SALIENCY DETECTION USING REGRESSION TREES ON HIERARCHICAL IMAGE SEGMENTS

Gökhan Yildirim, Appu Shaji, Sabine Süsstrunk

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
{gokhan.yildirim,appu.shaji,sabine.susstrunk}@epfl.ch

ABSTRACT

The currently best performing state-of-the-art saliency detection algorithms incorporate heuristic functions to evaluate saliency. They require parameter tuning, and the relationship between the parameter value and visual saliency is often not well understood.

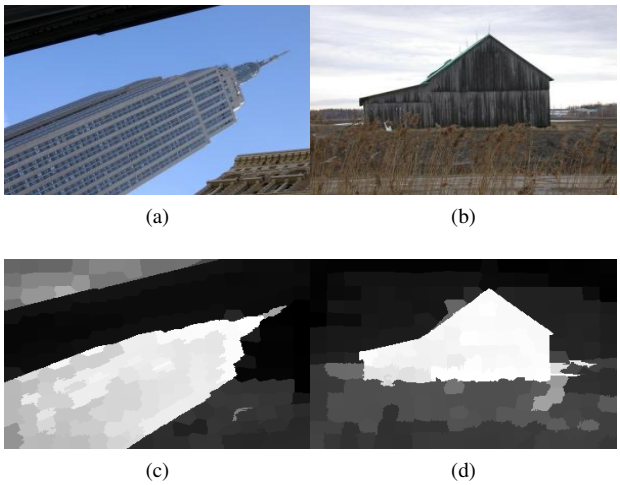
Instead of using parametric methods we follow a machine learning approach, which is parameter free, to estimate saliency. Our method learns data-driven saliency-estimation functions and exploits the contributions of visual properties on saliency. First, we over-segment the image into superpixels and iteratively connect them to form hierarchical image segments. Second, from these segments, we extract biologically-plausible visual features. Finally, we use regression trees to learn the relationship between the feature values and visual saliency. We show that our algorithm outperforms the most recent state-of-the-art methods on three public databases.

***Index Terms*—** saliency, superpixels, hierarchical regression, regression tree

1. INTRODUCTION

When we look at an image, distinctive objects in this visual stimulus instantly attract our attention. Visual saliency can thus be defined as the collection of the perceptual attributes that make these distinctive items stand out from their neighborhood and, consequently, be easily detectable.

Low-level visual saliency deals with the contrast between the fundamental properties of image regions, such as color and texture. The studies on this topic are inspired by the human visual system (HVS). One of the pioneering investigations in this field was done by Itti et al. [1], where center-to-surround differences in color, intensity, and orientations are combined at different scales to generate a final saliency map. As a consequence, the most recent and best-performing low-level saliency detection algorithms, summarized in Section 2, evaluate the final saliency by combining biologically-plausible [1] color contrast features. However, their performances depend highly on the heuristic function selection and the tuned parameters, because these features do not always imply saliency, and the relationship between the parameter values and saliency is not always straightforward. For example, in Figure 1(a), we should not penalize border pixels, because the salient object touches the image boundary. However, in Figure 1(b), non-salient sky pixels have high contrast with the rest of the image, and should be penalized. These handcrafted rules, as in this case, can contradict each other and can lead us to inaccurate saliency estimations.



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

Fig. 1: Image examples requiring contradicting saliency rules, and outputs of our saliency method.

Therefore, in Section 3, we follow a different approach and use visual features to *learn* which visual properties, individually and jointly, make an image region salient. This provides accurate estimations, as shown in Figure 1(c)-(d). Our data-driven method is able to model the visual saliency in difficult images with low object-background contrast, such as the images in the SED-100 [2] and the SOD [3] datasets. Consequently, in Section 4, we show that our method outperforms the most recent state-of-the-art methods on the MSRA-1000 [4] dataset and especially on the SED-100 and the SOD datasets. In Section 5, we discuss that our learning-based model is more favorable over parametric methods, because it has the potential to benefit from large-scale data.

This work was supported by the Swiss National Science Foundation under grant number 200021 143406 / 1.

2. RELATED WORK

The goal of salient-object detection is to develop techniques that detect the salient objects with pixel-level precision. For this purpose, researchers use object related information, such as contrast [5], context [6], concavity [7], objectness [8], composition [9], image edges, and borders [10].

Another method for finding salient objects is to structurally define the foreground and background regions. This can be achieved by using surround prior for background [11], employing global contrast based on histograms [12], exploiting uniqueness and distribution of the image segments [13], considering background as a low rank matrix [14], fusing hierarchical saliency maps [15], modeling the saliency detection problem as a Markov chain [16], and solving graph queries with different boundary conditions [17].

Visual saliency is intrinsically linked with the image content. However, the current state-of-the-art methods only use heuristics to combine feature values to estimate visual saliency. We propose a machine learning algorithm, which models the relationship between image segments at various scales, and is generalizable to images with different content.

3. SALIENCY DETECTION

Our method extracts superpixels of an image, forms hierarchical segments using superpixel average color vectors, computes visual features from these segments, uses them to learn and estimate visual saliency, and finally employs belief propagation to combine the saliency values of different levels.

3.1. Hierarchical Representation

We over-segment an input image into superpixels using Achanta et al.'s method [18]. These superpixels are then iteratively merged into larger segments. For the merging operation, we first find the spatial neighbors of each superpixel. If two spatially neigboring superpixels are each other's nearest neighbors (in terms of mean color vectors) in CIELAB space as well, these superpixels are merged into a larger segment. After all possible superpixel pairs are merged, their properties, such as mean color vector and segment size, are updated and the method moves onto the next level. The algorithm stops if there is nothing left to merge. The merging method is simple and non-parametric, yet it generates an accurate hierarchical representation of the input image. In Figure 2, hierarchical levels of an image are shown.

Fig. 2: Hierarchical representation of an image.

3.2. Image Segment Features

Based on the properties of the human visual system and the state-of-the-art methods explained in Section 2, we choose the visual features in Table 1. We extract a feature vector from each $\mathbf{S}_i^h$ (the i^{th} superpixel at hierarchy level h). Using $\mathbf{S}_i^h$, we compute the average color ($\mathbf{C}_i^h$) and color histogram ($\mathbf{H}_i^h$) in CIELAB color space, histogram of oriented gradients ($\mathbf{O}_i^h$) [19], the number of pixels in $\mathbf{S}_i^h$ normalized by the total number of pixels in the image (a_i^h), and the spatial variance in the horizontal and the vertical dimensions (X_i^h, Y_i^h). For the first four rows in Table 1, we use the functions $g(.) \in \{\min, \max, \text{mean}, \text{median}\}$, global ($\Omega^h$) and local ($\Omega_i^h$) superpixel sets to extract 8 features. Here, Ω^h is the set of all superpixels at hierarchy level h, and Ω_i^h is the set of spatially neighboring superpixels of $\mathbf{S}_i^h$. We have 41 features in total.

Table 1: The visual features that are used in regression trees.

Features for $\mathbf{S}_i^h$	Description	Global	Local
Color difference	$g(\|\|\mathbf{C}_i^h - \mathbf{C}_j^h\|\|_2)$	$j \in \Omega^h$	$j \in \Omega_i^h$
Histogram difference	$g(\|\|\mathbf{H}_i^h - \mathbf{H}_j^h\|\|_2)$		
HOGs difference	$g(\|\|\mathbf{O}_i^h - \mathbf{O}_j^h\|\|_2)$		
Size difference	$g(\|a_i^h - a_j^h\|)$		
Relative size	a_i^h		
Spatial variation	X_i^h, Y_i^h		
Color entropy	Entropy($\mathbf{H}_i^h$)		
HOGs entropy	Entropy($\mathbf{O}_i^h$)		
Element distribution	[13]		
Edge fit	[10]		
Border pixels	[10]		

3.3. Learning Regression Trees

A superpixel $\mathbf{S}_i^h$ is represented by its 41-dimensional feature vector. The precision of an individual superpixel is calculated as follows:

$$P(\mathbf{S}_i^h) = \frac{t_i^h}{t_i^h + f_i^h} \quad (1)$$

Here, $P(\mathbf{S}_i^h)$ is the precision of $\mathbf{S}_i^h$. t_i^h and f_i^h are the number of true and false positives, respectively, and are computed using $\mathbf{S}_i^h$ and the ground truth. In our method, we use the superpixel features in Table 1 as input and the $P(\mathbf{S}_i^h)$ as output in the gradient boosting tree implementation in [20]. A node of a tree compares a single feature value with a learned threshold. We test each image by employing leave-one-out cross-validation in its dataset. In order to avoid over-fitting, we use a shrinkage factor [21] and trees of depth 2. These trees perform at most 3 threshold operations (from root to leaf), which balances between a weak tree (decision stump) and a strong tree using all 41 features ($\log_2(41) \approx 5.4$). Despite the similarities between random forests and boosting trees, we choose the latter, because, its progressive nature enables us to better analyze the psychophysical aspects of visual saliency.

3.4. Feature Analyses

The importance of a feature is correlated to how frequently it is selected for thresholding on the nodes of regression trees. In Figure 3, the five most frequently selected features for the MSRA-1000, SED-100, and SOD datasets are illustrated. The "Border pixels" feature is one of the most frequent features. This implies that it is a powerful feature for distinguishing a salient object from a background region. In addition, it is more frequent in the MSRA-1000, because salient objects seldom reside on image boundaries in that dataset. In terms of "Element distribution", which measures the spatial variance of a superpixel color [13], the X dimension is frequently selected. This can be related to the fact that the field of view of humans is larger in horizontal dimensions, allowing them to easily identify the horizontal variations (or salient objects). Global "Color difference" features are more frequent than their local counterparts (not in Figure 3 due to its frequency), because they give plausible results on every hierarchy level, whereas the Local "Color difference" features are only powerful at the correct level (or scale), where the salient object is properly segmented.

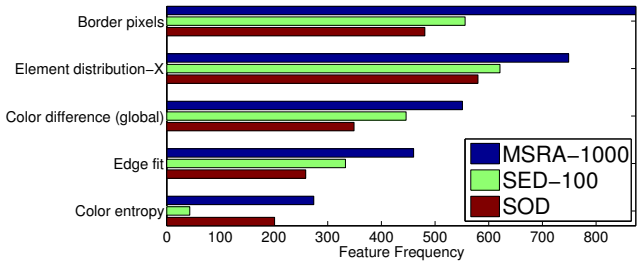


Fig. 3: The five most frequently selected features.

In Figure 4, we summarize the effect of certain features on the estimated saliency value. For example, as we mention in Section 1, color difference does not directly imply saliency in any of the datasets. The saliency value is correlated to the entropy of the color histogram. This means that segments with low entropy, i.e. regions with less colors, such as sky, sea, and grass, are less likely to be salient. The correlation decreases from the MSRA-1000 to the SED-100, which makes the MSRA-1000 an easier dataset compared to the others. This is also supported by the results in Section 4. Also, smaller distributions (variances) are favored to be salient, where the effect of the X dimension is significantly more prominent than the Y dimension.

3.5. Combining Levels with Belief Propagation

Naively combining the individual saliency maps of different levels might result in an inaccurate saliency estimation [15]. Therefore, we employ a belief propagation inference method [22]. This method effectively solves the following energy minimization problem:

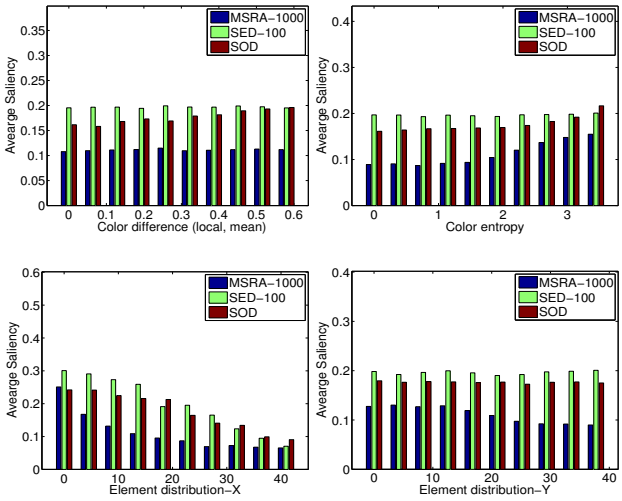


Fig. 4: The effect of some features to saliency value (see the text for further explanations).

$$E = \sum_{h=1} \sum_{i} ||l_i^h - s_i^h||_2^2 + \sum_{h=1} \sum_{i, \mathbf{S}_i^h \subseteq \mathbf{S}_j^{h+1}} ||l_j^{h+1} - l_i^h||_2^2 \quad (2)$$

Here, E is the energy, l_i^h is the saliency value after the inference, s_i^h is the initial saliency value that is estimated by the regression trees for $\mathbf{S}_i^h$. The two terms in the equation represent the data penalty and the smoothness penalty along hierarchy levels, respectively. The final saliency value $F(.)$ of a pixel at position $\mathbf{p}$ is calculated as:

$$F(\mathbf{p}) = \sum_{h=1} \sum_{j, \mathbf{p} \subseteq \mathbf{S}_j^h} l_j^h \quad (3)$$

4. RESULTS

We evaluate the performance of our method, HR (Hierarchical Regression), on the MSRA-1000, the SED-100, and the SOD datasets. The precision-recall curves of our method are compared to eight other techniques (the results are obtained using their codes), which are IT [1], FT [5], SF [13], LR [14], HSD [15], CH [10], GMR [17], and AMC [16] in Figure 5. We mainly focus our discussions on the best performing techniques, which are CH [10], GMR [17], and AMC [16]. On the MSRA-1000 dataset, we are marginally better than the other methods, as the performance of that dataset is overfitted and noisy labeling does not allow further significant improvements. On the other two datasets, our method outperforms the state-of-the-art techniques. As we discuss in Section 1, instead of defining heuristic rules, our algorithm learns the saliency from a group of images and successfully **transfers** this knowledge to other images. Regression trees effectively form complicated functions that are otherwise not so easy

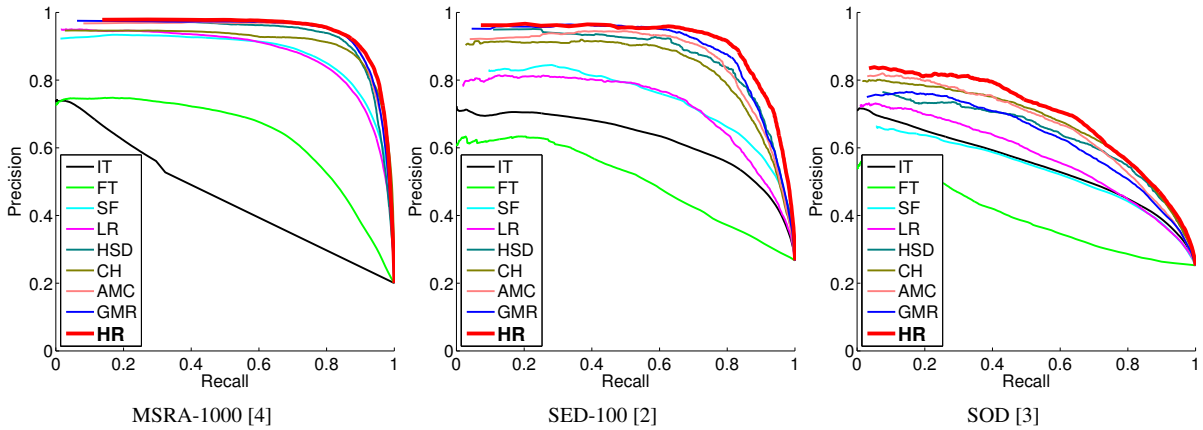


Fig. 5: The precision-recall curves calculated on different datasets.

to handcraft, even with the powerful structural methods using hierarchical context as in CH [10] or graph theory as in AMC [16] and GMR [17].

Some of the saliency detection results of our method and other methods are shown in Figure 6. There are two images from the MSRA-1000, SED-100, and SOD datasets in successive rows. In the first image, our method uniformly detects the salient object with a high saliency value, whereas for the other methods, local contrast between some image regions cause inaccurate saliency estimations.

Image-boundary penalty in CH [10], boundary node assumption in AMC [16], and different boundary conditions in GMR [17] prevent them to detect salient objects that are close to image boundary in the second and the third image. GMR [17] is less affected in the second image due to the combination of multiple boundary assumptions. On the same images, our method enforces a learned top-down contextual information to ensure these objects are uniformly highlighted.

In the fourth and fifth images, due to the low local color contrast, some parts of the salient objects are missed. HR is able to recover those regions by benefiting from "Edge Fit" feature. The last image is a difficult case where the saliency is very subjective. Therefore, all of the methods fail.

The methods, CH, AMC, GMR, [10, 16, 17] use color difference and contextual information as main features. However, there can also be textual contrast, spatial and chromo-spatial constraints for salient objects. As it is very complicated to build and tune a generalized function for all these features, our learning-based method outperforms them.

5. CONCLUSION

The simple saliency features, such as image border penalty and color contrast, when combined in heuristic functions with tuned parameters, do not yield the best saliency detection results. We show that our learning-based technique outperforms the parametric state-of-the-art methods by learning and transferring the saliency knowledge between images.

We abstract an image via hierarchical segmentation and analyze the relationship between the visual properties of the image segments and corresponding saliency values by using regression trees. We employ the same trees in saliency estimation and perform a scale-aware saliency inference via belief propagation.

Parametric methods works on a single image and have a constant performance, regardless of the number of images. Non-parametric methods utilizing large-scale datasets perform better in problems, varying from computational linguistics [23] to object recognition [24], where there is no analytical solution. As the image data are rapidly increasing every day, our data-driven method is preferable over parametric methods in terms of adapting to the new and more difficult saliency estimation problems.

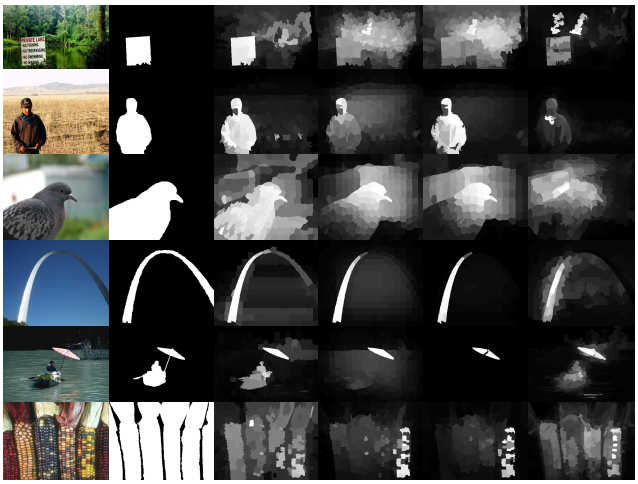


Fig. 6: Comparison of saliency maps produced by different methods (GT: Ground Truth).

6. REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proceedings of IEEE CVPR*, 2007, pp. 1–8.

[3] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proceedings of IEEE CVPR Workshops*, 2010, pp. 49–56.

[4] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proceedings of ICVS*, 2008, pp. 66–75.

[5] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proceedings of IEEE CVPR*, 2009, pp. 1597 – 1604.

[6] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proceedings of IEEE ICCV*, 2011, pp. 105–112.

[7] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *Proceedings of IEEE ICCV*, 2011, pp. 233–240.

[8] K. Chang, T. Liu, H. Chen, and S. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proceedings of IEEE ICCV*, 2011, pp. 914–921.

[9] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Proceedings of IEEE ICCV*, 2011, pp. 1028–1035.

[10] X. Li, Y. Li, C. Shen, A. Dick, and A. van den Hengel, "Contextual hypergraph modelling for salient object detection," *Proceedings of IEEE ICCV*, 2013.

[11] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of ECCV*, vol. 7574, pp. 29–42. 2012.

[12] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proceedings of IEEE CVPR*, 2011, pp. 409–416.

[13] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of IEEE CVPR*, 2012, pp. 733–740.

[14] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proceedings of IEEE CVPR*, 2012, pp. 853–860.

[15] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of IEEE CVPR*, 2013, pp. 1155–1162.

[16] B. Jiang, L. Zhang, H. Lu, M. Yang, and C. Yang, "Saliency detection via absorbing markov chain," *Proceedings of IEEE ICCV*, 2013.

[17] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of IEEE CVPR*, 2013, pp. 3166–3173.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE CVPR*, 2005, vol. 1, pp. 886–893 vol. 1.

[20] R. Sznitman, C. Becker, F. Fleuret, and P. Fua, "Fast object detection with entropy-driven evaluation," in *Proceedings of IEEE CVPR*, 2013, pp. 3270–3277.

[21] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The Elements of Statistical Learning*, pp. 337–387. 2009.

[22] M. Schmidt, K. Murphy, G. Fung, and R. Rosales, "Structure learning in random fields for heart motion abnormality detection," in *Proceedings of IEEE CVPR*, 2008, pp. 1–8.

[23] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Transactions on Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[24] A. Torralba, R. Fergus, and W.T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on PAMI*, vol. 30, no. 11, pp. 1958–1970, 2008.