Saliency map augmentation with facial detection

Julia Kucerova*
Supervised by: Elena Sikudova†

Faculty of Mathematics, Physics and Informatics
Comenius University
Bratislava / Slovakia

Abstract

Visual attention is very important in human visual perception. It is the ability of a vision system to detect salient objects in an observed scene. This scientific discipline has been studied for over a century. Nowadays it is involved in the disciplines of psychophysics, cognitive neuroscience and computer science.
This paper describes several visual attention models for detecting salient objects in complex scene and focuses on a model based on local context suppression of multiple cues. Although this model is useful to capture visual attention in images containing small objects, it fails in detecting faces as salient objects.
For this reason we improved the model by adding more attention cues. We propose a method for detecting salient objects based on texture, where face detection is used as an additional attention cue.

Keywords: Visual Attention, Texture attention cue, Salient object, Face detection

1 Introduction

"Everyone knows what attention is..."

William James, 1890

Humans cannot attend to all things at once. Their visual system has the ability to pay attention to some parts of the observed scene - salient objects. Visual attention models detect these salient objects in scene.

There are two general visual processes, called $bottom-up$ and $top-down$.
The bottom-up process is task-independent. This process tries to predict which parts of the observed scene could attract more attention and computes saliency map. It could be used in machine vision, automatic detection of goals in nature scenes, intelligent image compression, etc. Salient objects in scene are for example a burning candle in a dark room or the lips and eyes of a human face, because they are the most significant elements of the face. If there are many salient objects in the scene, they become obscure because of the big amount.
The top-down process is volition-controlled and task-dependent. It drives observer's attention on one or more objects that are relevant to the observers goal when studying the scene. For example the task could be to find red car on a car park, or to count particular objects in a scene. When the observer is concentrated to find some objects in the scene, he will fob off some salient objects. For that reason some objects that are salient in bottom-up process could not be found with top-down process.
In 1967 psychologist Yarbus recorded eye movements of participants watching an image. The subjects had task to observe Repin's picture "An Unexpected Visitor" and they were asked to answer a number of different questions (Figure 1).

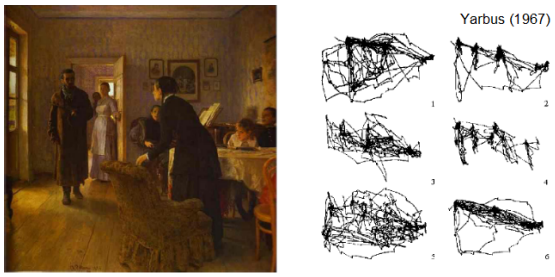

Figure 1: Repins picture was examined by subjects with different instructions; 1. Free viewing, 2. Judge their ages, 3. Guess what they had been doing before the unexpected visitors arrival, 4. Remember the clothes worn by the people, 5. Remember the position of the people and objects in the room, 6. Estimate how long the visitor had been away [1].

The motivation of our work is that visual attention is very important in human vision. We can use our knowledge about visual attention for many applications. In image compression we can compress background objects while salient objects stay untouched. Or we can use it in artefact removal algorithm to remove uninformative and distracting color boundaries [11].

Faces are very significant in human perception of the scene. This fact was studied in [5], where the authors investigated, whether faces are capable of capturing atten-

*kucerova.julia@gmail.com
†sikudova@sccg.sk

tion when compiting with other non-face objects. Their results suggest, that faces in fact attract attention.

The approach proposed in this paper is using face detection as an additional attention cue. It combines color, intensity and texture features with face detection map to get saliency map. Our work presents a model based on local context suppression of multiple cues presented by Hu [6].

This paper is organized in the following way: The work of other authors in the area of the visual attention is discussed in section 2. In section 3 the model based on local context suppression of multiple cues is described. In section 4 we describe face detection system used here and in section 5 the feature combination. In section 6 the paper presents the comparison of different methods and section 7 concludes the paper.

2 Related work

Visual attention has been studied for over a century. Early studies of visual attention were simple ocular observations. Since then the field has grown and nowadays it is involved in many scientific disciplines.

Scientists have observed human visual system, visual attention and many computational models have been proposed to predict what will attract our visual attention [2]. Human visual system is sensitive to features like changes in color, shapes, intensity etc. In some models low level features like color, intensity and orientation are used as attention cues.

Itti et al. [2] developed a visual attention model based on the behavior and the neural architecture of the early primate visual system. Authors used low level features like color, intensity and orientation as attention cues. They implemented linear center surround operation on multi-scaled feature images. This images are created using Gaussian pyramids. After normalization all feature images are combined into a single saliency map. 2D winner-take-all algorithm is used for detection saliency regions in an image. Ma and Zhang [9] proposed a new approach for obtaining the saliency map. They used contrast analysis and developed a fuzzy growing technique in the visual attention model to extract salient regions. Bergum et al. proposed mathematical framework of visual attention for robotic system. In [3] they integrated object- and space-based models of visual attention.

Visual attention models have a wide use. Nowadays we find them in robotic systems, image compression, commercial industry etc. There are many approaches and systems for detecting salient objects and they are still improving.

3 Model

In this section we describe model presented by Hu [6], which is used as a base model for our approach.

Hu's model is based on local context suppression. The authors used texture as an additional attention cue for salient region detection. They also developed feature combination strategy that suppresses regions in contrast maps. This strategy uses local context information to suppress spurious attention regions and enhance the true attention regions.

Texture is very useful to capture visual attention in images containing small objects. Texture Attention Cue used in this model was obtained as follows. Image was divided into blocks, called *texture patches*. By taking the Gabor Wavelet Transform at different scales each texture patch is represented by the mean and the standard deviation. In this way mean maps and standard deviation maps were obtained. Consequently Average Mean Difference (AMD) and Average Standard Deviation Difference (ASDD) were created. Texture contrast at a patch (i, j) at any scale s and orientation k was calculated as

$$TC_{s,k}(i,j) = AMD_{s,k}(i,j) \times ASDD_{s,k}(i,j). \qquad (1)$$

Consequently the final Texture contrast at patch (i, j) was calculated as

$$TC(i,j) = \sum_s \sum_k TC_{s,k}(i,j). \qquad (2)$$

Local context suppression strategy for adaptive combination of multiple attention cues like intensity, color and texture is describe here. Consider an image divided into blocks, called an *Attention Patches*, each containing $p \times q$ pixels. The contrast of particular feature at a patch centered at (i, j) is calculated as

$$FV(i,j) = \frac{1}{N} \sum_{u,v} |MF(i,j) - MF(i+u, j+v)|, \qquad (3)$$

where $MF(i, j)$ is the mean of the feature in patch (i, j) and N is the number of patches in its neighborhood. The contrasts at patch (i, j) for n features/attention cues are normalized to lie between $[0, 1]$. Each patch is now represented by the n dimensional feature contrast vector which is compared with other feature contrast vectors in its neighborhood and its contrast measure is suppressed if the patch and its neighbors are 'similar'. This similarity is estimated by the variance of data along eigen vectors of an $n \times n$ covariance matrix. This matrix is formed from the feature contrast vectors at a patch (i, j) and its neighborhood. The eigen values $\bar{\lambda}$ of this matrix represent the extent of similarity or dissimilarity among the attention cues. For example a large eigen value indicates large variance along the direction of its corresponding eigen vector, which implies higher discriminating power.

The suppression factor (SF) for patch (i, j) is obtained as $\tau(i,j) = \prod_{u=1}^{p} \bar{\lambda}_u$, where the $\bar{\lambda}$'s are sorted in ascending order and the parameter p controls the degree of suppression. For obtaining the saliency $S(i, j)$ for patch (i, j) the

multiple attention cues are linearly combined and the result is modulated by the *SF* as

$$S(i,j) = \tau(i,j) \times \sum_{u=1}^{k} FV_u(i,j). \quad (4)$$

The product of the combined map and the SF yields the final saliency map which contains the true Attention Regions. In the combined map there are spurious attention regions. Using Suppression Factor, these regions have been successfully removed [6].

4 Face detection

Detecting faces in the scene is a difficult problem. There are wide variety of faces to match, variations in lighting and shadows, presence of facial hair, possibility of scaling, angular and dimensional variances. Face detection is important in many human-computer interaction systems. There are many different approaches for detecting faces in the images: knowledge-based methods, feature invariant approaches, template matching, appearance-based methods.

In this paper we use face detection as an additional attention cue. We use two different systems for face detection and after comparison we decide for one of them. The first system is based on Rowley-Baluja-Kanade neural network. In order to better detect the faces this system was combined with skin detection. The second one is based on Viola/Jones' algorithm.

4.1 Rowley-Baluja-Kanade Face Detector

In this system we use combination of skin detection and Rowley-Baluja-Kanade (*RBK*) face detector. Skin color distribution used in this paper is modeled using a single 2D Gaussian distribution [12]. For face detection we used a software [10] that implements the Rowley-Baluja-Kanade *RBK* neural net face detector with some enhancements for training and recognition. Rowley's et al. face detection system is neural network-based system and authors present a straightforward procedure for aligning positive face examples for training.

As an input we have image in HSV color space. To get a faster face detection we used skin probability maps. Face detector is then applied only in the regions, where skin was detected. As an output we get regions of possible skin patches. We will label non-skin regions of the image (usually background) with 0 and possible skin regions with 1. Consequently we multiply this map with the input image. This results in set of possible face candidates that is used as input for face detection system. Consequently, after face detection, we used threshold to get binary map (Figure 2 c)).

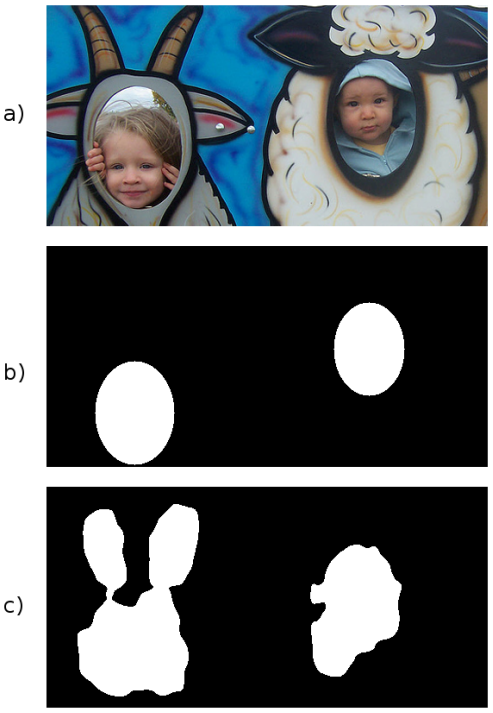

Figure 2: a) Input Image b) face detection based on 4.2 c) face detection based on 4.1

Face detector	Overlap	Left Out
Viola/Jones	80.24%	37.9%
Rowley	89.18%	79.2%

Table 1: Compare Face detectors

4.2 Viola/Jones' Face Detector

This system is used for real-time object detection. Training in this face detection system is slow, but detection is very fast. Key ideas of this face detector are integral images for fast feature evaluation, boosting for feature selection, attentional cascade for fast rejection of non-face windows.

We used the implementation of Viola/Jones' system (*VJ*) found in [8]. This system uses mid cumulative probability distribution point as threshold for weak classifiers.

We compared these two face detectors. As you can see in Table1-Overlap, Viola/Jones' system detected less faces as *RBK* Face Detector because of frontal detection. However *VJ* system has much less false positive detections than *RBK* Face Detector (Table1-Left Out). In our system we need good face detection with the least possible false positive detections. For that reason we decided to use *VJ* system.

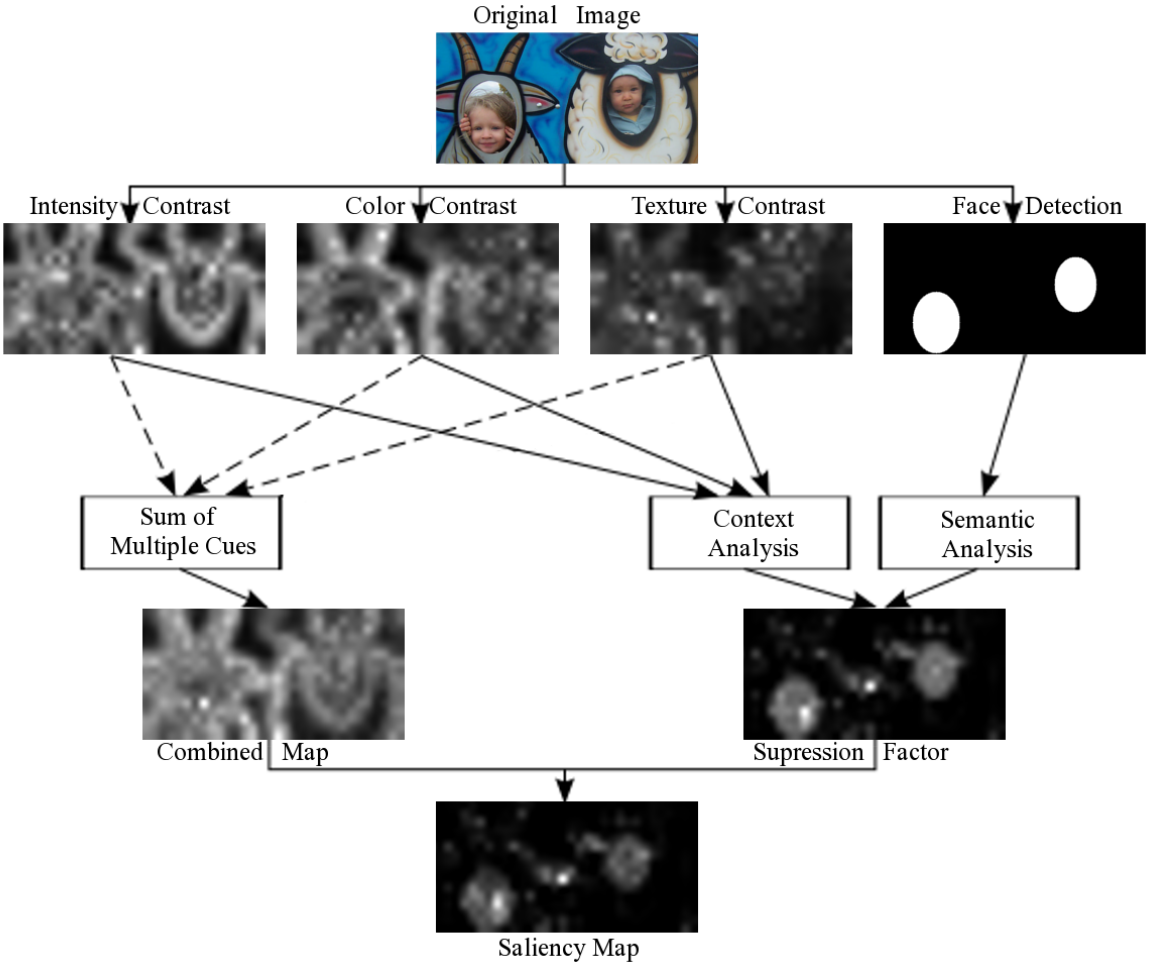

Figure 3: Features combination

5 Features combination

The combination of features that yields the final saliency map that includes only the true attention regions is a hard problem. Some approaches suggesting linear combination [2], other suggest some post-process, weighted combination etc.

In this paper we used and modified feature combination proposed in [6]. As shown in Figure 3, we have four features: color, intensity, texture and face detection maps. Contrast maps for intensity, color and texture are obtained the same way as in [6] and face detection map is obtained by Viola/Jones' Face Detector.

As a first step of feature combination we sum together and normalize three contrast maps (color, intensity, texture) to get the Combined map.
We derive the suppression factor by building up the suppression map from color, intensity and texture as proposed in [6]. We combine this map with the map for face detection to get the suppression factor, which highlights significant regions as well as faces.

Suppression factor is a map consisting of darker regions representing high suppression factor and brighter regions representing low suppression factor. That means, that brighter regions are more significant than darker regions. Consequently we multiply this map with Combined map. With this process we get final Saliency map for input image.

6 Results

This section summarizes the results of the proposed approach.

We used images from Visual Object Classes database [4] for testing, which is a benchmark in visual object category recognition and detection. It contains standard dataset of images and annotation, and standard evaluation procedures and significant variability on terms of object size, orientation, illumination etc. In this dataset images are sorted in many different classes as persons, animals, indoor images, vehicles etc.

We compare our system with the system based on Itti et

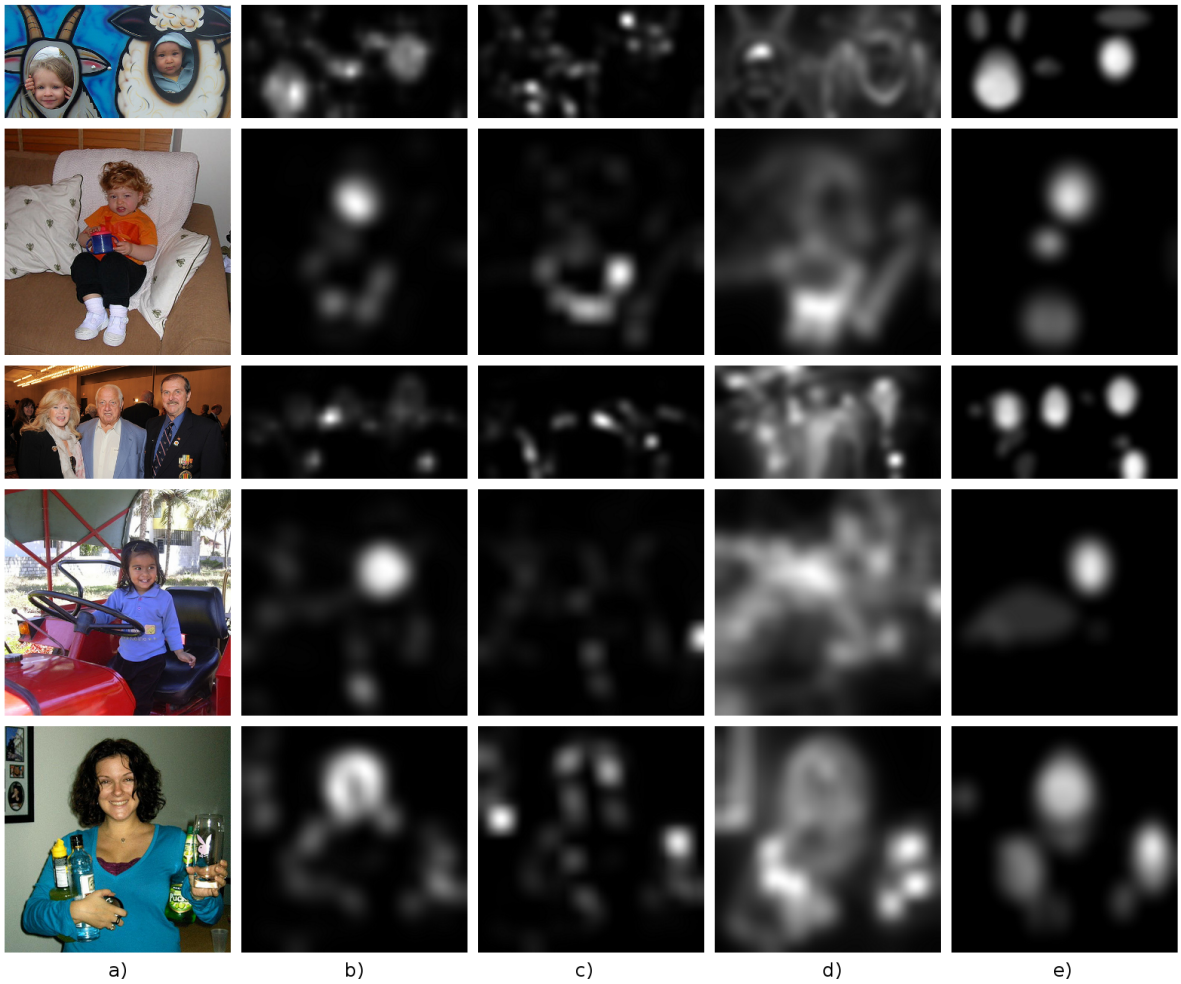

Figure 4: Experiment results a) Original Image; b) Saliency map using proposed method c) Hu‘s model [6] d) Itti‘s model [7] e) Manual combination

al. [7] and model proposed by Hu et al. [6]. For comparison we used salient regions obtained by manual inspection of the images. We asked several(11) observers to highlight significant regions. These maps were than summed together, normalized and then thresholded. In the next phase of our work, we will use data from eye tracking system to obtain real saliency data and compare them with our results.

For comparison our results we used symetric Kullback-Leibler divergence

$$KLD(P,Q) = \sum_i (P(i) - Q(i)) * \log \frac{P(i)}{Q(i)}, \qquad (5)$$

where P is saliency map obtained by Itti‘s model [2], Hu‘s model [6] or our model, and Q is the manual map. When the two probability densities are identical, KLD is null. The lower KLD, the better model.

As you can see in Table 2 our approach has the lowest KLD. Salient regions detected using Itti‘s model contain faces, but cover a significant portion of the input image.

Visual Attention Model	KLD
Proposed method	1.1377
HU	2.2807
ITTI	1.6642

Table 2: Compare Visual Attention Models

Hu‘s model is very useful in images obtaining for example texture foreground in non-texture background, but as you can see, it is less successful in images containing faces and bigger objects. Our approach detects salient regions of various sizes as well as faces.

Noticeable, that regions, which contain faces are more salient than other parts in the image.
Although the true attention regions are very subjective for each observer, they could be detected to a large extent. By using our model we can detect salient objects more accurately.

7 Conclusions and Future work

As a conclusion the best results achieved using this model are comparable with other visual attention models as Itti's model [2], Hu's model [6]. Our approach is based on the idea, that faces take more attention in the observed scene. We adapt input model [6] by adding face detection as an additional attention cue.

Data obtained with this approach are very useful. Detection of saliency regions in the observed scene is being used in image compression. Compression using visual attention rests in fact, that salient regions could be less compressed than non-salient regions.

In the next phase of our work the face detection process will be improved by exploring different feature combination and/or more color spaces. We will also improve the suppression factor for better results.

As a future work we planed to use eye tracking of subjects to obtain real saliency data and compare them with the proposed method.

8 Acknowledgments

The author wish to thank Elena Sikudova, PhD. for her support and the excellent leadership in this project.

References

[1] K. Cater, A. Chalmers, and G. Ward. Exploiting visual tasks for selective rendering. In *Eurographics Symposium on Rendering*, pages 270–280. Eurographics, 2003.

[2] L. Itti et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11).

[3] M. Begum et al. Object- and space-based visual attention: An integrated framework for autonomous robots. pages 301–306. IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2008.

[4] M. Everingham et al. Visual object classes database. [Online] `http://www.pascallin.ecs.soton.ac.uk/challenges/VOC/`, 2006-2009.

[5] S.R.H. Langton et al. Attention capture by faces. *Cognition*, 107:330–342, 2008.

[6] Y. Hu et al. Adaptive local context suppression of multiple cues for salient visual attention detection. In *IEEE International Conference on Multimedia and Expo*, pages 1–4, 2005.

[7] J. Harel. A saliency implementation in matlab. [Online] `http://www.klab.caltech.edu/~harel/share/gbvs.php`, 2010.

[8] V. Kazemi. Face detector (boosting haar features). [Online] `http://www.mathworks.com/matlabcentral/fileexchange/27150-face-detector-boosting-haar-features`, 2010.

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

[10] S. Sanner. Rowley-baluja-kanade face detector. [Online] `http://users.cecs.anu.edu.au/~ssanner/Software/Vision/Project.html`, 2005.

[11] F. Stentiford. A visual attention estimator applied to image subject enhancement and colour and grey level compression. In *International conference on Pattern Recognition (ICPR(3))*, pages 638–641. IEEE, 2004.

[12] E. Šikudová. *On some possibilities of automatic image data classification*. PhD thesis, Comenius University, Bratislava, Slovakia, March 2006.