WHICH COLORS BEST CATCH YOUR EYES: A SUBJECTIVE STUDY OF COLOR SALIENCY

Elisa Drelie Gelasca, Danko Tomasic, Touradj Ebrahimi

Ecole Polytechnique Fédérale de Lausanne
EPFL
CH-1015 Lausanne, Switzerland

ABSTRACT

To determine Regions of Interest (ROI) in a scene, perceptual saliency of regions has to be measured. When scenes are viewed with the same context and motivation, these ROIs are often highly correlated among different people. As a result, it is possible to develop a computational model of visual attention that can analyze a scene and accurately estimate the location of viewers' ROIs. Color saliency is investigated in this paper. In particular, a subjective experiment has been carried out to estimate which hues attract more human attention. The performance of the visual attention model including color saliency are assessed in the context of a segmentation evaluation application.

1 INTRODUCTION

When an observer concentrates on a particular scene, the Human Visual System does not process equally all the information available to the observer. The observer, rather, selectively attends to different aspects of the scene, at different times. Sometimes, the observer globally views the entire scene. At other times, he/she focuses on a selected object or set of objects. The observer may even concentrate on a specific part of an object or its various properties such as its color or texture. Our ability to engage in such flexible strategies for processing of visual information is generally referred to as *visual attention*. Visual attention is an important component of vision and recent experiments suggest that such an attention may be the prerequisite to consciously perceive anything at all [1]. Two different yet complementary notions are to be distinguished in visual perception, namely *capacity* and *selectivity*. Capacity is the amount of perceptual resources available for a given task or process. Attentional capacity can vary with a number of factors, such as alertness, motivation or the time of the day. For any given capacity, the amount of attention paid to different subsets of visual information can vary in a flexible manner. This ability allows attention to be selective in terms of what is processed and what is not. The latter is referred to as selectivity. Complex scenes such as those normally observed on our daily lives contain so much information that they cannot be immediately processed by human perception mechanism. As a result, humans need to sample the visual information in a series of distinct perceptual acts which are inherently selective. Voluntary eye movements are among these selective perceptual acts. But even when eyes become stationary the processing of the retinal image representing the scene is performed in a selective manner, because of the special structure of human retina and visual cortex. Visual search experiments, eye movement studies and other psychophysical and psychological tests have identified a number of factors which influence visual attention and eye movements. These factors are typically classified as either top-down (task or motivation driven) or bottom-up (stimulus driven), even though for some factors this distinction may not be so clearly defined. Bottom-up factors often have a stronger impact in the visual selection process when compared to top-down. For instance, it is difficult to not to attend to highly salient parts of a scene during a search task. Top-down instructions usually cannot override the influence of stronger bottom-up salient objects [2].

The term *salience* is used to refer to bottom-up task-independent factors [3]. *Relevance* concerns mostly the top-down volition-controlled and task-dependent behavior of the human attention. Salience is connected with observer- external objects or properties, while relevance is related to observer- internal factors such as goals and motivation.

According to various studies [4, 5, 3], *low-level* or *bottom -up factors*, such as motion, position, size, brightness, color, contrast and shape of objects as well as their orientation influence rapid and task-independent scanning of an image by a human observer.

In many previous works [4, 5, 6], authors have emphasized the importance of color as a visual attractor. Osberger [5] suggests that some particular colors (e.g. red) attract our attention more than others, or induce higher amount of masking. Color can also be used for identification of *high-level* or *top-down* factors in visual attention. For instance, face and hands can be detected by means of color-based skin detectors [7].

In this paper, we focus on the study of color as visual

attractor. The aim is that of identifying which colors attract more attention in terms of saliency. We intend to extend past observations with regard to color saliency and especially better quantify them by means of carefully designed subjective experiments.

A new visual attention model taking into account both bottom-up and top-down factors is also presented. In particular, this model makes use of the color saliency proposed in this paper. Many different applications can make use of this model. Examples include image and video quality metrics, image and video compression, image and video segmentation, search and retrieval from image and video databases and digital watermarking. The application of our model for estimation of annoyance level in segmentation is also presented and discussed.

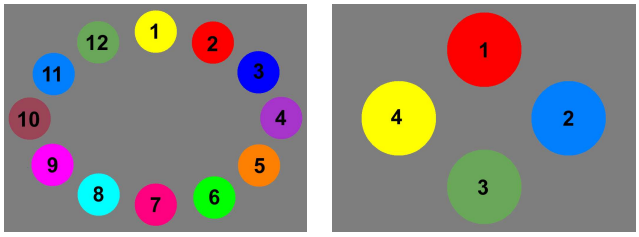
2 COLOR EXPERIMENT

In order to understand which specific colors or groups of colors influence more visual attention, a psychophysical experiment was designed. The goal of this experiment was to quantify the color saliency and to provide a ranking among some of the most common colors.

The experiment subjects consisted of 11 inexperienced persons, 3 females and 8 males, aged between 19 and 28. In order to have reliable results, ideally a high number of different colors should be considered. However, this would make the task too tiring for the subjects. A good compromise was found by selecting 12 colors. The tested colors were chosen in the CIELab color space because of the perceptual uniformity of this color space. In order to cover the whole color space in an approximately well spread manner, colors were chosen as sparse as possible in CIELab. The experiment was divided in two cycles and was carried out in a dark room.

During the first cycle, the subjects were shown 20 synthetic images containing 12 colored disks as shown in Figure 1(a). The background was gray with luminance (L) of 120 in HSL, where L varies between 0 and 240. The colored disks were disposed symmetrically along the contour of an ellipse to reduce the eventual influence of their position. The relative positions of the circles were changed during the tests to reduce an eventual influence of position in subjects' responses. In this first cycle, the task was to choose at first 3 or 4 colors which subjects considered the most salient among the displayed colors. Afterwards, the same images were shown but subjects were asked to choose only 1 or 2 colored circles which attracted most their attention. The subjects were asked to refer their first impression. Before each cycle, they were shown some examples and the first 4 images were considered as training and their results were not recorded.

The second cycle was performed with the aim to reinforce the results of the first cycle and to gather a more extensive analysis on color saliency. The subjects were shown twice 48 synthetic images containing only 4 colored disks as shown in Figure 1(b). Due to the impossibility to test all the possible combinations of colors it was decided to test 12 different combinations, each with four different dispositions of colored disks. Four different combinations of colors were tested with four different relative positions, which means that overall, each color was compared in 16 images. In the second cycle, the task was to choose at first 2 or 3 colors. Afterwards, subjects were asked to choose the most salient colored disk. In this cycle, they were also allowed to respond that all colors had the same importance.



(a) First cycle (b) Second cycle

Fig. 1. (a) First cycle: 12 colors test. Colors are sparsely chosen in CIELab color space: 1 yellow, 2 red, 3 blue, 4 violet, 5 orange, 6 green, 7 magenta, 8 cyan, 9 pink, 10 maroon, 11 light blue, 12 dark green (b) Second cycle: 4 colors test. One particular combination: 1 red, 2 light blue, 3 dark green, 4 yellow.

3 RESULTS ON COLOR

The results of the first cycle of the experiment, as presented in Figure 2 show a clear distinction in subjective importance between 12 tested colors. Especially, it was possible to divide colors in two big groups. The colors that had much more hits were red, yellow, green and pink. Those of lower saliency seemed to be light blue, maroon, violet and dark green. The results of the second cycle confirmed the results of the first cycle, as groups of colors having similar saliency were obtained. Analyzing these results, it is possible to conclude that subjects retained that light blue, violet, dark green and maroon have similar importance although we can still arrange them in order of importance. The same can be concluded for the group red, yellow, green and blue. It is very interesting to underline that in all 12 combinations of colors, the order of importance was always the same when compared to the first cycle. This is an indication that the ordering of color latency as found in the first cycle is rather robust and does not change with alternative tests as those in the second cycle of this experiment.

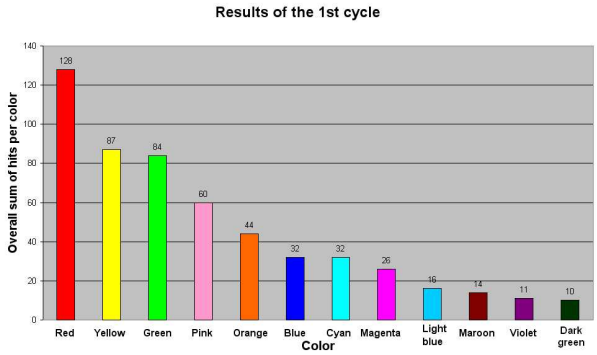


Fig. 2. Results of the first cycle: red had more hits (128).

4 VISUAL ATTENTION MODEL

The knowledge of bottom-up and top-down factors which influence visual attention provides the framework for the development of computational models for human visual attention. These models aim to detect the salient areas in a scene (i.e the areas where viewers are likely to focus their attention) in an unsupervised manner. This section presents a method that automatically predicts where the regions of interest are located in a natural scene. Various properties of human attention discussed in previous sections are used to obtain a final importance map for any scene provided as input. This model is a variant of previous works [4, 5, 6] and includes the color saliency discussed above. Figure 3 depicts the general block-diagram of the proposed model. It exploits the following bottom-up factors: color, size, position, contrast and motion. The first four factors are estimated on the regions obtained by a prior segmentation step. Motion is computed from an optical flow estimation between a current and a previous instantiation of the scene. Our model also takes advantage of a top-down factor by identifying skin regions which are assumed to correspond to face and hands regions. In the following, we will provide more details about the approach used to extract each of the above mentioned factors. The strategy used to fuse these factors based on their importance is then discussed.

Segmentation of the scene is a crucial point in this model. Regions in the segmentation should represent semantic regions either individually or collectively. The algorithm used for segmentation was an extension of watershed by Vincent and Soille [8] where the more appropriate color space CIELab was selected for a final region merging.

Regions exhibiting a high contrast in either their luminance or color exert a strong influence on visual attention. The contrast used in the proposed model is given by Eq.1, which is an extention of a similar definition in [5].

$$C_i = \frac{\sum_{j=1}^{N} (D_{ij} \cdot K_{ij})}{\sum_{j=1}^{N} K_{ij}} \quad (1)$$

$$D_{ij} = \sqrt{(L_i - L_{ij})^2 + (a_i - a_{ij})^2 + (b_i - b_{ij})^2} \quad (2)$$

$$K_{ij} = min\,(k \cdot B_{ij}, size\,(R_{i,j})) \quad (3)$$

In the above equations, $R_{i,j}$ represent regions sharing a 4-connected border with R_i for which the contrast C_i is calculated; k is a constant to limit the degree of influence of neighbors. A value of 10 was chosen for this constant in our experiments. B_{ij} is the number of pixels in $R_{i,j}$. L_i, a_i and b_i represent averaged values of CIELab color components for pixels in region R_i. L_j, a_j and b_j correspond to similar values for neighboring regions R_{ij}. The value of contrast, C_i is normalized to an interval between 0 and 255 in an adaptive manner. The importance of C_i of a region with a given contrast is reduced in presence of highly contrasted neighboring regions, and increased otherwise.

The color factor of region R_i is calculated according to the following process. The results of color experiment in Sec.2 and 3 are used to produce a weighting mechanism based on the saliency of the 12 colors selected in our tests. For each region R_i the average color coordinates in CIELab are computed. The nearest color among the 12 selected in Sec.2 is then identified using a simple euclidean distance between the average color of the region in CIELab coordinates and the 12 colors in our tests. The color factor for region R_i is then estimated as the weight of the closest color in Fig. 2 of our color experiments. These weights vary between 128 for red downto 10 for dark green (see Table 1). The value of color is then normalized to an interval between 0 and 255 in an adaptive manner.

color	**Red**	**Green**	**Blue**
weight	128	84	32
color	**Yellow**	**Magenta**	**Cyan**
weight	87	26	32
color	**Orange**	**Pink**	**Light bluen**
weight	44	60	16
color	**Violet**	**Dark green**	**Maroon**
weight	11	10	14

Table 1. Color weights

The size and position factors were estimated following the same approach as proposed by Osberger and Rohaly [5].

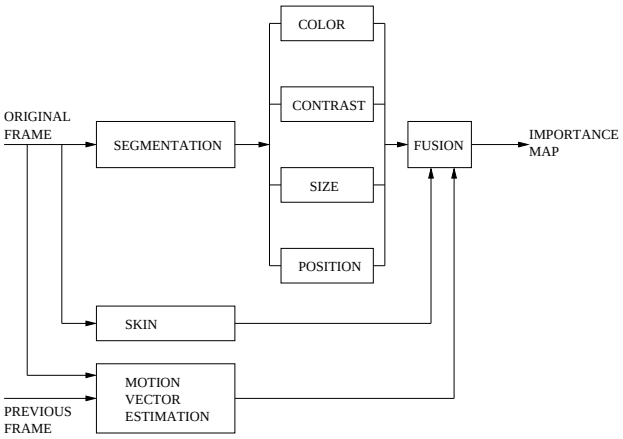


Fig. 3. Overall block diagram showing the operation of the proposed human attention model

In our model, we considered regions covering a surface between 10% and 20% of the frame to be most significant. The size factor importance reduces linearly for region of a surface between 20 % and 80 % to reach zero for sizes beyond 80 %. Likewise the importance of size factor for small regions covering up to 10 % of the frame surface increases linearly.

The Optical Flow estimation algorithm in [9] was used to compute the motion factor in our model.

The only top-down factor used in our model aims at identifying those pixels which could belong to a skin region. The skin detector approach proposed by Herodotou [7] was used for this purpose.

As mentioned above a very important step in the model is that of fusion between different visual attention factors. The following linear model was used to estimate the importance map of a scene:

$$I_{overall} = A \cdot \text{skin} + B \cdot movement + (1 - A - B) \cdot (position + size + contrast + C \cdot color) \quad (4)$$

where the weighting factors were obtained experimentally ($A = 0.35$, $B = 0.2$, $C = 1.2$). The reason behind the difference of weighting strategy for skin and movement was because their values were obtained on a per pixel basis as opposed to per region for the other visual attention factors.

5 ANNOYANCE LEVEL ESTIMATION IN SEGMENTATION

As already mentioned, various applications can benefit from a visual attention model. One such application is in the evaluation video object segmentation. Representation of a video content in terms of its constituing objects is useful

(a) The least salient region (b) The most salient region

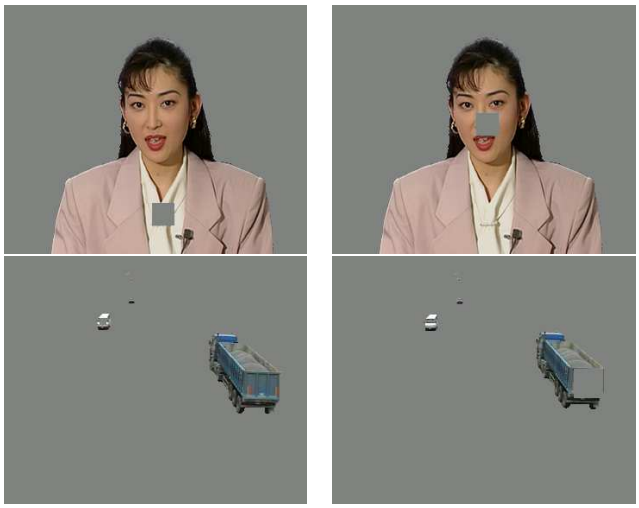


Fig. 4. Artifact is inserted in the least salient region (a) and in the most salient region (b) according to its size.

in applications such as object-based coding, video content search and retrieval, interactive video, and video surveillance. Segmentation is a key component to extract objects from a video content. In the past three decades, various video segmentation techniques have been proposed in literature. A proper assessment mechanism is needed to compare the performance of different segmentation techniques, or to tune their parameters for an optimal configuration. Some works along these lines have already been conducted [4, 10, 11]. For an efficient evaluation of segmentation results, the knowledge of the saliency of the region where an error appears can help to better correlate its impact to human perception. In addition to the saliency of the region where an error appears, the type of the artifact also impacts its perception to a human observer [10]. For instance, holes can appear in the segmentation masks due to imperfections or wrong estimation of appropriate parameters in many segmentation algorithms. To validate the visual attention model discussed in the previous section, we propose to correlate the level of annoyance introduced by a rectangular hole in different areas of a segmentation mask. The degree of annoyance of the artifact should be correlated to the visual saliency of the region where it is introduced.

The following experiments were performed to achieve this goal. Two different video sequences of 60 frames each were selected: Akiyo and Highway. Akiyo is a head and shoulder sequence of a human speaker presenting TV news. Highway contains various moving vehicules in different sizes, colors and shapes. Several versions of the above two video sequences were generated by introducing a rectangular shaped hole of varying sizes inside the most and the least salient regions according to the model of the previous section (see

Fig. 4).

Standard subjective evaluation methodologies for video segmentation quality are not yet available. We used the experimental method for subjective evaluation in [10] to evaluate the level of annoyance introduced by the artifact in the generated video sequences.

The experimental trials were performed with the complete set of test sequences presented in a random order. Our test subjects were drawn from a pool of 16 subjects (10 male, 6 female) aged between 21 and 30. The results of subjective evaluation are depicted in Fig.5.

All subjective evaluation results exhibit a higher level of perceived annoyance in those cases where the artifacts have been introduced in the region with a higher saliency according to the model proposed in this paper.

6 CONCLUSIONS

In this paper, we carried out a subjective experiment in order to find the ranking in terms of saliency among 12 typical colors. The colors were selected to spread as uniformly as possible in a CIELab color space. The subjective experiment was divided in two cycles aiming at the same objectives with two different approaches. The results obtained confirm previous results reported in literature stating that red is indeed the most salient color. This color is however closely followed by yellow and green. A visual attention model driven from a previous work in literature was then discussed taking into account bottom-up as well as top-down factors such as contrast, size, position, color, motion and skin. The color saliency in this model was estimated from the above mentioned experiment. This visual saliency model was validated in the context of annoyance level estimation in segmentation. To achieve this various video sequences were generated by applying a same amount of distortion to their most and the least salient regions. The level of annoyance of each sequence was then assessed using a subjective evaluation methodology. Results obtained confirmed that the model proposed is in concordance with the observations from subjective tests.

The work presented in this paper can be extended in various directions. From subjective evaluation viewpoint additional experiments are needed to assess the validity of the attention model for a larger pool of content and different types of artifacts. The color saliency itself can also be extended to cover more hues. An important point to verify is to quantify the impact of monitor calibration on the results of color importance ranking. Moreover, the validity of the attention model needs to be better quantified by means of more complex correlation approaches. Last but not least, the model presented in this paper can be used in applications other than annoyance level estimation in segmentation.

7 References

[1] Stephen E. Palmer, *Vision Science; Photons to Phenomenology*, The MIT Press, Cambridge, Massachussets, USA, 1999.

[2] W. Osberger, *Perceptual Vision Models for Picture Quality Assessment and Compression Applications*, Ph.D. thesis, Queensland University of Technology, Brisbane, Australia, March 1999.

[3] 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*, vol. 20, no. 11, November 1998.

[4] P. Correia and F. Pereira, "Estimation of video object's relevance," in *EUSIPCO 2000*, Tampere, Finland, September 2000.

[5] W. Osberger and A.M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *Proceedings of Human Vision and Electronic Imaging*, 2001, vol. 6, pp. 361–372.

[6] F. Birren, *Le Pouvoir de la Couleur*, Les Editions de l'Homme, 1998.

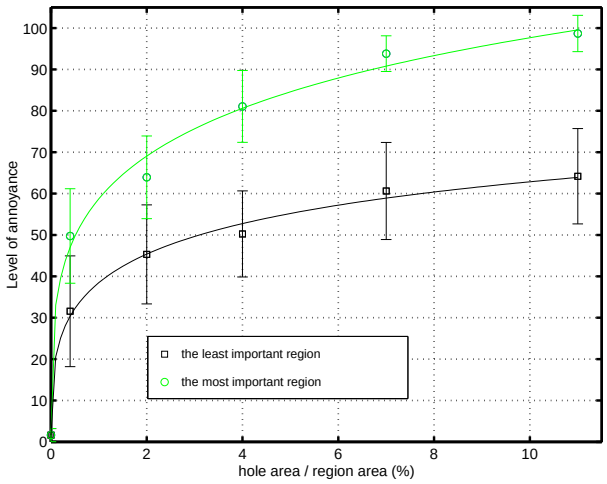
[7] N. Herodotou, K.N. Plataniotis, and A.N. Venetsanopoulos, "Automatic location and tracking of the facial region in color video sequences," *Signal Processing: Image Communication*, vol. 14, no. 5, pp. 359–388, March 1999.

[8] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–589, June 1991.

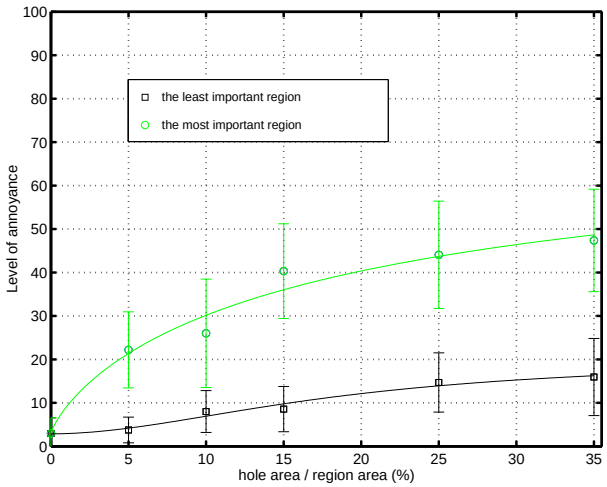
[9] J. Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker description of the algorithm," in *Intel Corporation*. Microprocessor Research Labs., OpenCV Documents 1999.

[10] E. Drelie Gelasca, T. Ebrahimi, M. Farias, M. Carli, and S. Mitra, "Annoyance of spatio-temporal artifacts in segmentation quality assessment," in *International Conference on Image Processing*. IEEE, October 2004.

[11] E. Drelie Gelasca, T. Ebrahimi, M. Farias, M. Carli, and S. Mitra, "Towards perceptually driven segmentation evaluation metrics," in *CVPR 2004 Workshop (Perceptual Organization in Computer Vision)*. IEEE, June 2004.



(a) Akiyo



(b) Highway

Fig. 5. Mean annoyance curves corresponding to different amount of segmentation artifact inserted in least and in the most salient region for the video sequence Akiyo (a) and Highway (b). The confidence intervals at 95% are plotted.