
Swiss Federal Institutes of Technology, Lausanne

Optional Project in Communication Systems

# Automatic Extraction of Interesting Image Content

*Author:*
Denis Filimonov

*Professor:*
Touradj Ebrahimi

*Supervisor:*
Ivan Ivanov

June 11, 2010

# Contents

# List of Figures

# 1 Introduction

Objects in the scene have different importance for the scene interpretation. The ability of humans to fixate on specific parts of an image which carry most of the useful information needed for scene interpretation is known as visual attention.

The task of this project is to examine how most salient image regions can be automatically extracted by content analysis techniques. Potential of using a computational bottom-up visual attention model for image tagging will be explored. More specifically, bottom-up attention model predicts the location of interesting objects in a given image. Although the bottom-up approach is considered as a very simple approximation of attention, it has been found to be quite successful in computer vision, where it has been modeled by the saliency map highlighting regions which "catch the eye" in the sense of low level image properties (color, intensity, orientation). Identifying the most salient regions in images is helpful in many applications as eye movements prediction, object detection and region based compression. One can assume that people spontaneously tag the most important objects in a picture.

More specifically, the goal of this project is to study different approaches for visual attention in still images and to assess and compare their performance. The following tasks will be performed:

- Study different approaches for visual attention analysis
- Experiment with the approach by Itti et al. "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis"
- Experiment with the approach by Harel et al. "Graph-Based Visual Saliency"
- Experiment with the approach by Achanta et al. "Frequency-tuned Salient Region Detection"
- Implement and validate a visual attention system developed in software running on a PC, which takes an image as input, performs the above mentioned approaches and suggests regions in the image which are most salient
- Assess and compare the performance of above-mentioned approaches.

Section 2 will present the theoretical insight of the three chosen approaches mentioned above. Section 3 will compare them from the theoretical side. Section 4 will present a

performance accessment and practical comparison of the chosen approaches. Section 5 will present an GUI developed to visualize the chosen approaches in use. Conclusions will be drawn in Section 6.

# 2 Chosen approaches, theoretical part

This section presents a theoretical description of the three chosen approaches for visual attention analysis in still images. A saliency map is a gray-scale image that represents visual saliency which is the quality of an object in an image to stand out from its neighbors and to catch peoples attention. It is used to guide the choice of interesting objects. Saliency map computation methods can be classified as biologically based, purely computational, or a combination of the two. Each of the chosen approaches falls into each of those three different categories. Itti, Achanta and Harel are biologically based, purely computational and combined respectively.

## 2.1 Itti

Itti approach is inspired by the behavior and the neuronal architecture of the early primate visual system. The overall structure of this approach is chosen in Figure 1. The input image is first decomposed into a set of topographic feature maps. 42 feature maps are computed in total: 6 for intensity, 12 for color and 24 for orientation. "All feature maps feed, in a purely bottom-up manner, into a master saliency map". The only difficulty with this approach is to recombine feature maps into a master saliency map. The feature maps differ in their ranges and extraction methods. One problem stands out in the fact that since the procedure recombines all the 42 feature maps, the maps where a salient object appears strongly can be masked by noise or less-salient objects [1]. That is why a normalization procedure was performed. This procedure either promotes or suppresses feature maps. Each map is normalized to the range $[0..M]$ where $M$ is the same for every map. The location of a global maxima $M$ is found and the mean of all local maxima is computed. Then the map is multiplied by $(M - \overline{m})^2$. If the difference between the maxima of the normalized map and its average is high then the map is promoted otherwise it is suppressed. Feature maps are then recombined into the saliency map. The implementation of this model by Dirk B. Walther and the California Institute of Technology can be found on $http://www.saliencytoolbox.net$.

The other implementation (the one that is easier and used in this work) can be found *http : //www.klab.caltech.edu/ harel/share/gbvs.php* [1].

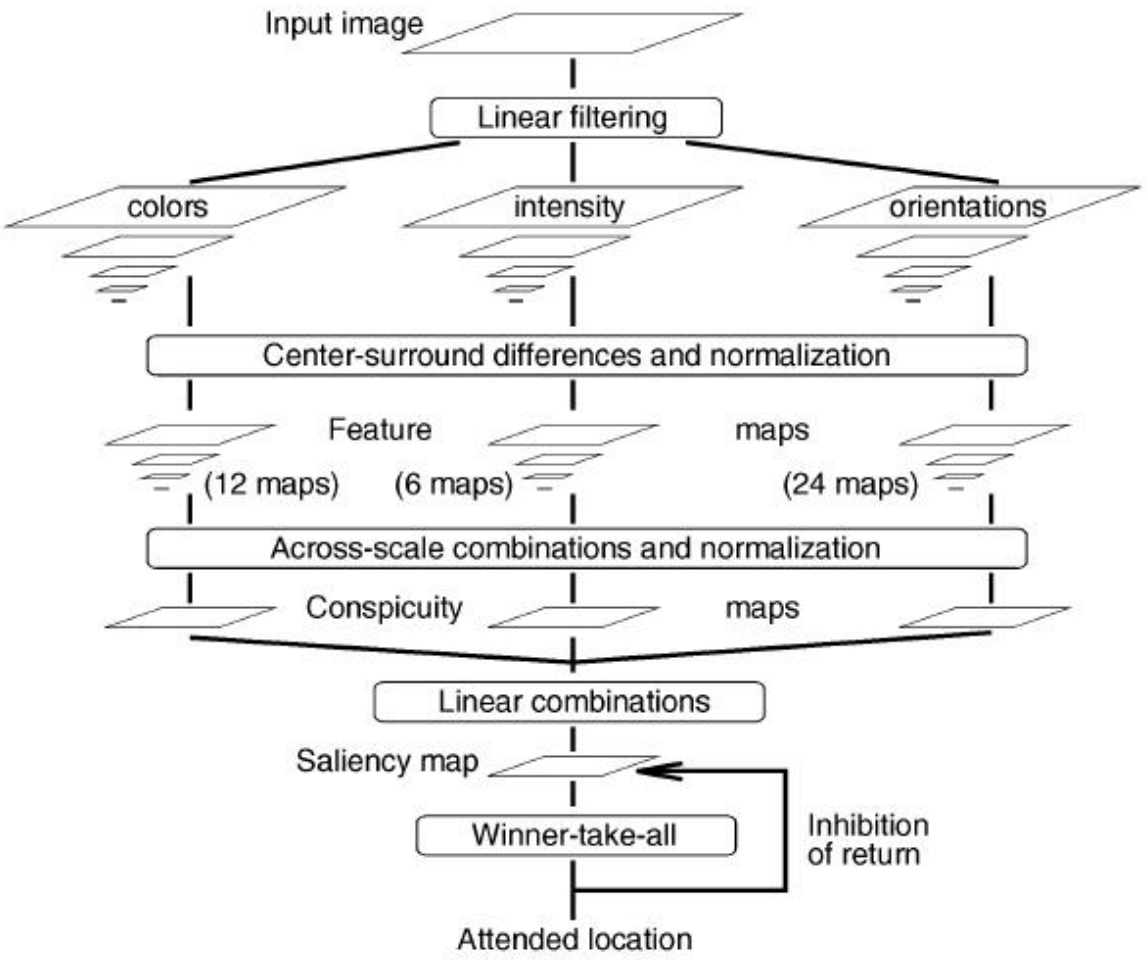



Figure 1: General architecture of Itti model

## 2.2 Achanta

Achanta "introduces a method for salient region detection that outputs full resolution saliency maps with well-defined boundaries of salient objects. These boundaries are preserved by retaining substantially more frequency content from the original image than other existing techniques" [3]. As mentioned previously this method is purely computational and it is not based on any biological model. The requirements for the saliency map generated by this model were

- Accentuate the largest salient objects

- Uniformly highlight whole salient regions
- Establish well-defined boundaries of salient objects
- Disregard high frequencies arising from texture, noise and blocking artifacts
- Efficiently output full resolution saliency maps

The saliency map is calculated as follows. Given the input image $I$, in Lab color space $S(x,y) = ||I_\mu - I_{w_{hc}}(x,y)||$ where $I_\mu$ is mean image feature vector and $I_{w_{hc}}(x,y)$ is the pixel at position $(x,y)$ of the Gaussian blurred version of the original image. This is computationally efficient. Figure 2 and 3 show the overall Achanta procedure [3].

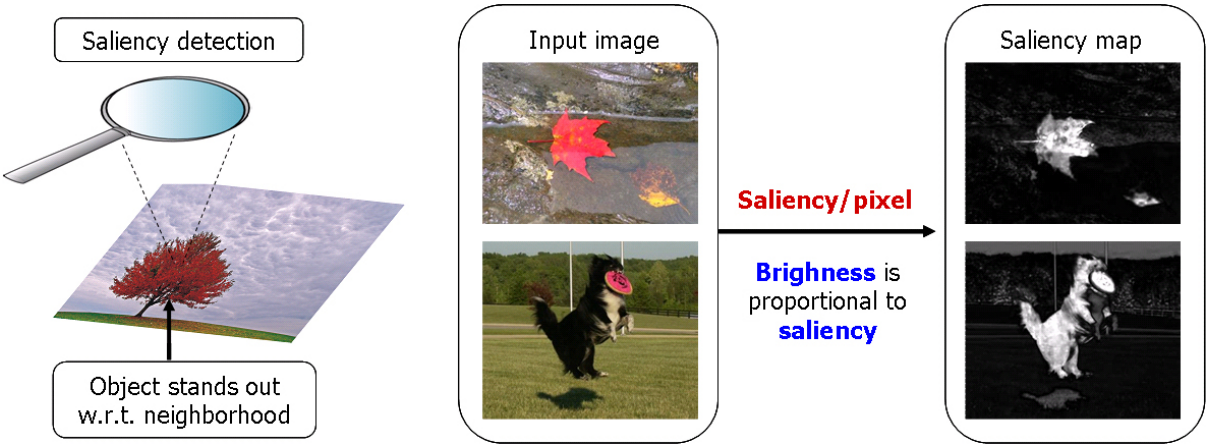



Figure 2: Introduction to Achanta algorithm

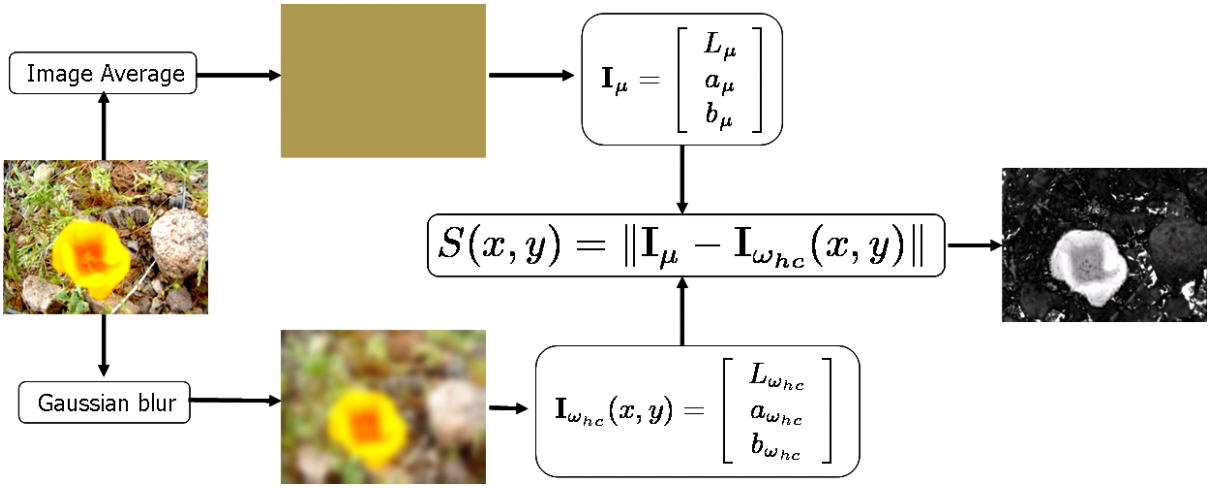



Figure 3: Achanta saliency detection algorithm

### 2.3 Harel

Harel approach is a new bottom-up visual saliency model. It is Graph-Based Visual Saliency (GBVS). It is of combined type (biologically based and computational). It has three steps:

- Extracting feature maps just like in Itti approach
- Creating "activation maps" using the feature vectors
- Normalizing those maps in different from Itti way and recombining them into the saliency map.

For step two Markov chains are defined over various image maps. A dissimilarity function is defined between each pair of nodes in the chain and the weight between those is proportional to the dissimilarity function. The weights are normalized to 1 and treated as the transition probability. Then the equilibrium distribution is calculated and treated over map locations as activation and saliency values. The implementation of this model by J. Harel (A Saliency Implementation in MATLAB) can be found on
*http* : *//www.klab.caltech.edu/ harel/share/gbvs.php* [2]

## 3 Chosen approaches, theoretical comparison

This section presents a theoretical comparison of the three chosen approaches for visual attention analysis in still images. The essential problems of current methods for saliency map calculation are that they produce low resolution maps with badly defined borders of salient objects. High complexity is also of a big concern. Another problem of nowadays methods is that they emphasize saliency of the edges of the objects but are not uniform over the entire object.

In general, methods determine the most salient objects by comparing the contrast of an object with its neighbors using low-level image properties like intensity, color and orientation.

As mentioned in the previous section the saliency computation methods are divided into three groups: biologically plausible (Itti), computational (Achanta) and combined (Harel).

Both Itti and Harel methods produce low resolution saliency maps. Itti's method outputs saliency map that is $\frac{1}{256}^{th}$ of the original image size and Harel has downscaling factor of 64. Both of those methods have badly defined salient object borders. This is due to the fact that these algorithms perform a critical resolution downscale which reduces the range of spatial frequencies in the original image [3]. The range of spatial frequencies retained by Achanta algorithm is more appropriate than the one by Itti and Harel. Itti method retains frequencies from the spectrum of the original image within $[\frac{\pi}{256}; \frac{\pi}{16}]$. Harel retains spatial frequencies within the range $[\frac{\pi}{128}; \frac{\pi}{8}]$. And the Achanta model retains spatial frequencies within the range $(0; \frac{\pi}{2.75}]$.

From complexity point of view Harel is the slowest one followed by Itti and Achanta. Table 1 resumes the comparison. $S$ is the input image size. $N$ is a small neighborhood of a pixel (could be for example 3x3 = 8 pixels). The operations per pixel vary: $k_{Ach} < k_{Itt} < k_{Har}$. K is the number of iterations in the Harel approach[3].

| Method | Freq. range | Resolution | Complexity |
|---|---|---|---|
| Itti | $[\frac{\pi}{256}; \frac{\pi}{16}]$ | $S/256$ | $O(k_{Itt}N)$ |
| Achanta | $(0; \frac{\pi}{2.75}]$ | $S$ | $O(k_{Ach}N)$ |
| Harel | $[\frac{\pi}{128}; \frac{\pi}{8}]$ | $S/64$ | $O(k_{Har}N^4K)$ |

Table 1: Comparison of the chosen approaches

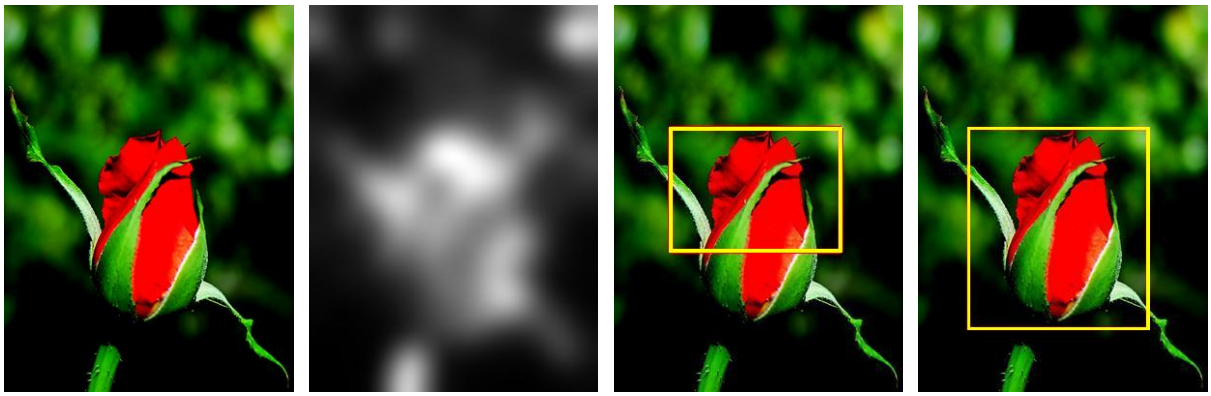
# 4 Chosen approaches, practical comparison

As mentioned in the Introduction saliency map can be used for many different purposes. The scope of this project was to use the saliency map for image segmentation to extract the most interesting objects. Saliency map is a gray scale image and so in order to differentiate between salient/interesting and non-salient object, the saliency map is binarized. Above a certain threshold the map values are set to 255 (white/salient) and values below are set to 0 (black/non-salient). In this Section two methods for image segmentation will be presented. After words the three chosen approaches for visual attention analysis will be compared in terms of their acceptance and accuracy rates for segmentation (considering only the most salient object, which was fixed as the one that contained the largest number of pixels) and also in terms of their calculation times.

Two tests were performed for comparison using the MSRA Salient Object Database from Microsoft Research Asia. The reason for the choice of this database was that it contained 5000 images with ground truth data (coordinates in the image) on the most salient object from a subjective test previously performed. Only 500 images were used for tests in this work due to computational time.

## 4.1 Example of segmented images

Figures 4, 5, 6 show the saliency maps and the most salient object of an image with a rose take from previously cited MSRA database for Itti, Achanta and Harel approaches respectively. In every figure (a) is the original image, (b) is the saliency map, (c) is the most salient object obtained with image-adaptive thresholding and (d) is the most salient object obtained with mean-shift segmentation.

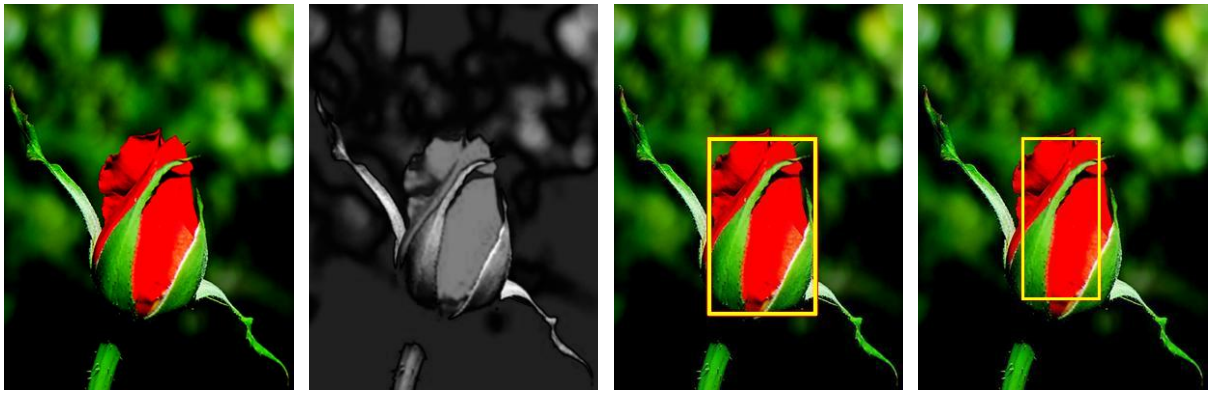


(a) (b) (c) (d)

Figure 4: Rose, Itti approach

## 4.2 Definition of acceptance and accuracy rates

Define $GTA$ as the area of the ground truth object, $PA$ as the area of the object predicted by one of the chosen approaches, $IN = intersection(GTA, PA)$, $UN = union(GTA, PA)$. For every image from the database for each of the three approaches calculate $IN$ and $UN$. If the acceptance value $IN/UN$ is bigger than a particular threshold then the object is accepted. The acceptance rate is calculated as the number of accepted objects divided by the total number of objects (of images as there is only one object per image (the most salient)). Then the particular threshold is fixed to 50% which seemed to



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

Figure 5: Rose, Achanta approach

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

Figure 6: Rose, Harel approach

be the most appropriate value. For every accepted object with acceptance value bigger than 50% an accuracy value $IN/GTA$ is calculated. This is done in order to define precision of accepted objects. Figure 7 shows a situation where two objects were accepted but one is more precise than the other. The green binding box represents the ground truth data, the red one represents the predicted area for one approach and the yellow one for some other approach (it is just fake example to see what can happen). One can clearly see that the acceptance values are approximatively the same but the red approach is more precise since it covers a bigger area of the ground truth therefore its accuracy value is higher.

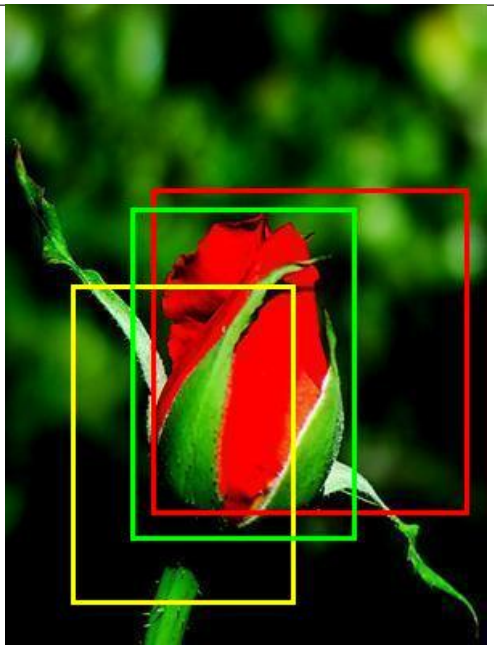


Figure 7: Example of acceptance and accuracy

## 4.3 Adaptive thresholding

In the first experiment, an image-adaptive binarization of saliency maps is performed. The threshold value for segmentation is equal to $T = \frac{2}{HW} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y)$ where $S$ is the saliency map and $H$x$W$ its size. Values of the saliency map bigger than $T$ are mapped to 255(white), 0 otherwise. Then the coordinates and the area of the most salient object are calculated. Figure 8 shows the acceptance rates of the image set (500 images from MSRA database) for the three approaches in terms of the acceptance threshold. Harel (blue bars) accepts more objects than other approaches for threshold from 50 to 65 %, Itti (green) and Achanta (red) evolve more or less equivalently and they accept more objects than Harel for threshold higher that 70 %. If the acceptance value grows then the accuracy value will grow also. Therefore this can suggest that more objects with Itti and Achanta have high accuracy value/more objects are precise than with Harel.

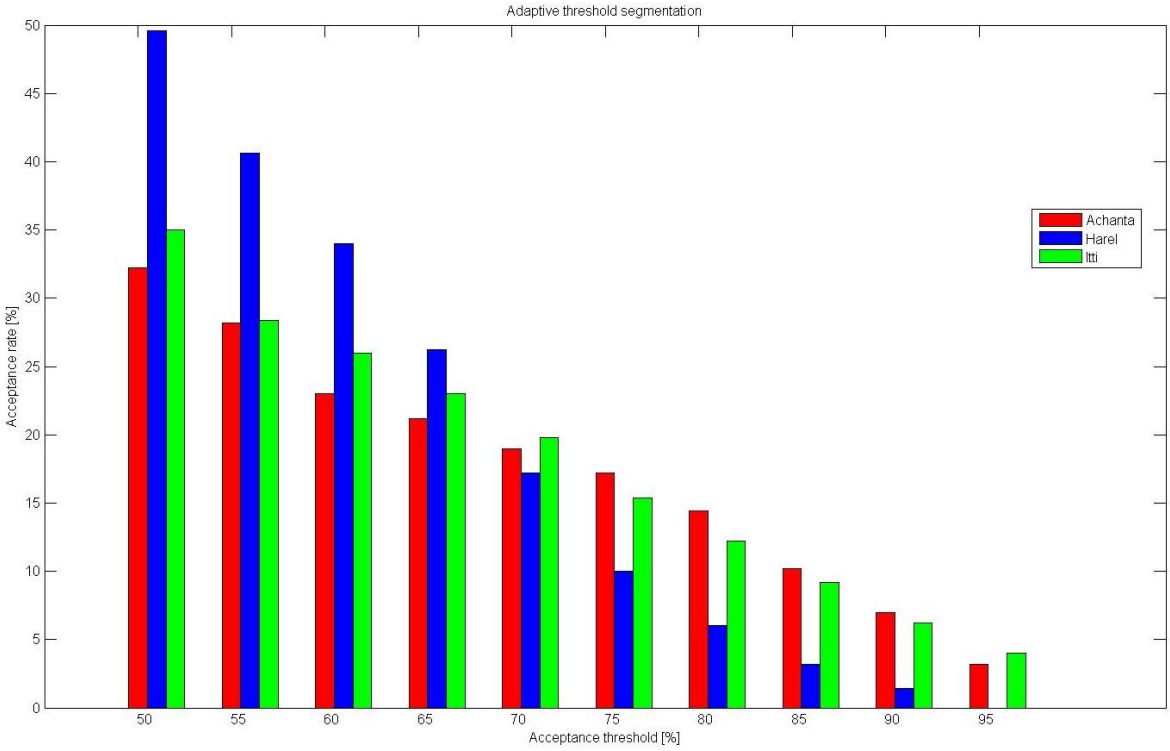



Figure 8: Adaptive thresholding: acceptance rate vs threshold

From now on acceptance threshold is fixed to 50% which seemed to be the appropriate value. For all objects that were accepted, accuracy value is calculated. Figure 9 shows the accuracy rates for the three approaches in terms of accuracy value intervals for the fixed acceptance threshold (number of object falling into an interval divided by the number of accepted objects). One can see that for accuracy value bands from 75 to 100 % adding all the rates together (multiplying by the number of accepted objects and obtaining the number of objects in (75...100]) Itti (green bars) and Achanta (red bars) have more objects than Harel (blue bars). Therefore Itti and Achanta hare higher number of precise objects than Harel which prooves the suggestion from Figure 8. Between Itti and Achanta it seems that Itti has larger number of precise objects.

On the other side for accuracy value bands form 50 to 75 % Harel contains the highest rates than other approaches therefore Harel accepts larger number of non-precise objects than Itti and Achanta. Achanta seems to have larger number of non-precise objects than Itti.

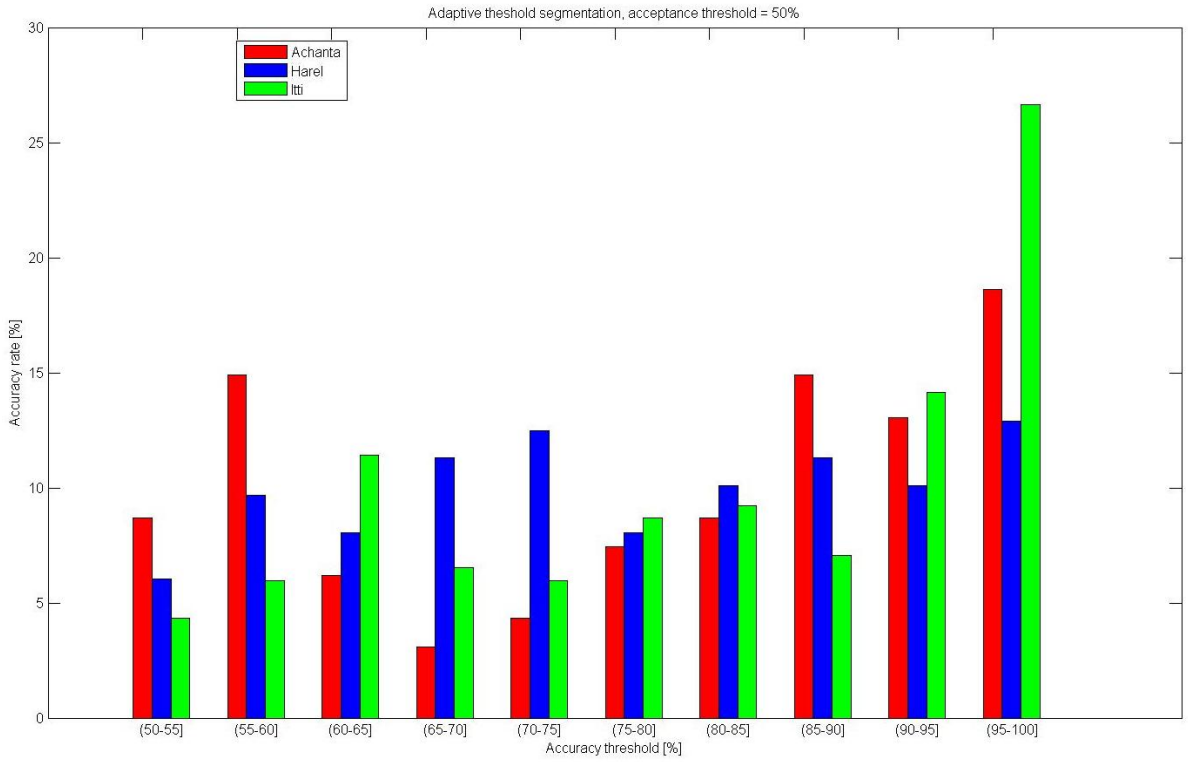



Figure 9: Adaptive thresholding: accuracy rate vs threshold bands

## 4.4 Mean-shift segmentation

In the second experiment mean-shift segmentation is performed. Mean-shift is an "image segmentation, decomposition of a gray level or color image into homogeneous tiles. Homogeneity is usually defined as similarity is pixels values"[8]. Figure 10 shows the segmentation algorithm using mean-shift. For each mean shift region the average saliency $S_k$ is calculated and compared to $2S_\mu$ where $S_\mu$ the mean saliency of the entire saliency map. If $S_k > 2S_\mu$ then the region is salient and all the pixels in this region are set to 255 (white) otherwise to 0 (non-salient/black). Then coordinates and the area of the most salient object are calculated.

Figure 11 shows the acceptance rates of the image set (500 images from MSRA database) for the three approaches in terms of the acceptance threshold. Here the story is more or less the same as in image-adaptive thresholding. The only difference is that Harel (blue bars) seems to have larger number of precise objects than before since it accepts more objects for high acceptance thresholds. Figure 12 shows the accuracy rates for the three approaches in terms of accuracy value intervals for the fixed acceptance

threshold (50 %). The suggestion from Figure 11 turnes out to be true. Harel accepts larger number of objects that are precise than for image-adaptive thresholding. Actually mean-shift segmentation for every approach accepts larger number of precise objects than for image-adaptive thresholding. Itti seems to accept the higher number of precise objects than other approaches.

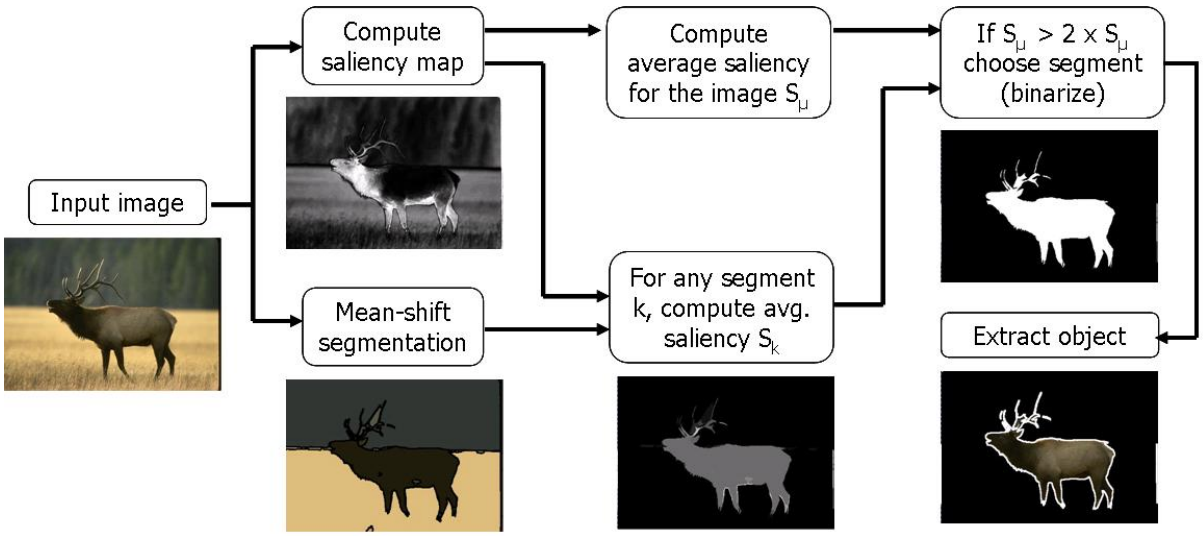



Figure 10: Salient Region Segmentation Algorithm using Mean-Shift

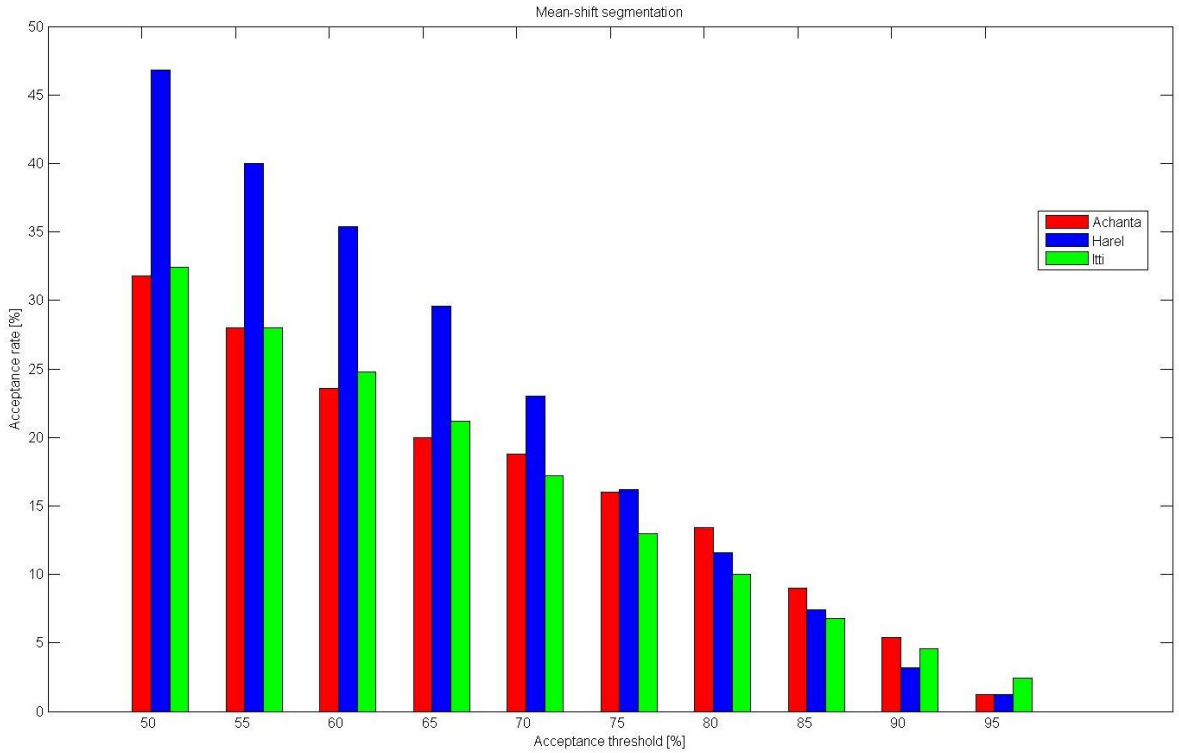



Figure 11: Mean-shift segmentation: acceptance rate vs threshold

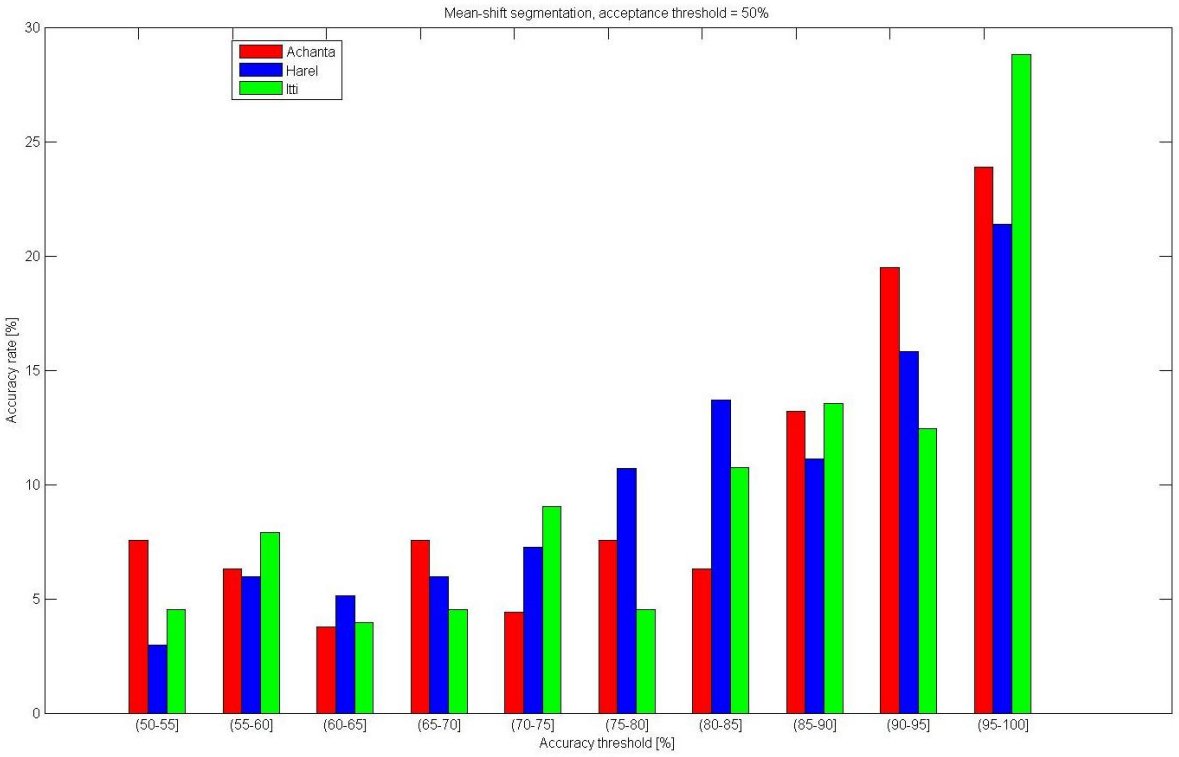



Figure 12: Mean-shift segmentation: accuracy rate vs threshold bands

## 4.5 Calculation time comparison

Figure 13 shows the average saliency map calculation time for 50 randomly chosen images in terms of the resolution of the image in pixels. Different resolutions were obtained by cropping the initial image (2271x1704) in the middle. One can see that for low resolution images Achanta (red bars) outperformes others. For resolutions from 1280x960 up to the highest, Itti (green bars) approach is the quickest. The Harel (blue bars) is the slowest one for all resolution except 1920x1440. These results are consistent with the complexity stated in Section 3. Figure 14 shows the average saliency map and mean-shift segmentation calculation time for the same 50 images as in Figure 13. The mean-shift procedure turned out to be very slow especially for high resolution images therefore it masks the difference in saliency map calculation time. So all the three approaches evolve similarly.

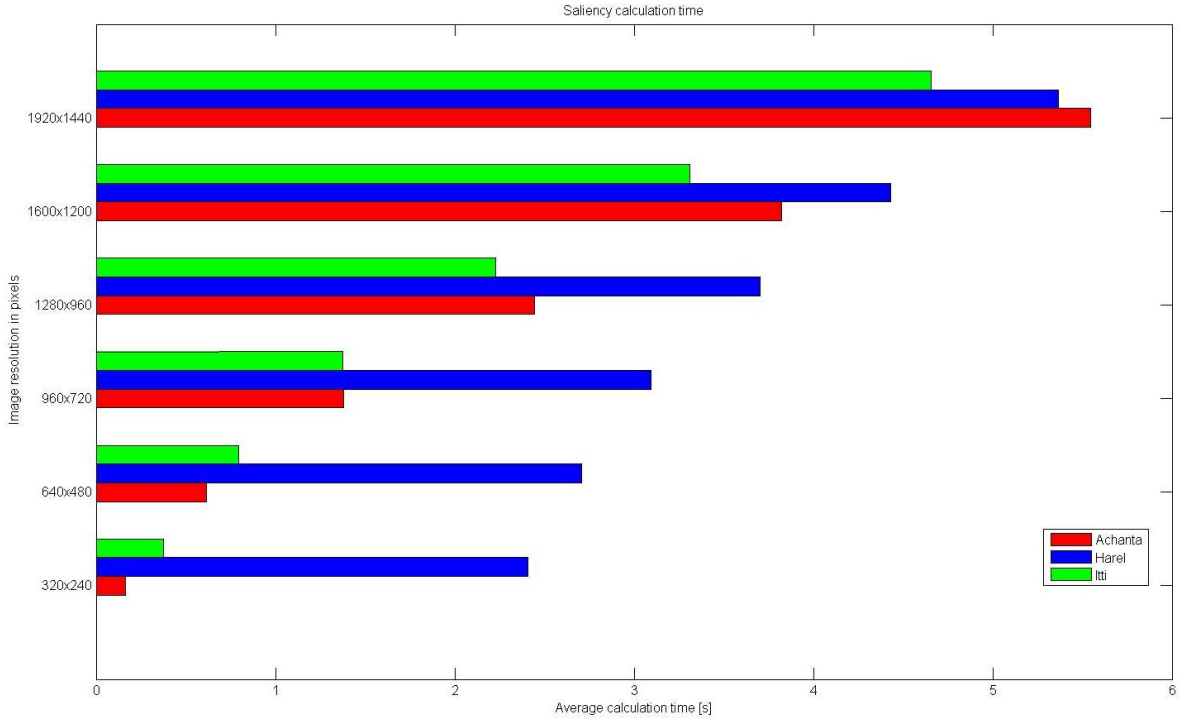



Figure 13: Average saliency calculation time

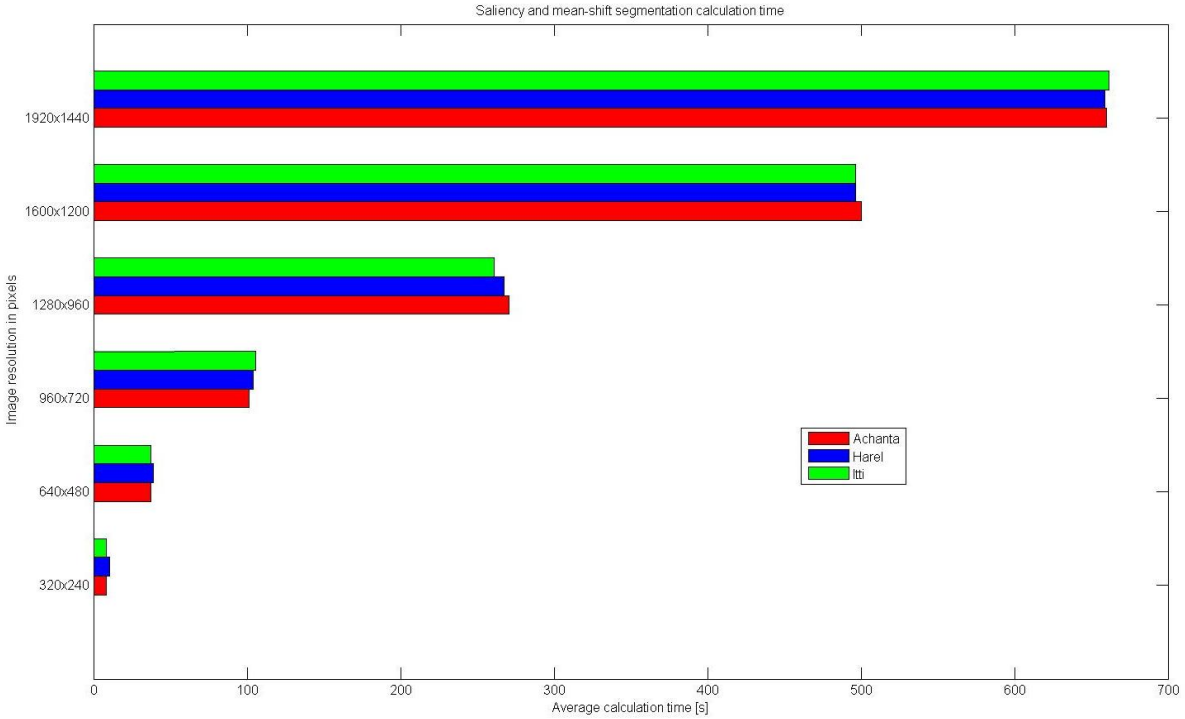



Figure 14: Average saliency + mean-shift segmentation calculation time

# 5 GUI in Java for interesting content extraction

The software that was developed for this project is presented in this section. It is a Graphical User Interface in Java. It allows to load any image ("Load Image" button at the left bottom corner), calculate saliency maps using the three chosen approaches ("Calculate Saliency Map" button at the bottom in the middle) and to show the segmented images ("Extract Content" button at the bottom on the right shows binding boxes around salient objects).

On the left side of the interface one can see two combo-boxes. They allow to switch from one approach to the other for comparison. There is a possibility to perform mean-shift segmentation (check the radio button in the middle on the right side near the label "Mean-shift segmentation" on the same level as the combo-box, if not checked performs image-adaptive thresholding) and also to show only the most salient object (check a radio button near the label "Most salient object" on the same level as the combo-box).

Figure 15 shows a snap-shot of the interface in use. In order to use the demo please refer to Readme.pdf

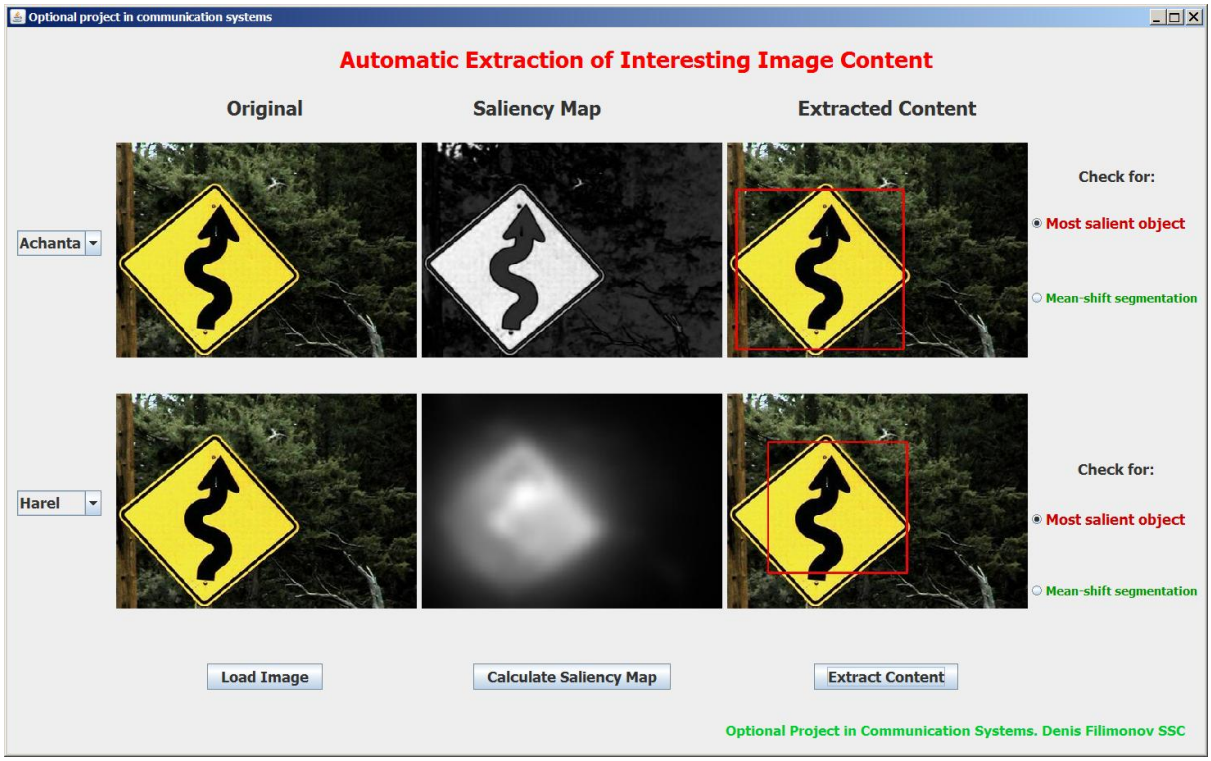



Figure 15: GUI in Java for interesting content extraction

# 6 Conclusion

The task of the project was at the first place to study different approaches for visual attention analysis and then to compare their performance. Three approaches that propose a method to calculate a saliency map were chosen. A test was run using a database with ground truth coordinates of the most salient object in an image. The saliency maps were calculated for each image and the image was segmented using these saliency maps and two different thresholding procedures (only the most salient object was kept). After words diagrams with acceptance and accuracy rates were produced. The evaluation of performance was done in terms of these diagrams. The first approach from [1] turned out to be the best in terms of the number of precise objects from the ones that were accepted. The approach from [2] was the worst one. [2] was also the worst one in terms of average saliency map calculation time for different image resolutions. The approach from [3] was almost as performant as the one from [1]. This concludes the study of methods for visual attention. Some future work that would be interesting to do is to study the approaches is more profound and deep level rather than theoretical and visual. It would also be

tempting to study other measures of precision and acceptance. The other point that is of a big concern may be to study images on the existence of salient objects.

# References

[1] 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, 1998.

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

[3] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, Frequency-tuned Salient Region Detection, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[4] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk. Salient region detection and segmentation. International Conference on Computer Vision Systems, 2008.

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

[6] Yiqun Hu, Xing Xie, Wei-Ying Ma, Liang-Tien Chia, and Deepu Rajan, Salient Region Detection Using Weighted Feature Maps Based on the Human Visual Attention Model, Advances in Multimedia Information Processing - PCM 2004

[7] C. Koch and S. Ullman, Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. Human Neurobiology, vol. 4, pp. 219-227, 1985.

[8] Dorin Comaniciu, Member, IEEE, and Peter Meer, Senior Member, IEEE, Mean Shift: A Robust Approach Toward Feature Space Analysis

[9] Tie Liu, Xi'an Jiatong University, P.R. China and Jian Sun, Miscrosoft Research Asia Beijing, P.R. China, Learning to Detect a Salient Object