Linear vs. Nonlinear Feature Combination for Saliency Computation: A Comparison with Human Vision

Nabil Ouerhani, Alexandre Bur, and Heinz Hügli

Institute of Microtechnology, University of Neuchâtel
Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland
{nabil.ouerhani, alexandre.bur, heinz.hugli}@unine.ch

Abstract. In the heart of the computer model of visual attention, an interest or saliency map is derived from an input image in a process that encompasses several data combination steps. While several combination strategies are possible and the choice of a method influences the final saliency substantially, there is a real need for a performance comparison for the purpose of model improvement. This paper presents contributing work in which model performances are measured by comparing saliency maps with human eye fixations. Four combination methods are compared in experiments involving the viewing of 40 images by 20 observers. Similarity is evaluated qualitatively by visual tests and quantitatively by use of a similarity score. With similarity scores lying 100% higher, non-linear combinations outperform linear methods. The comparison with human vision thus shows the superiority of non-linear over linear combination schemes and speaks for their preferred use in computer models.

1 Introduction

It is generally admitted today that the human vision system makes extensive use of visual attention mechanisms in order to select a reduced set of relevant information among the huge amount of visual input gathered by the retina. By reducing the amount of data to be transferred to cortical areas responsible for higher level tasks, visual attention speeds up the vision process and contributes to its efficiency. Like in human vision, visual attention represents a fundamental mechanism for computer vision where similar speed up of the processing can be envisaged. Thus, the paradigm of computational visual attention has been widely investigated during the last two decades. Numerous computational models have been therefore reported [1], [3]. Most of them rely on the feature integration theory [4]. The saliency-based model of Koch and Ullman was first presented in [5] and gave rise to numerous software and hardware implementations [6], [7]. Further, it has been used to solve numerous issues in various fields including mobile robotics [8], [9], color image segmentation [10] and object recognition [11].

The saliency-based model of visual attention generates, for each visual cue (color, intensity, orientation, etc), a conspicuity map, i.e. a map that highlights the scene locations that differ from their surroundings according to the specific visual cue. Then, the computed maps are integrated into a unique map, the saliency map which encodes the saliency of each scene location. Depending on the scene, visual cues may

K. Franke et al. (Eds.): DAGM 2006, LNCS 4174, pp. 314 – 323, 2006.

contribute differently to the final saliency and of course, some scene locations may have higher saliency values than others. Therefore, the integration process of the conspicuity maps into the saliency map should account optimally for these two aspects.

Note that the map integration process, described here for the purpose of fusing cues, is also available at earlier steps of the computational model, namely for the integration of multi-scale maps or integration of different features. Omnipresent in the model, the competitive map integration process plays an important role and deserves careful design. The question whether the map integration process is linear or non-linear, or more precisely which of the linear or non-linear model performs better in comparison to human eye movements motivated this research.

In [12] four methods are considered for performing the competitive map integration and the methods were evaluated with respect to the capability to detect reference locations, but no comparison with eye movements is performed. Specifically, the authors propose an interesting weighting method which will be considered here. Also a so-called iterative method is proposed which performs a non-linear transform of a map. Another feature integration scheme which comprises several masking mechanisms was also proposed in [18]. Leaving by side for the moment these two advanced non-linear approaches as well as other scaling like the long-term normalization proposed in [13], the present paper compares two simple linear and two simple exponential models.

The comparison of saliency maps with human eye fixations for the purpose of model evaluation has been performed previously. In [15] the authors propose the notion of chance-adjusted saliency for measuring the similarity of eye fixations and saliency. This requires the sampling of the saliency map at the points of fixations. In [17] the authors propose the reconstruction of a human saliency map or fixation map from the fixations and perform the comparison by evaluating the correlation coefficient between fixation and saliency maps. This method was also used in [18]. In the present work, the chance adjusted saliency method is used to define a similarity score.

The remainder of this paper is organized as follows. Section 2 gives a brief description of the saliency-based model of visual attention. Section 3 defines the tools used for comparing saliency and fixations. Section 4 is devoted to the selection and definition of the four map integration methods that are then evaluated by experiments described in section 5. Finally, section 6 concludes the paper.

2 The Saliency-Based Model of Visual Attention

The saliency-based model of visual attention was proposed by Koch and Ullman in [5]. It is based on three major principles: visual attention acts on a multi-featured input; saliency of locations is influenced by the surrounding context; the saliency of locations is represented on a scalar saliency map. Several works have dealt with the realization of this model [2], [6]. Although any number of features and cues can be considered, this paper describes the model used during in order to simplify the notation. In fact, the model generates a saliency map from 3 cues namely contrast, orientation and chromaticity and the cues stem from 7 features. The different steps of the model are detailed below.

2.1 Feature Maps

First, 7 features (j=1...7) are extracted from the scene by computing the so-called feature maps from an RGB color image. The features are:

- Intensity feature: $F_1 = I = 0.3\ R + 0.59\ G + 0.11\ B$
- Two chromatic features based on the two color opponency filters red-green and blue-yellow: $F_2 = (R-G)/I$ and $F_3 = (B-Y)/I$. Note that the normalization of the opponency signals by I decouples chromaticity from intensity.
- Four local orientation features $F_4...F_7$ according to the angles $\theta \in \{0^o;45^o;90^o;135^o\}$.

2.2 Conspicuity Maps

In a second step, each feature map is transformed into its conspicuity map. The computation of the conspicuity maps relies on three main components:

- The multiscale approach is aimed at detecting conspicuous features of different sizes and consists in the representation of each feature F_j at multiple resolution levels (k=1...6), producing a set of images $F_{j,k}$
- The center-surround mechanism is used to extract local activities and consists in a difference-of-Gaussians-filter DoG which applies at each resolution level and produces the multiscale maps:

$$M_{j,k} = \left| F_{j,k} * DoG \right|. \tag{1}$$

- The map integration scheme. At this level, the multiscale maps are combined, in a competitive way, into a single *feature conspicuity map* C_j in accordance with:

$$C_j = \sum_{k=1}^{K} \mathrm{N}(M_{j,k}), \tag{2}$$

where N(.) is a normalization function that simulates both intra-map competition and inter-map competition among the different scale maps.

In the third step, using the same competitive map integration scheme as above, the seven (j=1...7) features are then grouped, according to their nature, into the three cues intensity, color and orientation. Formally, the *cue conspicuity maps* are thus:

$$C_{\mathrm{int}} = C_1; \quad C_{orient} = \sum_{j \in \{2,3,4,5\}} \mathrm{N}(C_j); \quad C_{chrom} = \sum_{j \in \{6,7\}} \mathrm{N}(C_j). \tag{3}$$

2.3 Saliency Map

In the final step of the attention model, the cue conspicuity maps are integrated, by using the scheme as above, into a saliency map S, which formally is:

$$S = \sum_{cue \in \{int, orient, chrom\}} \mathrm{N}(C_{cue}). \tag{4}$$

3 Comparing Fixations and a Saliency Map

The idea is to design a computer model which is close to human visual attention and, here, our basic assumption is that human visual attention is tightly linked to eye movements. Thus, eye movement recording is a suitable means for studying the spatial deployment of human visual attention. More specifically, while the observer watches at the given image, the K successive fixation locations of his eyes

$$\mathbf{X}^i = (\mathbf{x}_1^i, \mathbf{x}_2^i, \mathbf{x}_3^i, .., \mathbf{x}_k^i, ...\mathbf{x}_K^i) \tag{5}$$

are recorded and then compared to the computer generated saliency map.

The degree of similarity of a set of successive fixations with the saliency map is evaluated qualitatively and quantitatively. For the qualitative comparison, the fixations are transformed in a so-called fixation map which resembles the saliency map and the similarity is evaluated by comparing them visually. For the quantitative comparison, a similarity score is used.

3.1 Fixation Map

The fixation map is computed under the assumption that it is an integral of weighted point spread functions $h(\boldsymbol{x})$ located at the positions of the successive fixations. It is assumed that each fixation $\mathbf{x}_k$ gives rise to a gaussian distributed activity. The width σ of the gaussian was chosen to approximate the size of the fovea. A weighting of $h(\boldsymbol{x})$ as a function of the fixation duration or position k in the eye trajectory was not considered. Formally, the human *fixation map is*:

$$H(x) = \frac{1}{K}\sum_{k=1}^{K} h(\mathbf{x}_k) \tag{6}$$

3.2 Score

In order to compare a computational saliency map and human fixation patterns quantitatively, we compute a score s, similar to the chance-adjusted saliency used in [15]. The idea is to define the score as the difference of average saliency $\bar{s}_{fix}$ obtained when sampling the saliency map S at the fixations points with respect to the average $\bar{s}$ obtained by a random sampling of S. In addition, the score used here is normalized and thus independent of the scale of the saliency map, as argued in [16]. Formally, the score s is thus defined as:

$$s = \frac{\bar{s}_{fix} - \bar{s}}{\bar{s}} \quad \text{, with} \qquad \bar{s}_{fix} = \frac{1}{K}\sum_{k=1}^{K} S(\mathbf{x}_k) \ . \tag{7}$$

4 Four Map Integration Methods

The summation in eq. 2, 3 and 4, which is supposed to perform the competitive map integration, uses the normalization function N(.) which will now be defined.

To perform intra-map competition, and for the purpose of linear and non-linear scaling, we choose a straightforward peak to peak linear normalization and the corresponding exponential normalization as follows:

$$C' = \frac{C - C_{\min}}{C_{\max} - C_{\min}} \; ; \qquad C'' = \left(\frac{C - C_{\min}}{C_{\max} - C_{\min}} \right)^{\gamma} . \tag{8}$$

The exponential form of this transformation promotes the higher conspicuity values and demotes the lower values; it therefore tends to suppress the lesser important values forming the background.

For the purpose of inter-map competition, most of the previous works dealing with saliency-based visual attention use a competition-based scheme for map combination [6]. We adopt the same scheme in this work and attribute a weight w to each conspicuity maps for expressing its contribution. The weight is computed from the conspicuity map itself and tends to catch the global interest of that map. We consider following weight definitions:

$$w_1 = (M - \overline{m})^2 \text{ and } \qquad w_2 = \frac{C_{\max}}{\overline{C}} . \tag{9}$$

In the first weight expression w_1 which stems from [6], M is the maximum value of the normalized conspicuity map and $\overline{m}$ is the mean value of its local maxima. This weight tends to promote maps with few dissimilar peaks and to demote maps with a lot of same peaks. In the second weight expression w_2, $C_{\max}$ and $\overline{C}$ are respectively the maximum and mean values of the conspicuity map. This weight tends to promote maps with few large peaks and demote maps with a lot of similar peaks.

Considering above alternatives, we come up with the following definition of N(.)

$$\begin{aligned} \mathrm{N}_{lin\,w1}(C) &= w_1 \cdot C' & \mathrm{N}_{\exp w1}(C) &= w_1 \cdot (C')^{\gamma} \\ \mathrm{N}_{lin\,w2}(C) &= w_2 \cdot C' & \mathrm{N}_{\exp w2}(C) &= w_2 \cdot (C')^{\gamma} \end{aligned} \tag{10}$$

where C' is the peak to peak normalized conspicuity C according to eq. 8. Four map integration methods are thus defined.

5 Comparison Results

This section presents comparisons between the four map integration methods. The basic idea consists in comparing, for a given set of color images, the saliency maps produced by the four methods with human eye movement patterns recorded while subjects are looking at the same color images [14].

Eye movements were recorded with an infrared video-based tracking system (EyeLink™, SensoMotoric Instruments GmbH, Teltow/Berlin). This system consists of a headset with a pair of infrared cameras tracking the eyes, and a third camera monitoring the screen position in order to compensate for any head movements. The images were presented in blocks of 10. The images were presented in a dimly lit room on a 1900 CRT display with a resolution of 800x600, 24 bit color depth, and a refresh rate of 85 Hz. Every image was shown for 5 seconds, preceded by a center fixation display

of 1.5 seconds. Image viewing was embedded in a recognition task. For every image and each subject, the measurements yielded a sequence of fixations according to eq. 5.

The experimental image data set consists in 40 color images of various types like natural scenes, fractals, and abstract art images. Most of the images (36) were shown to 20 subjects while the remaining were viewed by 7 subjects only. As stated above, these images were presented to the subjects for a duration of 5 seconds per image, resulting in an average of 290 fixations per image. Regarding the fixation maps, they were computed according to eq. 6 using, for a given image, all fixations from all subjects. The four map integration methods were used, the value of γ is set to 2.

Figure 1 provides for image #3 a visual comparison of fixation map and the saliency maps computed by the four different methods. We note only small differences between the w_1 and w_2 alternatives, but significant differences between the linear and non-linear methods. Comparing later methods with the fixation map, we observe good similarity at the higher intensity values, but at the lower intensity values, the linear methods provide a lot of energy where there is none in the fixation map. This illustrates the advantage of the non-linear methods, which tend to keep only the highest peaks at each map integration step and accumulate thus less background signal in comparison to linear methods.

Figure 2 provides another illustration of the same comparison. Unlike previous figure where each saliency map is individually scaled to the full intensity range for best viewing purposes, here all saliency maps are scaled to the same average intensity, as this is the way a universal comparison can be performed with the fixation map. The motivation for this is the fact that all fixation maps have a constant average by construction and that they should also be compared with saliency maps with the same constant averages. This figure illustrates even better the higher similarity of the fixation map with saliencies for non-linear methods. Note that the score definition in eq. 7 reflects quantitatively the comparison illustrated here. Another example is provided in figure 3 with image #7.

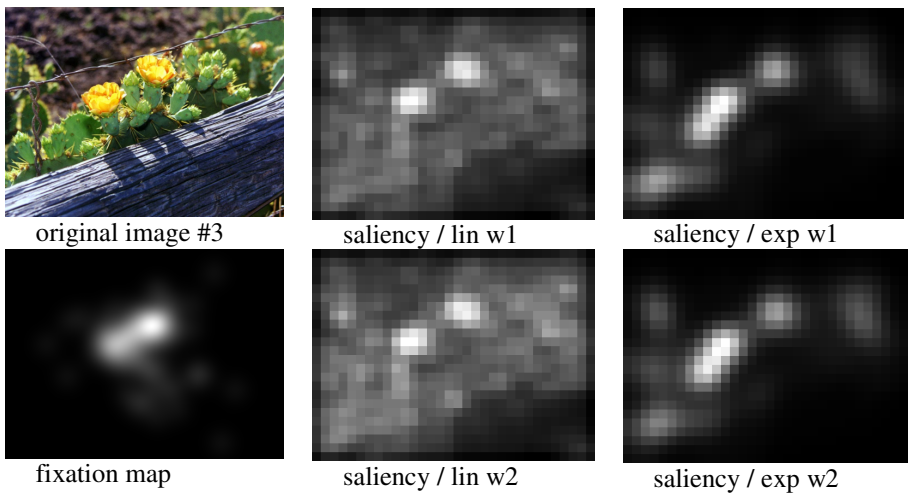

Fig. 1. For image #3, comparison of fixation map with saliency maps obtained by 4 different methods. Each saliency map is represented on the full scale image intensity.

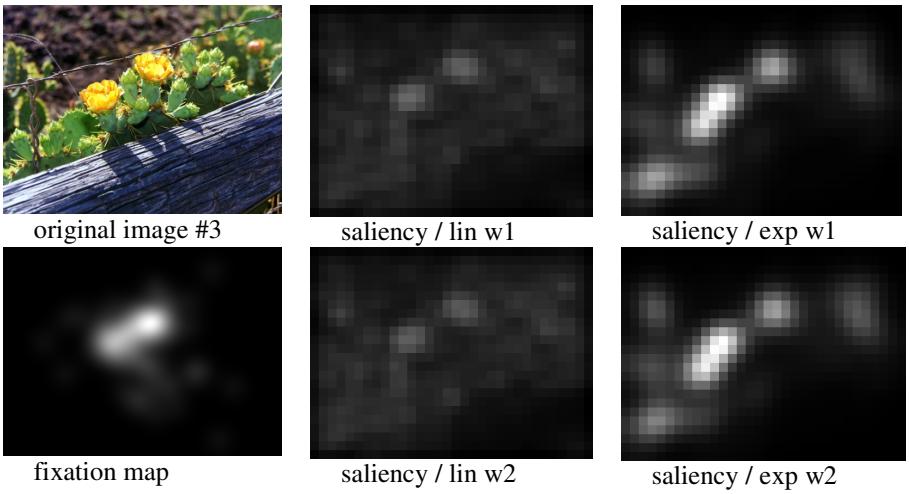

Fig. 2. For image #3, comparison of fixation map with saliency maps obtained by 4 different methods. All saliency maps are represented with the same average intensity.

The result of the quantitative comparison is given in figures 4 and 5. Figure 4 shows the average score over all subjects obtained by each method and each individual image. More precisely, the presented values reflect the average score over the first 5 fixations, but other numbers of fixations look similar. The plot illustrates the relatively large individual variations; detailed analysis shows that the non-linear methods outperform linear methods in more than 80% of the images.

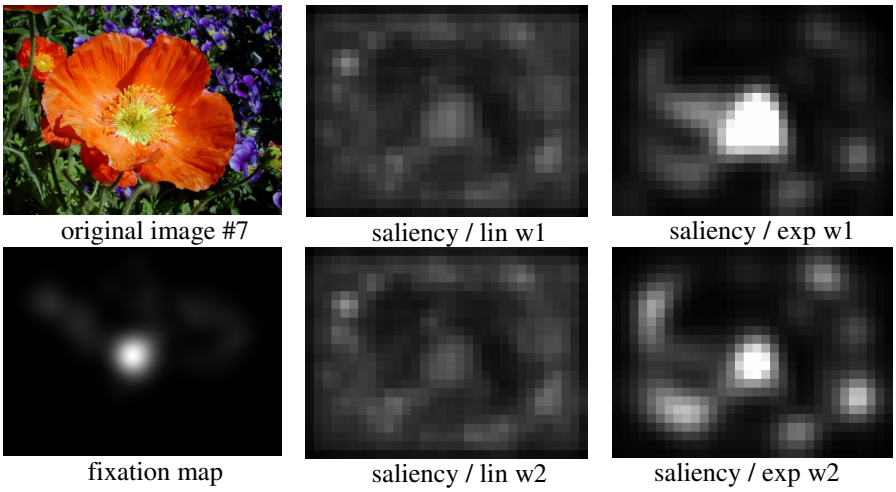

Fig. 3. For image #7, comparison of fixation map with saliency maps obtained by 4 different methods. All saliency maps are represented with the same average intensity.

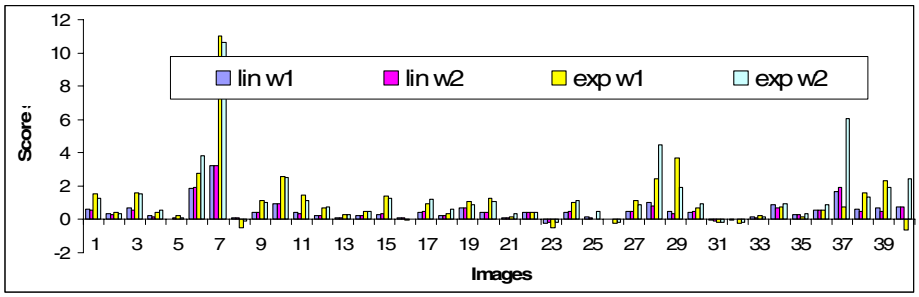

Fig. 4. Scores of the four methods for the 40 individual images

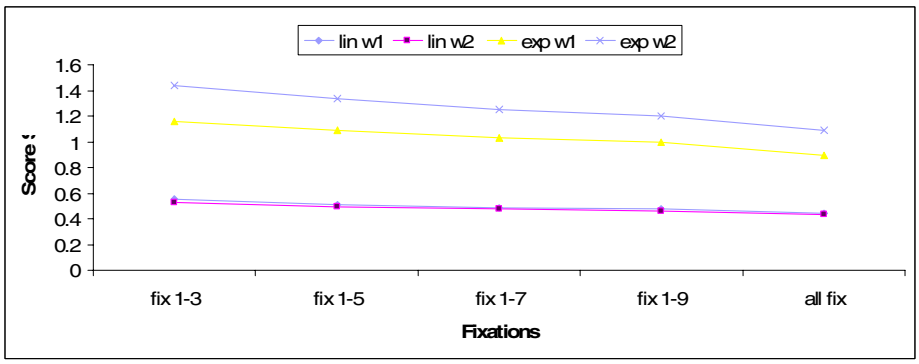

Fig. 5. Scores of the four methods for different viewing durations

Figure 5 shows the results of the comparisons of the four methods. The represented score is the average score over all subjects and all images considering a varying number of fixations. The successive values represent the first 3, 5, 7, etc fixations recorded during the viewing of a single image and illustrate the influence of viewing duration. It is noteworthy that for all cases, the model of visual attention using non-linear methods fares better in predicting where human observers foveate than the model using linear competition method. Quantitatively, the non-linear methods model yields an average score over 100% higher than the linear model. Regarding the weighting methods, w_2 performs better than w_1 with the non-linear method but both perform similarly with the linear methods. Here, differences are not very significant for a general preference of a method.

6 Conclusions

This paper presents a contribution to the design of models for visual attention computation by measuring the performance of selected methods. Performance is evaluated under the assumption that human visual attention is tightly linked to eye movements and that best similarity between the eye fixations and the saliency maps reflects also best performance. Motivated by visual comparisons of a large number of

fixation maps and corresponding saliency maps, we selected four different map integration methods and conducted a number of experiments to assess their performance. The four methods differ in their intra-map normalization scheme and inter-map weighting scheme. The normalization is either linear or exponential and there are two weighting schemes. The experiments refer to the evaluation of the collective and individual scores obtained with 40 images and from measurements of the eye movement by 20 subjects. For each image, the fixation map was visually compared to the saliency maps generated according to the different methods, and also, the relative score was computed in order to assess the performance quantitatively. The alternate weighting schemes do not differ very much in performance. The normalization methods however do, and the exponential method exhibits a score value more than twice as large as the linear method score, clearly showing the advantage of the non-linear approach. The advantage of the non-linear approach seems to be bound to the reduction of the background noise which tends to accumulate with the linear scheme. Further work is planned that will analyze this question and also consider integration schemes for evaluation.

Acknowledgments. The presented work was supported by the Swiss National Science Foundation under project number FN-108060.

References

1. S. Ahmed. VISIT: An Efficient Computational Model of Human Visual Attention. PhD thesis, University of Illinois at Urbana-Champaign, 1991.
2. R. Milanese: Detecting Salient Regions in an Image: from Biological Evidence to Computer implementation. PhD thesis, Dept. Computer Science, Univ. of Geneva, Switzerland, 1993.
3. J.K. Tsotsos: Toward a computational model of visual attention. In T. V. Papathomas, C. Chubb, A. Gorea & E. Kowler, Early vision and beyond, MIT Press, pp. 207-226, 1995.
4. A.M. Treisman and G. Gelade: A feature-integration theory of attention. Cognitive Psychology, pp. 97-136, 1980.
5. Ch. Koch and S. Ullman: Shifts in selective visual attention: Towards the under- lying neural circuitry. Human Neurobiology, Vol. 4, pp. 219-227, 1985.
6. 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 (PAMI), Vol. 20, No. 11, pp. 1254-1259, 1998.
7. N. Ouerhani and H. Hügli: Real-time visual attention on a massively parallel SIMD architecture. International Journal of Real Time Imaging, Vol. 9, No. 3, pp. 189-196, 2003.
8. J.J. Clark and N.J. Ferrier: Control of visual attention in mobile robots. IEEE Conference on Robotics and Automation, pp. 826-831, 1989.
9. N. Ouerhani, A. Bur, and H. Hügli: Visual attention-based robot self-localization. European Conference on Mobile Robotics (ECMR 2005), September 7-10, 2005, Ancona, Italy, pp. 8-13, 2005.
10. N. Ouerhani and H. Hügli: MAPS: Multiscale attention-based presegmentation of color images. 4th International Conference on Scale-Space theories in Computer Vision, Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2695, pp. 537-549, 2003.

11. D. Walther, U. Rutishauser, Ch. Koch, and P. Perona: Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Computer Vision and Image Understanding, Vol. 100 (1-2), pp. 41-63, 2005.
12. L. Itti and Ch. Koch: A comparison of feature combination strategies for saliency- based visual attention systems. SPIE Human Vision and Electronic Imaging IV (HVEI'99), Vol. 3644, pp. 373-382, 1999.
13. N. Ouerhani, T. Jost, A. Bur, and H. Hügli: Cue normalization schemes in saliency- based visual attention models. Proc. International Cognitive Vision Workshop,Graz, Austria, 2006.
14. T. Jost, N. Ouerhani, R. von Wartburg, R. Mueri, and H. Hügli: Assessing the contribution of color in visual attention. International Journal of Computer Vision and Image Understanding (CVIU), Vol. 100, pp. 107-123, 2005.
15. D. Parkhurst, K. Law, and E. Niebur: Modeling the role of salience in the allocation of overt visual attention. Vision Research, Vol. 42, No. 1, pp. 107-123, 2002.
16. L. Itti. Quantitative modeling of perceptual salience at human eye position. Visual Cognition, in press, 2005.
17. N. Ouerhani, R. von Wartburg, H. Hügli, R.M. Müri: Empirical validation of Saliency-based model of visual attention, Electronic Letters on Computer Vision and Image Analysis 3(1): 13-24, 2003.
18. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau: A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol.28, No. 5, May 2006.