Using Machines to Improve Human Saliency Detection

Nikhil Rao
Department of Electrical and Computer Engineering
University of Wisconsin - Madison, Madison, Wisconsin 53706
Email: nrao2@wisc.edu

December 9, 2010

Contents

Acknowledgment

I would like to thank my advisor, Prof. Robert Nowak for his great support and encouragement throughout the project. The insight into the subject and the ideas he proposed went a long way in maiking the project a success. I would also like to thank my colleague Tyler Karrels, without whom this work would be impossible to finish in the time frame. The cognitive science aspect of the project was wonderfully overseen by Prof. Timothy Rogers, and his students Lang Chen and Joseph Harrison. My special thanks to Joe who helped in collecting data for studies conducted on human subjects. This work was partially supported by the Raytheon Company. I thank Nitesh Shah and Visar Berisha for their support and helpful feedback from time to time.

Abstract

Humans are adept at identifying informative regions in individual images, but it is a slow and often tedious task to identify the salient parts of every image in a large corpus. A machine, on the other hand, can sift through a large amount of data quickly, but machine methods for identifying salient regions are unreliable. In this work, we develop a new method for identifying salient regions in images and compare this to two previously reported approaches. We then consider how such machine-saliency methods can be used to improve human performance in a realistic target-detection task.

1 Introduction

Consider the task of an analyst searching for signs of human habitation amongst images taken by drone planes over some vast and sparsely occupied region. The drones are capable of providing masses of data, but only a few will contain items of interest. The human analysts are skilled at detecting the interesting images when they appear, but the data is generated much more rapidly than it can be processed: the analyst is overwhelmed with an ever-increasing number of images, most containing nothing of interest. How can machines help the analyst to find the interesting images?

One idea might be to apply pattern recognition methods to the image data, but such approaches are both resource-intensive – requiring large amounts of human-labeled training data – and notoriously unreliable. In this paper we instead consider a salience-based approach to target detection. Salience describes the tendency for certain subregions of a scene to "stand out" from their surroundings in human perception. In detection problems like that described above, the target images often contain highly salient items viz. items that are dissimilar to their backgrounds, such as a house in the desert or a boat in the water, whereas the uninteresting images often do not. In such cases, salience might provide a useful cue for finding target images. Machines could aid in the detection task by automatically computing the maximal salience of each image in the corpus and prioritizing highly salient images for presentation to the analyst.

Empirical research in human perception suggests that salient subregions of a scene are those that carry high information relative to other subregions [2]. There are, however, at least two different ways of thinking about the informativeness of a given region. The first proposes that a region is informative if its features cannot be predicted by the features present in its immediate spatial surroundings [1]. We will refer to this as the center-surround approach to saliency. The second approach stipulates that a region carries high information if it contains features that are dissimilar to those appearing in the scene globally. We will refer to this as the similarity-based approach to saliency. Human saliency detection appears to be influenced by both the spatial distribution of features in a scene (consistent with the center-surround approach; see [4,5,7]) and by the global similarity of features (consistent with the similarity-based view; see [8]).

Computer models based on the center-surround approach have been developed by [7], but we are unaware of prior work proposing a strict similarity-based model of salience. After reviewing existing methods for salience in Section 2, we will develop a novel similarity-based approach in Section 3. We outline the ranking schemes we developed in Section 4. We will then compare this model with the others reviewed in Section 5. The different models will be used to rank order the items in a corpus of 500 satellite images by maximal salience. Fifteen percent of the images contain evidence of human habitation, and we will investigate how many of these "target" items appear among the highly-ranked images for each algorithm. We will see that the models vary significantly in their performance on this task. In an empirical study, we will then consider how the machine rankings can be used to dramatically improve human performance in the target detection task. In Section 6 we will present conclusions and areas for future research.

2 Existing Methods

Recently, robotic vision, anomaly detection, and vast image databases have spurred interest in computer algorithms for visual saliency. Many methods have been developed to find salient objects in images, but much of the work involves knowledge of the target object. Here we consider only "bottom-up" methods that do not require such knowledge. In [4, 6, 7], the authors develop a model inspired by the center surround architecture of neurons in early visual cortex. The final saliency map recombines normalized activation maps across several spatial scales. In [8], the saliency map reflects the stationary distribution on a graph where the nodes are pixels and the weights correspond to a similarity measure that captures both spatial proximity and feature similarity. An information theoretic approach is taken by [9], wherein peaks in the entropy distribution correspond to the location of regions of interest. An extremely simple and fast method using the Gaussian pyramid is developed in [12], but it trades off accuracy for speed. The authors in [13] use image windows to find the saliency map of a given image, and propose a method to efficiently place the windows.

It can be seen that all methods can be broadly classified into two categories: local and global. Local methods use the biologically motivated principle of finding objects that are different from their immediate surroundings (scale is incorporated to vary the "size" of surroundings). Global methods, on the other hand, tend to look for objects that are most different in the entire image; they look for outliers in the image.

3 Cluster Based Saliency

We now describe a novel technique to find salient regions in images. Like some of methods above, our approach uses unsupervised clustering and outlier detection to find salient regions. Most outlier detection schemes detect point outliers in the data, but we are interested in finding an *outlier cluster*: data that is an essential part of the image, and is thus not an outlier in the true sense of the word. Related work was reported by [11], wherein the authors use the spatial scan statistic [10] to find spatially anomalous clusters.

The algorithm extracts features from every pixel in the image, including color (red, green, blue, yellow), orientation ($0°, 45°, 90°, 135°$), scale and intensity. Note here that for grayscale images, no color information is extracted and for color images, the intensity would just be the average of the three (RGB) color planes. This yields a feature vector for every pixel in the image, which are then clustered using the scheme developed in [3]. (Other methods for clustering were tried, but were either not suitable for a very large amount of data, or did not give satisfactory clustering performance.) To measure the "salience" of each cluster, we first construct normalized probability distributions associated with the cluster itself, and with a supercluster formed by combining *all* other clusters (Figure 3). We then take the Kullback- Leibler divergence between the two distributions thus formed as a measure of cluster saliency: the cluster that is most distant in KL divergence from all other clusters will be deemed most salient. Figure 3 shows some results obtained on natural images.

Though similar in some ways to the approach of [8], our approach differs in two respects. First, it does not take the spatial proximity of different pixels into account–thus it represents a pure feature-similarity-based approach to saliency. The contrast of our approach to that of [8] thus allows us to investigate to what extent the spatial distribution of features contributes to salience detection. Second, [8] employ graph-based methods for clustering whereas we apply the method of [3] to non-graphical data.

Pseudocode

- **input** image I of size m x n
- $\forall$ pixel i $\in I$, extract features f_i
  - cluster f_i into k clusters $\mathcal{C}_1 \cdots \mathcal{C}_k$
- $\forall j \in 1, 2, , k$
  - normalize $\mathcal{C}_j$ and $\mathcal{C}_{j^c} = \cup_i \mathcal{C}_i \backslash \mathcal{C}_j$
  - $d_j = KL(\mathcal{C}_j || \mathcal{C}_{j^c})$
- find $d = max_j d_j$, and $s = argmax_j d_j$
- **output** $\mathcal{C}_s$, d

Note that this method varies from some others mentioned in this paper, in that the saliency map is not a distribution of saliency values of individual pixels but groups of pixels (corresponding to specific clusters). In the raw version of the algorithm, only the most salient cluster is returned. If the saliency value of every pixel in the image is to be returned, then the algorithm can be modified to return all clusters in decreasing order of the pairwise distances returned.

To assess the face validity of our approach, we found the most salient cluster for a database of 225 images created using the images used by [7] to assess the center-surround method. Figure 3 shows some results that we obtain using the clustering method to identify salient regions in images. We found that the method identified the salient object in approximately 79.6 % of the 225 test images, which is comparable to other existing algorithms (see results in [7, 8] for comparison).

Figure 1: Results of the cluster based saliency method. The images on the left are the original images, and the ones on the right display the salient object occluded by an opaque patch.

4 Ranking Schemes

Our main concern in the current work, however, is not the overall accuracy of the algorithm, but its potential for helping to aid human performance in target detection tasks. How can the salience measures returned by different models be used toward this goal? The key idea is that images with targets should contain a highly salient region whereas images without targets should not (under the assumption that targets tend to be more salient than non-target regions). What is needed, then, is a method for ranking images according to the salience of their most salient region.

The ranking scheme for the center-surround and graph-based methods is inspired from the normalization scheme used in [7]. We use the difference between the peak value of the saliency map and its mean value to determine the "saliency rank" of that image (Figure 2). So, an image whose saliency map has an isolated large peak but is otherwise flat, will have a large max-mean difference as opposed to in image whose saliency map has many peaks of varying heights. Consequently, the former image will be deemed more important.

To derive a ranking for the cluster-based approach, recall that this method uses the KL divergence between clusters (after normalization) to determine the important cluster. Naturally, we can use the distance between the salient cluster and the remaining clusters as the metric used to rank images. Hence, a salient cluster that is well separated from the combination of all other clusters will be deemed more important that one that is not that well separated. The idea here is that a larger degree of separation (in the KL divergence sense) implies that the salient object is "more different" than the other objects in the image.

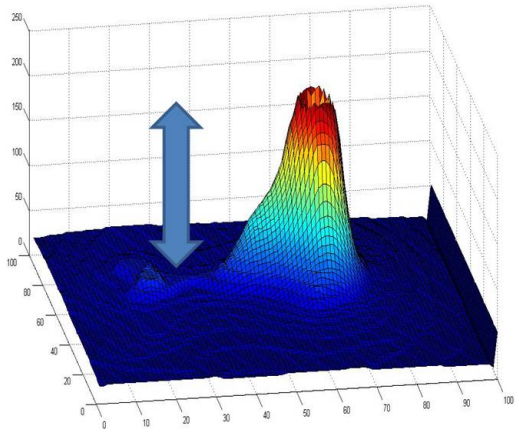

Figure 2: Ranking for center surround and graph based methods. The length of the double headed arrow indicates the difference between max and mean in the final saliency map. A higher length indicates the particular image is ranked higher, or is more important

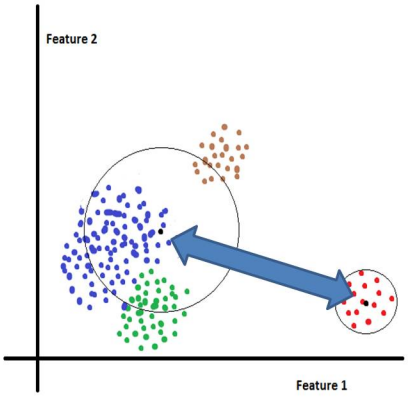

Figure 3: Cluster based saliency. The ellipse around the clusters on the left indicates that they are all considered to be a single cluster, and the KL divergence is calculated between this supercluster and the isolated cluster on the right. Note that this happens in a much higher dimensional setting in our case. The length of the double headed arrow indicates the KL divergence between the supercluster indicated by the ellipse and the smaller cluster on the right. A higher length indicates the particular image is ranked higher, or is more important.

5 Comparing Saliency Models for Target Detection

To test the ranking schemes, we formed a database of 500 aerial images taken over various terrain including forest, sea, desert, snow and fields. Of these, 75 images contained man-made objects such as boats, houses, warehouses, roads, and so on. These comprised the "salient" images. Figure 4 and Figure 5 show a small sample of the images we used. The models were compared by using them to rank-order the 500 images by maximal salience. We then considered how many of the 75 target images appeared in the top 75 ranked images for each model.

All three of the models we considered–center surround, graph based and cluster based saliency–employ the same features extracted from the given image. These features are essentially those introduced by [7], which were motivated by properties of neurons in early visual cortex. To make all algorithms comparable, we modified the *multiscale* center surround method to incorporate scale as a separate feature. Scale of a pixel is determined by entropy peaks as in [9].

Figure 4: sample of target images.

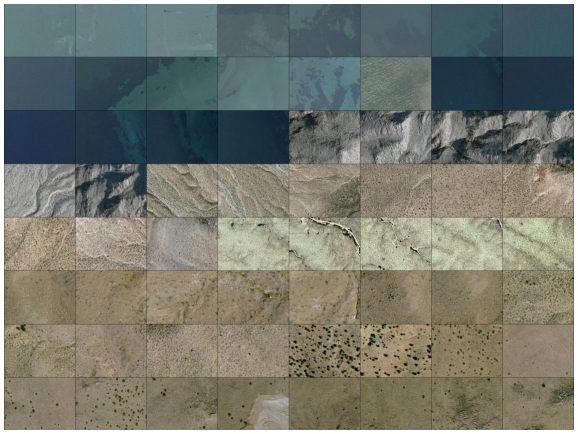

Figure 5: sample of images without targets.

Figure 6 shows how many of the 75 target images appear within the top *n-ranked images, with* n plotted along the abscissa. All three models produce rankings that exceed random performance, but there are clear differences among them. Notably, the graph-based and clustering models–both of which take feature similarity into account–strongly outperform the center-surround approach. Taken on their own, however,

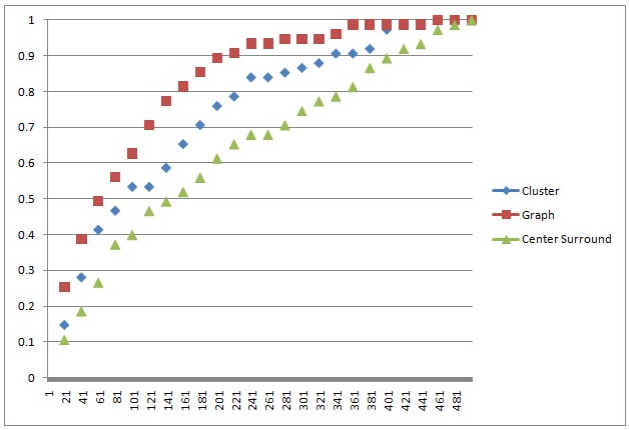

Figure 6: Comparison of various ranking schemes. The y axis corresponds to the proportion of salient images (out of 75) identified and ranked. The x axis is the rank assigned. An optimal method would have all the salient images in the first 75 after ordering, and a random ordering will result in a 45° line

the models are far from optimal. Table 1 shows, for each model, how many of the 75 target images appear within the top-75 ranked images (Hits), and how many non-targets (false alarms). The table also shows the hit rate (hits / total number of targets) and false alarm rate (false alarms / total number of non-targets) associated with each model, and the corresponding d'(hit rate - false alarm rate) for discriminating targets and non-targets. Even the best-peforming graph-based model detects just above 50% of the targets, and achieves a d' of just 1.57. The clustering model is next-best, followed by the center-surround model.

Table 1: hit rates for the methods without humans. 'Hits' corresponds to the number of important images in the top 75 images as ranked by the schemes (ideally hits will be 75).

Method	Hits	False Alarms	Hit Rate	FA Rate	d'
Graph	42	33	0.56	0.08	1.57
Cluster	36	39	0.48	0.09	1.28
CS	28	47	0.37	0.11	0.89
Random	11	64	0.15	0.15	0.00

Given the mediocre performance of these different models, we next inquired whether they could be of use in improving human performance on the target-detection task. We conducted a behavioral study that was a simple analog of the detection task: participants were given 3 minutes to identify among the 500 items as many items containing evidence of human habitation as possible. Images were presented sequentially to the participant, who pressed a button to indicate whether any man made objects were visible in the scene. Participants had as long as they liked to inspect each image, but were told they had only 3 minutes to find as many targets as possible. As soon as a decision for a given image was made, a new image appeared. There were four different conditions varying only in the order in which the images were presented to the participants: images could be ordered at random or according to the rankings generated by one of the three models. We then considered, for each condition, how many targets were detected and how many false-alarms were generated.

Table 2 shows the mean number of hits, false alarms, hit and false alarm rates achieved by human participants in the four different conditions. The random condition provides baseline performance against which the models can be assessed. The orderings generated by all three models were dramatic improvements over the random ordering, showing that even mediocre ranking algorithms can significantly improve human

Table 2: Hit rates and false alarm rates for the methods. The hit rates are calculated for the images seen, whereas the false alarms for the entire database. FA stands for False Alarm, HR stands for Hit Rate and FA R, for False Alarm Rate

Method	Images seen	Hits	F A	HR	FA R
Graph	115	45	11	0.61	0.03
cluster	133	39	4	0.52	0.01
center surround	141	34	9	0.46	0.02
Random	145	26	36	0.34	0.09

performance. However, there were strong differences across the algorithms as well. The clustering and graph-based methods both generated superior performance compared to the center-surround model. Though the graph-based orderings produced the largest number of hits, it also generated substantially more false alarms, leading to an overall smaller d' compared to the clustering method.

Why does the rank-ordering improve performance so substantially relative to random ordering? The reason is that, among the subset of images they can view in the limited time permitted, the participants view a larger number of targets. They are not "wasting" their time on images that are very unlikely to contain interesting regions. Instead, their limited resources are optimized–the machine can weed out obviously uninteresting images, and the analyst can spend her more valuable time sorting among images that are more likely to contain targets. Of course, the better the algorithm, the better the human-machine pair will perform overall.

Table 3: Number of important images in the total number of images analyzed by human subjects.

Method	**images seen**	**target images**
random	145	29
center surround	141	39
graph	115	51
cluster	133	44

6 Conclusions and Future Work

We have shown that automatic computation of saliency in an image can be used as a cue for detecting targets in tasks where the targets differ in important ways from their surroundings. Using maximal salience to detect images with targets, and using current state-of-the-art saliency models, machines alone will achieve only mediocre performance. Our research shows, however, that even this level of performance can dramatically improve the performance of human analysts working without machine aid by de-prioritizing images with no highly-salient regions.

Our results also suggest that approaches to saliency that include featural similarity as a major cue will achieve superior performance to those based solely upon center-surround contrast: the graph-based method and the cluster-based approach introduced here both out-performed a center-surround method introduced by [7].

It is possible that these salience measures could be further improved by adding feature selection steps to the algorithm. There may also exist other metrics to determine the salient cluster. For instance, one might consider the clustering scheme itself to determine the salient cluster. Specifically, if the scheme chooses k clusters, we then force it to use k - *1* clusters, and determine what cluster when "absorbed" into the other clusters provide the worst fit to the data. We leave these developments for future work.

Interestingly, we also found that human participants show higher false-alarm rates when the images are presented in random order. We anticipate that this is a consequence of the target sparsity–in a given time, participants encounter very few targets, and so may be more inclined to "jump the gun" for questionable images. The reasons for this are interesting from a psychological point of view, and can be investigated in future work.

References

[1] F. Attneave. Some informational aspects of visual perception. In *Psychological Review*, volume 61, 1954.

[2] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. In *Journal of Vision*, volume 9, pages 1–24, 2009.

[3] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, pages 381 – 396, March 2002.

[4] L. Itti, C. Gold, and C. Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793, Sep 2001.

[5] L. Itti and C. Koch. Target detection using saliency-based attention. In *Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified), Utrecht, The Netherlands, RTO-MP-45 AC/323(SCI)TP/19*, pages 3.1–3.10, Jun 1999.

[6] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, Jan 2001.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[8] C. Koch J. Harel and P. Perona. Graph-based visual saliency. In *Proceedings of Neural Information Processing Systems*, 2007.

[9] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, V45(2):83–105, November 2001.

[10] M.Kulldorf. A spatial scan statistic. In *Communications in Statistics: Theory and Methods*, volume 26, pages 1481–1496, 1997.

[11] D. Neill and A. Moore. Anomalous spatial cluster detection. In *AD-KDD05*, August 2005.

[12] P. Rosin. A simple method for detecting salient regions. volume 42, pages 2363–2371, November 2009.

[13] T. Toriu and S. Nakajima. A method of calculating image saliency and of optimizing efficient distribution of image windows. In *Proceedings of the First International Conference on Innovative Computing, Information and Control*, 2006.