "less popular" in the sense that some baselines do not even release detection results and source-code. We only reproduced 10 methods on iCoSeg thanks to their existing source-code. The visual comparison of saliency maps generated from our method and different baselines are demonstrated in Figure 6. Our results are close to ground truth and focus on the main salient objects. We first evaluate our methods using a precision/recall curve which is shown in Figure 7a, b. Our method outperforms all other baselines in both two settings, namely fixed threshold and adaptive threshold. As shown in Figure 7c, our method achieves the best performance in terms of MAE. Our work outperforms other methods by a large margin, 25%.

3.4 Computational Efficiency

It is also worth investigating the computational efficiency of different methods. In Table 1, we compare the average running time of our approach to the currently best performing methods on the benchmark images. We compare the performance of our method in terms of speed with methods with most competitive accuracy (GC [Cheng *et al.*, 2013], SF [Perazzi *et al.*, 2012]). The average time of each method is measured on a PC with Intel i7 3.3 GHz CPU and 8GB RAM. Performance of all the methods compared in this table are based on implementations in C++ and MATLAB. The CA method the slowest one because it requires an exhaustive nearest-neighbor search among patches. Meanwhile, our method is

Table 1: Comparison of running times in the MSRA 1000 benchmark [Achanta *et al.*, 2009].

Method	CA	RC	SF	GC	Ours
Time (s)	51.2	0.14	0.15	0.09	0.07
Code	Matlab	C++	C++	C++	C++

able to run in a real-time manner. Our procedure spends most of the computation time on generating the objectness map (about 35%) and forming the compactness map (about 50%). From the experimental results, we find that our algorithm is effective and computationally efficient.

4 Conclusion and Future Work

In this paper, we have presented a novel method, *augmented hypotheses* (AH), which adopts the object hypotheses in order to rapidly detect *salient objects*. To this end, three maps are derived from object hypotheses: superimposed hypotheses form an objectness map, a foreground map is computed from deviations in color from the background, and a compactness map emerges from propagating saliency labels in the oversegmented image. These three maps are fused together to detect salient objects with sharp boundaries. Experimental results on two challenging datasets show that our results are 24% - 25% better than the previous best results (compared against 10+ methods in two different datasets), in terms of mean absolute error while also being faster.

For future work, we aim to investigate more sophisticated techniques for objectness measures and integrate more cues, i.e., depth [Lang *et al.*, 2012] and audio [Chen *et al.*, 2014] information. Also, we would like to study the impact of salient object detection into the object hypothesis process.

5 Acknowledgments

This work was supported by Singapore Ministry of Education under research Grants MOE2012-TIF-2-G-016 and MOE2014-TIF-1-G-007.

References

[Achanta and Süsstrunk, 2010] Radhakrishna Achanta and Sabine Süsstrunk. Saliency detection using maximum symmetric surround. In *ICIP*, pages 2653–2656, 2010.

[Achanta *et al.*, 2008] Radhakrishna Achanta, Francisco J. Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In *International Conference of Computer Vision Systems*, pages 66–75, 2008.

[Achanta *et al.*, 2009] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[Achanta *et al.*, 2012] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *T-PAMI*, 34(11):2274–2282, 2012.

[Alexe *et al.*, 2012] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *T-PAMI*, 34(11):2189–2202, 2012.

[Batra *et al.*, 2010] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010.

[Bruce and Tsotsos, 2005] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *NIPS*, 2005.

[Chen *et al.*, 2012] Qiang Chen, Zheng Song, Yang Hua, ZhongYang Huang, and Shuicheng Yan. Hierarchical matching with side information for image classification. In *CVPR*, pages 3426–3433, 2012.

[Chen *et al.*, 2014] Yanxiang Chen, Tam V. Nguyen, Mohan S. Kankanhalli, Jun Yuan, Shuicheng Yan, and Meng Wang. Audio matters in visual attention. *T-CSVT*, 24(11):1992–2003, 2014.

[Cheng *et al.*, 2011] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[Cheng *et al.*, 2013] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *CVPR*, pages 1529–1536, 2013.

[Cheng *et al.*, 2014] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[Goferman *et al.*, 2010] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010.

[Harel *et al.*, 2006] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[Hou and Zhang, 2007] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *T-PAMI*, 20(11):1254–1259, 1998.

[Koch and Ullman, 1985] C Koch and S Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 1985.

[Lampert *et al.*, 2008] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[Lang *et al.*, 2012] Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan S. Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, pages 101–115, 2012.

[Ma and Zhang, 2003] Yu-Fei Ma and HongJiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM MM*, pages 374–381, 2003.

[Murray *et al.*, 2011] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Párraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR*, pages 433–440, 2011.

[Nguyen *et al.*, 2013a] Tam V. Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan Kankanhalli, and Shuicheng Yan. Image re-attentionizing. *Multimedia, IEEE Transactions on*, 15(8):1910–1919, 2013.

[Nguyen *et al.*, 2013b] Tam V. Nguyen, Mengdi Xu, Guangyu Gao, Mohan Kankanhalli, Qi Tian, and Shuicheng Yan. Static saliency vs. dynamic saliency: a comparative study. In *ACM MM*, pages 987–996, 2013.

[Nguyen *et al.*, 2015] Tam V. Nguyen, Zheng Song, and Shuicheng Yan. STAP: Spatial-temporal attention-aware pooling for action recognition. *T-CSVT*, 2015.

[Perazzi *et al.*, 2012] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012.

[Uijlings *et al.*, 2013] Jasper Uijlings, Koen van de Sande, Theo Gevers, and Arnold Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[Viola and Jones, 2001] Paul A. Viola and Michael J. Jones. Robust real-time face detection. In *ICCV*, page 747, 2001.

[Zhai and Shah, 2006] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM MM*, pages 815–824, 2006.

[Zhang *et al.*, 2008] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.