This CVPR2013 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Hierarchical Saliency Detection

Qiong Yan Li Xu Jianping Shi Jiaya Jia
The Chinese University of Hong Kong
{qyan,xuli,jpshi,leojia}@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/

Abstract

When dealing with objects with complex structures, saliency detection confronts a critical problem – namely that detection accuracy could be adversely affected if salient foreground or background in an image contains small-scale high-contrast patterns. This issue is common in natural images and forms a fundamental challenge for prior methods. We tackle it from a scale point of view and propose a multi-layer approach to analyze saliency cues. The final saliency map is produced in a hierarchical model. Different from varying patch sizes or downsizing images, our scale-based region handling is by finding saliency values optimally in a tree model. Our approach improves saliency detection on many images that cannot be handled well traditionally. A new dataset is also constructed.

1. Introduction

Saliency detection, which is closely related to selective processing in human visual system [22], aims to locate important regions or objects in images. It gains much attention recently [2, 8, 4, 15, 25, 23, 30]. Knowing where important regions are broadly benefits applications, including classification [24], retrieval [11] and object co-segmentation [3], for optimally allocating computation.

Stemming from psychological science [28, 22], the commonly adopted saliency definition is based on how pixels/regions stand out and is dependent of what kind of visual stimuli human respond to most. By defining pixel/region uniqueness in either local or global context, existing methods can be classified to two streams. Local methods [13, 10, 1, 15] rely on pixel/region difference in the vicinity, while global methods [2, 4, 23, 30] rely mainly on color uniqueness in terms of global statistics.

Albeit many methods have been proposed, a few commonly noticeable and critically influencing issues still endure. They are related to complexity of patterns in natural images. A few examples are shown in Fig. 1. For the first two examples, the boards containing characters are salient foreground objects. But the results in (b), produced by a previous local method, only highlight a few edges that scatter in the image. The global method results in (c) also cannot clearly distinguish among regions. Similar challenge arises when the background is with complex patterns, as shown in the last example of Fig. 1. The yellow flowers lying on grass stand out. But they are actually part of the background when viewing the picture as a whole, confusing saliency detection.

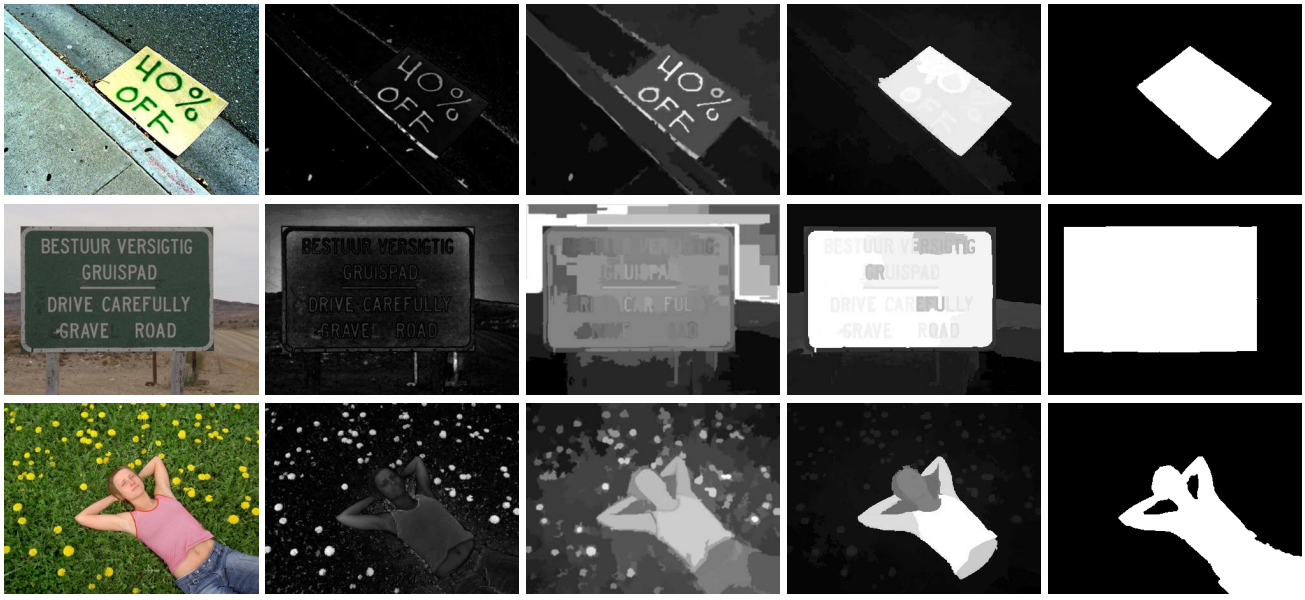
These examples are not special, and exhibit one common problem – that is, *when objects contain salient small-scale patterns, saliency could generally be misled by their complexity*. Given texture existing in many natural images, this problem cannot be escaped. It easily turns extracting salient objects to finding cluttered fragments of local details, complicating detection and making results not usable in, for example, object recognition [29], where connected regions with reasonable sizes are favored.

Aiming to solve this notorious and universal problem, we propose a hierarchical model, to analyze saliency cues from multiple levels of structure, and then integrate them to infer the final saliency map. Our model finds foundation from studies in psychology [20, 17], which show the selection process in human attention system operates from more than one levels, and the interaction between levels is more complex than a feed-forward scheme. With our multi-level analysis and hierarchical inference, the model is able to deal with salient small-scale structure, so that salient objects are labeled more uniformly.

In addition, contributions in this paper also include 1) a new measure of region scales, which is compatible with human perception on object scales, and 2) construction of a new scene dataset, which contains challenging natural images for saliency detection. Our method yields much improvement over others on the new dataset as well as other benchmarking data.

2. Related Work

Bottom-up saliency analysis generally follows location- and object-based attention formation [22]. Location meth-



(a) Input (b) AC [1] (c) RC [4] (d) Ours (e) Ground truth

Figure 1. Saliency detection with structure confusion. Small-scale strong details easily influence the process and cause erroneous results.

ods physically obtain human attention shift continuously with eye tracking, while the latter set of approaches aim to find salient objects from images. Both of them are important and benefit different applications in high-level scene analysis. A survey of human attention and saliency detection is provided in [27]. Below we briefly review a few.

The early local method [13] used an image pyramid to calculate pixel contrast based on color and orientation features. Ma and Zhang [19] directly computed center-surrounding color difference in a fixed neighborhood for each pixel. Harel *et al.* [10] proposed a method to non-linearly combine local uniqueness maps from different feature channels to concentrate conspicuity. These methods detect only high-contrast edges and attenuate smooth object interior. Patch-based center-surrounding difference was used in [18, 1] to remedy this issue. The accompanying problem is to choose an appropriate surrounding patch size. Besides, high-contrast edges are not necessarily in the foreground, as illustrated in Fig. 1.

Global methods mostly consider color statistics. Zhai and Shah [31] introduced image histograms to calculate color saliency. To deal with RGB color, Achanta *et al.* [2] provided an approximate by subtracting the average color from the low-pass filtered input. Cheng *et al.* [4] extended the histogram to 3D color space. These methods find pixels/regions with colors much different from the dominant one, but do not consider spatial locations. To compensate the lost spatial information, Perazzi *et al.* [23] measured the variance of spatial distribution for each color. Global methods have their difficulty in distinguishing among similar colors in both foreground and background. A few recent methods exploit background smoothness [25, 30]. Note that assuming background is smooth could be invalid for many natural images, as explained in Section 1.

High-level priors are also commonly used based on common knowledge and experience. Face detector was adopted in [8, 25]. The concept of center bias – that is, image center is more likely to contain salient objects than other regions – was employed in [18, 14, 25, 30]. In [25], it is assumed that warm colors are more attractive to human.

Prior work does not consider the situation that locally smooth regions could be inside a salient object and globally salient color, contrarily, could be from the background. These difficulties boil down to the same type of problems and indicate that saliency is ambiguous in one single scale. As image structures exhibit different characteristics when varying resolutions, they should be treated differently to embody diversity. Our hierarchical framework is a unified one to address these issues.

3. Hierarchical Model

Our method is as follows. First, three image layers of different scales are extracted from the input. Saliency cues are computed for each layer. They are finally fused into one single map using a graphical model. These steps are described in Sections 3.1 – 3.3 respectively. The framework is illustrated in Fig. 2.

3.1. Image Layer Extraction

Image layers, as shown in Fig. 2(c), are coarse representation of the input with different degrees of details, balanc-

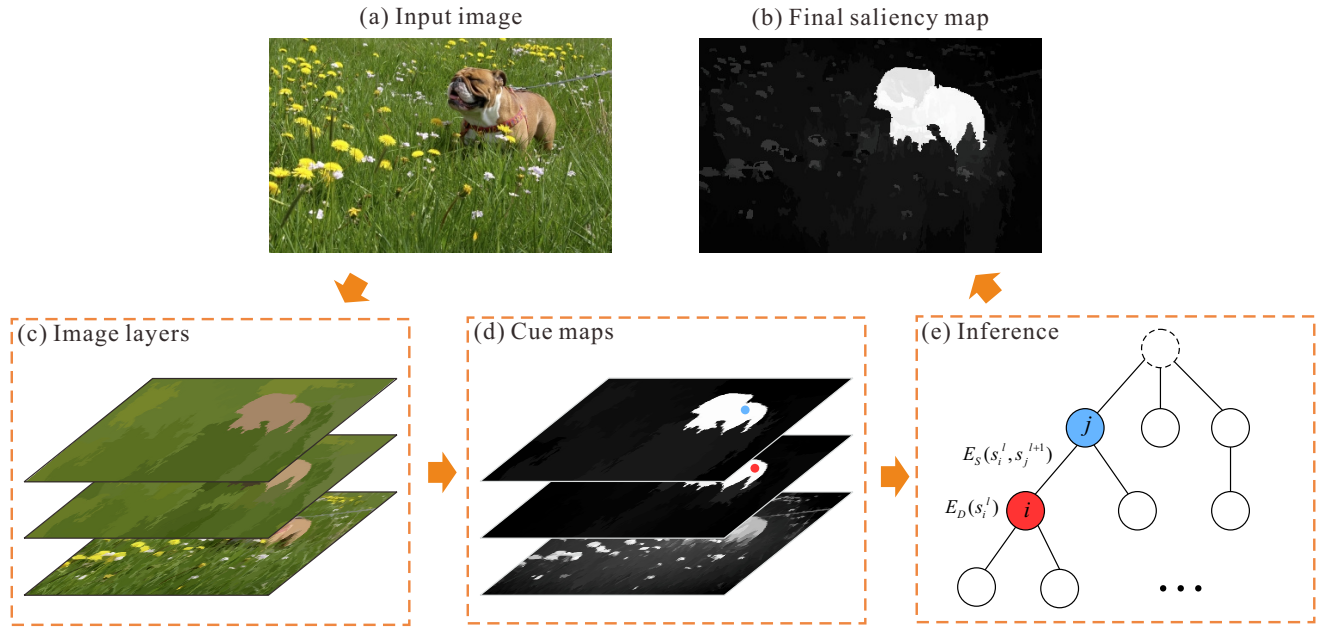


Figure 2. An overview of our hierarchical framework. We extract three image layers from the input, and then compute saliency cues from each of these layers. They are finally fed into a hierarchical model to get the final results.

(a) Input (b) Over-segmentation (c) Layer $\mathcal{L}^1$ (d) Layer $\mathcal{L}^2$ (e) Layer $\mathcal{L}^3$

Figure 3. Region-merge results under different scales.

ing between expression capability and structure complexity. The layer number is fixed to 3 in our experiments. In the bottom level, finest details such as flower are retained, while in the top level large-scale structures are produced.

3.1.1 Layer Generation

To produce the three layers, we first generate an initial over-segmentation as illustrated in Fig. 3(b) by the watershed-like method [9]. For each segmented region, we compute a scale value, where the process is elaborated on in the next subsection. They enable us to apply an iterative process to merge neighboring segments. Specifically, we sort all regions in the initial map according to their scales in an ascending order. If a region scale is below 3, we merge it to its nearest region, in terms of average CIELUV color distance, and update its scale. We also update the color of the region as their average color. After all regions are processed, we take the resulting region map as the bottom layer $\mathcal{L}^1$. The middle and top layers $\mathcal{L}^2$ and $\mathcal{L}^3$ are generated similarly from $\mathcal{L}^1$ and $\mathcal{L}^2$ with larger scale thresholds. In our experiment, we set thresholds for the three layers as $\{3, 17, 33\}$ for typical 400×300 images. Three layers are shown in Fig. 3(c)-(e). More details in this process can be found in our supplementary file in the project website. Note a region in the middle or top layer embraces corresponding ones in the lower levels. We use it for saliency inference described in Section 3.3.

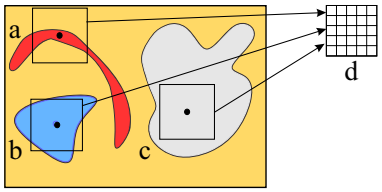


Figure 4. Our region scale is defined as the largest square that a region can contain. In this illustration, the scales of regions a and b are less than 5, and that of c is larger than 5.

3.1.2 Region Scale Definition

In methods of [5, 7] and many others, the region size is measured by the number of pixels. Our research and exten-

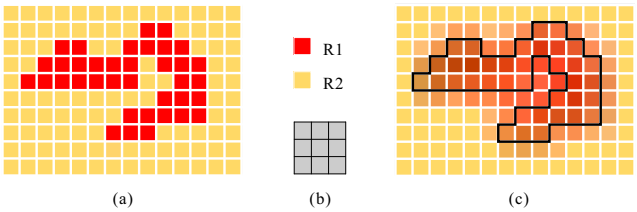


Figure 5. Efficient computation of scale transform. (a) Initial region map. (b) Map labels and the box filter. (c) Filtered region map. As shown in (c), all colors in R1 are updated compared to the input, indicating a scale smaller than 3.

sive experiments suggest this measure could be wildly inappropriate for processing and understanding general natural images. In fact, a large pixel number does *not* necessarily correspond to a large-scale region in human perception.

An example is shown in Fig. 4. Long curved region a contains many pixels. But it is not regarded as a large region in human perception due to its high inhomogeneity. Region b could look bigger although its pixel number is not larger. With this fact, we define a new *encompassment* scale measure based on shape uniformities and use it to obtain region sizes in the merging process.

Definition *Region R encompassing region R' means there exists at least one location to put R' completely inside R, denoted as $R' \subseteq R$.*

With this relation, we define the scale of region R as

$$\text{scale}(R) = \arg\max_t \{R_{t\times t} | R_{t\times t} \subseteq R\}, \tag{1}$$

where $R_{t\times t}$ is a $t \times t$ square region. In Fig. 4, the scales of regions a and b are smaller than 5 while the scale of c is above it.

3.1.3 Efficient Algorithm to Compute Region Scale

To determine the scale for a region, naive computation following the definition in Eq. (1) needs exhaustive search and comparison, which could be costly. In fact, in the merging process in a level, we only need to know whether the scale of a region is below the given threshold t or not. This enables a fast algorithm by applying a box filter with $t \times t$ kernel to the segment and checking if all pixel values inside the segment are changed during filtering. A positive output means the segment scale is below t. The process is illustrated in Fig. 5. Proof that this simple algorithm works is provided in the supplementary file.

3.2. Single-Layer Saliency Cues

For each layer we extract, saliency cues are applied to find important pixels from the perspectives of color, position and size. We present two cues that are particularly useful.

Local contrast Image regions contrasting their surroundings are general eye-catching [4]. We define the local contrast saliency cue for R_i in an image with a total of n regions as a weighed sum of color difference from other regions:

$$C_i = \sum_{j=1}^{n} w(R_j)\phi(i,j)||\mathrm{c}_i - \mathrm{c}_j||_2, \tag{2}$$

where c_i and c_j are colors of regions R_i and R_j respectively. $w(R_j)$ counts the number of pixels in R_j. Regions with more pixels contribute higher local-contrast weights than those containing only a few pixels. $\phi(i,j)$ is set to $\exp\{-D(R_i,R_j)/\sigma^2\}$ controlling the spatial distance influence between two regions i and j, where $D(R_i,R_j)$ is a square of Euclidean distances between region centers of R_i and R_j. With the $\phi(i,j)$ term, close regions have larger impact than distant ones. Hence, Eq. (2) measures color contrast mainly to surroundings. Parameter σ^2 controls how large the neighborhood is. It is set to the product of $(0.2)^2$ and the particular scale threshold for the current layer. In the top layer, σ^2 is large, making all regions be compared in a near-global manner.

Location heuristic Psychophysical study shows that human attention favors central regions [26]. So pixels close to a natural image center could be salient in many cases. Our location heuristic is thus written as

$$H_i = \frac{1}{w(R_i)} \sum_{x_i \in R_i} \exp\{-\lambda \|x_i - x_c\|^2\}, \tag{3}$$

where $\{x_0, x_1 \cdots\}$ is the set of pixel coordinates in region R_i, and x_c is the coordinate of the image center. H_i makes regions close to image center have large weights. λ is a parameter used when H_i is combined with C_i, expressed as

$$\bar{s}_i = C_i \cdot H_i. \tag{4}$$

Since the local contrast and location cues have been normalized to range $[0, 1)$, their importance is balanced by λ, set to 9 in general. After computing $\bar{s}_i$ for all layers, we obtain initial saliency maps separately, as demonstrated in Fig. 6(b)-(d). We propose a hierarchical inference procedure to fuse them for multi-scale saliency detection.

3.3. Hierarchical Inference

Cue maps reveal saliency in different scales and could be quite different. At the bottom level, small regions are produced while top layers contain large-scale structures. Due to possible diversity, none of the single layer information is guaranteed to be perfect. Also, it is hard to determine which layer is the best by heuristics.

Multi-layer fusion by naively averaging all maps is not a good choice, considering possibly complex background

(a) Input (b) Cue map at Layer $\mathcal{L}^1$ (c) Cue map at Layer $\mathcal{L}^2$ (d) Cue map at Layer $\mathcal{L}^3$ (e) Final saliency map

Figure 6. Saliency cue maps in three layers and our final saliency map.

and/or foreground. On the other hand, in our region merging steps, a segment is guaranteed to be encompassed by the corresponding ones in upper levels. We therefore resort to hierarchical inference based on a tree-structure graphical model. An example is shown in Fig. 2(e), where nodes represent regions in their corresponding layers. For instance, the blue node j corresponds to the region marked in blue in (d). It contains two segments in the lower level and thus introduces two children nodes. The root maps to the entire image in the coarsest representation.

For a node corresponding to region i in layer $\mathcal{L}^l$, we define a saliency variable s_i^l. Set $\mathcal{S}$ contains all of them. We minimize the following energy function

$$E(\mathcal{S}) = \sum_l \sum_i E_D(s_i^l) + \sum_l \sum_{i, R_i^l \subseteq R_j^{l+1}} E_S(s_i^l, s_j^{l+1}) \quad (5)$$

The energy consists of two parts. Data term $E_D(s_i^l)$ is to gather separate saliency confidence, and hence is defined, for every node, as

$$E_D(s_i^l) = \beta^l ||s_i^l - \bar{s}_i^l||_2^2, \quad (6)$$

where β^l controls the layer confidence and $\bar{s}_i^l$ is the initial saliency value calculated in Eq. (4). The data term follows a common definition.

The hierarchy term $E_S(s_i^l, s_j^{l+1})$ enforces consistency between corresponding regions in different layers. If R_i^l and R_j^{l+1} are corresponding in two layers, we must have $R_i^l \subseteq R_j^{l+1}$ based on our encompassment definition and the segment generation procedure. E_S is defined on them as

$$E_S(s_i^l, s_j^{l+1}) = \lambda^l ||s_i^l - s_j^{l+1}||_2^2, \quad (7)$$

where λ^l controls the strength of consistency between layers. The hierarchical term makes saliency assignment for corresponding regions in different levels similar, beneficial to effectively correcting initial saliency errors.

To minimize Eq. (5), exact inference can be achieved via a belief propagation [16]. It can reach the global optimum due to the tree model. The propagation procedure includes bottom-up and top-down steps. The bottom-up step updates variables s_i^l in two neighboring layers by minimizing Eq. (5), resulting in new saliency s_i^l representation with regard to the initial values $\bar{s}_i^l$ and those of parent nodes s_j^{l+1}. This step brings information up in the tree model by progressively substituting high-level variables for low-level ones.

Once results are obtained in the top layer, a top-down step is performed. In each layer, since there is already a minimum energy representation obtained in the previous step, we optimize it to get new saliency values. After all variables s_j^l are updated in a top-down fashion, we obtain the final saliency map in $\mathcal{L}^1$. An example is shown in Fig. 6 where separate layers in (b)-(d) miss out either large- or small-scale structures. Our result in (e) contains information from all scales, making the saliency map better than any of the single-layer ones.

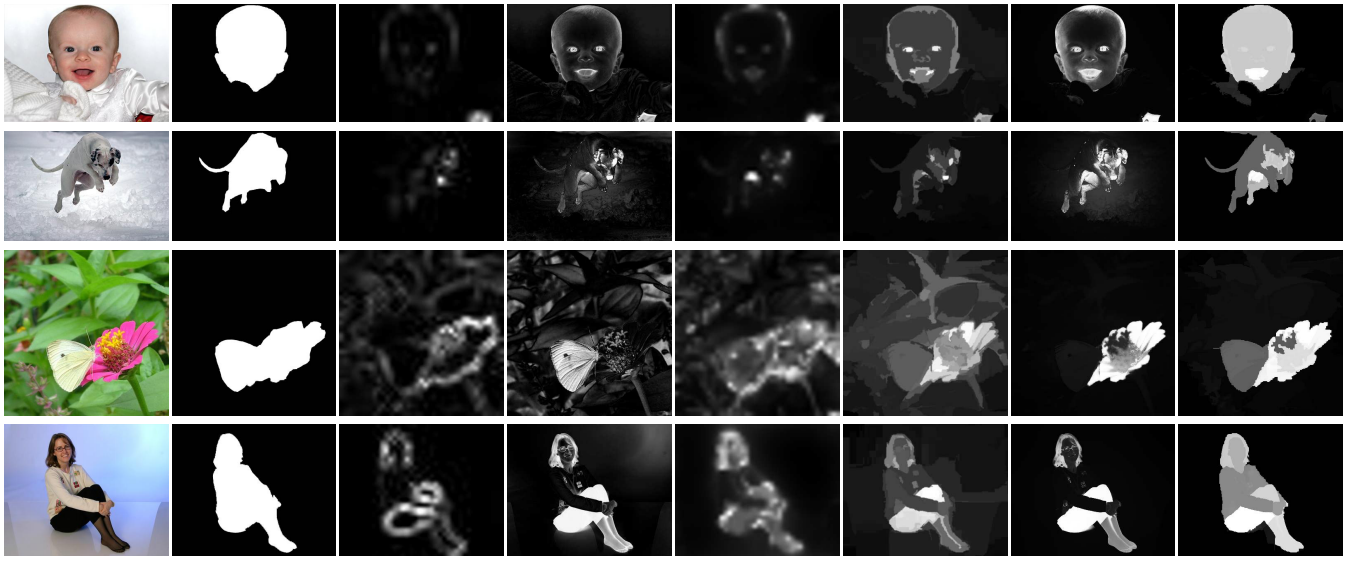
In fact, solving the three layer hierarchical model via belief propagation is equivalent to applying a weighted average to all single-layer saliency cue maps. Our method differs from naive multi-layer fusion by selecting weights optimally for each region in hierarchical inference instead of global weighting. The proposed solution theoretically and empirically performs better than simply averaging all layers, while not scarifying much computation efficiency.

4. Experiments

Currently, our un-optimized C++ implementation takes on average $0.28s$ to process one image with resolution 400×300 in the benchmark data on a PC equipped with a 3.40GHz CPU and 8GB memory. The computationally most expensive part is extraction of image layers with different scale parameters, which is also the core of our algorithm.

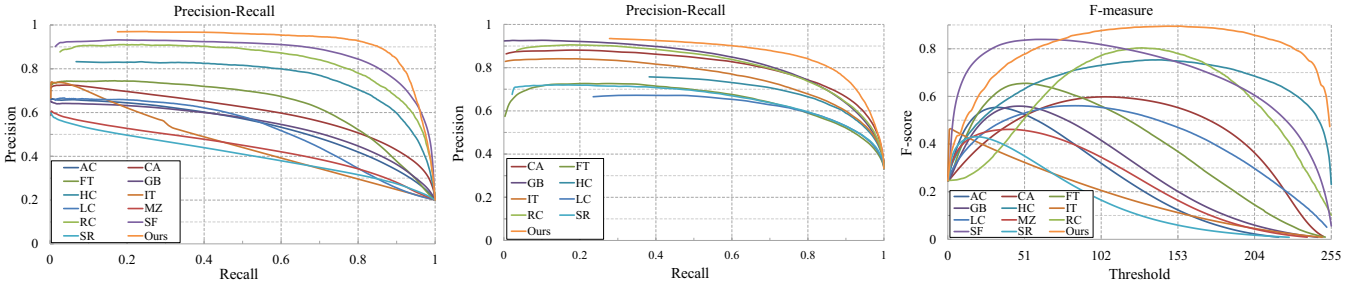
4.1. MSRA-1000 [2] and 5000 Datasets [18]

We first test our method on the saliency datasets MSRA-1000 [2] and MSRA-5000 [18] where MSRA-1000 is a subset of MSRA-5000 and contains 1000 natural images with their corresponding ground truth masks. MSRA-5000, contrarily, contains only the bounding box labels. We compare our method with several prior ones, including local methods – IT [13], MZ [19], GB [10], and global methods – LC [31], FT [2], CA [8] HC [4], RC [4], SF [23], LR [25], SR [12]. The abbreviations are the same as those in [23], except for LR, which represents the low rank method of [25]. For HC,



(a) Input (b) GT (c) MZ (d) LC (e) GB (f) RC (g) SF (h) Ours

Figure 7. Visual comparison on MSRA-1000 [2].



(a) (b) (c)

Figure 8. Statistical comparison.

RC and SR, we use the implementation of [4]. For IT, GB, FT and CA, we run authors' codes. For LC, MZ, SF and LR, we directly use author-provided saliency results.

The visual comparison is given in Fig. 7. Our method can handle complex foreground and background with different details. More results are available on our website.

In quantitative evaluation, we plot the precision-recall curves for the MSRA-1000 and 5000 datasets in Figs. 8(a) and 8(b) respectively. Our experiment follows the setting in [2, 4], where saliency maps are binarized at each possible threshold within range $[0, 255]$. Our method achieves the highest precision in almost the entire recall range $[0, 1]$. It is because combining saliency information from three scales makes background generally have low saliency values. Only sufficiently salient objects can be detected in this case.

In many applications, high precision and high recall are both required. We thus estimate the F-Measure [2] as

$$F_\beta = \frac{(1+\beta^2) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}. \quad (8)$$

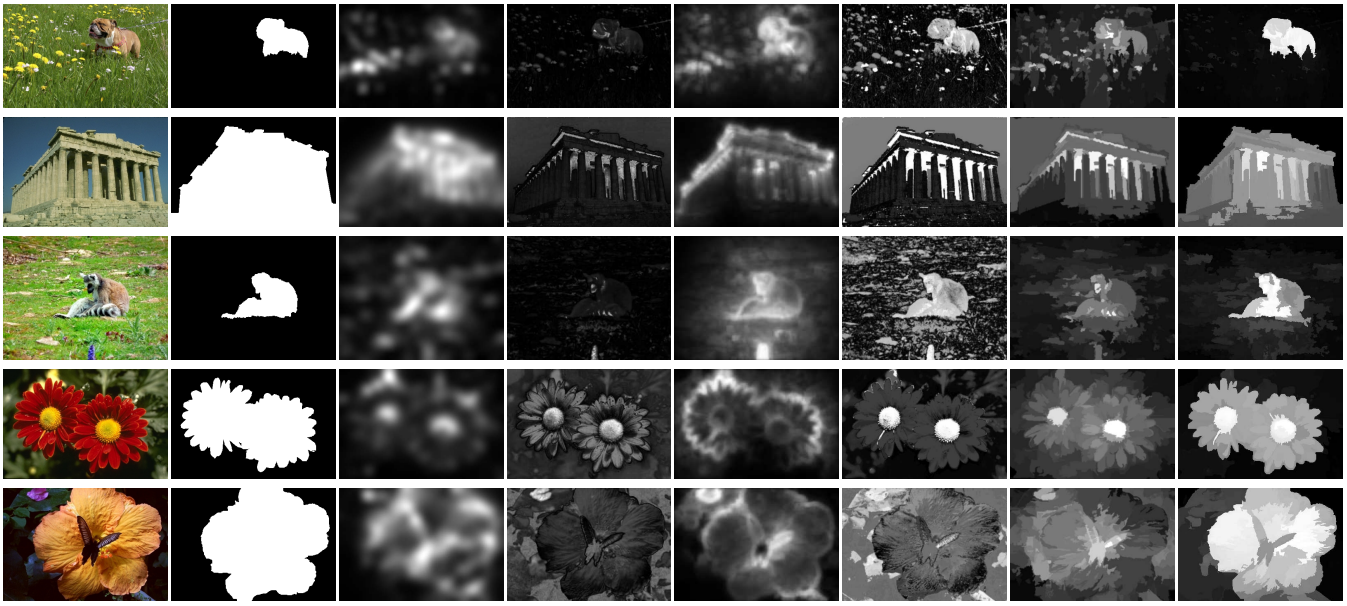
Thresholding is applied and β^2 is set to 0.3 as suggested in [2]. The F-measures for MSRA-1000 are plotted in Fig. 8(c). Our method has high F-scores in a wide range, indicating less sensitivity to picking a threshold.

4.2. Evaluation on Complex Scene Saliency Dataset

Although images from MSRA-1000 [2] have a large variety in their content, background structures are primarily simple and smooth. To represent more general situations that natural images fall into, we construct a Complex Scene Saliency Dataset (CSSD) with 200 images. They contain diversified patterns in both foreground and background. Ground truth masks are produced by 5 helpers. These images are collected from BSD300 [21], VOC dataset [6] and internet. Our dataset is now publicly available.

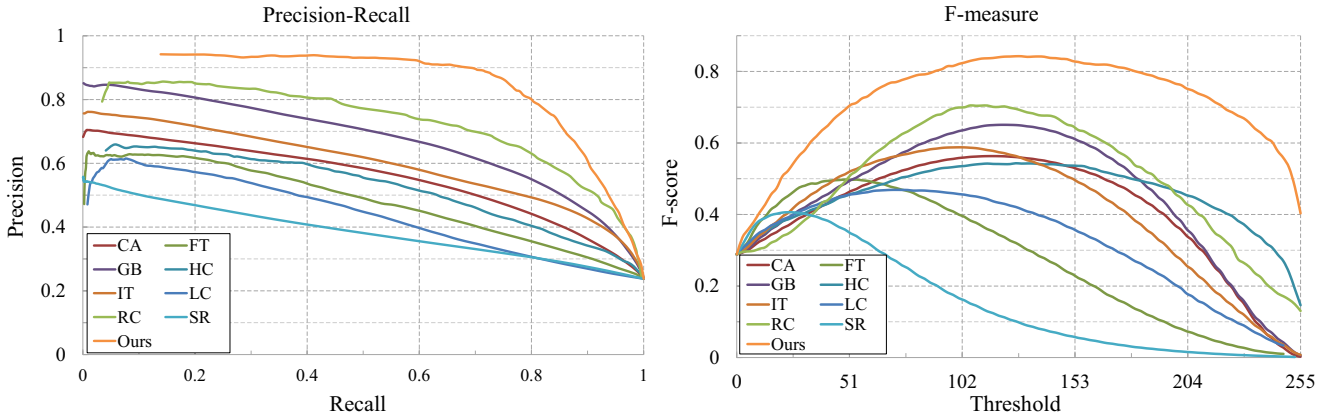
Visual comparison of results are shown in Fig. 9. On these difficult examples, our method can still produce reasonable saliency maps. More results are available online.

Follow previous settings, we also quantitatively compare our method with several others with publicly avail-



(a) Input (b) GT (c) IT (d) FT (e) CA (f) HC (g) RC (h) Ours

Figure 9. Visual Comparison on CSSD.



(a) (b)

Figure 10. Quantitative comparison on dataset CSSD.

able implementation. We plot the precision-recall and F-score curves in Fig. 10(a)-(b). The difference between our method and others is clear, manifesting the importance to capture hierarchical saliency in a computationally feasible framework.

4.3. Comparison with Single-Layer

Our method utilizes information from multiple image layers, gaining special benefit. Single-layer saliency computation does not work similarly well. To validate it, we take $\bar{s}_i$ in Eq. (4) in different layers, as well as the average of them, as the saliency values and evaluate how they work respectively when applied to our CSSD image data. The precision-recall curves are plotted in Fig. 11. Result from layer $\mathcal{L}^1$ is the worst since it contains many small structures. Results in the other two layer with larger-scale regions perform better, but still contain various problems related to scale determination. The result by naively averaging the three single-layer maps is also worse than our final one produced by optimal inference.

5. Concluding Remarks

We have tackled a fundamental problem that small-scale structures would adversely affect salient detection. This problem is ubiquitous in natural images due to common texture. In order to obtain a uniformly high-response saliency map, we propose a hierarchical framework that infers importance values from three image layers in different scales. Our proposed method achieves high performance and broadens the feasibility to apply saliency detection to more

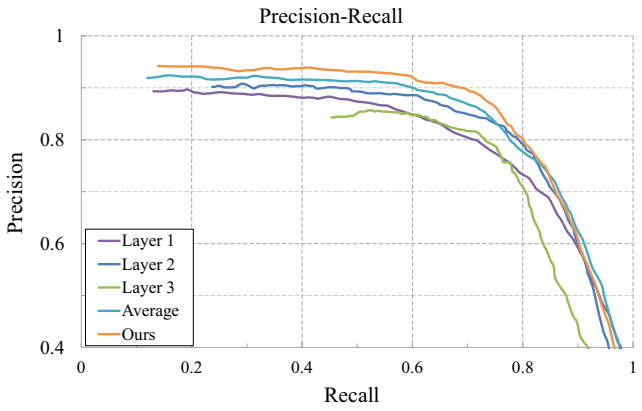


Figure 11. Single-layer vs. multi-layer.

applications handling different natural images.

Acknowledgements

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. 412911).

References

[1] R. Achanta, F. J. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *ICVS*, pages 66–75, 2008.

[2] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[3] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, pages 2129–2136, 2011.

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010.

[9] R. Gonzalez and R. Woods. *Digital image processing*. Prentice Hall Press, ISBN 0-201-18075-8, 2002.

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[11] P. Hiremath and J. Pujari. Content based image retrieval using color boosted salient points and shape features of an image. *International Journal of Image Processing*, 2(1):10–17, 2008.

[12] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[14] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[15] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011.

[16] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[17] T. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[18] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR*, 2007.

[19] Y.-F. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*, pages 374–381, 2003.

[20] E. Macaluso, C. Frith, and J. Driver. Directing attention to locations and to sensory modalities: Multiple levels of selective processing revealed with pet. *Cerebral Cortex*, 12(4):357–368, 2002.

[21] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 2, pages 416–423, July 2001.

[22] S. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999.

[23] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[24] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, pages 3506–3513, 2012.

[25] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012.

[26] B. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[27] A. Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2131–2146, 2011.

[28] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, Jan. 1980.

[29] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognitionła gentle way. In *Biologically Motivated Computer Vision*, pages 251–267. Springer, 2002.

[30] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, 2012.

[31] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM Multimedia*, pages 815–824, 2006.