Saliency Map for Human Gaze Prediction in Still Images

Puneet Sharma, Faouzi Alaya Cheikh and Jon Yngve Hardeberg
Norwegian Color Research Laboratory,
Faculty of Computer Science and Media Technology,
Gjøvik University College, Gjøvik, Norway
er.puneetsharma@gmail.com

Abstract

Under natural viewing conditions humans tend to fixate on specific parts of the image that interests them naturally. Understanding the mechanisms of the human visual attention may benefit numerous applications in a various fields of engineering, marketing and art such as image quality evaluation, label design, human computer interaction, etc.

Saliency map, proposed by Dirk Walther et al., represents the regions which are more prominent than other regions in terms of low level image properties such as intensity, color and orientation. We propose in this paper to modify this saliency map algorithm to account for one high-level feature, human faces, to better mimic the natural human attention and compare it to gaze maps obtained experimentally. The fixations of the gaze map are compared with the most salient regions of the saliency map. The factors that influence the relationship between the saliency maps and gaze maps are analyzed. Gaze map analysis was done for 20 test subjects using eye tracking device as they were shown a set of 190 images.

Introduction

Understanding visual attention is a challenge due to the variability of human visual perception. Human eyes when looking on an image tend to fixate on some important parts of image because the complexity of visual world exceeds the processing capacity of the human brain [1]. Attention implements an information-processing bottleneck that allows only a small part of incoming sensory information to reach short term memory thus understanding a complex scene is a series of computationally less demanding, local visual analysis problems [2]. These regions are valuable for understanding the dynamics of human visual system and the development of applications in various fields. The selection of these regions depends on stimulus and goal driven objectives. The subjects selectively direct attention to objects in a scene using both bottom-up, image-based cues and top-down, task-dependent cues [2]. Human visual perception is task specific to objects in a scene for e.g. in a busy restaurant when we are looking for an empty table then we ignore all the other details like people, decorations, background etc. However if we are free to look we will pay attention to all the details like lights, decorations etc. of the same restaurant. So human vision is the perception of the scene depending on stimulus (saliency) and task assigned visual attention is the ability of a vision system, biological or artificial, to rapidly detect potentially relevant parts of a visual scene, on which higher level vision tasks, such as object recognition, can focus. It is generally accepted nowadays that under normal circumstances human eye movements are tightly coupled to visual attention [3]. This can be partially explained by the anatomical structure of the human retina, which is composed of a high resolution central part, the fovea, and a low resolution peripheral one. Visual attention guides eye movements in order to place the fovea on the interesting parts of the scene [3].

Saliency is the quality of an object or item to stand out from rest of the objects or items. There are many different physical qualities that can make an object more salient than other objects in the scene such as its color, orientation, size, shape, movement or unique onset [1].

Saliency Map

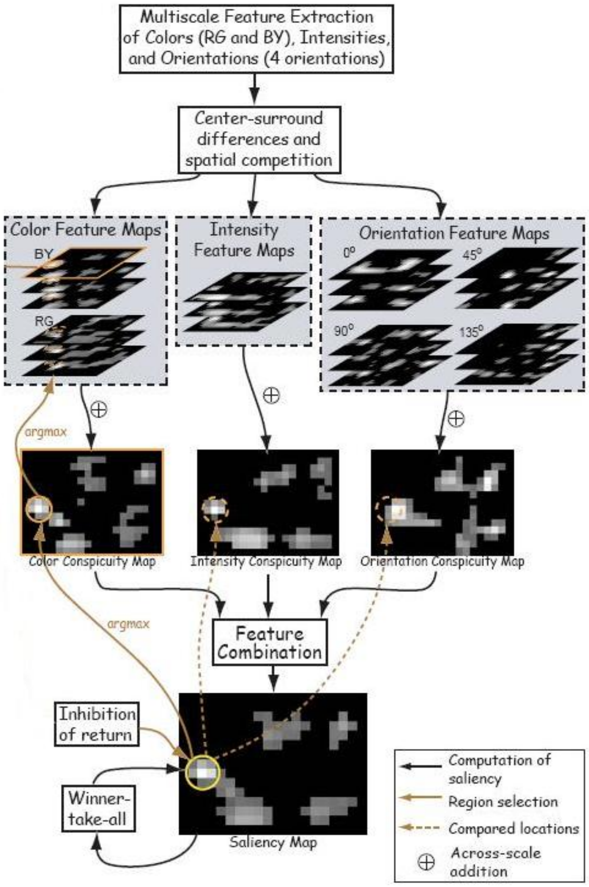

Figure 1. *Saliency map algorithm (Dirk Walther).*

Saliency at a given location is determined primarily by how different this location is from its surroundings in color, orientation, motion, depth etc. It refers to physical bottom-up distinctiveness of an object in relation to other objects in the scene [3]. The saliency map was designed as input to the control mechanism for covert selective attention. Koch and Ullman [4] state that the most salient location (in the sense defined above) in a visual scene would be a good candidate for attentional selection. Once a topographic map of saliency is established, this location is obtained by computing the position of the maximum in this map by a Winner-Take-All (WTA) mechanism. After the selection is made, suppression of activity at the selected location (which may correspond to the psychophysically observed "inhibition of return" mechanism) leads to selection of the next location at the location of the second highest value in the saliency map and a succession of these events generates a sequential scan of the visual scene.

As shown in figure 1 the procedure for saliency map generation [5] is described as follows:

- Input image is sub sampled and 6 different scales of the image are obtained.
- Intensity map is calculated from the R, G, and B components of the image.
- Red-Green and Blue-Yellow maps of the image pyramid are calculated.
- Local orientation maps are obtained from the intensity pyramid levels for the angles of 0, 45, 90 and 135 degrees.
- Each iteration step consists of self excitation and neighbor induced inhibition, implemented by convolution with a "difference of Gaussians" filter followed by rectification.
- Feature maps are summed over the center surround combinations using across scale addition and the sums are normalized again.
- Conspicuity maps corresponding to color, intensity and orientation are obtained.
- All Conspicuity maps are linearly combined into single saliency map.

A combined model of face detection and low-level saliency outperforms a low-level model in predicting locations humans fixate on [6]. Viola and Jones [7] feature-based template matching algorithm combined with bottom-up saliency map model of Itti [8] was used. Seven subjects viewed a set of 250 images in a three phase experiment. Overall in both of the experimental conditions i.e. "free viewing" and "search" faces were powerful attractors of attention, accounting for a strong majority of early fixations when present. The findings pointed towards a specialized "face channel" in our vision system, which is subject to current debate in the attention literature. Inspired by biological understanding of human attention allocation to meaningful objects - faces - a new model for computing an improved saliency map which is more consistent with gaze deployment in natural images containing faces than previously studied models was developed. Results suggested that faces always attract attention and gaze, relatively independent of the task. It should therefore be considered as part of the bottom-up saliency pathway [6].

In [9] the authors used a coherent computational approach for the modeling of the bottom-up visual saliency. Contrast sensivity functions, perceptual decomposition, visual masking and center surround interactions were used in the model. Ten natural color images with various contents were selected. The quality of these pictures was degraded using different techniques (spatial filtering, JPEG, JPEG2000 coding etc.). Forty-six pictures were finally obtained. Each image was viewed in random order by up to 40 observers for 15 seconds each in a task-free viewing mode. Qualitative or subjective evaluation showed that similarity between the predictions and the experimental results was good. The architecture of the proposed model was similar in spirit to the Koch and Ullman [4] architecture. The fundamental difference was the normalization of all the early visual features. The visibility threshold was modified by the context, and was incorporated by the modeling of visual masking. Linear correlation coefficient and the Kullback-Leibler divergence were used to conduct the qualitative comparison. These coefficients were 0.71 and 0.46, respectively. The proposed model outperformed the model of Itti [8] in all the tested configurations [9].

Experiment Setup

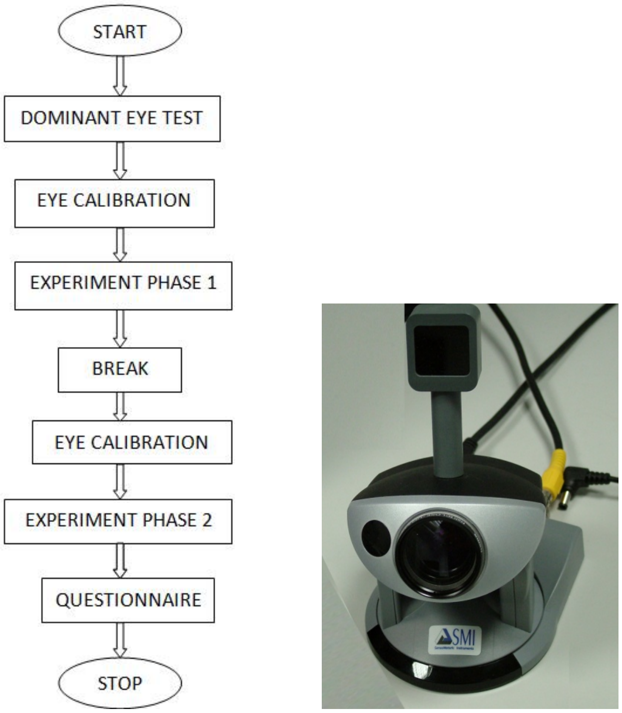

Figure 2(a). ***Figure2(b).***

Figure 2. (a) *Experiment sequence,* ***(b)*** *RED eye tracking camera.*

In our experiments subjects were shown a set of 190 images and data was recorded as they freely viewed on each image. Each image was shown for 2 sec only similar to the experiment performed [5]. The experiment was divided into two phases in order to provide adequate rest to the test subjects. As shown in figure 2(a) the experiment was performed in 7 steps. First of all the dominant eye of the test subject was determined by using *Porta* test [10]. Next the eye tracker was calibrated to the dominant eye. Performing the calibration on the dominant eye yields accurate results which is an important issue in this project [11].

After calibration experiment phase 1 was started. In this phase each subject was shown a set of 90 images for a total period of 180 seconds. Each image was shown for a period of 2 seconds only. After this phase the subject was given a break to rest his or her eyes. The dominant eye is recalibrated before entering the experiment phase 2. In experiment phase 2 each subject was shown a set of 100 images for a total period of 200 seconds. Each image was shown for a period of 2 seconds only. After the completion of this step subject was given a questionnaire to fill. The trails were performed under free view conditions on the dominant eye of the subject using SMI eye tracker equipment available in the lab. The eye tracker shown in figure 2(b) was used to measure the fixation of the user's eye at important parts of the image. It is a contact free gaze measurement device. Remote Eye Tracking Device was used for the experiment. Observers felt worse with Head Mounted Eye Tracking Device (HED) than Remote Eye Tracking Device (RED) because of the size and weight of HED. The difference of precision between HED and RED is about 10-16 pixels for printed images [12]. These reasons favored the use of RED for this experiment. The chair used had 4 legs, armrests and backrest, this type of chair was chosen to minimize observer movement. All the experiments were performed with same lighting conditions and the distance between the observers head and display monitor was approximately 70 cm giving a viewing angle of 30x23 degrees. The viewing angle is calculated by measuring the width of the monitor and distance between the viewing location and the monitor. The intensity of light at the front of display, back of the display was 159 lux and 146 lux respectively as measured from i1Display device. All the images were of same resolution i.e. 1024x768 pixel and they were displayed on a monitor of same resolution. Matlab program was used to show full screen images to the subjects and communicate the recording of the eye fixations with the computer connected to eye tracker.

Proposed Saliency Map Model

We propose a modified saliency map model by adding a top-down face detection procedure as shown in figure 3. The conventional saliency map model is based on low level features only. Saliency is computed by multi scale feature extraction and center surround differences to obtain color, intensity and orientation maps. Across the scale addition of these maps yields conspicuity maps of color, intensity and orientation. Next an equal weight combination of the three normalized conspicuity maps of color, intensity, orientation with face detected region will generate saliency map by WTA network. We used FaceOnIt library [13] for face detection procedure. The Face detection algorithm [13] is able to detect 189 correct faces out of the total 212 faces in the 190 images. The number of false faces detected are 30, and 18 faces present in the images are not detected at all. The face regions are represented as 1 in a binary image to generate the face map. This face map is added to the normalized saliency map and the resulting map is normalized again. The modified saliency map is multiplied by a factor of 255 in order to represent the original image in the background.

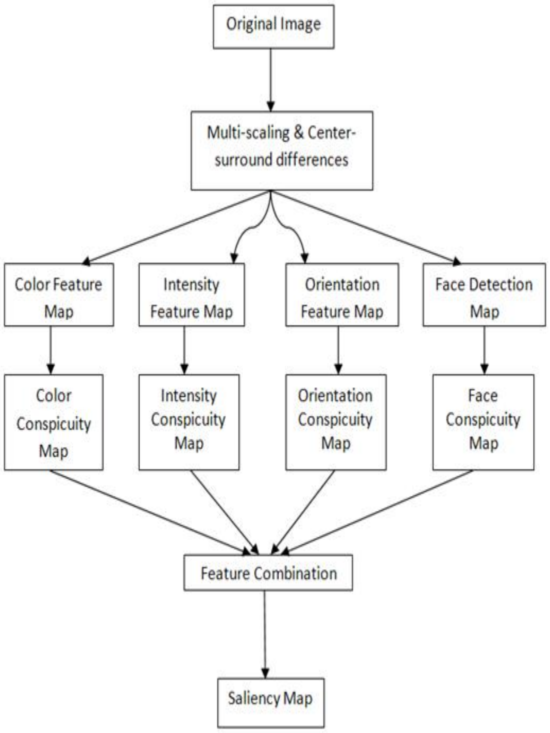

***Figure 3**. Modified Saliency map algorithm.*

Results

The set of images used for this experiment consisted of 190 images chosen from the database of images used by [6]. The subjects were not given any instructions to judge the images or look for a particular object in it. The gaze maps obtained for 20 users were summed and normalized to get 190 gaze maps corresponding to 190 images. There were 190 images with 212 faces in them, 17 grayscale images and 39 images with no faces. Subjects fixated on 194 of the total 212 faces in the database which validate one of the observations made by [5]. Subjects fixated on the faces in all the grayscale images also. Objects including faces at center of the image were fixated more as compared to faces or objects at other parts of the image. But in a few images faces at the other parts of images were fixated more than faces at the center. This was possibly because subjects fixated more on the changes i.e. new faces or new objects appearing in the images. The area of face regions in the image was also a factor as larger face (as a result of person standing near to the camera) was fixated more as compared to smaller faces. After the faces the most prominent regions were objects like toy car, toy banana, mobile phones, magic cube, books etc. Numbers, alphabets on posters were also strong in attracting attention of the subjects.

Figure 4 shows the gaze map in the form of heat map with the original image superimposed in the background. The regions fixated by the subjects are two faces, book at left-bottom, toy at top-left and some other random regions. The regions that are fixated appear red or yellow depending on the fixation or gaze time spent by the subject on those regions within the 2 second time of each image view. The maps are filtered using Gaussian filter to enhance the representation of the map [14].

Figure 5 shows the regions obtained by modified saliency map. Face in the image appears salient because of face detection procedure added to the saliency map. Book and toy appear as salient regions in the map. Figure 6 shows the gaze map and the regions fixated by subjects are two faces in the image and toy banana at the center of the image and some fixations around the text region at top-center of the image. Figure 7 shows the regions obtained by modified saliency map as two faces and toy banana at the center of the image. Figure 8 shows the gaze map for another image with fixations around the face, bookshelf on the top-left of the image, objects on the top of bookshelf and some other random fixations. Figure 9 shows the regions obtained by modified saliency map as the face at bottom right of the image and objects at the top left of the image. The results for these images show high correlation between the saliency map and gaze map. As we have four common regions in figure 3 and 4, three common regions in figures 5 and 6 and two common regions in figures 7 and 8. The database [6] of images was arranged in such a manner that none of the two consecutive images had the same background although similar people were appearing in the images. Some random images were added (in database) to break the sequence feeling during viewing of similar images and to maintain the level of interest among the subjects.

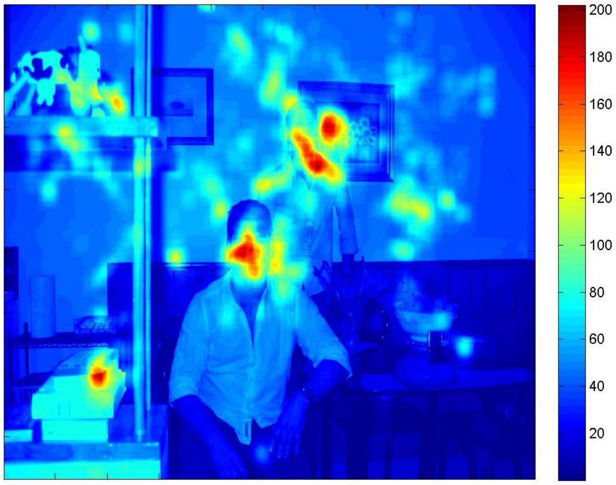

Figure 4. *Gaze map.*

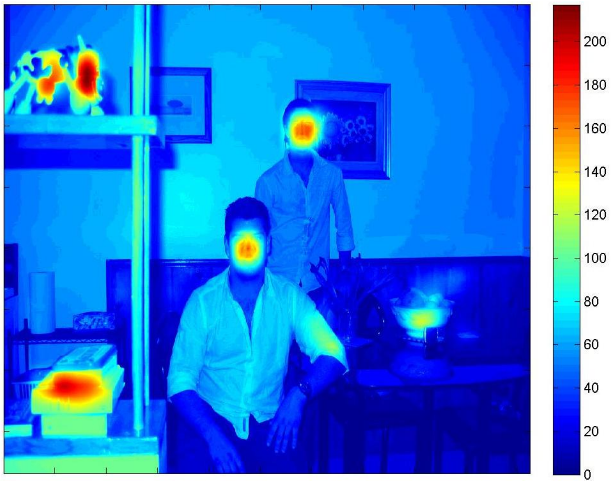

Figure 5. *Modified Saliency map.*

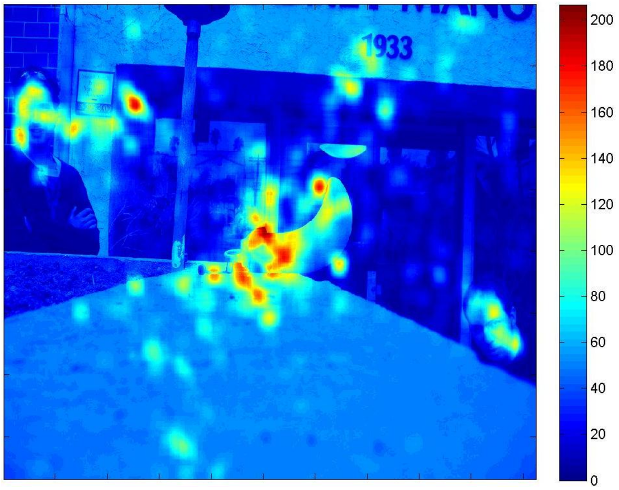

Figure 6. *Gaze map.*

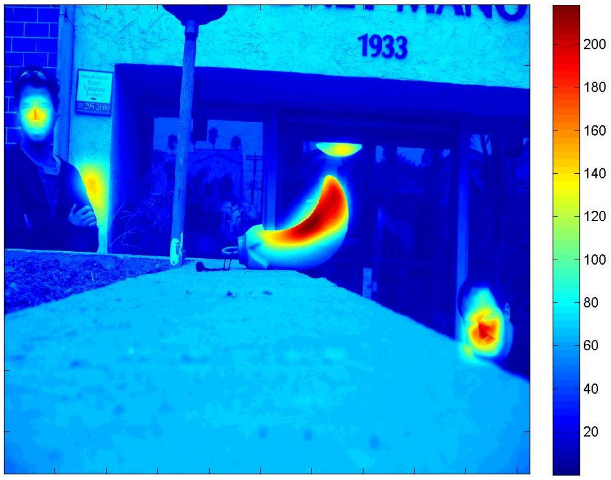

Figure 7. *Modified Saliency map.*

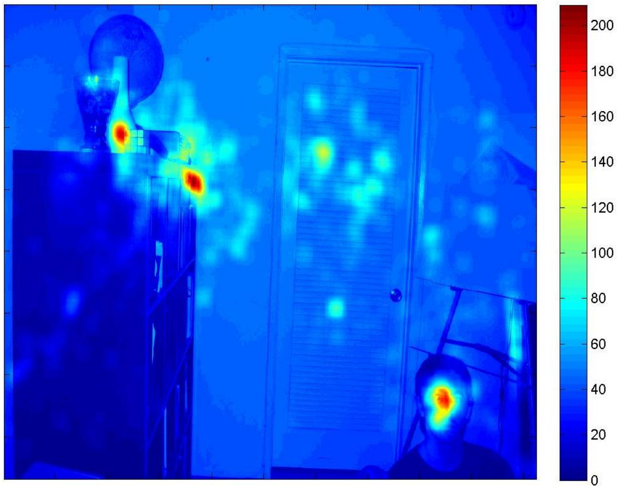

Figure 8. *Gaze map.*

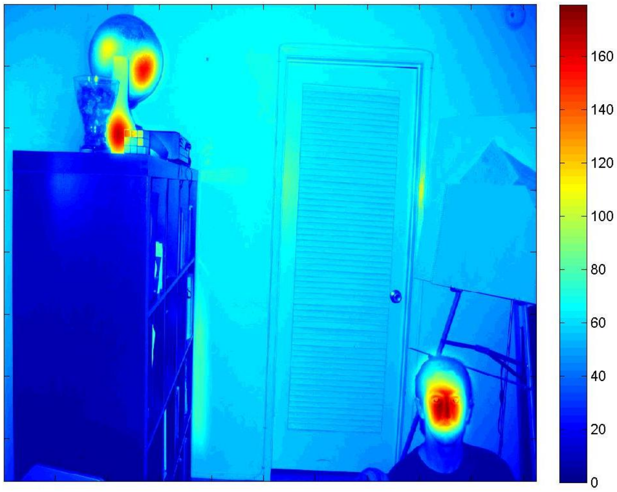

Figure 9. Modified Saliency map.

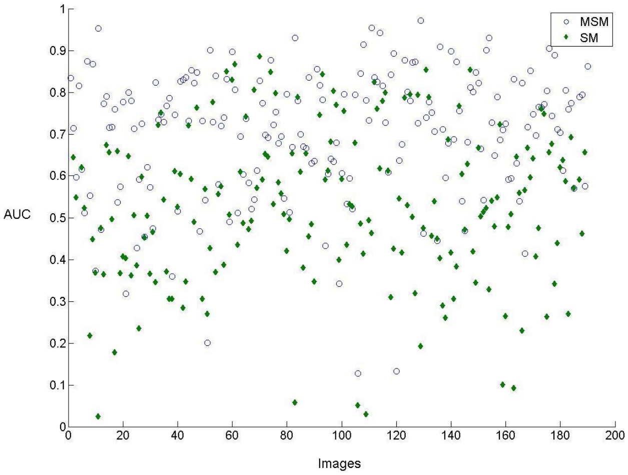

Figure 10. Scatter plot of AUC for MSM and SM.

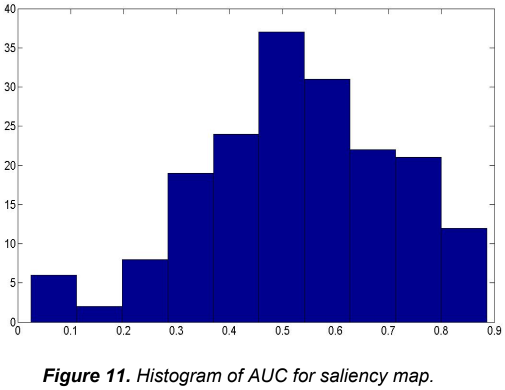

Figure 11. Histogram of AUC for saliency map.

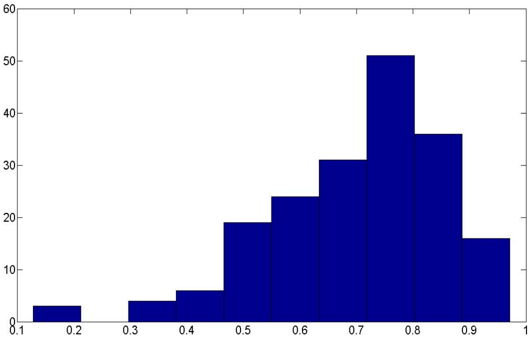

Figure 12. Histogram of AUC for modified saliency map.

A Receiver Operating Characteristic (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance [15]. ROC curves provide a visual tool for examining the trade off between the ability of a classifier to correctly identify positive cases and the number of negative cases that are incorrectly classified. An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we can reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC. Area under the receiver operating characteristic graph (AUC) [15] was used as measure to classify the performance of saliency map and modified saliency map algorithm. Figure 10 shows the scatter plot of AUC for saliency map(SM) and modified saliency map (MSM). Clearly the AUC for MSM lies above the AUC for SM. Figure 11 shows the histogram for AUC of SM. Most of the values lie around the range of 0.5. Figure 12 shows the histogram for AUC of MSM. It shows that most of the values lie in the range 0.7 to 0.8. The mean AUC for MSM is 33 percent better than mean AUC for SM. Hence the performance of MSM is better than that of SM for prediction of gaze maps.

Conclusions and Future Work

Based on the observations we can conclude that there is high correlation between the modified saliency map and gaze map. The performance of saliency map algorithm is improved by 33% with the addition of face detection procedure. The number of test subjects i.e. 20 in the experiment is good, as compared to 7 used by [6], for analyzing the perception of human visual system more accurately. The addition of face detection procedure in the saliency map corresponds better to the observations from gaze map obtained during the experiment. Objects seen in daily life like mobile phones, computers, books, balloons etc. may contribute to early visual fixation as analyzed from the results of gaze map. Under normal conditions also we tend to fixate more often on known objects which may or may not be salient. The conditions in the experiments (as described in experiment setup) gave the subjects the natural environment for attention selection in images. The concept of saliency must be broadened to include top down approach like face detection and possibly other objects detection in future.

References:

[1] J. H. Fecteau, D. P. Munoz, “Salience, relevance, and firing: a priority map for target selection,” Trends in Cognitive Sciences, Elsevier, 10:382-390(2006).

[2] L. Itti, C. Koch, “Computational modelling of visual attention,” Neuroscience, 2(3), 194-203 (2001).

[3] T. Jost, N. Ouerhani, R. V. Wartburg, R. Muri, H. Hugli, “Assessing the contribution of color in visual attention,” Computer Vision and Image Understanding, Elsevier, 100, 107-123 (2005).

[4] C. Koch, S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” Human Neurobiology, 4, 219-227 (1985).

[5] Dirk Walther, Interactions of Visual attention and Object Recognition: Computational Modeling, Algorithms and Psychophysics. (PhD thesis, California Institute of Technology, Pasadena, California, 2006).

[6] M. Cerf, J. Harel, W. Einhauser, and C. Koch, Predicting human gaze using low-level saliency combined with face detection, In Advances in Neural Information Processing Systems (NIPS 2007), pg. 241-248, (2007).

[7] P. Viola, M. Jones, Rapid object detection using boosted cascade of simple features, CVPR 2001, IEEE, 2, pg. 589-592 (2001).

[8] Laurent Itti, Models of Bottom-Up and Top-Down Visual Attention. (PhD thesis, California Institute of Technology, Pasadena, California, 2000).

[9] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, “A coherent computational approach to model bottom-up visual attention,” IEEE Transcations on Pattern Analysis and Machine Intelligence, 28, 802-817 (2006).

[10] Clare Porac and Stanley Coren, The dominant eye (Psychological Bulletin, 83(5), 1976) pg. 880–897.

[11] H. L. Roth, A. N. Lora and K. M. Heilman, “Effects of monocular viewing and eye dominance on spatial attention,” Brain, 125, 2023-2035 (2002).

[12] B. Kominkova, M. Pedersen, J. Y. Hardeberg and M. Kaplanova, Comparison of eye tracking devices used on printed images, Human Vision and Electronic Imaging VIII (HVEI-08), pg. 68061I-68061I-12 (2008).

[13] Tiffany Sauquet, Yann Rodriguez, Sebastien Marcel. Multiview face detection (IDIAP-RR 49, IDIAP, 2005).

[14] Marius Pedersen, Importance of region-of-interest on image difference metrics (Master's thesis, Department of Computer Science and Media Technology, Gjovik University College, 2007).

[15] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, 27, Elsevier, pg. 861-874 (2006).

Author Biography

Puneet Sharma received his B.Tech in Electronics & Communications Engg. from the Punjab Technical University, India in 2006 and his M.Tech from National Institute of Technology, Jalandhar, India in 2008. He worked at Norwegian Color Research Laboratory, Norway on the project ‘Perceptual Image Difference Metrics – Saliency Maps & Eye Tracking’. His research interests include image processing and Digital Signal Processing.

Dr. Faouzi Alaya Cheikh received his Ph.D. in Information Technology from Tampere Univ. of Technology, Finland in 2004; where he worked with the Signal Processing Algorithm Group since 1994. Since 2006 he works as associate professor with the department of computer science and media technology at Gjøvik Univ. College, Norway. His research interests include image and video processing and analysis and content-based retrieval. He holds over 40 papers and is IEEE, EURASIP and NOBIM member.

Dr. Jon Y. Hardeberg received his Ph.D. from the Ecole Nationale Sup´erieure des T´el´ecommunications in Paris, France in 1999. His Ph.D. research concerned color image acquisition and reproduction, using both colorimetric and multispectral approaches. He then worked for 2.5 years as a color scientist with ViewAhead Technology in Bellevue, Washington, USA. He is currently professor with Gjøvik University College in Norway, where he is teaching and researching in the field of color imaging science.