Efficient Salient Region Detection with Soft Image Abstraction

Ming-Ming Cheng Jonathan Warrell Wen-Yan Lin Shuai Zheng Vibhav Vineet Nigel Crook
Vision Group, Oxford Brookes University

Abstract

Detecting visually salient regions in images is one of the fundamental problems in computer vision. We propose a novel method to decompose an image into large scale perceptually homogeneous elements for efficient salient region detection, using a soft image abstraction representation. By considering both appearance similarity and spatial distribution of image pixels, the proposed representation abstracts out unnecessary image details, allowing the assignment of comparable saliency values across similar regions, and producing perceptually accurate salient region detection. We evaluate our salient region detection approach on the largest publicly available dataset with pixel accurate annotations. The experimental results show that the proposed method outperforms 18 alternate methods, reducing the mean absolute error by 25.2% compared to the previous best result, while being computationally more efficient.

1. Introduction

The automatic detection of salient object regions in images involves a soft decomposition of foreground and background image elements [7]. This kind of decomposition is a key component of many computer vision and graphics tasks. Rather than focusing on predicting human fixation points [6, 32] (another major research direction of visual attention modeling), salient region detection methods aim at uniformly highlighting entire salient object regions, thus benefiting a large number of applications, including object-of-interest image segmentation [19], adaptive compression [17], object recognition [44], content aware image editing [51], object level image manipulation [12, 15, 53], and internet visual media retrieval [10, 11, 13, 29, 24, 23].

In terms of improving salient region detection, there are two emerging trends:

- Global cues: which enable the assignment of comparable saliency values across similar image regions and which are preferred to local cues [2, 14, 16, 26, 31, 42].
- Image abstraction: where an image is decomposed into perceptually homogeneous element, a process which abstracts out unnecessary detail and which is important for high quality saliency detection [42].

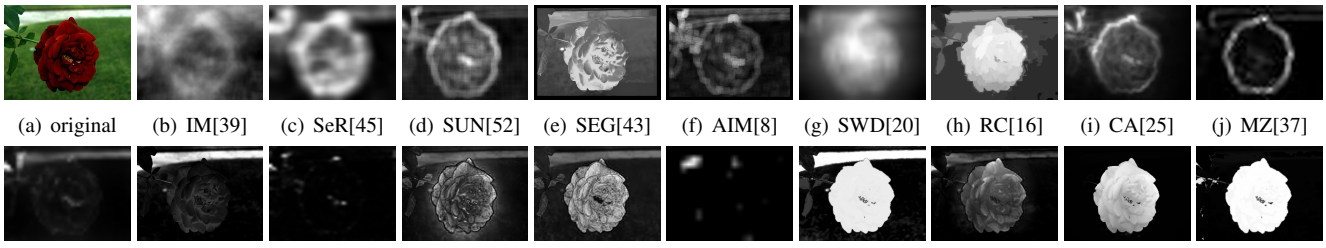
(a) Source image (b) Our result (c) Ground truth

Figure 1. We use soft image abstraction to decompose an image into large scale perceptually homogeneous elements (see Fig. 3), which abstract unnecessary details, assign comparable saliency values across similar image regions, and produce perceptually accurate salient regions detection results (b).

In this paper, we propose a novel *soft image abstraction* approach that captures large scale perceptually homogeneous elements, thus enabling effective estimation of global saliency cues. Unlike previous techniques that rely on super-pixels for image abstraction [42], we use histogram quantization to collect appearance samples for a global Gaussian Mixture Model (GMM) based decomposition. Components sharing the same spatial support are further grouped to provide a more compact and meaningful presentation. This soft abstraction avoids the hard decision boundaries of super pixels, allowing abstraction components with very large spatial support. This allows the subsequent global saliency cues to uniformly highlight entire salient object regions. Finally, we integrate the two global saliency cues, Global Uniqueness (GU) and Color Spatial Distribution (CSD), by automatically identifying which one is more likely to provide the correct identification of the salient region.

We extensively evaluate our salient object region detection method on the largest publicly available dataset with 1000 images containing pixel accurate salient region annotations [2]. The evaluation results show that each of our individual measures (GU and CSD) significantly outperforms existing 18 alternate approaches, and the final Global Cues (GC) saliency map reduces the mean absolute error by 25.2% compared to the previous best results (see Fig. 2 for visual comparisons) [1], while requiring substantially less running times.

[1] Results for these methods on the entire dataset and our prototype software can be found in our project page: http://mmcheng.net/effisalobj/.



(a) original (b) IM[39] (c) SeR[45] (d) SUN[52] (e) SEG[43] (f) AIM[8] (g) SWD[20] (h) RC[16] (i) CA[25] (j) MZ[37]

(k) GB[27] (l) LC[50] (m) SR[28] (n) AC[1] (o) FT[2] (p) IT[30] (q) HC[16] (r) MSS[3] (s) SF[42] (t) Our GC

Figure 2. Saliency maps computed by different state-of-the-art methods (b-s), and with our proposed GC method. Most results highlight edges, or are of low resolution. See also Fig. 4 and the supplementary.

2. Related work

While often treated as an image processing operation, saliency has its roots within human perception. When observing a scene, it has been noticed that humans focus on selected regions, for efficient recognition and scene understanding. A considerable amount of research in cognitive psychology [48] and neurobiology [18] has been devoted to discovering the mechanisms of visual attention in humans [38]. These regions where attention is focused are termed *salient regions*. Insights from psycho-visual research have influenced computational saliency detection methods, resulting in significant improvements in performance [5].

Our research is situated in the highly active field of visual attention modelling. A comprehensive discussion of this field is beyond the scope of this paper. We refer interested readers to recent survey papers for a detailed discussion of 65 models [5], as well as quantitative analysis of different methods in the two major research directions: salient object region detection [7] and human fixation prediction [6, 32]. Here, we mainly focus on discussing bottom-up, low-level salient object region detection methods.

Inspired by the *early representation* model of Koch and Ullman [35], Itti et al. [30] proposed highly influential computational methods, which use local centre-surrounded differences across multi-scale image features to detect image saliency. A large number of methods have been proposed to extend this method, including the fuzzy growing method by Ma and Zhang [37], and graph-based visual saliency detection by Harel *et al*. [27]. Later, Hou and Zhang [28] proposed an interesting spectral-based method, which finds differential components in the spectral domain. Zhang *et al*. [52] find salient image regions using information theory. Extensive evaluation results in [16], however, show that these methods tend to overemphasize small and local features, making them less suitable for important applications such as image segmentation, object detection, etc.

Methods modeling global properties have become popular recently as they enable the assignment of comparable saliency values across similar image regions, and thus can uniformly highlight the entire object regions [16]. Goferman *et al*. [25] use a patch based approach to incorporate global properties. Wang *et al*. [47] estimate saliency over the whole image relative to a large dictionary of images. Liu *et al*. [36] measure center-surrounded histograms over windows of various sizes and aspect ratios in a sliding window manner, and learn the combination weights relative to other saliency cues. While these algorithms are generally better at preserving global image structures and are able to highlight entire salient object regions, they suffer from high computational complexity. Finding efficient and compact representations has been shown to be a promising way of modeling global considerations. Initial efforts tried to adopt only luminance [50] or first-order average color [2] to effectively estimate consistent results. However, they ignored complex color variations in natural images and spatial relationships across image parts. Recently, Cheng *et al*. [16] proposed a region contrast-based method to model global contrast, showing significantly improved performance. However, due to the use of image segments, saliency cues like spatial distribution cannot be easily formulated.

More recently, Perazzi *et al*. [42] made the important observation that decomposing an image into perceptually homogeneous elements, which abstract unnecessary details, is important for high quality salient object detection. They used superpixels to abstract the image into perceptually uniform regions and efficient N-D Gaussian filtering to estimate global saliency cues. As detailed in §3, we propose a GMM based abstract representation, to capture large scale perceptually homogeneous elements, resulting in the efficient evaluation of global cues and improved salient object region detection accuracy.

3. Soft Image Abstraction via Global Components Representation

3.1. Histogram based efficient GMM decomposition

In order to get an abstract global representation which effectively captures perceptually homogeneous elements, we cluster image colors and represent them using Gaussian Mixture Models (GMM). Each pixel color I_x is represented

as a weighted combination of several GMM components, with its probability of belonging to a component c given by:

$$p(c|I_x) = \frac{\omega_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}{\sum_c \omega_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}, \quad (1)$$

where ω_c, μ_c, and Σ_c represent respectively the weight, mean color, and covariance matrix of the c^{th} component.

We use the GMM to decompose an image in to perceptually homogenous elements. These elements are structurally representative and abstract away unnecessary details. Fig. 3(a) shows an example of such a decomposition. Notice that our GMM-based representation better captures large scale perceptually homogeneous elements than superpixel representations (as in Fig. 3(d)) which can only capture local homogeneous elements. We will discuss how our global homogeneous components representation benefits global saliency cue estimation in §4.

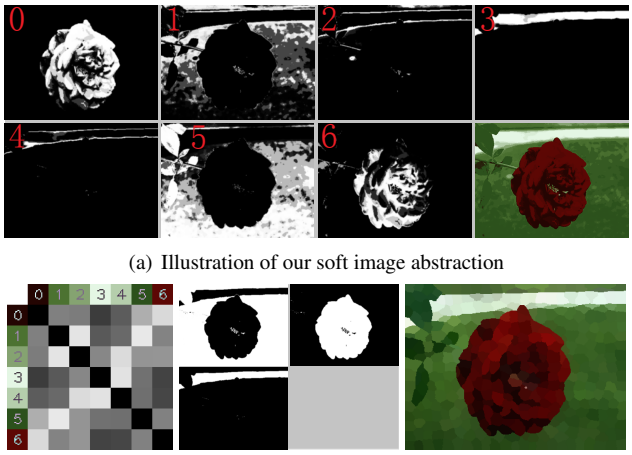
A time consuming step of building the GMM-based representation is clustering pixel colors and fitting them to each GMM component. Such clustering can be achieved using Orchard and Bouman's algorithm [40], which starts with all pixels in a single cluster and iteratively uses the eigenvalues and eigenvector of the covariance matrix to decide which cluster to split and the splitting point. Inspired by [16], we first run color quantization in RGB color space with each color channel divided in to 12 uniform parts and choose the most frequently occurring colors which account for 95% of the image pixels. This typically result in a histogram based representation with N bins (on average $N = 85$ for 1000 images dataset [2] as reported by [16]). We take each histogram bin as a weighted color sample to build the color covariance matrix and learn the remaining parameters of the GMM (the means and probabilities for belonging to each component) from the weighted bins. We use the indexing table (detailed in §3.3) to associate image pixels with histogram bins for computational efficiency.

3.2. Spatial overlap based components clustering

Direct GMM based color clustering ignores valuable spatial correlations in images. As the example shown in Fig. 3(a), the 0th and 6th GMM components have similar spatial supports, and thus have high probability of belonging to the same object, even when their colors (shown in side of Fig. 3(b)) are quit dissimilar. We explore the potential of such spatial relations to build pairwise correlation between GMM components as illustrated in Fig. 3(b), where the correlation of two GMM components c_i and c_j is defined as their spatial agreement:

$$\mathbb{C}(c_i, c_j) = \frac{\sum_{I_x} \min(P(c_i|I_x), P(c_j|I_x))}{\min\left(\sum_{I_x} P(c_i|I_x), \sum_{I_x} P(c_j|I_x)\right)}. \quad (2)$$

In GMM representations, the probability vector of a pixel I_x belonging to each GMM components is typically sparse



(a) Illustration of our soft image abstraction

(b) Correlation (c) Clustered GMM (d) Abstraction by [42]

Figure 3. Example of global components representation for the source image shown in Fig. 1. Components in our GMM based representation (a) are further clustered according to the their spatial correlations (b) to get a more meaningful global representation (c), which better capture perceptually homogeneous regions. In (a), the 0-6th sub-images represent the probabilities of image pixels belonging to each GMM component, while the last sub-image shows a reconstruction using these GMM components. An abstract representation by [42] using superpixels is shown in (d).

(with a very high probability of belonging to one of the top two components). This allows us to scan the image once and find all the pairwise component correlations simultaneously. For every pixel, we only choose the top two components with the highest probability and make this pixel only contribute to these two components. In the implementation, we blur the probability maps of the image pixels belonging to each component by a 3×3 uniform kernel to allow the correlation calculation to consider a small surrounding neighborhood.

The correlation matrix of these GMM components is taken as their similarity for message-passing based clustering [22]. We use message-passing based clustering as it does not need a predefined cluster number, making it applicable to an unknown underlying distribution. After such clustering, the probability of each pixel color I_x belonging to each cluster $\mathcal{C}$ is the sum of its probabilities for belonging to all GMM components c in the cluster:

$$p(\mathcal{C}|I_x) = p(\mathcal{C}|I_b) = \sum_{c \in \mathcal{C}} p(c|I_b), \quad (3)$$

where I_b is the quantized histogram bin color of I_x.

In Fig. 3(c), we demonstrate an example of clustering a GMM of 7 initial components to 3 clusters with more homogenous semantic relations. In this example, although the red flower and its dark shadows have quite dissimilar colors, they are successfully grouped together since they cover approximately the same spatial region. Notice that Fig. 3

is a toy example for easier illustration, and we use 15 initial GMM components in order to capture images with more complicated structure in our final implementation.

3.3. Hierarchical representation and indexing

The proposed representation forms a 4-layer hierarchical structure with an index table to associate cross-layer relations efficiently. The 0^{th} layer contains all the image pixels, thus allowing us to generate full resolution saliency maps. During the construction of the subsequent layers, including the histogram representation in the 1^{st} layer, the GMM representation in the 2^{nd} layer, and the clustered representation in the 3^{rd} layer, we record the index table associating the lower layer with higher layer so that the cross layer associations can be achieved highly efficiently. When estimating global saliency cues in §4, we mainly work at the higher layer when feasible in order to allow *large scale perceptually homogenous elements* to receive similar saliency values, and to speed up the computation time. In the hierarchical representation, only the 0^{th} layer contains a large number of elements (the same as image pixels). The number of elements in the subsequent layers are much smaller: ≈ 85, 15, and < 15 in our experiments. Since only a few analysis or assignment steps in our global saliency cue estimation algorithm will go to the bottom layer, finding full resolution saliency maps only requires a computational complexity linear to the number of pixels.

4. Efficient Global Saliency Cues Estimation

4.1. Global uniqueness (GU)

Visual uniqueness in terms of high contrast to other image regions is believed to be the most important indicator of low-level visual saliency [21, 41]. The uniqueness of a global component c_i is defined as its weighted color contrast to all other components:

$$U(c_i) = \sum_{c_j \neq c_i} \exp\left(\frac{D(c_i, c_j)}{-\sigma^2}\right) \cdot \omega_{c_j} \cdot \|\mu_{c_i} - \mu_{c_j}\|, \quad (4)$$

where $D(c_i, c_j)$ is the spatial distance between centroid of the two GMM components c_i and c_j, and we use $\sigma^2 = 0.4$ as in [16] to allow distant regions to also contribute to the global uniqueness.

Notice that the global uniqueness in §4.1 is defined for GMM components in layer 2, thus we only need to consider relations between this very small number of elements, making the estimation very efficient. The mean colors of the abstracted components are needed when estimating the GU saliency. We do not directly work at layer 3 here as the mean color of this top layer cannot capture its potentially complicated color distribution accurately enough. To further incorporate the important spatial overlap correlation, the uniqueness based saliency of GMM components belonging to the same cluster are finally averaged, to encourage semantically correlated regions to receive similar saliency.

4.2. Color spatial distribution (CSD)

While saliency implies uniqueness, the opposite might not always be true [33, 42]. A spatially compact distribution is another important saliency indicator which is an important complementary cue to contrast [25, 36]. Our semantically consistent representation in layer 3 naturally supports the color spatial distribution suggested in [36], while the efficient representation here significantly improves its run time performance and better capture the true spatial distribution of objects.

Referring to [36], we define the horizontal spatial variance of our clustered component $\mathcal{C}$ as:

$$V_h(\mathcal{C}) = \frac{1}{|X|_{\mathcal{C}}} \sum_x p(\mathcal{C}|I_x) \cdot |x_h - M_h(\mathcal{C})|^2, \quad (5)$$

$$M_h(\mathcal{C}) = \frac{1}{|X|_{\mathcal{C}}} \sum_x p(\mathcal{C}|I_x) \cdot x_h, \quad (6)$$

where x_h is the x-coordinate of the pixel x, and $|X|_{\mathcal{C}} = \sum_x p(\mathcal{C}|I_x)$. The spatial variance of a clustered component $\mathcal{C}$ is

$$V(\mathcal{C}) = V_h(\mathcal{C}) + V_v(\mathcal{C}), \quad (7)$$

where the vertical variance $V_v(\mathcal{C})$ is defined similarly to the horizontal variance. We finally define the CSD of $\mathcal{C}$ as:

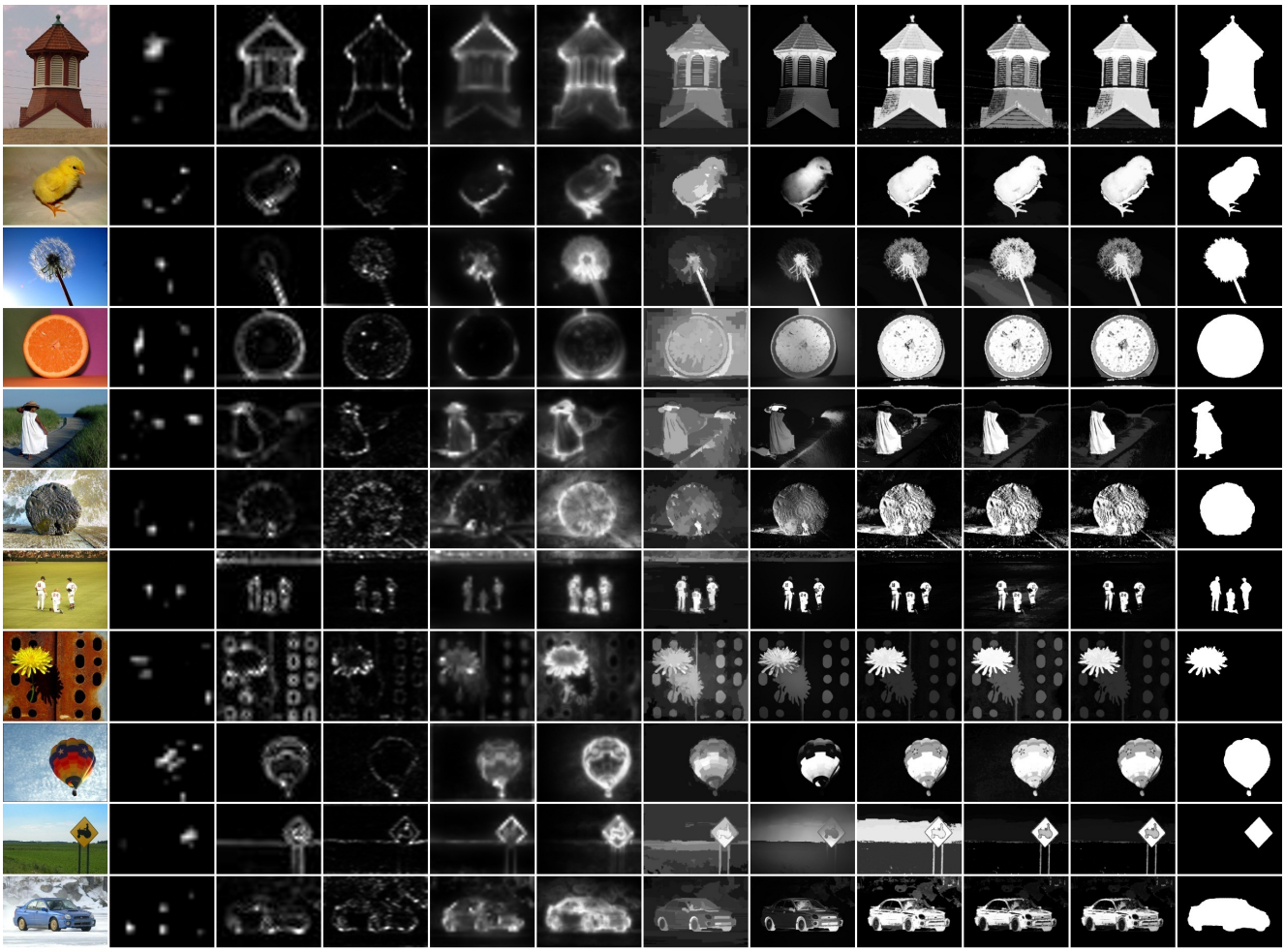
$$S(\mathcal{C}) = (1 - V(\mathcal{C})) \cdot (1 - D(\mathcal{C})), \quad (8)$$

where $D(\mathcal{C}) = \sum_x p(\mathcal{C}|I_x) d_x$ is a center-weighted normalization term [36] to balance the border cropping effect, and d_x is the distance from pixel x to the image center. Both $V(\mathcal{C})$ and $D(\mathcal{C})$ are normalized to [0, 1] for all $\mathcal{C}$ before combining them as in Equ. (8). In the implementation, we set the saliency value for each histogram color as $S(I_b) = \sum_{\mathcal{C}} p(\mathcal{C}|I_b) S(\mathcal{C})$. The saliency of each pixel is efficiently assigned using the indexing scheme between pixels and histogram bins as discussed above.

Since we further consider spatial overlapping relations, the clustered representation (as demonstrated in Fig. 3(c)) better captures semantically homogenous regions as a whole. Due to the nature of the CSD definition, improperly separating a semantically homogenous component will significantly change the spatial variance value of the regions, producing suboptimal results. For instance, the spatial variance of the flower region in Fig. 3(c) increases to about two times of its actual values, if divided to two parts by the GMM based representation as demonstrated in Fig. 3(a).

4.3. Saliency cues integration

The global saliency cues estimation efficiently produces two different saliency maps, where each is a complementary



(a) Source (b) IT[30] (c) MZ[37] (d) SR[28] (e) GB[27] (f) CA[25] (g) RC[16] (h) SF[42] (i) Our CSD (j) Our GU (k) Our GC (l) G-Truth

Figure 4. Visual comparison of previous approaches to our two saliency cues (GU and CSD), final results (GC), and ground truth (GT). Here we compare with visual attention measure (IT), fuzzy growing (MZ), spectral residual saliency (SR), graph based saliency (GB), context aware saliency (CA), region contrast saliency (RC), and saliency filters (SF). See supplementary for results of the entire benchmark.

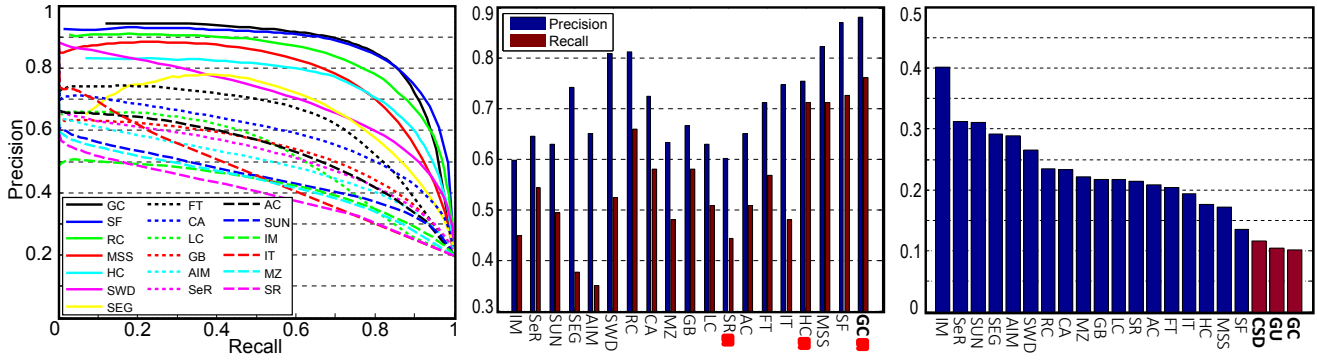
to the other. As also discussed in [26], combining individual saliency maps using weights may not be a good choice, since better individual saliency maps may become worse after they are combined with others. We automatically select between the two saliency maps as a whole to integrate the two cues and generate a final saliency map according to the compactness measure in [26], which uses the compact assumption to select the saliency map with smaller spatial variance. This is achieved by considering the saliency maps as a probability distribution function and evaluating their spatial variance using Equ. (7).

5. Experiments

We exhaustively compare our algorithms' global uniqueness (GU), color spatial distribution (CSD), and global cues (GC) on the largest public available dataset with pixel accurate ground truth annotations [2], and compare with 18 alternate methods. Results of the alternative methods are obtained by one of the following ways: i) results for this famous dataset provided by the original authors (FT[2], SF[42], AC[1], HC[16], RC[16]), ii) running the authors' publicly available source code (GB[27], SR[28], CA[25], AIM[8], IM[39], MSS[3], SEG[43], SeR[45], SUN[52], SWD[20]), and iii) from saliency maps provided by [16] (IT[30], MZ[37], LC[50]). Comparisons with other methods [34, 9, 31, 46] on the same benchmark could be found in the survey paper [7]. Fig. 4 gives a visual comparison of different methods. For objective evaluation, we first use a precision recall analysis, and then discuss the limitations of such measure. Mean absolute errors, as suggested in [42], are further used for objective comparison.

5.1. Precision and recall

Following [2, 16, 42], we first evaluate our methods using precision recall analysis. *Precision* measures the per-



(a) Fixed thresholding (b) Adaptive thresholding [2] (c) Mean absolute error (MAE)

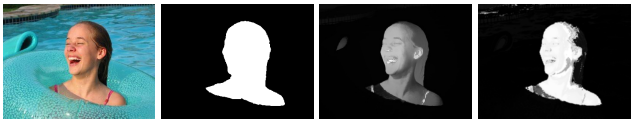
Figure 5. Statistical comparison with 18 alternative saliency detection methods using all the 1000 images from the largest public available benchmark [2] with pixel accuracy saliency region annotation: (a) average precision recall curve by segmenting saliency maps using fixed thresholds, (b) average precision recall by adaptive thresholding (using the same method as in FT[2], SF[42], etc), (c) average MAE. Notice that while our algorithm significantly outperforms other methods, it only achieve similar performance as SF[42] in terms of the precision-recall curve. However, our method achieved 25% improvement over the best previous method in terms of MAE, and the SF paper itself suggested that MAE is a better metric then precision recall analysis for this problem. Results for all methods on the full benchmark as well as our prototype software can be found in the supplementary materials. We highly recommend the readers to see the results in the supplementary as the visual advantages of our results is much more significant than the statistic numbers shows here. Note that other than quantitative accuracy, our results are also perceptually accurate, an important consideration for image processing tasks.

centage of salient pixels correctly assigned, while recall measures the percentage of salient pixel detected. Binary saliency maps are generated from each method using a number of fixed thresholds in $[0, 1, \cdots, 255]$, each resulting in a precision and recall value. The resulting precision recall curves is shown in Fig. 5(a).

While our algorithm significantly outperforms other methods, it only achieved similar performance as SF[42] in terms of the precision-recall curve. However, as discussed in the SF[42] paper itself, neither the precision nor recall measure considers the true negative counts. These measures favors methods which successfully assign saliency to salient pixels but fail to detect non-salient regions over methods that successfully do the opposite. Fig. 6 demonstrates an example which shows the limitation of precision recall analysis. The GC saliency map is better at uniformly highlighting the entire salient object region but its precision recall values are worse. One may argue that a simple boosting of saliency values for SF[42] results would improve it. However, a boosting of saliency values could easily result in the boosting of low saliency values related to background (see the small middle left regions in Fig. 6(c) and more examples in Fig. 4). To further clarify this concern quantitatively, we tried using gamma correlation to refine SF maps for the entire dataset. For gamma values $[0.1, 0.2, \cdots, 0.9]$, we consistently observed worse average MAE (see §5.2) values $[0.46, 0.34, \cdots, 0.14]$.

5.2. Mean absolute error

For a more balanced comparison, we follow Perazzi *et al.* [42] to evaluate the *mean absolute error* (MAE) between a



(a) original (b) GT (c) SF[42] (d) Our GC

Figure 6. Example saliency detection results to demonstrate the limitation of precision recall analysis. When using precision recall analysis, the saliency map in (c) continually achieves near 100% precision for a wide range of recall values, while the saliency map in (d) performs worse because of the small number of false alarm foreground pixels in the upper left corner. However, the saliency map in (d) is closer to the ground truth (b) and better reflects the true salient region in the original image (a).

continuous saliency map $\mathbb{S}$ and the binary ground truth $\mathbb{G}$ for all image pixels I_x, defined as:

$$MAE = \frac{1}{|I|} \sum_{x} |\mathbb{S}(I_x) - \mathbb{G}(I_x)|, \tag{9}$$

where $|I|$ is the number of image pixels.

Fig. 5(c) shows that our individual global saliency cues (GU and CSD) already outperform existing methods in terms of MAE, which provides a better estimate of dissimilarity between the saliency map and ground truth. Our final GC saliency maps successfully reduces the MAE by 25% compared to the previous best reported result (SF[42]).

We note that MAE does not measure discrete classification errors, which is better represented by segmentation performance. When using fixed thresholding, our segmentation performance (see Fig. 5(a)) is not significant better

Method	HC[16]	RC[16]	SF[42]	Our GC
Time (s)	0.01	0.14	0.15	0.09

Table 1. Average time taken to compute a saliency map for images in the benchmark [2] (most images have resolution 300×400).

then the state-of-the-art method [42]. However, when using smarter adaptive thresholding for segmentation [2, 42], the segmentation performance of our method is significantly better than all other methods as evaluated in Fig. 5(b). Moreover, in some application scenarios the quality of the weighted, continuous saliency maps may be of higher importance than the binary masks [42].

5.3. Computational Efficiency

We compare the performance of our method in terms of speed with methods with most competitive accuracy (SF[42]) or similarity to ours (RC[16], HC[16], SF[42]). Tab. 1 compares the average time taken by each method on a laptop with Intel i7 2.6 GHz CUP and 4GB RAM. Performance of all the methods compared in this table are based on implementations in C++. Our method runs in linear complexity with small constant. The two saliency cues based on our abstract global representation already significantly outperform existing methods, while still maintaining faster running times. Our method spends most of the computation time on generating the abstract global representation and indexing tables (about 45%) and finding the CSD saliency (about 51%). Note that our method is more time efficient than concurrent alternatives [49].

6. Limitations

Saliency cues integration based on the compactness measure may not always be effective. e.g., the third example in Fig. 4. Currently we only use this simple measure for convenience, leaving this largely unexplored area as future research. We believe that investigating more sophisticated methods to integrate these complementary saliency cues would be beneficial. Our proposed representation is generally composed of semantically meaningful components from images. It would be also interesting to investigate other saliency cues using this representation, e.g. [4].

As correctly pointed out by recent survey papers, there are two major branches of visual attention modeling: saliency region detection [7] and human fixation prediction [6, 32]. Our method aims at the first problem: finding the most salient and attention-grabbing object in a scene. Our design concept of uniformly highlighting perceptually homogeneous elements might not be suitable for prediction human fixation points, which frequently correspond to sharp local features. We currently only test our algorithm on the most widely used benchmark [2] for saliency region detection so that comparison with other methods are straight forward. Although this dataset only contains images with non-ambiguous salient objects, we argue that efficiently and effectively finding saliency object region for such images is already very important for many important computer graphics and computer vision applications, especially for automatically processing large amount of internet images.

7. Conclusion and Future Work

We have presented a global components representation which decomposes the image into large scale perceptually homogeneous elements. Our representation considers both appearance similarity and spatial overlap, leading to a decomposition that better approximates the semantic regions in images and that can be used for reliable global saliency cues estimation. The nature of the hierarchical indexing mechanism of our representations allows efficient global saliency cue estimation, with complexity linear in the number of image pixels, resulting in high quality full resolution saliency maps. Experimental results on the largest public available dataset show that our salient object region detection results are 25% better than the previous best results (compared against 18 alternate methods), in terms of mean absolute error while also being faster.

Acknowledgement This research was funded by the EPSRC (EP/I001107/1).

References

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

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

[3] R. Achanta and S. Süsstrunk. Saliency detection using maximum symmetric surround. In *IEEE ICIP*, 2010.

[4] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE TPAMI*, 34(11), 2012.

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 2012.

[6] A. Borji, D. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE TIP*, 2012.

[7] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, 2012.

[8] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5:1–24, 2009.

[9] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *IEEE ICCV*, pages 914–921, 2011.

[10] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: Internet image montage. *ACM TOG*, 28(5):124:1–10, 2009.

[11] T. Chen, P. Tan, L.-Q. Ma, M.-M. Cheng, A. Shamir, and S.-M. Hu. Poseshop: Human image database construction and personalized content synthesis. *IEEE TVCG*, 19(5), 2013.

[12] M.-M. Cheng. *Saliency and Similarity Detection for Image Scene Analysis*. PhD thesis, Tsinghua University, 2012.

[13] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu. Salientshape: Group saliency in image collections. *The Visual Computer*, pages 1–10, 2013.

[14] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Salient object detection and segmentation. Technical report, TPAMI-2011-10-0753, 2011.

[15] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. RepFinder: Finding Approximately Repeated Scene Elements for Image Editing. *ACM TOG*, 29(4):83:1–8, 2010.

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

[17] C. Christopoulos, A. Skodras, and T. Ebrahimi. The JPEG2000 still image coding system: an overview. *IEEE T CONSUM ELECTR*, 46(4):1103–1127, 2002.

[18] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 1995.

[19] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof. Saliency driven total variation segmentation. In *IEEE ICCV*, pages 817–824, 2009.

[20] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. In *IEEE CVPR*, pages 473–480, 2011.

[21] W. Eihhauser and P. Konig. Does luminance-constrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, 17:1089–1097, 2003.

[22] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[23] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, and T.-S. Chua. Camera constraint-free view-based 3-d object retrieval. *IEEE TIP*, 21(4):2269–2281, 2012.

[24] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE TIP*, 21(9):4290–4303, 2012.

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

[26] V. Gopalakrishnan, Y. Hu, and D. Rajan. Salient region detection by modeling distributions of color and orientation. *IEEE Trans. Multimedia*, 11(5):892–905, 2009.

[27] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2007.

[28] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE CVPR*, pages 1–8, 2007.

[29] S.-M. Hu, T. Chen, K. Xu, M.-M. Cheng, and R. R. Martin. Internet visual media processing: a survey with graphics and vision applications. *The Visual Computer*, pages 1–13, 2013.

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

[31] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, pages 1–12, 2011.

[32] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, MIT, 2012.

[33] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, 2001.

[34] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *IEEE ICCV*, pages 2214–2219, 2011.

[35] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurbiology*, 4:219–227, 1985.

[36] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, T. X., and S. H.Y. Learning to detect a salient object. *IEEE TPAMI*, 33, 2011.

[37] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*, 2003.

[38] G. Medioni and P. Mordohai. Saliency in computer vision. In L. Itti, G. Rees, and J. Tsotsos, editors, *Neurobiology of Attention*. Elsevier Science, 2005.

[39] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Saliency estimation using a non-parametric low-level vision model. In *IEEE CVPR*, pages 433–440, 2011.

[40] M. Orchard and C. Bouman. Color quantization of images. *IEEE T SIGNAL PROCES*, 39(12):2677–2690, 1991.

[41] D. Parkhurst, K. Law, E. Niebur, et al. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–124, 2002.

[42] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, pages 733–740, 2012.

[43] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila. Segmenting salient objects from images and videos. In *ECCV*, 2010.

[44] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *IEEE CVPR*, pages 37–44, 2004.

[45] H. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 2009.

[46] L. Wang, J. Xue, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue. In *IEEE ICCV*, 2011.

[47] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley. Image saliency: From intrinsic to extrinsic context. In *IEEE CVPR*, pages 417–424, 2011.

[48] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, pages 5:1–7, 2004.

[49] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. CVPR, 2013.

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

[51] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin. A shape-preserving approach to image resizing. *Computer Graphics Forum*, 28(7):1897–1906, 2009.

[52] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32:1–20, 2008.

[53] Y. Zheng, X. Chen, M.-M. Cheng, K. Zhou, S.-M. Hu, and N. J. Mitra. Interactive images: Cuboid proxies for smart image manipulation. *ACM TOG*, 31(4):99:1–11, 2012.