Saliency Aggregation: A Data-driven Approach

Long Mai Yuzhen Niu Feng Liu
Department of Computer Science, Portland State University
Portland, OR, 97207 USA
{mtlong,yuzhen,fliu}@cs.pdx.edu

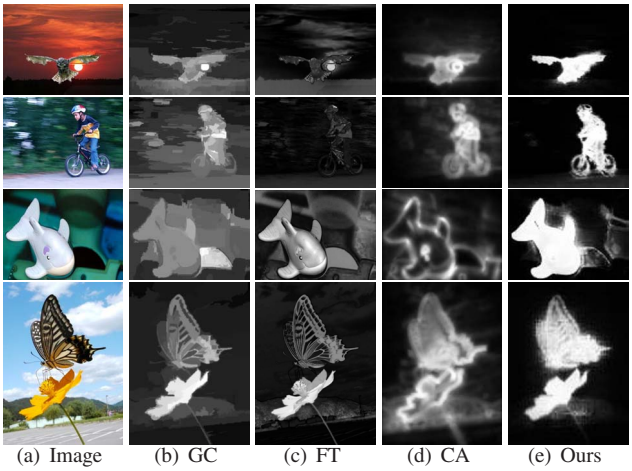
Abstract

A variety of methods have been developed for visual saliency analysis. These methods often complement each other. This paper addresses the problem of aggregating various saliency analysis methods such that the aggregation result outperforms each individual one. We have two major observations. First, different methods perform differently in saliency analysis. Second, the performance of a saliency analysis method varies with individual images. Our idea is to use data-driven approaches to saliency aggregation that appropriately consider the performance gaps among individual methods and the performance dependence of each method on individual images. This paper discusses various data-driven approaches and finds that the image-dependent aggregation method works best. Specifically, our method uses a Conditional Random Field (CRF) framework for saliency aggregation that not only models the contribution from individual saliency map but also the interaction between neighboring pixels. To account for the dependence of aggregation on an individual image, our approach selects a subset of images similar to the input image from a training data set and trains the CRF aggregation model only using this subset instead of the whole training set. Our experiments on public saliency benchmarks show that our aggregation method outperforms each individual saliency method and is robust with the selection of aggregated methods.

1. Introduction

Visual saliency measures low-level stimuli to the human vision system that grabs a viewer's attention in the early stage of visual processing [17]. It has been used in a wide range of computer vision, multimedia, and graphics applications, such as automatic object detection [16], image retrieval [23], video summarization [25], adaptive image compression [7], and content-aware image/video resizing [32].

There is a rich literature on image saliency analysis [1, 2, 4–6, 8–13, 15, 17–20, 22, 24–31, 33, 35–45]. These methods design a variety of biologically plausible models or use data-driven approaches to compute a saliency map from an image. While these methods achieve good results statistically on public benchmarks, each of these methods has its own advantages and disadvantages. As shown in Figure 1, each method works well for some images (or some part of the image), but none of them can handle all the images. More interestingly, different saliency methods can often complement each other. Therefore, the aggregation of these saliency analysis results can likely outperform each individual one, as reported in a recent study [5]. Their study shows that the combination of a few best-performed saliency analysis methods using pre-defined functions, such as averaging, can improve each individual one.



(a) Image (b) GC (c) FT (d) CA (e) Ours

Figure 1. Saliency aggregation. Individual saliency methods, such as GC [6], FT [2], and CA [12], often complement each other. Saliency aggregation can effectively combine their results and perform better than each of them.

In this paper, we present a data-driven approach to saliency aggregation. Our method combines saliency maps from various methods with diverged properties and large performance gaps. To effectively combine these saliency maps, our approach uses machine learning methods to learn an aggregation model that appropriately determines the contribution of each individual method. Specifically, we use a Conditional Random Field (CRF) framework [21] for

saliency aggregation that not only models the contribution from individual saliency map but also the interaction between neighboring pixels. It has been observed that the performance of each individual method varies over images. Therefore, saliency aggregation should be customized to each individual image. To account for the dependence of aggregation on individual image, our approach first selects from a training dataset a subset of similar images to the input image and trains the CRF aggregation model only using this subset instead of the whole training set.

Compared to standard aggregation methods that use predefined combination functions and treat each individual method equally, our data-driven method has the following advantages. First, our method considers the performance gaps among individual saliency analysis methods and better determines their contribution in aggregation. Second, our method considers that the performance of each individual saliency analysis method varies over images and is able to customize an appropriate aggregation model to each input image. As more and more saliency analysis methods have been developed recently, our research provides a way to best use the existing and forthcoming saliency methods and allows the possibility for pushing forward the state-the-art results in saliency analysis.

2. Saliency Aggregation

Our method starts from running a set of m saliency analysis algorithms, $\{M_i \| 1 \leq i \leq m\}$, on a given image I, and produces m saliency maps, $\{S_i \| 1 \leq i \leq m\}$, one for each algorithm. Each element $S_i(p)$ in a saliency map encodes the saliency value at pixel p. The saliency value in each map is normalized to [0, 1]. Our goal is to take these m saliency maps as input and produce a final saliency map S. This section begins with the standard aggregation methods from previous work that use pre-defined combination functions and then elaborates our data-driven saliency aggregation approaches.

2.1. Standard Saliency Aggregation

To serve as our baseline method, we first apply the combination strategies from [5] to saliency aggregation. We then discuss their performance to motivate our data-driven aggregation approaches.

Given a set of m saliency maps $\{S_i \| 1 \leq i \leq m\}$ computed from an image I, the aggregated saliency value $S(p)$ at pixel p of I is modeled as the probability

$$\begin{aligned} S(p) &= P(y_p = 1 | S_1(p), S_2(p), .., S_m(p)) \\ &\propto \frac{1}{Z} \sum_{i=1}^{m} \zeta(S_i(p)), \end{aligned} \tag{1}$$

where $S_i(p)$ represents the saliency value of pixel p in the saliency map S_i, y_p is a binary random variable taking the value 1 if p is a salient pixel and 0 otherwise, and Z is a constant. Following [5], we implemented three different options for the function ζ in Equation 1, including

$$\zeta_1(x) = x,\ \zeta_2(x) = \exp(x),\ \text{and}\ \zeta_3(x) = \frac{-1}{\log(x)}. \tag{2}$$

We used these standard aggregation methods to combine a range of saliency analysis methods and tested them on two public saliency benchmarks FT [2] and SS [31]. Figure 2 (a) shows that when these methods are used to aggregate three best-performed methods, they can produce encouraging results. On the FT benchmark, the aggregation methods produce comparable results to the best individual method, and on the SS benchmark, they outperform each individual one. When the individual methods have large performance gaps, these standard aggregation methods produce less successful results, as shown in Figure 2 (b). The main reason is that they do not consider the performance difference among individual methods and treat them equally. Therefore, the low-performance individual methods compromise the aggregation result. This happens even when only the best-performed methods are aggregated.

2.2. Data-driven Saliency Aggregation

We observe that while various saliency analysis methods often complement each other, there are performance gaps among them. Moreover, the performance of each method varies over individual images. Therefore, saliency aggregation should be individual method-aware and individual image-aware. We design data-driven approaches to achieve such saliency aggregation.

2.2.1 Pixel-wise Aggregation

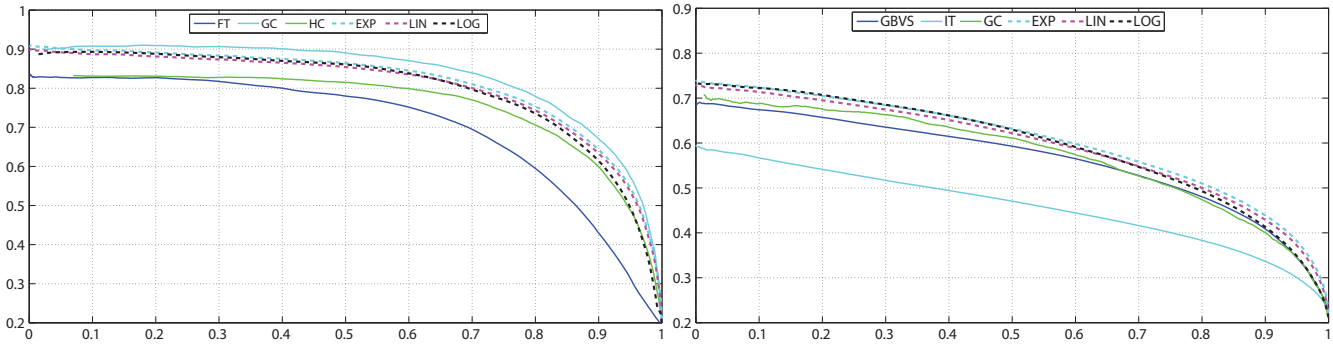
Our first method associates each pixel p with a feature vector $\mathbf{x}(p) = (S_1(p), S_2(p), \cdots, S_m(p))$, where $S_i(p)$ is the saliency value at p in the saliency map S_i. We also assign a binary random variable y_p, which indicates whether the pixel is salient or not. It takes value 1 if p is salient; otherwise 0. We compute the final saliency value $S(p)$ as the posterior probability $P(y_p = 1|x(p))$. Specifically, we model $P(y_p = 1|x(p))$ using the logistic model [3],

$$P(y_p = 1 | x(p); \lambda) = \sigma(\sum_{i=1..m} \lambda_i S_i(p) + \lambda_{m+1}) \tag{3}$$

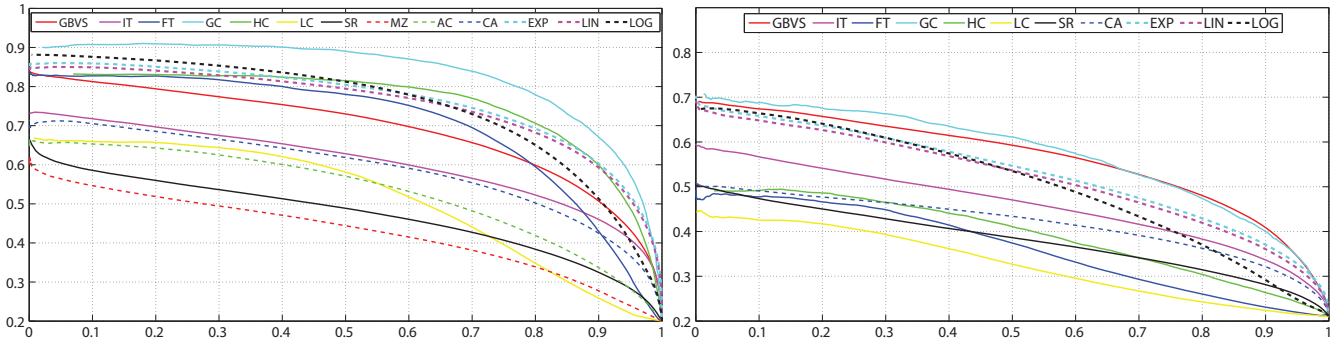
where $\lambda = \{\lambda_i | i = 1..m+1\}$ is the set of model parameters which weigh the contribution of each individual saliency map. Here $\sigma(.)$ denotes the sigmoid function

$$\sigma(z) = \frac{1}{1 + exp(-z)} \tag{4}$$

The parameter λ can be learned using a standard logistic regression technique on the training data. The learnt model



(a) Standard saliency aggregation using the best-performed three individual methods on the FT benchmark (left) and the SS benchmark (right)



(b) Standard saliency aggregation using ten individual methods on the FT benchmark (left) and eight ones on the SS benchmark (right)

Figure 2. Standard saliency aggregation using pre-defined combination functions. LIN, EXP, and LOG refer to the three ζ functions in Equation 2 that are used in the standard combination function defined in Equation 1, respectively.

can then appropriately account for the performance gaps among individual saliency methods.

2.2.2 Aggregation using Conditional Random Field

One potential problem with estimating the saliency value for each pixel individually is its ignorance of the interaction between neighboring pixels. Our second method addresses this problem by modeling saliency estimation using binary Conditional Random Field (CRF) [21]. We use CRF to capture the relation between neighboring pixels. CRF was also used in the previous method by Liu *et al.* for saliency analysis [24]. Their method estimates saliency map directly using image features. In contrast, our method uses CRF to aggregate saliency analysis results from multiple methods.

We model each pixel as a node. Like the pixel-wise aggregation method, we associate each node with a saliency feature vector $\mathbf{x}(p) = (S_1(p), S_2(p), \cdots, S_m(p))$ and a binary random label y_p, 1 for salient and 0 for non-salient. The saliency label of each pixel depends not only on its feature vector, but also the labels of neighboring pixels. The interactions within the pixels also depend on the features. We use a grid-shaped CRF to model the relationship between the label and feature and the feature-dependent relationship between the labels of neighboring pixels. We define the conditional distribution of labels $Y = \{y_p | p \in I\}$ on the features $X = \{x_p | p \in I\}$ as follows,

$$P(Y|X;\theta) = \frac{1}{Z}\exp\Big(\sum_{p\in I} f_d(\mathbf{x}_p, y_p) + \sum_{p\in I}\sum_{q\in N_p} f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)\Big) \quad (5)$$

where p is a pixel in image I, $\mathbf{x}_p$ is its feature, and y_p is its saliency label. θ is the CRF model parameters. $f_d(\mathbf{x}_p, y_p)$ is the feature function that defines the relationship between the feature and label. $f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ is another feature function that defines the feature-dependent relationship between the labels of neighboring pixels p and q. N_p is the set of pixels that are directly connected to p. We consider the 8-connection neighborhood here. Z is a constant.

We define the feature function $f_d(\mathbf{x}_p, y_p)$ based on only the input saliency maps S_i.

$$f_d(\mathbf{x}_p, y_p) = \sum_{i=1}^{m} \lambda_i S_i(p) y_p + \lambda_{m+1} y_p \quad (6)$$

where $\{\lambda_i\}$ is a subset of the CRF model parameters and $S_i(p)$ is the saliency value at pixel p in the saliency map S_i.

The feature function $f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ has two components to model the data-dependent relationship between the

labels of neighboring pixels.

$$f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) = f_e(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) + f_c(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) \quad (7)$$

The first component $f_e(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ encodes the observation that if two pixels have different saliency values according to an individual saliency method, they are likely to have different saliency labels in the aggregation result. Particularly, if a pixel takes a high saliency value than its neighbor in an individual saliency map, it is also likely to take a more salient label after aggregation.

$$f_e(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) = \sum_{i=1}^{m} \alpha_i (\mathbf{1}(y_p = 1, y_q = 0) - \mathbf{1}(y_p = 0, y_q = 1))(S_i(p) - S_i(q)) \quad (8)$$

where α_i are CRF model parameters in this feature function. $\mathbf{1}(.)$ is an indicator function.

$f_c(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ follows the idea from [24] to incorporate the observation that neighboring pixels with similar colors should have similar saliency labels.

$$f_c(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) = -\mathbf{1}(y_p \neq y_q) \exp(-\eta \|I(p) - I(q)\|) \quad (9)$$

where $\|I(p) - I(q)\|$ is the color difference between pixel p and q in the RGB color space. η is set as $(2 < ||I(p) - I(q)||^2 >)^{-1}$, where $< . >$ denotes the expectation operator. We follow the idea from [24] to assign a constant model parameter 1 for this feature function to penalize two neighboring pixels taking different labels if they are similar in color.

The CRF aggregation model parameters $\Theta = \{\lambda, \alpha\}$ are optimized to maximize the likelihood on the training data. The saliency aggregation result for each pixel is the posterior probability of being labeled as salient, which is computed using a standard inference procedure according to the trained CRF aggregation model [21]. Our method uses the UGM CRF toolkit from Mark Schmidt for both training and inferencing[1].

2.2.3 Image-Dependent Saliency Aggregation

The above CRF-based saliency aggregation model considers the performance gaps among individual methods and captures the interaction between neighboring pixels. It, however, does not consider the fact that the performance of a saliency analysis method varies over images. In practice, once the CRF aggregation model parameters are learnt, they are applied to all the new images consistently without considering the performance variation of a saliency analysis method on individual images.

We can further improve the above global saliency aggregation method. Our idea is to train an aggregation model for each individual image. Mathematically, we upgrade the aggregation model from $P(Y|X;\theta)$ into $P(Y|X;\theta(I))$ for each image I. Here, $\theta(I)$ indicates that the model parameters are customized to image I. We train such an image-dependent saliency aggregation model based on the observation that a saliency analysis method has similar performances on similar images. Specifically, given an input image, our method first finds its k nearest neighbors in the training set and then trains a saliency aggregation model using these k images.

Our method uses the GIST descriptor to find similar images. The GIST descriptor has been shown effective in measuring image similarities in computer vision [34]. We compute the distance between two images using the L_2 distance between their GIST descriptors, as suggested in [34].

Figure 3 shows a given image and its nearest neighbors in the SS benchmark. We can see that each individual method performs consistently on these images. For this set of images, the HC method performs the best and it contributes the most to the final aggregation result shown in Figure 3 (h). If we use the whole SS dataset to train the CRF model, the result is heavily affected by the GBVS and GC methods and is less successful, as shown in Figure 3 (g).

3. Experiments

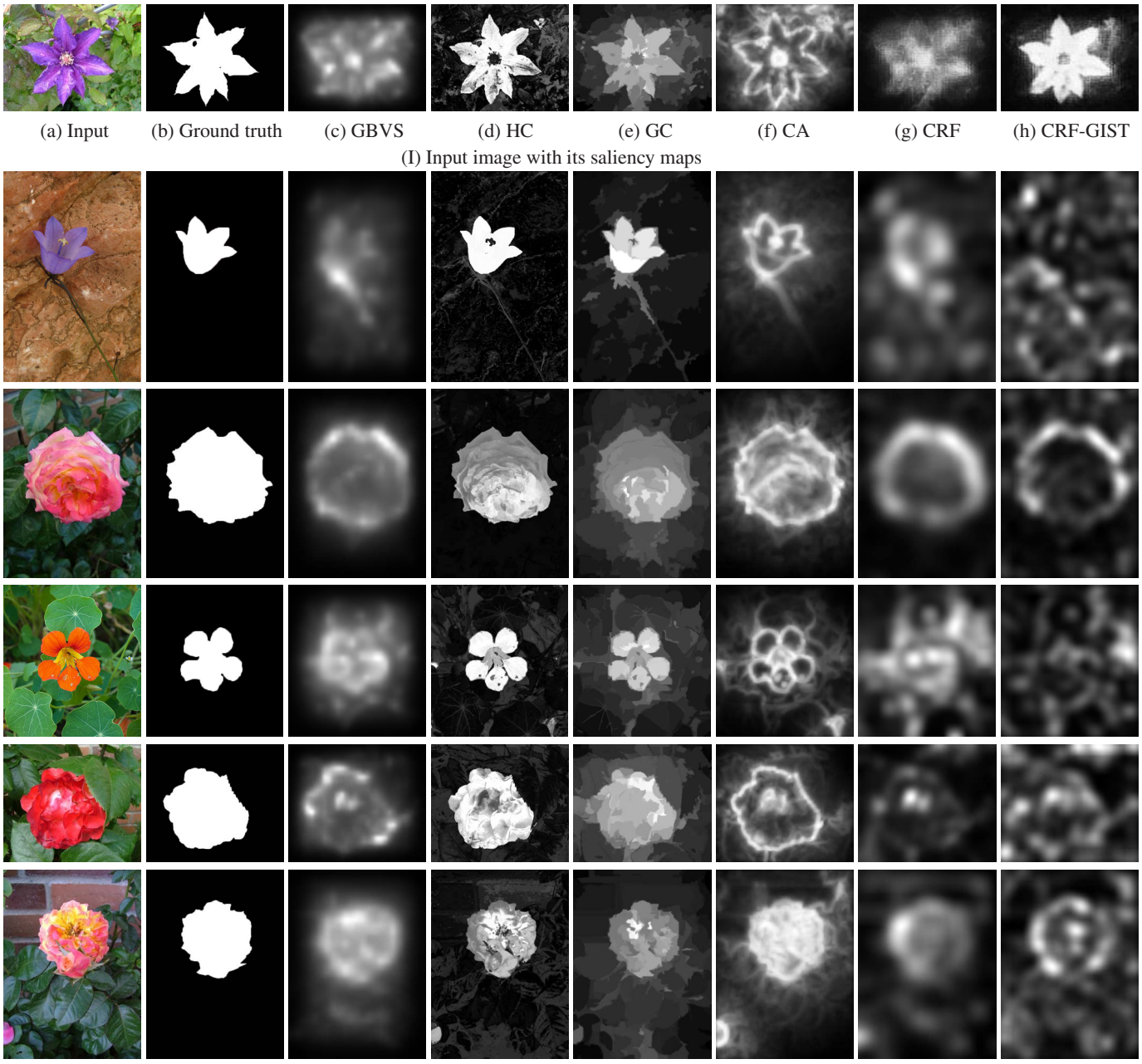
We experimented with our saliency aggregation approaches on two public image saliency benchmarks. The first one is the FT image saliency benchmark from [2]. This dataset contains 1000 images from [24] and includes a manually segmented saliency object mask for each image. The second dataset is the Stereo Saliency dataset (SS) from [31]. This dataset has 1000 stereoscopic images along with the manually segmented saliency masks. In our experiments, we extract all the left images from the stereoscopic images along with their saliency masks and use them in the same way as the FT dataset.

3.1. Aggregation Performance

We first examine the overall performance of our aggregation approaches. For each image in the FT benchmark, we obtained ten saliency maps using saliency analysis methods, including IT [18], MZ [26], LC [46], GBVS [14], SR [15], AC [1], FT [2], HC [6], GC [6], and CA [12]. Specifically, the saliency maps for MZ and AC methods were downloaded together with the FT benchmark [2]. All the others were created using the recent versions of the author-provided implementations. For the SS benchmark, we created eight saliency maps using the same set of methods as the FT benchmark except MZ and AC as their implementations are not available.

We evaluated our methods in a leave-one-out way. Specifically, for each image in the dataset, we use our pixel-wise saliency aggregation approach (PW) described

[1] http://www.di.ens.fr/~mschmidt/Software/UGM.html



(a) Input (b) Ground truth (c) GBVS (d) HC (e) GC (f) CA (g) CRF (h) CRF-GIST

(I) Input image with its saliency maps

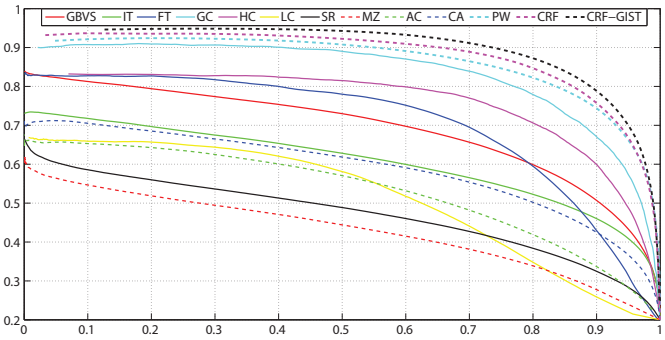
(i) Input (j) Ground truth (k) GBVS (l) HC (m) GC (n) CA (o) IT (p) SR

(II) Nearest neighbors and their saliency maps

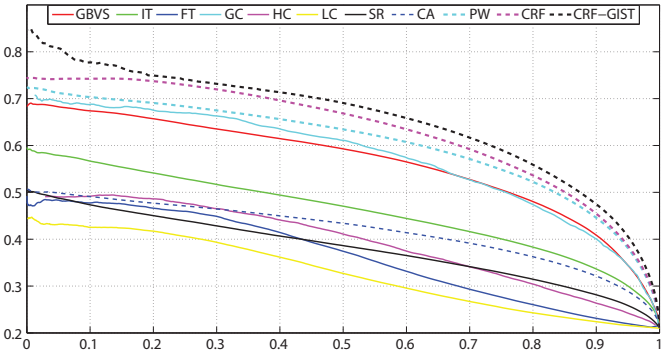
Figure 3. Image-dependent saliency aggregation. Given an input image (a), our approach finds its k nearest neighbors and uses these neighbors to customize a CRF aggregation model for this image. This customized CRF aggregation model produces a better aggregation result (h) than a generic CRF aggregation model (g).

in Section 2.2.1 to train the corresponding saliency aggregation model using the rest of the images in the dataset. The trained model is then used to produce the aggregation saliency map. We test our CRF-based aggregation approach described in Section 2.2.2 in the same way. For the image-dependent saliency aggregation method (CRF-GIST) in Section 2.2.3, our method first finds an image its $k = 50$ nearest neighbors in the rest of the dataset, and then uses these neighbors to train the CRF-based model. We show the precision-recall curves of these approaches on the FT and SS benchmarks in Figure 4 (a) and (b), respectively.

Figure 4 shows that all of our three aggregation methods consistently outperform each individual saliency analysis method. Compared to the baseline aggregation results that are created using pre-defined aggregation functions shown in Figure 2, our data-driven approaches can appropriately



(a) Precision-recall curves on the FT dataset



(b) Precision-recall curves on the SS dataset

Figure 4. Precision-recall curves of our saliency aggregation approaches, including PW, CRF, and CRF-GIST.

consider the performance gaps among individual methods and produce better aggregation results.

There are two categories of our aggregation approaches, pixel-wise aggregation (PW) and CRF-based aggregation (CRF and CRF-GIST). The former computes the saliency value for each pixel independently and the latter considers the interaction between the neighboring pixels. As salient pixels are often grouped together in an image, the CRF-based aggregation approaches perform better than the pixel-wise aggregation, as shown in Figure 4. Meanwhile, our pixel-wise aggregation method still consistently performs better than each individual method. One important reason is that some saliency analysis methods already consider the smoothness of saliency map. For example, the GC method computes saliency values for regions, instead of pixels.

Between our two CRF-based methods, the image-dependent aggregation method (CRF-GIST) performs better than the other (CRF). This confirms that the performance of each individual method varies over images and aggregation should be customized to each individual image. Some representative aggregation results are shown in Figure 5.

3.2. Robustness of Saliency Aggregation

We now examine how the individual saliency methods that are aggregated together affect our aggregation approaches. We performed the following tests. In our first test, we selected the three best-performed saliency methods and aggregated them together using our image-dependent CRF aggregation method. As shown in the first column of Figure 6, our approach produces significantly better results using all the individual methods than the three best-performed ones on the FT dataset and at least comparable results on the SS dataset. In our second test, at each time we removed one individual method from the set of methods that were aggregated together. We find that using all the individual methods will at least not hurt the final aggregation result, as shown in the second column of Figure 6. In fact, removing some individual method will downgrade the aggregation result. These two tests show the capability of our aggregation approach in making the best use of individual saliency analysis methods.

To further examine the robustness of our approach, we add a random map as one of the basic saliency maps used in aggregation. This "faked" saliency detector randomly assigns each pixel a saliency value in the range $[0, 1]$ according to a uniform distribution. We find that our aggregation results with/without this random saliency map are almost the same, as shown in the last column of Figure 6. This demonstrates that our method is robust against such a noisy "contributor".

3.3. Discussions

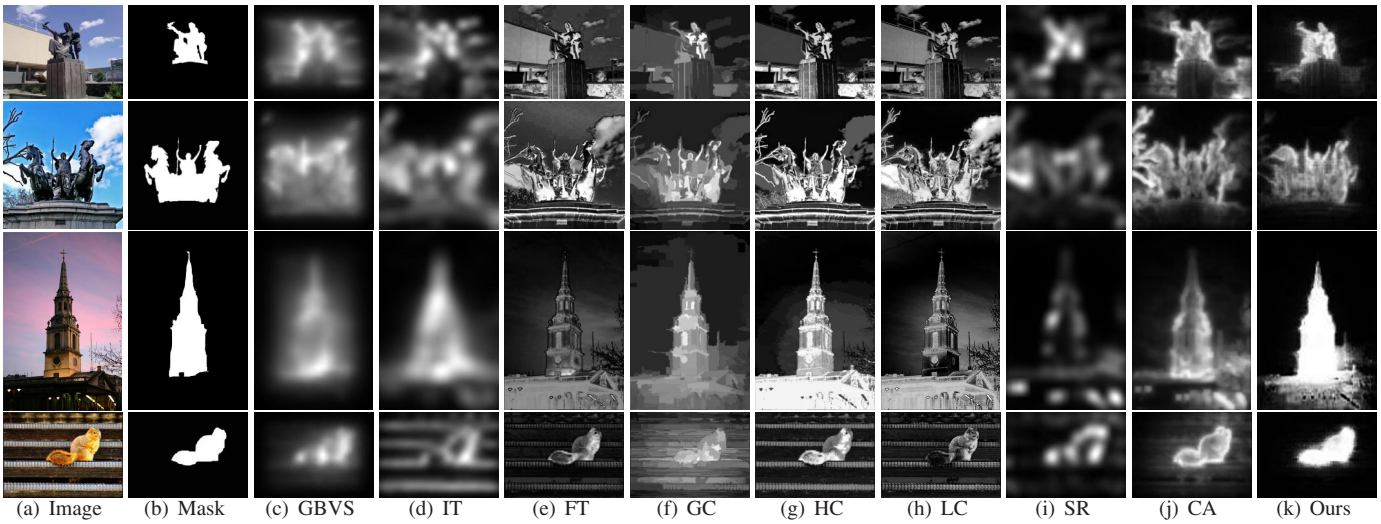
Our experiments show that saliency aggregation can consistently improve the performance of each individual saliency analysis method. However, there is a limit of improvement. Because aggregation is based solely on the saliency maps from individual methods, when all the individual methods fail to identify a salient region in an image, saliency aggregation will usually fail too. On the other hand, our aggregation result can benefit from the progress of the research on individual saliency analysis methods.

Our image-dependent saliency aggregation method currently uses the GIST descriptor to find similar images to an input one. Its performance will sometimes be affected if the GIST method does not find similar images. This problem can be addressed by incorporating better image similarity measurements.

As our method requires results from all the individual methods, it is slower than each individual one. Besides, our aggregation method (CRF-GIST), including the CRF model training and inference steps, takes about 40 seconds on a desktop machine with an i7 3.40 GHz CPU for k=50.

4. Conclusion

In this paper, we presented data-driven approaches to saliency aggregation that integrate saliency analysis results from multiple individual saliency analysis methods. We designed and discussed three saliency aggregation approaches. All our approaches can consistently perform bet-



(a) Image (b) Mask (c) GBVS (d) IT (e) FT (f) GC (g) HC (h) LC (i) SR (j) CA (k) Ours

Figure 5. Saliency aggregation examples. We show the input images (a), the ground-truth (b), individual saliency maps (c $\sim$ j), and our aggregation results using our image-dependent CRF aggregation method.

ter than each individual saliency method. Among these methods, our image-dependent CRF-based approach that considers the interaction among pixels, the performance gaps among individual saliency analysis methods, and the dependent of saliency analysis on individual image, works the best. Our work provides a robust way to combine individual saliency analysis methods into a more powerful one.

Acknowledgements. We would like to thank Tetsuya Shimizu, Wayne Karberg, Romas Sniegeckas and Flickr users, including voxel123, Marius C., and Balliolman for letting us use their photos under a Creative Commons license or with their permissions. This work was supported by NSF CNS-1205746 and CNS-1218589.

References

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *International Conf. on 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] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[4] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *IEEE CVPR*, 2012.

[5] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *Proc. European Conference on Computer Vision (ECCV), Florence, Italy*, Oct 2012.

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

[7] C. Christopoulos, A. Skodras, and T. Ebrahimi. The jpeg2000 still image coding system: an overview. *IEEE Trans. on Consumer Electronics*, 46(4):1103–1127, 2000.

[8] Y. Cohen and R. Basri. Inferring region salience from binary and gray-level images. *Pattern recognition*, 36:2349–2362, 2003.

[9] L. Congyan, T. Nguyen, harish Katti, K. Yadati, S. Yan, and M. Kankanhalli. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, 2012.

[10] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *IEEE ICCV*, 2011.

[11] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8:1–18, 2008.

[12] S. Goferman, L. Zelnik-manor, and A. Tal. Context-aware saliency detection. In *IEEE CVPR*, 2010.

[13] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *IEEE CVPR*, pages 1698–1705, 2009.

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

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

[16] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

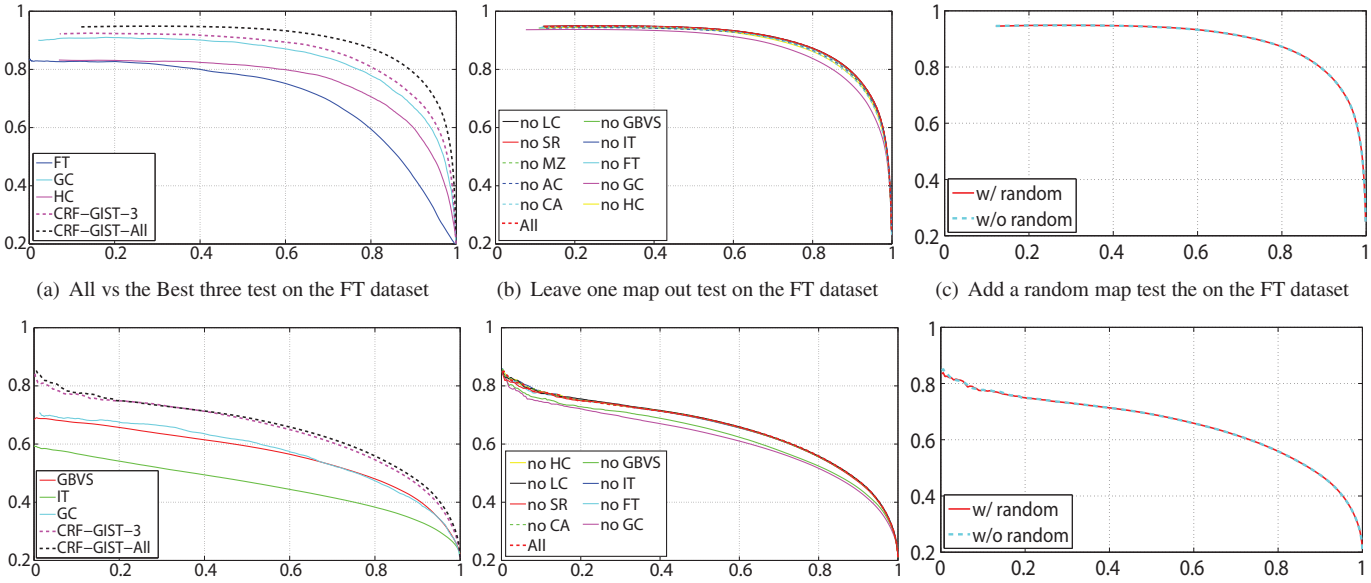
[17] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2:194–203, 2001.

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

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE ICCV*, 2009.

[20] G. Kim, D. Huber, and M. Hebert. Segmentation of salient regions in outdoor scenes using imagery and 3-d data. In *IEEE WACV*, 2008.

[21] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.



(a) All vs the Best three test on the FT dataset (b) Leave one map out test on the FT dataset (c) Add a random map test the on the FT dataset

(d) All vs the Best three test on the SS dataset (e) Leave one map out test on the SS dataset (f) Add a random map test the on the SS dataset

Figure 6. Effect of individual saliency maps on saliency aggregation. The first column shows the precision-recall curves from aggregations using all the saliency maps and the best three ones. The second column shows the aggregation results when one saliency map is removed at each time. The third column shows the test when a random map is added into the aggregation process as a faked saliency map.

[22] Z. Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6:9–16, 2002.

[23] H. Liu, X. Xie, X. Tang, Z.-W. Li, and W.-Y. Ma. Effective browsing of web image search results. In *ACM International Conf. on Multimedia Information Retrieval*, 2004.

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

[25] Y. Ma, L. Lu, H. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings ACM Multimedia 2002*, pages 533–542, 2002.

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

[27] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in highly dynamic scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:171–177, 2010.

[28] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *IEEE ICCV*, 2009.

[29] R. Margolin, L. Zelnik-Manor, and A. Tal. Saliency for image manipulation. *The Visual Computer*, pages 1–12, 2012.

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

[31] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. In *IEEE CVPR*, 2012.

[32] Y. Niu, F. Liu, X. Li, and M. Gleicher. Warp propagation for video resizing. In *IEEE CVPR*, pages 537–544, 2010.

[33] H. Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40(10-12):1183–1201, 2000.

[34] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

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

[36] G. Sharma, F. Jurie, and C. Schmid. Discriminative Spatial Saliency for Image Classification. In *IEEE CVPR*, 2012.

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

[38] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *IEEE Trans Pattern Anal Mach Intell*, 2012.

[39] X. Sun, H. Yao, and R. Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *IEEE CVPR*, 2012.

[40] A. Torralba. Contextual influences on saliency. *Neurobiology of attention*, pages 586–593, 2005.

[41] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[42] 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.

[43] P. Wang, J. Wang, G. Zeng, J. Feng, and H. Zha. Salient object detection for searched web images via global saliency. In *IEEE CVPR*, 2012.

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

[45] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *IEEE CVPR*, 2012.

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