# HUMAN VISUAL PERCEPTION CONCEPTS AS MECHANISMS FOR SALIENCY DETECTION

Oana Loredana BUZATU
*“Gheorghe Asachi” Technical University of Iasi, 11, Carol I Boulevard, 700506 Iasi, Romania*
*lbuzatu@etti.tuiasi.ro*

Abstract: **Predicting human eye fixations is a difficult task that usually involves complex knowledge about the human visual perception pathways. This paper proposes a method to detect salient locations using a psycho visual space based on color perception, perceptual decomposition of visual information in multiple processing channels and contrast sensitivity as basic concepts. Using a linear support vector machine to train a saliency detector directly from human eye fixation data, salient locations were predicted as a linear combination of features. Testing results have proved that the proposed method performs better than other state-of-the-art methods with certain common concepts.**

***Keywords:*** *visual attention, saliency, color opponency, steerable pyramids, contrast sensitivity, support vector machine.*

## I. INTRODUCTION

For many applications in design, graphics and human computer interaction, processing a large amount of visual information from a given scene is a challenging and complex task, especially when dealing with natural images. In image analysis, complexity reduction is based on selecting only the most interesting regions, applying biologically inspired strategies. It is known that the Human Visual System (HVS) continuously removes redundant information in favor of the most salient features or objects, by means of a highly selective mechanism- known as visual attention. This is why attention became an adequate bio-inspired solution which can solve the problem of complexity reduction. Visual attention results both from fast, pre-attentive bottom-up mechanisms, which accounts for features that stand out from the context, as well as from slower top-down attention which is knowledge-based and task-dependent. Although, neurophysiologic results suggested that these two mechanisms of attention take place in different areas of the brain [1, 2], they interact to each other and a complete simulation of both of them results in a tremendously complex and time-consuming algorithm.

As far as bottom-up attention is concerned, it is well known that there are psychophysical and neuropsychological experiments [3, 4] which support the idea that in a certain part of the brain a topological image-based saliency representation exists, named saliency map. Hence, understanding and modeling of bottom-up saliency may not only help to elucidate the mechanisms of attention, but also reduce the amount of information to process. Thus, most biologically inspired models of saliency are based on a bottom-up computational approach. In most of the cases, these methods follow the feature integration theory (FIT) developed by Treisman [3]. This theory states that distinct features from a scene are automatically registered by the visual system and processed in parallel channels, before the items in the image are actually identified by the human observer.

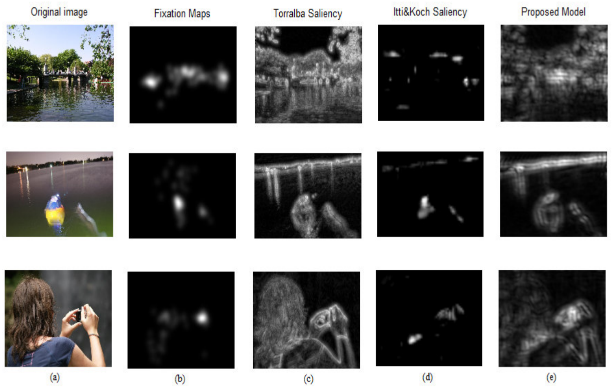

*Figure 1. Most of the saliency models do not match accurately actual human saccades from eye-tracking data. (a) Original images; (b) fixation maps from human eye-tracking data base; (c)* ***Torralba*** *saliency map; (d)* ***Itti & Koch*** *saliency map; (e) resulted saliency map from SVM (see Section III).*

Methods such as those proposed in [4, 5], which are inspired by the human visual system, extract low-level visual features such as color, intensity, motion direction, texture or orientation and aggregate them into salient regions. Although they perform well, these methods have limited appli-

*Manuscript received January 25, 2012; revised February 22, 2012*

cability, because in most cases they do not match with human eye-fixation from eye-tracking data. In figure 1 it is shown how the low-level models select mostly the strong edges or other features, which do not match all the times with human eye fixations. However, this mismatch is generally accepted, given the fact that human visual system modeling requires accurate knowledge about the entire pathways and, at present, only certain aspects of vision are well understood. Therefore these methods are often improved by tuning specific design parameters.

This paper proposes a method based on biologically plausible concepts which take into account human vision principles such as the color perception mechanism, the perceptual decomposition of visual information in multiple processing channels and contrast sensitivity. Following Treisman theory [3], the method exploits low-level features simultaneously through computation and composition to form a saliency map which is aiming to predict human eye-fixations. In order to fine tune and train the model a linear support vector machine (SVM) was used. Each feature performance was tested and then combined to each other to obtain the best results. Also, this paper provides objective evaluation of the accuracy of the proposed method against two state-of-the-art models using as ground truth the eye-tracking dataset developed by Judd *et al.* in [6].

## II. PREVIOUS WORK

Saliency estimation methods can be broadly classified as biologically inspired [4, 5], purely computational [7, 8] or a combination of both [6, 9]. Generally, most models employ a bottom-up approach by determining which stimuli are clearly different from the homogenous surrounding and attract human's glance without the need to search in the scene or regardless of the number of nearby objects acting as distractors [3].

Itti *et al.* [5] derive their method from the biologically plausible architecture developed by Koch & Ullman [4]. The computational saliency-based architecture conceived in [4] gathers the ideas of Treisman [3] and includes a Winner Takes All (WTA) network to determine the next salient location and a mechanism of Inhibition of Return (IOR) to allow a dynamic selection of different regions in a scene. The similar approach of Itti *et al.* [5], made use of contrast, colour and orientation features, which were computed in a centre-surround fashion using Difference of Gaussians (DOG) filters applied at different scales.

The purely computational model of Gao and Vasconcelos [7] estimates saliency by maximizing the mutual information between the distributions of local features such as centering and surrounding at a given point of the image, while Liu *et al.* [8] calculates pixel saliency considering multi-scale contrast, center-surround color histograms and color spatial distribution. All features are combined in a Conditional Random Field (CRF), resulting in a binary label map which separates the salient objects from the background.

The third category of methods incorporates ideas that are partly biologically inspired and partly computationally based. For instance, Judd *et al.* [6] combines a set of low, middle and high-level image features to define a saliency map and uses a linear support vector machine to train a model of saliency. Most of the low-level feature extractions are based on biological vision concepts. Such examples are the colour contrast, intensity and orientation conspicuity maps which are computed using either Itti & Koch [5] saliency method or the local energy of the physiologically plausible steerable pyramid filters, first proposed in [10], which have been shown to correlate with visual attention.

## III. THE PROPOSED METHOD

Based on biological observations regarding the visual bottom-up attention task and following FIT theory, this model aims to estimate salient locations in a psycho visual space, justified by psychophysical experiments, and from which low-level features (intensity, color, orientation, spatial frequencies) are extracted and combined into salient locations.

The model described in this framework is built-up from three stages, according to the most important pathways of the HVS seen as a passive selector. The first stage is dedicated to color perception in the retina, taking into account the wavelength of the visible light. The photoreceptor cells from the retina process the chromatic information according to the Young-Helmholtz trichromatic theory [11, 12], and then the color data is encoded following the opponent colors theory [13]. There are two types of photoreceptor cells: rods and cones, each of them having different functions. Rods are found primarily in the periphery of the retina and are used to perceive low levels of light. Furthermore, rods are not sensitive to color, only to light/dark or to white/black. Cones, which are less sensitive to intense light, are concentrated at the outer edges of the retina being used in the peripheral vision. Depending on the absorbed light wavelength, cones are called short or blue (S), middle or green (M), and long or red (L). Cones are being used to distinguish colors at a normal level of light.

From the three primaries given by cones and the intensity given by rods, the visual information is encoded as one luminance channel and two chrominance channels, one for red-green cones, and other for blue-yellow cones, respectively.

Following this theoretical approach, the proposed method computes the three color channels following the definitions of luminance (*I*), red-green (*RG*) and blue-yellow (*BY*) color opponencies used in [14]. With the trichromatic RGB pixel values *(r, g, b)*, the three color opponent channels are defined as:

$$I = \frac{r+g+b}{3} \tag{1}$$

$$RG = \frac{r-g}{\max(r,g,b)} \tag{2}$$

$$BY = \frac{b - \min(r, g)}{\max(r, g, b)} \tag{3}$$

This approach addresses two main computational problems of the model proposed by Itti in [15]: the definition of yellow and normalization with brightness. Yellow is perceived as an overlap of red and green in equal parts, so the amount of yellow contained in an RGB pixel is given by $\min(r, g)$. Only amounts of red or green exceeding this value should be counted towards red-green opponency in order to ensure independence of the *RG* and *BY* opponency channels.

At the next stage of the proposed method, it is taken into account the perceptual decomposition of visual information in multiple processing channels. It is well known that every stimulus from the visual field is separately processed and that many cells in the HVS, and mainly in the visual cortex, have been proven to be selectively sensitive to certain frequencies or orientations. The neuronal tuning mechanism present in the primary visual cortex (V1) area of the visual cortex provides decomposition of the visual information into a number of channels, this area being more sensitive to vertical lines or textures with particular spatial frequencies. To create a similar behavior, in this framework a spatial mechanism modeled by means of steerable pyramid decomposition, proposed by Simoncelli [10] is employed. The algorithm used to create the pyramid is based on recursive application of filtering and downsampling operations. The linear decomposition subdivides an image into a collection of subbands localized in both scale and orientation.

Similar approaches to compute saliency are presented in [16], where each color channel is decomposed using a steerable pyramid tuned to 6 orientations and 4 scales. For each subband, saliency is computed as proposed by Rosenholtz [17]. In [6] steerable pyramid filters are used to compute the local energy for the responses at every filter of the bank.

In this framework, the steerable pyramid bank filters is used to decompose the three color opponent channels, an image being recursively downsampled by a factor of two and filtered with band-pass filters tuned to orientations of 0, 45, 90 and 135 degrees. The four orientations are computed at three scales, obtaining an architecture which agrees with the primary visual cortex mechanisms regarding color, spatial frequency and orientation perception. In figure 2 a sample image is decomposed using this multichannel decomposition. The figure outputs the thirteen subbands corresponding to the luminance channel of a sample image.

The third stage in the proposed method is to determine which stimuli presented in the three given pyramids and for each subband are more salient, given the fact that there are biological evidences which have shown that visual cells respond to stimuli above a certain contrast. The contrast value above which the stimuli are visible is called visibility threshold and it can vary with many parameters such as spatial frequency, orientation, viewing distance, etc. Such a threshold is suggested by the Contrast Sensitivity Function (CSF), sometimes called visual acuity, which is a measure of the sensitivity to various frequencies of the visual stimuli.

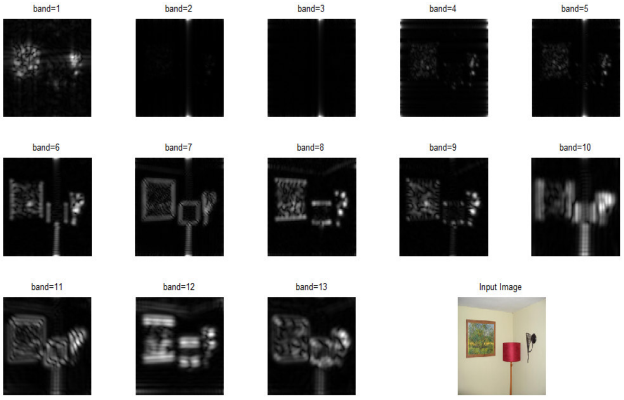

*Figure 2. Output from the steerable pyramid transform applied to the luminance channel (four orientations and three decomposition levels and the low-passed version of the input image).*

Most of the actual models of these functions were built from experimental measurements. Besides the classical luminance contrast sensitivity function, now are available chromatic contrast sensitivity functions for the two color opponent channels: RG and BY. Typically, these functions describe the envelope of the visible gratings for a given channel. Figure 3 shows the human eye sensitivity to luminance and chrominance components. For the achromatic channel, CSF shows a band-pass shape peaking around 8 cycles per degree (cpd), with sensitivity dropping either sides of the peak and a cut-off frequency of about 60 cpd, meaning that the visual system ability to resolve details is optically limited. On the other hand, chromatic CSFs have a low-pass shape, sensitivity being constant at low spatial frequencies and decreasing at medium and high spatial frequencies. The two low-pass filters corresponding to the RG and BY channels have cut-off frequencies of about 5.5 cpd, and 4.1 cpd, respectively. The shape difference between the spatial contrast sensitivity functions for chrominance and luminance gratings has been suggested to be caused by the lack of lateral inhibition in the visual system for chromatic stimuli [18].

In the proposed method, 2D anisotropic CSFs were applied as filtering operations in the frequency domain of each subband of the three luminance and chrominance pyramids. For the luminance channel a function proposed by Mannos and Sakrison [20] was used. According to their method, for each subband of the luminance pyramid, the intensity matrix $I$ is first normalized by the mean intensity value $I_m$, and afterward, the nonlinearity in perception is accounted for by taking the cube root of the normalized intensity. The FFT is then computed for the resulting matrix, the magnitude at frequencies in the horizontal and vertical directions $(u, v)$

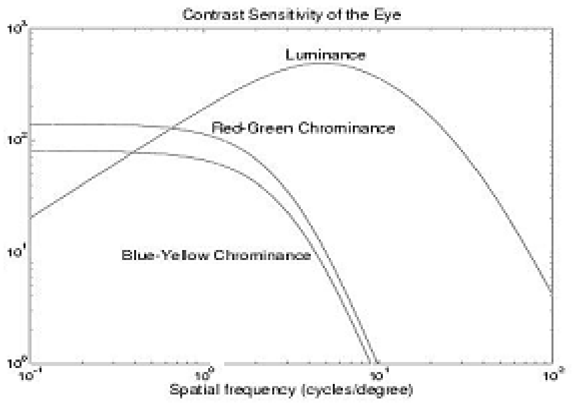

*Figure 3. Sensitivity to luminance and chrominance as a function of spatial frequency [19].*

being denoted as $f_{uv}\left(\left(\frac{I}{I_m}\right)^{.333}\right)$. Furthermore, the required spatial sensitivity $g_{uv}$ given by equation (5) is obtained by filtering $f_{uv}$ with $A_I\left(r(u,v)\right) = A_I\left(r\right)$ filter proposed in [20] (equation (4), where the radial pulsation $r = u^2 + v^2$ is expressed in cpd).

$$A_I\left(r\right) = 2.6 \times \left[0.0192 + 0.144 \times \sqrt{r}\right] \times e^{-\left(0.144 \times \sqrt{r}\right)^{1.1}} \qquad (4)$$

$$g_{uv} = f_{uv}\left(\left(\frac{I}{I_m}\right)^{.333}\right) \cdot A_I\left(r\right) \qquad (5)$$

Following the same procedure, the FFT transforms for each subband of the corresponding pyramids for the two chromatic channels were computed. The resulting amplitudes $f_{uv}\left(RG\right)$ and $f_{uv}\left(BY\right)$ were filtered using the CSFs proposed in [18]. The filtering operation results are suggested by equations (6) and (7). This time, the contrast sensitivity is computed as a function of the radial pulsation $r$ and orientation $\theta$, as it may be seen in equations (8) and (9), where the contrast sensitivity functions of RG channel and BY channel were given as they were proposed in [21].

$$g_{uv}\left(RG\right) = f_{uv}\left(RG\right) \cdot A_{RG}\left(r,\theta\right) \qquad (6)$$

$$g_{uv}\left(BY\right) = f_{uv}\left(BY\right) \cdot A_{BY}\left(r,\theta\right) \qquad (7)$$

$$A_{RG}\left(r,\theta\right) = \frac{33}{1+\left(\frac{r}{5.52}\right)^{1.72}} \cdot \left(1 - 0.27 \cdot \sin\left(2 \cdot \theta\right)\right) \qquad (8)$$

$$A_{BY}\left(r,\theta\right) = \frac{5}{1+\left(\frac{r}{4.12}\right)^{1.64}} \cdot \left(1 - 0.24 \cdot \sin\left(2 \cdot \theta\right)\right) \qquad (9)$$

To this end, the method computes three pyramids for every opponent color channel; each pyramid includes thirteen subbands which were filtered using CSF based filters. The thirteen subbands were obtained as it follows: the low-pass filtered input image was divided into four band-pass versions, each tuned to a certain orientation previously given in this section. This procedure was applied three times as a recursive process, each time, the input image consisting on the low-passed version obtain in the previous stage and down-sampled by a factor of two.

The next step of the algorithm was to determine the optimal linear weights for each subband of a pyramid. In order to do that, it was used a linear support vector machine to train and test a classifier. Each pyramid was processed with the liblinear SVM application proposed in [22]. The final model was obtained as a combination of the best results determined for the three channels.

## IV. RESULTS

Against other methods which combine biologically plausible filters to estimate saliency, this method uses a learning approach to train a linear classifier directly from the human eye tracking data created by Judd *et al.* [6]. The 1003 natural images of the data set were divided into 900 training images and 103 testing images. From each image were randomly chosen 10 positively labeled samples form the 10% most salient locations of the human ground truth saliency maps and another 10 negatively labeled samples from the 70% less salient locations. All needed settings for lunching the liblinear SVM application were the same as those proposed in [6].

For each channel, 10 training and testing trails were run, from which the best performances were chosen. To assess every result a ROC curve was used. Thus, after selecting the optimal linear combination of weights for each channel, a new testing trail was performed not only for each channel but also for the three channels which were gathered together. Tests were applied for the entire dataset, and performances were appreciated as average value of the area under the ROC curves. In figure 4, three sample images are represented. Furthermore, the sensitivity predictions for luminance and color opponent channels and also the final saliency maps, all computed using the corresponding learned model, were depicted.

In Table 1 the performances of each channel and for the proposed model (as a linear combination of the three channels) are given. It may be noticed that when the three chan-

| *Intensity Channel* | *RG Channel* | *BY Channel* | *SVM Model* | ***Itti&Koch** Saliency* | ***Torralba** Saliency* |
|---|---|---|---|---|---|
| *0.6687* | *0.7162* | *0.7003* | ***0.7653*** | *0.6112* | *0.7095* |

***Table 1***
*Area under ROC curve*

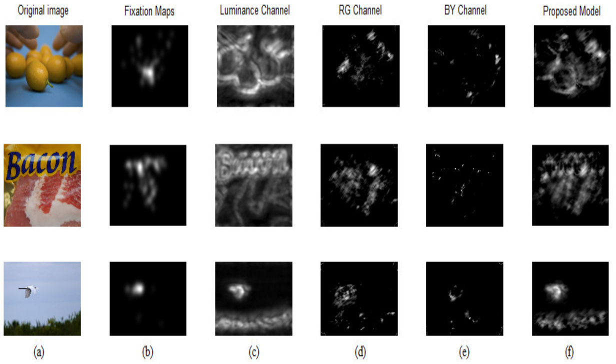

*Figure 4. Channels sensitivities predictions and final saliency map of the proposed model. (a) original images; (b) fixation maps as predicted by the human observers; (c) luminance sensitivity predictions; (d) RG sensitivity predictions; (e) BY sensitivity predictions; (f) proposed saliency map. All prediction maps were computed using the corresponding SVM model.*

nels were gathered together the results have improved. Analyzing the performances of the two chromatic channels, is straightforward that, as the chromatic sensitivity theory states, the RG chromatic channel is more sensitive than the BY one. Also, these two channels outperform the luminance sensitivity. Although this last observation has no theoretical background, it must be mentioned that in this framework the three CSF-based filters were applied to the low-frequencies of the images. In addition to this approach, the chromatic sensitivities have a low-pass allure, whereas the achromatic sensitivity a band-pass one, and hence, the low performance of the luminance channel is justified.

In order to review this framework method, the data set was tested against another two state-of-the-art algorithms proposed by Itti & Koch and Torralba, respectively. Both methods use similar theoretical or computational concepts with those applied in this framework. Itti & Koch method computes saliency following the FIT theory by combining several features, among which colour and intensity contrasts which are computed in a centre-surround fashion by applying DOG filters at different scales. Torralba also uses steerable pyramids, but computes saliency as a local propriety. Following the Rosenholtz [17] approach, the hypothesis underlying Torralba model is that locations with different proprieties from their neighboring regions are considered more informative and therefore will initially attract attention and eye movements.

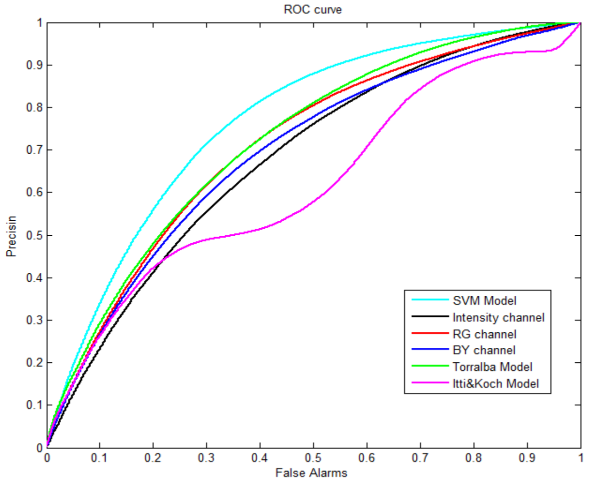

*Figure 5. The ROC curve of performances for SVMs trained on each channel individually and combined together, and for the two reference models.*

In figure 5 the ROC curves describing the performances of the proposed SVMs methods and of the two reference models averaged over the entire dataset are given. Itti & Koch model was computed using the Saliency Toolbox application proposed in [23], whereas Torralba Saliency was computed as proposed in [6] using a steerable pyramid. The pyramid subbands were found in four orientations and three scales, and saliency was computed as a distribution of the local features in the image. Resulting Saliency maps of these two methods can be seen in figure1 from Section I of this paper. In Table 1 the ROC performances for the two methods are also given. It can be noticed that even if Torralba's approach outperforms Itti & Koch Saliency estimation, none of them performs better than the method proposed in this paper.

## V. CONCLUSIONS

In this work a biologically inspired method to predict human eye fixation was proposed. The method simulates only the behavior of the primary visual cortex (V1), which is necessary for conscious vision. Concepts like color perceptions, perceptual decomposition of visual information in multiple processing channels and contrast sensitivity were the basis of this framework. Decomposing the three color opponent channels with physiologically plausible steerable pyramids, a set of features tuned to certain orientations and scales were obtained. Each of these features was then filtered using CSF-based thresholds. In order to conceive optimal results, a linear support vector machine was used to learn a model of saliency directly from human eye movement data. Thus, for each color opponent channel training and testing trails to determine the best linear combination of features were run. Furthermore, the results with the highest performance were combined and formed the SVM proposed model.

The testing results of the two chromatic channels sensitivities have shown the expected results: human eye perception is more sensitive to RG opponency than to BY opponency. Another observation is the importance of considering these two contrast sensitivities. Most of the similar models which use CSF in predicting human eye fixations, take into account only the achromatic contrast sensitivity. In this framework it was proved that when dealing with low-level, rare features like the salient ones, the luminance sensitivity is outperformed by the chromatic ones. Also, the results can be improved by adding the chromatic sensitivities to the luminance.

For accurate assessment of the proposed method, another two state-of-the-art methods were used for benchmarking. Although these two methods have certain common particularities with the SVM model, testing results have showed that the proposed method performs better in predicting human eye fixation. This result proved the fact that using human visual perception concepts in building saliency prediction models is more efficient than purely computational or combined methods. Even if the task of predicting human eye fixations still remains a complex and difficult challenge, the proposed SVM model and, especially the perceptual concepts taken into account, represent a step forward in improving the biologically inspired saliency models.

## ACKNOLEDGEMENTS

This paper was realised with the support of POSDRU CUANTUMDOC “DOCTORAL STUDIES FOR EUROPEAN PERFORMANCES IN RESEARCH AND INNOVATION” ID79407 project funded by the European Social Fund and Romanian Government.

## REFERENCES

[1] M. Corbetta, and G. L. Shulman, ”Control of Global-directed and Stimulus-driven Attenion in the Brain”, *Nature Reviews*, 3(3), 201-215, 2002.

[2] T. J. Buschman, E. K. Miller, „Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices”, 315(5820), pp. 1860-1862, 2007.

[3] A. Treisman, and G. A. Gelade, „Feature-Integration Theory of Attention”, *Cognitive Psychology*, 12, pp. 97- 136, 1980.

[4] Ch. Koch and S. Ullman, „Shifts in Selective Visual Attention: towards the underlying neural circuitry”, *Human Neurobiology*, 4(4), 219-227, 1985.

[5] L. Itti, Ch. 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.

[6] T. Judd, K. Ehinger, F. Durand, A. Torralba,” Learning to Predict Where Humans Look”, *In Proceedings of ICCV*, Kyoto, pp. 1-8, 2009.

[7] V. Gao and N. Vasconcelo, „Bottom-up saliency as a discriminat process”, *IEEE Conference on Computer Vision*, 2007.

[8] T. Liu, J. Sun, N. N. Zheng, X. Tang, H. Y. Shum, „Learning to Detect a Salient Object”, *in Proceeding of IEEE Computer Society Conference on CVPR*, February 2007.

[9] N. Bruce and J. Tsotsos, „Attention based on information maximization”, *Journal of Vision*, 7(9): 950-950, 2007.

[10] E. P. Simoncelli and W. T. Freeman, „The steerable pyramid: a flexible architecture for multi-scale derivative computation”, *2nd IEEE International Conference on Image Processing*, Washington, DC. 3, pp. 444-447, October, 1995.

[11] T. Young, “On the Theory of Light and Colors”, *Philosophical Transactions of the Royal Society of London*, 92: 20–71 (1802).

[12] H. von Helmholtz, “Handbuch der Physiologischen Optik”, *Hamburg and Leipzig*, Voss, 1867.

[13] Hering E., “Outlines of a Theory of the Light Sense”. *Harvard University Press*, Cambridge, Mass, 1964.

[14] Dirk Walther, “Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics”, PhD thesis, *California Institute of Technology*, Pasadena, CA, 23th February 2006, Available: http://resolver.caltech.edu/CaltechETD:etd-03072006-135433 [Accessed: December, 2011].

[15] L. Itti and Ch. Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention”. *Vision Research*, 40(10-12), pp. 1489-1506, 2000.

[16] A. Torralba, A. Oliva, M. S. Castelhano, J. M. Henderson, “Contextual Guidance of eye movements and attention in real-world scenes: The role of global features on object search”, *Psychological Review*, 113(4), pp. 766-786, October (2006).

[17] R. Rosenholtz, “A simple saliency model predicts a number of motion popout phenomena”, *Vision Research*, 39(1999), pp. 3157-3163.

[18] J. M. Rovamo, M. I. Kankaanpaa, H. Kukkonen, “Modeling spatial contrast sensitivity functions for chromatic and luminance-modulated gratings”, Vision *Research,* 39 (1999), pp. 2387-2398.

[19] Available: http://cnx.org/content/m11084/latest/ [Accessed: January, 2012].

[20] J. L. Mannos and D. J. Sakrison, “The Effect of a Visual Fidelity Criterion on the Encoding of Images”, *IEEE Transactions on Information Theory*, IT-20(4), July 1974.

[21] P. Le Callet, A. Saadane, D. Barba, “Interactions of Chromatic Components on the Perceptual Quantization of the Achromatic Component”, *SPIE Human Vision and Electronic Imaging*, 3644, 1999.

[22] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, “LIBLINEAR: A library for large linear classification”, *Journal of Machine Learning Research*, 9(2008), pp. 1871-1874, Available: http://www.csie.ntu.edu.tw/~cjlin/liblinear/ [Accessed: November, 2011].

[23] D. Walther and Ch. Koch, ”Modeling attention to salient proto-objects”. *Neural Networks* 19, pp. 1395-1407, (2006), Available: http://www.saliencytoolbox.net/index.html [Accessed: October, 2011].