Published in Proceeding. Second International Cognitive Vision Workshop (ICVW), 2006
which should be used for any reference to this work

Cue Normalization Schemes in Saliency-based Visual Attention Models

Nabil Ouerhani, Timothée Jost, Alexandre Bur, and Heinz Hügli *
Institute of Microtechnology
University of Neuchâtel
Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland
{Nabil.Ouerhani, Heinz.Hugli}@unine.ch

Abstract

Saliency-based visual attention models provide visual saliency by combining the conspicuity maps relative to various visual cues. Because the cues are of different nature, the maps to be combined show distinct dynamic ranges and a normalization scheme is therefore required. The normalization scheme used traditionally is an instantaneous peak-to-peak normalization. It appears however that this scheme performs poorly in cases where the relative contribution of the cues varies significantly, for instance when the kind of scene changes, like when the scene under study becomes unsaturated or worse, when it looses any chromaticity. To remedy this drawback, this paper proposes an alternative normalization scheme that scales each conspicuity map with respect to a long-term estimate of its maximum, a value which is learned initially from a large number of images. The advantage of the new method is first illustrated by several examples where both normalization schemes are compared. Then, the paper presents the results of an evaluation where the computed visual saliency of a set of 40 images is compared to the respective human attention as derived from the eye movements by a population of 20 subjects. The better performance of the new normalization scheme demonstrates its capability to deal with scenes of varying type, where cue contributions vary a lot. The proposed scheme seems thus preferable in any general purpose model of visual attention.

1. Introduction

Visual attention refers to the ability of a vision system to rapidly select the most salient visual information on which higher level tasks, like object recognition, can focus. It is generally accepted today that human vision relies extensively on a visual attention mechanism in order to process the huge amount of visual information gathered by the retinas, which partially explains the efficiency of our vision system.

Like in human vision, visual attention represents a fundamental tool for computer vision since the rapid selection of relevant visual information can be benefic for many computer vision applications. Thus, the paradigm of computational visual attention has been widely investigated during the last two decades. Numerous computational models have been therefore reported [1, 10, 16]. Most of them rely on the feature integration theory presented by Treisman *et al.* in [15]. The saliency-based model of Koch and Ullman was first presented in [7] and gave rise to numerous software and hardware implementations [5, 13]. Further, it has been used to solve numerous issues in various fields including mobile robotics [2, 11], color image segmentation [12] and object recognition [17].

The saliency-based model of visual attention operates on a multi-cue visual input and computes feature and conspicuity maps at various scales. In order to compute the final map of attention, the saliency map, the model has to combine maps provided by different visual cues and computed at various scales. Since the maps to be combined show distinct dynamic ranges, a normalization of these maps is necessary before integrating them together into the saliency map. Various normalization methods have been reported in previous works [4]. However, these approaches use a simple peak-to-peak normalization in order to scale the different maps to comparable dynamic ranges. Indeed, the activity of each conspicuity map is divided by its instantaneous maximum value. This instantaneous normalization scales all conspicuity maps to exactly the same dynamic range, regardless of the initial amplitudes of the conspicuity signals. It is obvious that the instantaneous normalization method systematically amplifies the importance of a priori low conspicuity signals as illustrated in Figure 1(d), which represents a major drawback of the method.

Aiming at overcoming the described drawback of instantaneous normalization, this paper reports a new normalization method that uses cue-related long-term average values

*Corresponding author: Heinz.Hugli@unine.ch

to scale the activity of the different conspicuity maps into comparable ranges. The new long-term (as opposed to instantaneous) normalization method first learns the typical maximum response of a conspicuity map related to a given cue. This step is achieved by computing the average, over a large set of training images of various types, of the maximum response of a conspicuity map: the average of maxima . During the combination process, each cue-related conspicuity map is scaled by the corresponding average of maxima. This scaling step permits to remove the intrinsic across-modality amplitude differences while preserving the relative importance of the conspicuity signal. In addition, the long-term normalization method yields a saliency map whose values give insights about the overall conspicuousness of images, which permits to attribute different significance rates for different images. Further, we validate the new normalization method by comparing the so produced saliency maps with a human map of attention as computed from eye movement patterns.

The remainder of this paper is organized as follows. Section 2 gives a brief description of the saliency-based model of visual attention. In Section 3, the long-term normalization strategy for map combination is presented. Comparison results between the long-term and the instantaneous normalization methods, taking human saliency as reference, are reported in Section 4. Finally, the conclusions of our work are stated in Section 5.

2. The saliency-based model of visual attention

The saliency-based model of visual attention was proposed by Koch and Ullman in [7]. 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 map (the saliency map). Several works have dealt with the realization of this model [10, 5]. In this work, we adopt an implementation of the model which computes a saliency map from three cues namely contrast, orientation and chromaticity [5]. The different steps of the model are detailed below.

2.1. Feature maps

First, a set of 7 of features (1..j..7) is extracted from the scene by computing the so-called feature maps from an RGB color image.

- Intensity feature $F_1 = I = 0.3R + 0.59G + 0.11B$
- 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 hue from intensity.
- Four local orientation features $F_{4..7}$ 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 which highlights the parts of the scene that strongly differ, according to the feature specificity, from their surroundings. The computation of the conspicuity maps relies on three main components:

- The center-surround mechanism, implemented with a difference-of-Gaussians-filter, $\mathcal{D}o\mathcal{G}_k$ (k is the scale of the filter), is used to extract local activities for each feature type.
- A multiscale approach permits to detect conspicuous regions of different sizes. The solution proposed in [5] is based on a multi-resolution representation of images and computes, for each feature j, a set of conspicuity maps $\mathcal{M}_{j,k}$ at different resolutions k, according to:

$$\mathcal{M}_{j,k} = |F_j * \mathcal{D}o\mathcal{G}_k| \tag{1}$$

- A normalization and summation step during which, for each feature j, the multiscale maps $\mathcal{M}_{j,k}$ are combined, in a competitive way, into a unique feature-related conspicuity map C_j in accordance with Equation 2.

$$C_j = \sum_{k=1}^{K} \mathcal{N}_1(\mathcal{M}_{j,k}) \cdot w_k \tag{2}$$

where $\mathcal{N}_1(.)$ is a normalization operator that aims at scaling the different maps $\mathcal{M}_{j,k}$ to comparable dynamic ranges as will be explained in Section 2.6. w_k is a map-intrinsic weighting factor which allows competition between the different conspicuity maps.

2.3. Cue maps

The seven (1..j..7) features described above can be grouped into three cues J_{cue} (intensity, color, orientation): $J_{int} = \{1\}$, $J_{col} = \{2, 3\}$, and $J_{orient} = \{4, 5, 6, 7\}$. Therefore, the different feature-related conspicuity maps, computed sofar, are grouped into cue-based conspicuity maps $\hat{C}_{cue}$. Each cue conpspicuity map is the combination of feature-based conspicuity maps that stem from features belonging to the same visual cue J_{cue}, according to Equation 3.

$$\hat{C}_{cue} = \sum_{j \in J_{cue}} \mathcal{N}_2(C_j) \cdot w_j \tag{3}$$

where $\mathcal{N}_2(.)$ and w_j are defined similarly to Equation 2.

2.4. Saliency map

In the third step of the attention model, the cue-related conspicuity maps $\hat{C}_{cue}$ are integrated , in a competitive manner, into a saliency map S in accordance with Equation 4.

$$S = \sum_{cue=1}^{m} \mathcal{N}_3(\hat{C}_{cue}) \cdot w_{cue} \quad (4)$$

where m is the number of the considered cues, $\mathcal{N}_3(.)$ is a normalization operator and w_{cue} is a cue map-intrinsic weighting factor which will be described below.

2.5. Competition-based map combination

Most of the previous works dealing with saliency-based visual attention use a competition-based scheme for map combination [5]. We adopt the same scheme in this work. Indeed the integration is conceived as a weighted sum of the various maps, where the weights w simulate the competition between the conspicuity maps. They are computed in a bottom-up manner from the different maps, so that conspicuity maps that contain only few peak responses are promoted, whereas maps that contain numerous comparable responses are suppressed. For the three combination steps (Equation 2..4), the weighting factor w is computed from a conspicuity map C according to Equation 5.

$$w = (M - \overline{m})^2 \quad (5)$$

where M is the maximum value of the normalized conspicuity map ($\mathcal{N}_i(\hat{C})$, with $i \in \{1, 2, 3\}$) and $\overline{m}$ the mean value of its local maxima. It is obvious that w is large for conspicuity maps where only few peaks are detected and small for maps where several peaks of comparable amplitudes are detected.

2.6. The instantaneous normalization scheme

As argued above, the model of visual attention computes various conspicuity maps from visual features of different nature and using dissimilar extraction mechanisms, which yields a set of maps that have distinct dynamic ranges. Therefore, the combination of these maps must be preceded by a normalization step, which aims at scaling the activities of the conspicuity maps into comparable ranges.

Most of the previous works dealing with saliency-based visual attention [5] normalize the set of conspicuity maps to be integrated using a peak-to-peak procedure. Indeed, each conspicuity map is divided by its instantaneous maximum value, which leads to a scaling of its activities to the range $[0..1]$. This instantaneous normalization scheme is applied to the three map combination steps of Equation 2..4. Formally, this normalization scheme is described by Equation 6.

$$\mathcal{N}_i(C) = \frac{C - C_{min}}{C_{max} - C_{min}} \quad (6)$$

where $i \in \{1, 2, 3\}$, C_{min} and C_{max} are the minimum and maximum values of the map C, respectively. Note that the peak-to-peak scaling procedure has been also used in other normalization schemes like the non-linear normalization [4].

3. The long-term normalization scheme

Despite the fact that the instantaneous normalization described above is widely used in saliency-based visual attention modeling, this normalization scheme has an undesirable drawback. Indeed, all the conspicuity maps to be integrated are scaled to exactly the same dynamic range, regardless of the relative importance of the conspicuity signals. This drawback is illustrated in Figure 1(d), where two objects of significantly different contrasts (according to the corresponding features) have been assigned the same saliency value.

In this section we describe a new map normalization scheme, the long-term normalization, that scales various maps of different nature to comparable dynamic ranges while preserving the relative importance of each conspicuity map.

3.1. Intra-cue maps have comparable dynamic ranges

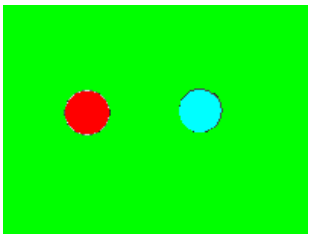
In order to combine conspicuity maps stemming from the same visual cue as computed by Equation 2 and Equation 3, no prior dynamic-range scaling is necessary.

In particular, the multiscale conspicuity maps $\mathcal{M}_{j,k}$ are scale-normalized and do not need any dynamic range modification. It has been proven in [9, 3] that difference-of-Gaussian filters closely approximates the scale-normalized Laplacian of Gaussian as studied by Lindeberg in [8].

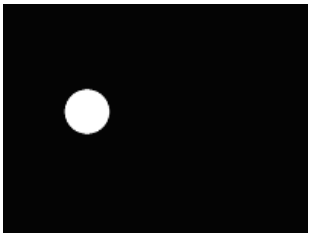
As far as for the integration of the feature-related conspicuity maps into a single cue-based conspicuity map (Equation 3), no dynamic range changes are necessary. Since feature maps of the same cue have the same dynamic ranges and since the same extraction mechanisms are used to compute the conspicuity maps related to these features, the resulting conspicuity maps have comparable dynamic ranges. Therefore, the new normalization operators $\mathcal{N}_1(.)$ and $\mathcal{N}_2(.)$ are formally described by Equation 7.

$$\mathcal{N}_1(.) = \mathcal{N}_2(.) = Id(.) \quad (7)$$

where $Id(.)$ is the identity function.

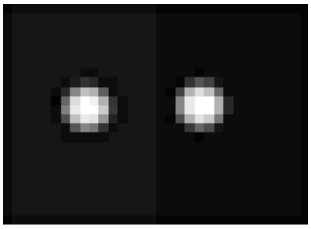


(a) Color image

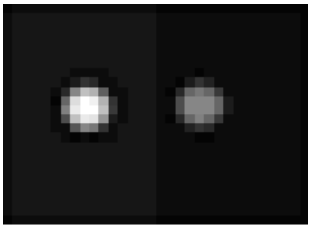


(b) (R-G) feature map

(c) (B-Y) feature map



(d) Saliency: Instantaneous



(e) Saliency: Long-term

Figure 1. Long-term normalization preserves the original saliency differences between features. (a) original color image containing two differently color-contrasted objects (discs). (b) and (c) are feature maps computed as red/green and blue/yellow opponency, respectively. (d) is the saliency map computed from both color features (R-G) and (B-Y) using the instantaneous (peak-to-peak) normalization scheme. It can be seen that both discs are assigned the same saliency values. (e) is the saliency map computed from the same features and using the long-term normalization scheme. Here the original contrast difference between the two objects is preserved in the saliency map.

Figure 1 shows an advantage of the new normalization method over the instantaneous normalization model. It can be seen that the two differently color-contrasted objects are assigned the same saliency values using the latter model, whereas the new normalization method permits to maintain the original contrast differences in the saliency map.

3.2. Long-term normalization for inter-cue map combination

Now, if we deal with conspicuity maps provided by different cues of different nature and computed by dissimilar extraction mechanisms, then we need a normalization method that accounts for the intrinsic dynamic range dissimilarity. As pointed above, the instantaneous normalization scheme uses the maximum response of each map to scale the activities of all maps to same range. Given, the drawbacks of this scheme (as argued above), a more appropriate normalization method is needed. The long-term (as opposed to instantaneous or peak-to-peak) is proposed.

The basic idea behind the long-term normalization scheme is to scale the dynamic range of each conspicuity map by a long-term average value computed for each visual cue. This long-term average can be seen as the typical response of a conspicuity map related to a cue.

In this work the long-term normalization value is computed as the average, over a large set of training images, of the maximum response of the conspicuity map: the average of maxima $\overline{M}_{cue}$. Indeed, this value is the typical maximum response of a conspicuity map stemming from a certain visual cue. Formally, $\overline{M}_{cue}$ is computed in accordance with Equation 8.

$$\overline{M}_{cue} = \frac{1}{n}\sum_{q=1}^{n} max(\hat{C}^{q}_{cue}) \tag{8}$$

where n is the number of training images, $\hat{C}^{q}_{cue}$ is a cue conspicuity map computed from image q and $max(C)$ computes the maximum value of a map C. In this work over 500 images of different types (fractals, landscape scenes, traffic scenes, art images, ...) have been used to compute $\overline{M}_{cue}$.

Thus, the long-term normalization scheme transforms a cue-related conspicuity map C_{cue} according to Equation 9.

$$\mathcal{N}_3(\hat{C}_{cue}) = \frac{\hat{C}_{cue}}{\overline{M}_{cue}} \tag{9}$$

a further and major advantage of the long-term normalization scheme is that the corresponding saliency maps quantify the overall saliency of images . Unlike, the instantaneous normalization scheme which produces saliency maps of comparable maximum values for all images, the maximum value of the saliency maps produced by the new scheme is a direct indicator of the overall saliency of each

image. Figure 2 clearly illustrates this advantage. It can be observed that, using the instantaneous normalization method, Image 1 which does not contain clearly outstanding objects produces similar saliency values to those produced by Image 2 that contains two extremely salient objects (yellow fishes). On the other hand, the long-term normalization scheme produces nearly flat (zero) saliency map for the first image and remarkably high saliency values for Image 2, reflecting, thus, the overall-saliency differences between the two images. It is noteworthy that the long-term normalization method is also applicable to other conspicuity maps combination method like the non-linear combination presented in [4].

4. Comparison results

This section presents comparison results between the long-term and the instantaneous normalization schemes, taking as criteria their plausibility with human visual attention. The basic idea consists in comparing the saliency maps produced by the two models from color images with human eye movement patterns recorded while subjects are looking at the same color images [6]. Our assumption is that human visual attention is tightly linked to eye movements. Thus, eye movement recording is a suitable means for studying the temporal and spatial deployment of human visual attention in most situations.

Eye movements were recorded with a infrared video-based tracking system (EyeLinkTM, 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 $19''$ CRT display with a resolution of 800×600, 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 i, the measurements yielded an eye trajectory T^i composed of the coordinates of the successive fixations f_k, expressed as image coordinates (x_k, y_k):

$$T^i = (f_1^i, f_2^i, f_3^i, ...) \tag{10}$$

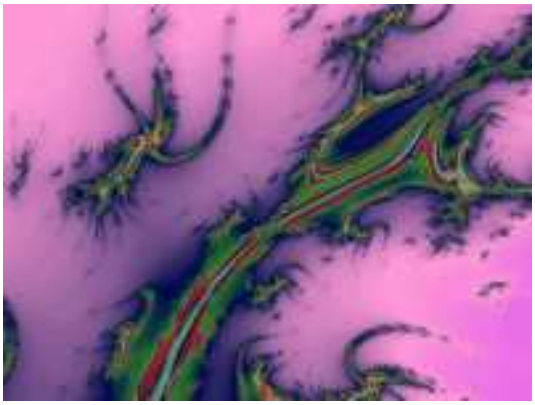
In order to quantitatively compare a computational saliency map and human fixation patterns, we compute a score s, also known as chance-adjusted saliency in [14]. Formally, this score is defined as $s = \overline{s}_{fix} - \overline{s}_{ran}$ and corresponds to the difference of average values of two sets of samples from the computer saliency map $S(x)$; $\overline{s}_{fix}$ refers to the set of N samples taken at the recorded human fixation locations, while $\overline{s}_{ran}$ refers to N random samples.

The experimental image data set consisted in 41 color images of various types like natural scenes, fractals, and abstract art images. Note that this test image set is different from the training image set used in Section 3.2. Most of the images (36) were shown to 20 subjects; the remaining 5 were viewed by 7 subjects. As stated above, these images were presented to the subjects for 5 seconds apiece, resulting in an average of 290 fixations per image.

Figure 3 shows the results of the comparisons of the two normalization schemes with human fixations. In this figure, we represent the mean score (over all images and all subjects) for each normalization scheme and taking different numbers of fixations into account: the first fixation of each subject, the three first fixations, the first five fixations, and all fixations. It is noteworthy that for all cases, the model of visual attention using long-term normalization fares batter in predicting where human observers foveate than the model using instantaneous normalization. More precisely, the long-term normalization model yields an average score over 22% higher than the instantaneous normalization model.

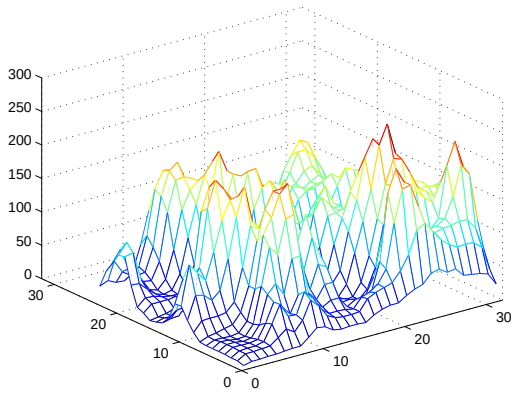
5. Conclusions

This paper reports a new normalization scheme for feature integration in order to compute a saliency map from a multi-cue input: the long-term normalization. During a training step, the new normalization scheme learns the typical maximum response (the average of maxima) of conspicuity maps related to a given visual cue. Before combining various conspicuity maps provided by different visual cues, the long-term normalization scheme scales the activity of each map by the corresponding long-term average of maxima. This scaling yields a set of conspicuity maps with comparable dynamic ranges, while preserving the relative importance of each map. The use of over 500 images of different natures to learn the long-term average of maxima reinforces the universality of this scaling factor. In addition, the long-term normalization method, unlike instantaneous normalization schemes, provides saliency values that quantify the overall saliency of different images. Further, comparisons between computational saliency maps and human eye movement patterns show that the long-term normalization fares better in predicting the human visual attention than the instantaneous normalization over a large set of images of various types. Therefore, the proposed normalization scheme seems very suitable for general purpose saliency-based models of attention that apply to images of different nature.

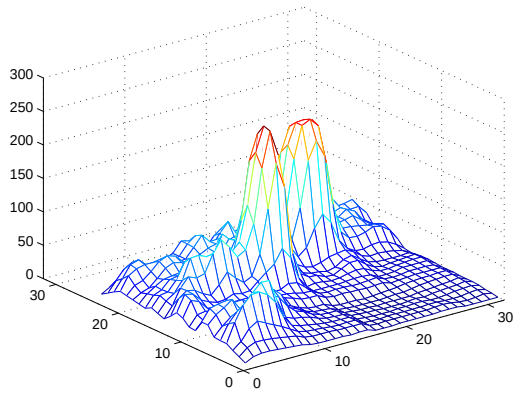


(a) Image 1

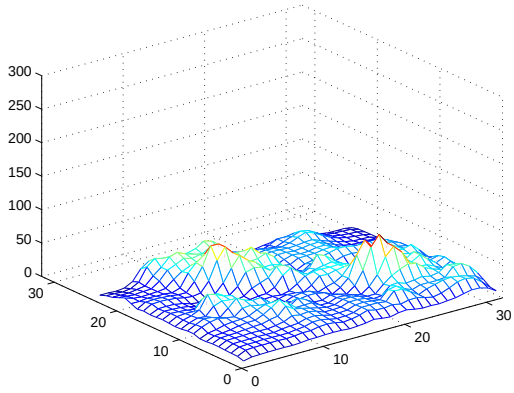
(b) Image 2



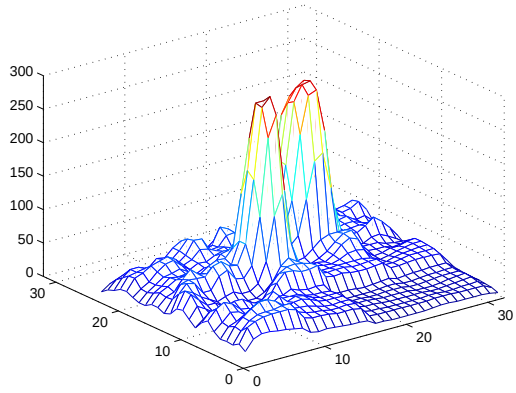
(c) Saliency map 1 (instantaneous)



(d) Saliency map 2 (instantaneous)



(e) Saliency map 1 (long-term)



(f) Saliency map 2 (long-term)

Figure 2. Advantage of the long-term normalization method. Instantaneous normalization method produces comparable saliency values for both images ((c) and (d)), whereas, the long-term normalization produces significantly higher saliency values for the Image 2 than for Image 1.

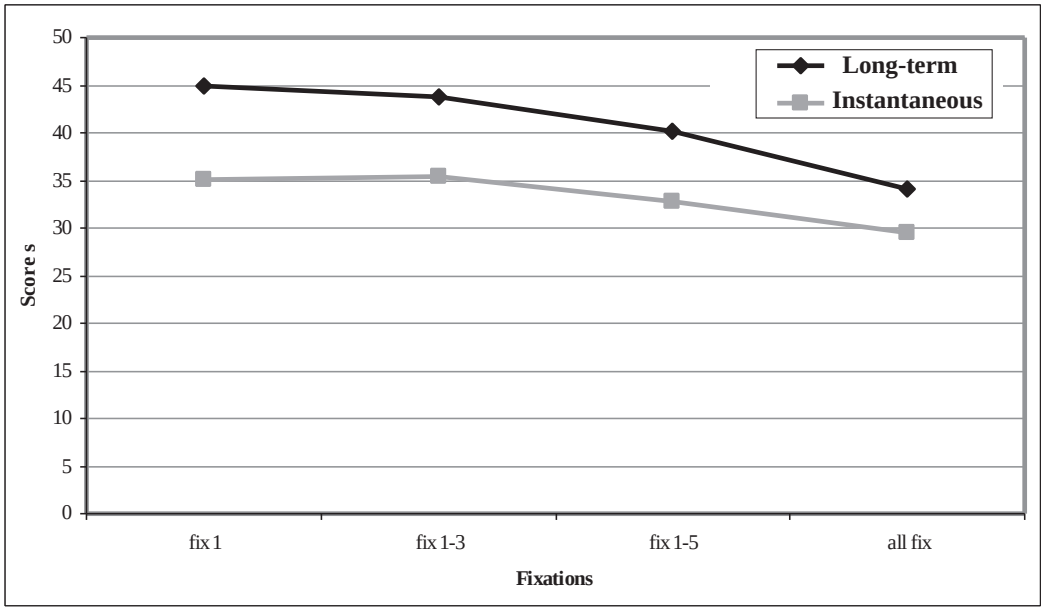


Figure 3. Long-term versus instantaneous normalization: comparison with human visual attention

References

[1] S. Ahmed. *VISIT: An Efficient Computational Model of Human Visual Attention*. PhD thesis, University of Illinois at Urbana-Champaign, 1991.

[2] J. Clark and N. Ferrier. Control of visual attention in mobile robots. *IEEE Conference on Robotics and Automation, pp. 826-831*, 1989.

[3] J. Crowley and O. Riff. Fast computation of scale normalized gaussian receptive fields. *Scale Space 03, Springer Verlag, Lecture Notes in Computer science (LNCS 2695), pp. 584-598*, 2003.

[4] L. Itti and C. 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.

[5] 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 (PAMI), Vol. 20, No. 11, pp. 1254-1259*, 1998.

[6] T. Jost, N. Ouerhani, R. von Wartburg, R. Mueri, and H. Hugli. Assessing the contribution of color in visual attention. *International Journal of Computer Vision and Image Understanding (CVIU), Vol. 100, pp. 107-123*, 2005.

[7] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, Vol. 4, pp. 219-227*, 1985.

[8] T. Lineberg. Scale-space theory: A baic tool for analyzing structures at different scales. *Journal of Applied Statistics, Vol. 21 (2), pp. 224-270*, 1994.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, Vol. 60, (2), pp. 91-110*, 2004.

[10] R. Milanese. *Detecting Salient Regions in an Image: from Biological Evidence to Computer implementation*. PhD thesis, Dept. of Computer Science, University of Geneva, Switzerland, 1993.

[11] N. Ouerhani, A. Bur, and H. Hugli. Visual attention-based robot self-localization. *European Conference on Mobile Robotics (ECMR 2005), September 7-10, 2005, Ancona, Italy, pp. 8-13*, 2005.

[12] N. Ouerhani and H. Hugli. 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.

[13] N. Ouerhani and H. Hugli. Real-time visual attention on a massively parallel SIMD architecture. *International Journal of Real Time Imaging, Vol. 9, No. 3, pp. 189-196*, 2003.

[14] 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.

[15] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology, pp. 97-136*, 1980.

[16] J. Tsotsos. Toward 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.

[17] D. Walther, U. Rutishauser, C. 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.