Modified Methods of Generating Saliency Maps Based on Superpixels

Veronika Olešová*
Supervised by: Ing. Vanda Benešová, PhD.

Institute of Applied Informatics
Faculty of Informatics and Information Technologies STU
Bratislava

Abstract

Various types of approaches that can model a human visual attention have been already proposed. However, a model that could perfectly simulate the human perception and methods of computer prediction of human visual attention belongs to one of the high focused research areas.

Our work is aimed at methods of generating a saliency map which will detect the areas in the picture that could most likely attract a human attention. The basis of our work is method of saliency detection using superpixels published by authors Z. Liu, O. Meur and S. Luo. Several modifications of this method with the aim to improve the results have been proposed, tested and evaluated in our experiments. In this paper, we present our proposed modification based on border prior and statistical evaluation of the saliency central position in the used dataset. This center position will be expressed using a fitted Gaussian function. Results of all experiments are evaluated and presented in the following paper.

Keywords: saliency map, superpixels, border prior

1 Introduction

In our daily lives we are surrounded by incredible amount of information, which we are not able to process all at once. We need to restrict our attention only on certain area or objects at a time so we can process this information one after another. Scientists have been examining what underlies our attention to help us avoid information overload. They came up with the idea to create a saliency map for a given image that represents information about human visual attention of this image.

The saliency map is a topographically arranged map to represent the saliency of the visual scene and it gives us information about where in the image the areas that attract our attention are. It can reflect several salient objects or areas which are sorted by their saliency.

Saliency map is often used as a prior for a classification system to detect objects. These maps are useful for many applications such as image compression, predicting eye movements, autofocus and visualization.

The main problem of existing models generating saliency maps is that they usually work with specific cases and are not able to cover all of them.

2 State Of The Art

There are a lot of differently oriented models to creating a saliency map that have achieved good performance in predicting human eye fixations. The most common models that are often used for comparison are A Model of Saliency-based Visual Attention for Rapid Scene Analysis [5], Graph-Based Visual Saliency [4] and SUN: A Bayesian Framework for Saliency Using Natural Statistics [10]. We will describe the main ideas of these models in this section. In more detail we will analyze another model, called Superpixel-based saliency detection [8], which is the basis of our work.

2.1 Model of Saliency-based Visual Attention for Rapid Scene Analysis

Itti proposed a model [5] which is inspired by the architecture proposed by Koch and Ullman, who came up with the idea that the different visual features should be combined into one single topographically oriented map. Most of the later works use Ittis model for comparison since it is the earliest model of a saliency map.

Visual preprocessing of this model consists of creating five Gaussian pyramids that are generated from intensity image and four color channels - red, green, blue and yellow. Image is then decomposed into a set of topographic feature maps. Each feature is computed by a set of linear center-surround operations. These maps are normalized and combined into three conspicuity maps. The final saliency map is the result of the normalization followed by a summation of the three conspicuity maps.

This architecture is not designed to detect conjunctions of features; it can only recognize a target which is different from surrounding by its intensity, color, size and orientation, and will fail once the salient object has another

*v.nika.olesova@gmail.com

feature. The salient object has to be represented in at least one feature map in order to pop out.

2.2 SUN: A Bayesian Framework for Saliency Using Natural Statistics

The title of the second mentioned approach [10] is SUN because it depends on the statistics of natural images. The saliency map of this framework can be generated either by bottom-up, top-down or a combination of those approaches. By choosing bottom-up approach, saliency is represented by self-information and by choosing top-down, it is defined as log-likehood. In this model, the features are calculated in two ways. The first approach calculates the features as responses of linear filter known as DoG and the second as the responses to filters learned from natural images using independent component analysis ICA.

2.3 Graph-Based Visual Saliency

Graph-Based Visual Saliency [4] consists of three main steps. First, feature maps need to be extracted at multiple spatial scales. To do that, a scale-space pyramid is obtained from image features: intensity, color and orientation, which is similar to model of Itti. The second step is to form an activation map using these feature maps. In the final step the activation map is normalized to emphasize the most important information and then combined into a single saliency map.

This model assigns greater saliency to locations situated in the middle of the image. The reason is that most of nodes are closer to a few center nodes than to any point located near the image boundary. The described process is computationally quite expensive and the resulting saliency map has ill-defined object boundaries, which can restrict the usefulness in certain applications.

2.4 Superpixel-based saliency detection

This model [8] consists of three major steps. At the beginning it is important to simplify the input image by using superpixel segmentation and color quantization. Then, similarity between each superpixel has to be found. Finally, the global contrast and spatial sparsity is computed for each superpixel.

A superpixel should contain pixels that are similar in color and texture, and therefore are likely to belong to the same object. This assumption leads to the advantage of superpixel primitives over pixel primitives. Another advantage of this representation is that computational elements are greatly reduced and the segmentation result will be better since superpixels preserve the objects shape information and are more robust to noise.

2.4.1 Image simplification

The image is converted to the CIE L*a*b*, perceptual uniform color space, which is designed to approximate human vision. The first simplification consists of creating superpixels using SLIC algorithm [2]. This divides a picture into approximately 200 smaller regions, which is sufficient to preserve different boundaries in the used dataset well. The result of superpixel segmentation using SLIC algorithm can be seen in Figure 1. Then, the number of distinct colors has to be reduced by applying the color quantization. The image histogram is created by quantizing each color into *qxqxq* bins. For each bin, mean color and number of pixels belonging to this bin is computed. Bins that cover more than certain number of pixels are preserved and the rest are merged into ones that have the smallest difference between their quantized colors.

Figure 1: Superpixel segmentation

2.4.2 Superpixel similarity

Each superpixel is assigned to a color histogram which is calculated based on the one created in the previous step. The histogram is normalized so that the summation of values in each histogram is equal to 1. Two types of similarities for each pair of superpixels are computed.

The color similarity of two superpixels is computed as the sum of intersection of their histograms:

$$Sim_c(i,j) = \sum_{k=1}^{m} min\left\{H_i(k), H_j(k)\right\} \tag{1}$$

The spatial similarity is defined as:

$$Sim_d(i,j) = 1 - \frac{\left\|\mu_i - \mu_j\right\|}{d} \tag{2}$$

where d is the diagonal length of the image and μ is the center of the superpixel.

By combining those similarities, the resulting similarity is obtained:

$$Sim(i,j) = Sim_c(i,j) * Sim_d(i,j) \tag{3}$$

2.4.3 Superpixel saliency

Authors [8] suggested that color contrast can be easily seen between the salient object and the background. They also noticed that spatial distribution of salient object superpixels is sparser than background superpixels. Because of this, global contrast of each superpixel and their spatial sparsity are evaluated for measuring the final saliency.

Global contrast of each superpixel is defined as:

$$GC(i) = \sum_{j=1}^{n} W(i,j) \cdot \left\| mc_i - mc_j \right\| \tag{4}$$

where *mc* is the mean color of superpixel and the weight is defined as:

$$W(i,j) = \mid SP_j \mid \cdot Sim_d(i,j) \tag{5}$$

where $\mid SP_j \mid$ stands for the number of pixels in the superpixel. We have to normalize this global contrast so that the values map to the range from 0 to 1:

$$NGC(i) = \frac{GC(i) - GC_{min}}{GC_{max} - GC_{min}} \tag{6}$$

where GC_{max} is the maximum value of global contrast among all the superpixels.

The spatial sparsity of a superpixel is computed as:

$$SS(i) = \frac{\sum_{j=1}^{n} Sim(i,j) \cdot D(j)}{\sum_{j=1}^{n} Sim(i,j)} \tag{7}$$

where $D(j)$ is a distance between the center of image and the superpixel j. This is also normalized, but this time inversely:

$$NGC(i) = \frac{GC(i) - GC_{min}}{GC_{max} - GC_{min}} \tag{8}$$

We have refined the normalized global contrast and spatial sparsity so that superpixels with higher similarity have more similar values:

$$RGC(i) = \frac{\sum_{j=1}^{n} Sim(i,j) \cdot NGC(j)}{\sum_{j=1}^{n} Sim(i,j)} \tag{9}$$

$$RSS(i) = \frac{\sum_{j=1}^{n} Sim(i,j) \cdot NSS(j)}{\sum_{j=1}^{n} Sim(i,j)} \tag{10}$$

The final saliency value for each superpixel is defined as the multiplication between refined global contrast and spatial spread:

$$Sal(i) = RGC(i) * RSS(i); \tag{11}$$

3 Our Contribution

In this section we present our experiments that include border prior, its update and central position modification.

3.1 Border Prior

We have extended the original model by adding the border prior, which achieves better results. This prior comes from the basic rule of photographic composition, that is, most photographers will not crop salient objects along the view frame. In other words, the image boundary is mostly background [9]. However, this only applies to photographs that are intentionally taken by humans and it is not general.

Huaizu Jiang and others [6] made the following survey: "we made a simple survey on the MSRA-B data set with 5000 images and found that 98% of pixels in the border area belong to the background."

In our algorithm we label the superpixels that touch any of the image borders as background and find other superpixels that are very similar to them. Each of these superpixels is considered background and its saliency is automatically zero. In the Figure 2 we can see the difference between the saliency map which uses this prior and the one that does not.

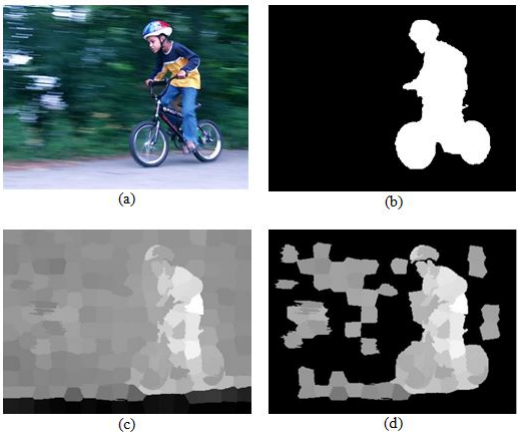

Figure 2: (a) Original image, (b) ground truth, (c) saliency map without border prior, (d) saliency map with border prior.

However, if there is a salient object that only slightly touches the boundary, the whole object could be missed. In order to prevent such situation, we compute global contrast for the group of superpixels that touch the boundary and remove the first 10% whitest of them. These superpixels could be a part of the object and it would be wrong to mark them as background. The chosen percentage is only an estimation based on observation of the used dataset of images. An example is shown in Figure 3 where we can see that in the image c) a man is missed because he touches the boundary and in the image d) we see that the most contrast superpixels help us identify this man.

3.2 Central Position Modification

The original distance shown in Equation 7, which is used for computing the spatial sparsity, did not seem accurate to us. The result of the function $D(j)$ is simply a distance

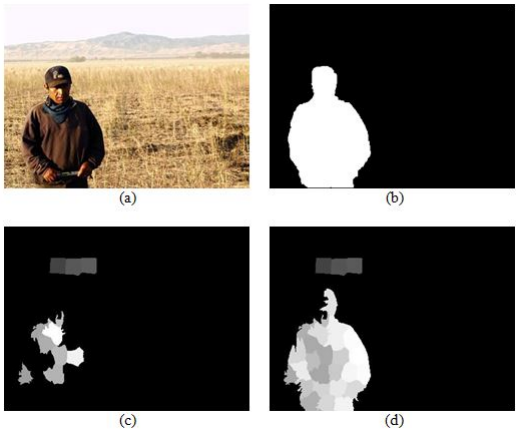

Figure 3: (a) Original image, (b) ground truth, (c) saliency map border prior, (d) saliency map with updated border prior.

between a superpixel j and center of image. It does not take into account the most probable distance to which the salient object could occur. We decided to statistically evaluate the central position in the dataset and create a new function that could be used instead of the original distance. This modification is also not general and applies only to the used dataset.

Firstly, our intention was to get a histogram for each ground truth image in the dataset, which would indicate how far is the salient object from the center. Ground truth is human-segmented image dataset used to compare image segmentation algorithms. Basically, it is a binary image whose white pixels belong to the salient object and black pixels to background. To create a histogram, we calculated the number of white pixels that fall within each distance from the center position of the image. Then we summed all the histograms of each image and divided each value of the resulting histogram by the length of the corresponding circle. This histogram was then fitted to Gaussian function using matlab:

$$[fy, god] = fit(x, y,' gauss1'); \tag{12}$$

where x is a vector of distances from the center image and y is a vector of number of pixels.

The plot of the resulting function is in the Figure 4 where we see that most of the pixels belonging to the object are situated near the center of the image and the output is the Gaussian function in the following form:

$$726.9 * exp(-(\frac{x-6.692}{97.7})^2) \tag{13}$$

The distance $D(j)$ has been replaced by this exponential function.

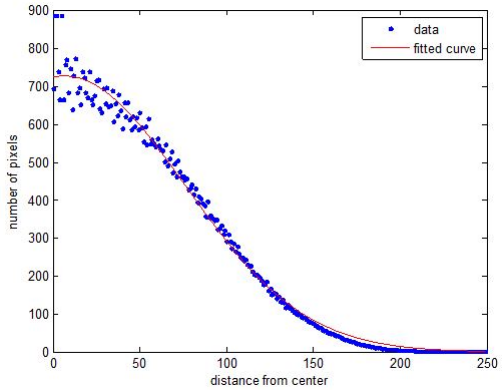

Figure 4: Fitted gaussian function.

4 Tests and Results

We came up with two types of evaluation. In each of them we use images from the MSRA[1] dataset, which is the largest object dataset containing 20 000 images in set A and 5 000 images in set B. Achanta [1] has created the dataset[2] containing 1 000 manually segmented ground truths corresponding to 1 000 images from the set B.

4.1 Precision and Recall

The first type of evaluation is used to test a precision and recall of a border prior, its update and a center modification. Precision and recall are statistical measures that are very often used to measure how well the saliency model is able to predict human eye fixations. Precision is a measure of accuracy and recall is a measure of completeness.

At first we have to generate a saliency map for each of 1000 images from MSRA dataset. To get a segmented image we simply threshold the map by assigning the pixels above the given threshold as salient (white background) and below the threshold as non-salient (black background). Then we compare the resulting image to its ground truth. From this comparison we are able to get statistics like precision and recall rate by using the following pseudo-code:

```
if (value_of_saliency_map > threshold)
   {
        segmented_foregound_pixels++;
        if (value_of_ground_truth != 0)
            hit++;
   }

if (value_of_ground_truth != 0)
    ground_truth_foreground_pixels++;

precision = hit / segmented_foregound_pixels;
recall = hit / ground_truth_foreground_pixels;
```

By sliding the threshold from minimum to maximum

[1]Downloaded from http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm

[2]Downloaded from http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/

value, we achieved the precision-recall curves that we use for the comparison between various methods.

The graph of comparison between the algorithm without and with the border prior implemented (SB and BP) is in Figure 5. We can see that our algorithm updated with the border prior achieves better results in precision. There is no saliency map that would have the precision smaller than 0.55. In addition, this graph shows the difference between another 3 models including Graph-Based Visual Saliency (GB) [4], A Model of Saliency-based Visual Attention for Rapid Scene Analysis (IT) [5] and Frequency-tuned Salient Region Detection (IG) [1]. To compare these methods subjectively, we created a table of few images that can be found in the Figure 6. We can see that the background is most suppressed using the border prior. In this comparison we used datasets containing 1000 saliency maps for each model created by Achanta et al.

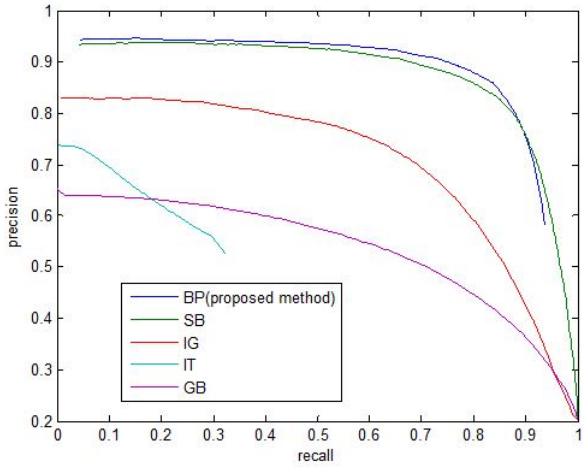

Figure 5: Comparison between different saliency models.

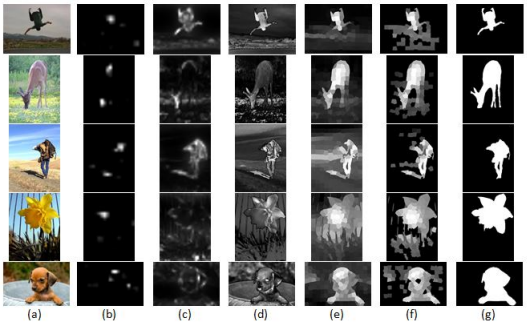

Figure 6: (a) Original image, (b) IT, (c) GB, (d) IG, (e) original, (f) border prior, (g) ground truth.

In the second graph represented by Figure 7, we can see a precision-recall curve between border prior (BP) and its updated version (UBP). Unfortunatly, precision of this method has regressed but the recall has improved. Images with the object touching the boundary were succesfully identified, however, this dataset contains a lot of pictures without such objects. In those images, by removing 10% of superpixels from background we removed superpixels that were actually background.

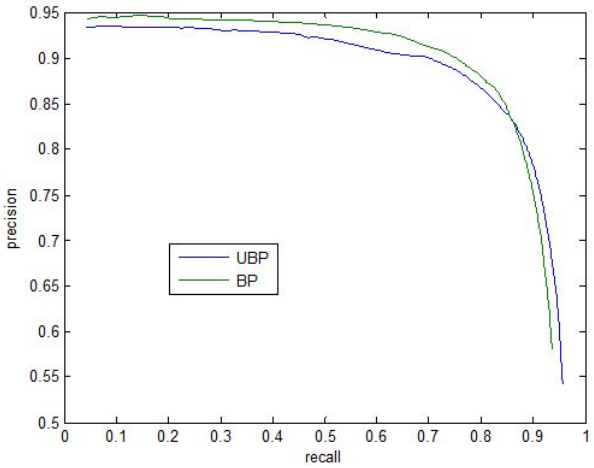

Figure 7: Comparison between border prior (BP) and updated border prior (UBP).

We have also evaluated the modification of center position (BPCM) which can be observed in Figure 8. The recall rate of this modification is the same as the unmodified border prior but the precision has decreased. We assume that its because of the images that do not fit into our gaussian function.

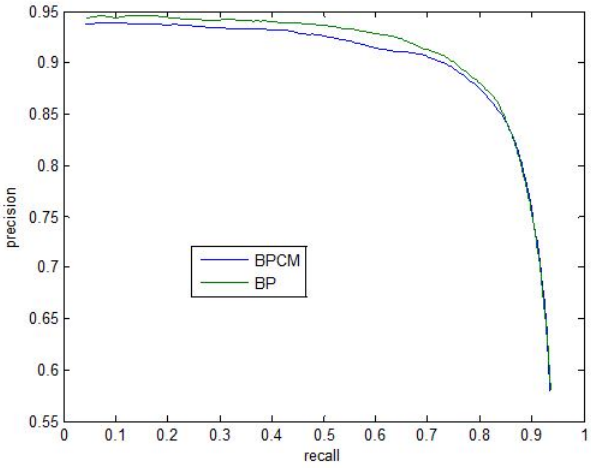

Figure 8: Comparison between border prior (BP) and border prior with center position modification (BPCM).

4.2 Histograms

The second evaluation is implemented in matlab and is aimed at any of the modification but we used it to test the updated border prior. The key is to create a histogram by which we would be able to see how many pixels and what shades of gray from our saliency map belong to the object

and how many to background. This is done by comparing our saliency map to the ground truth. The number of pixels that belong to salient object and to background are computed individually. We divided grayscale into 10 intervals and assigned a number of corresponding pixels from our saliency map to each of them. Therefore each bar of histogram is an interval of size 25 and holds a number of pixels.

An example of such histogram is in the Figure 9, which evaluates the images (c) and (d) in the Figure 3. A symbol TP in this histogram stands for the true positive (number of pixels belonging to the object) and FP is false positive (number of pixels belonging to background). We can see that TP - original (image (c) in the mentioned figure) bar with the pixel value of 1 is bigger than bar TP - modification (image (d)) next to it. This means, that the image with only border prior implemented (TP - original) has more pixels in the range between values 0-25 belonging to object. The other method d(TP - modification) has this number lower, which is good, because we do not want black pixels in the object.

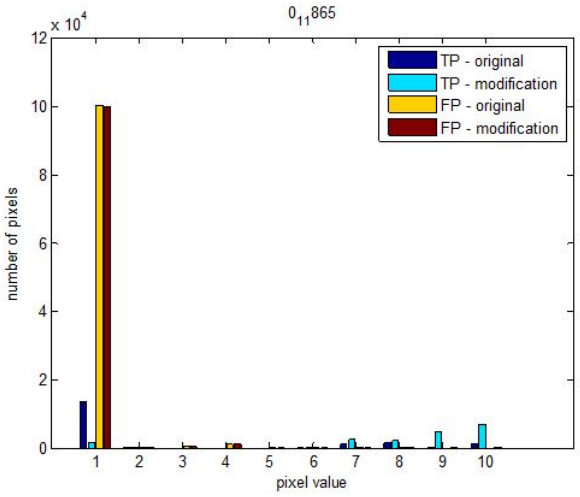

Figure 9: Comparison between border prior (original) and its updated version (modification) by histogram.

5 Conclusion and Future Work

We have presented a few modifications to the existing method [8] to creating a saliency map. These modifications are customized to the used dataset and therefore are not general. Comparing to other models using the same dataset we were able to see that our modification of border prior is better at precision but slightly worse in recall. The update to this method slightly downgrades the precision but improves recall and modification of center position does not change the recall of the border prior but decreases precision.

However, results provided by this method are still not perfect and other modifications are required. We assume that using only color contrast, spatial distribution and border prior is not enough and it would be vital to use higher features such as face detection. Our next goal is to implement a center surround method adjusted to superpixels. Authors suppose that the salient object is enclosed by a rectangle R and they construct a surrounding contour R_s with the same area of R. Then the distance between R and R_s can be measured using various features such as intensity, color, and texture. By this technique it is possible to measure how distinct the salient object in the rectangle is with respect to its surroundings. In our case we would use groups of superpixels instead of rectangles and measure a distance between these groups and their surrounding superpixels by color.

Images in the MSRA dataset contain only a single salient object and most of them are large and near the image center. For the future work it would be appropriate to use more challenging images in a combination with a dataset containing human eye fixations.

References

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

[2] R. Achanta, A. Shaji., K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels. *EPFL Technical Report*, (149300), June 2010.

[3] R. Gonzalez and R. Woods. *Digital Image Processing.* Number 2. 2001.

[4] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems 19*, pages 545–552, 2007.

[5] 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), 1998.

[6] H. Jiang, J. Wang., Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, 2013.

[7] T. Liu and Z. Yuan. Learning to detect a salient object. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):353–367, February 2011.

[8] Z. Liu, O. Meur, and S. Luo. Superpixel-based saliency detection. *International Workshop on Image and Audio Analysis for Multimedia Interactive services*, pages 1–4, July 2013.

[9] Y. Wei, F. Wen., W. Zhu, and J. Sun. Geodesic saliency using background priors. *Proceedings of the 12th European Conference on Computer Vision*, pages 29–42, 2012.

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