Stochastic Attentional Selection and Shift on the Visual Attention Pyramid

Masayasu Atsumi

Dept. of Information Systems Science, Faculty of Eng., Soka University,
1-236 Tangi-cho, Hachioji-shi, Tokyo 192-8577, Japan
matsumi@t.soka.ac.jp

Abstract. This paper proposes a computational model of visual attention which performs stochastic attentional selection and shift on the visual attention pyramid that is computed for each frame of a video sequence. In this model, the visual attention pyramid is generated according to the rareness criteria by using intensity contrast, saturation contrast, hue contrast, orientation and motion energy on a Gaussian resolution pyramid. On this attention pyramid, stochastic attentional selection and shift is performed on mechanisms of the dynamic maintenance of IOR(Inhibition Of Return), the bottom-up spatial modulation of attention and the adaptive competitive filtering of attention. Experimental results show that this model achieves stochastic visual pop-out to artificial pop-out targets and stochastic attentional selection and shift, especially the whole-part attention shift and the motion-follow attention, in daily scenes.

1 Introduction

Visual attention selects spatially circumscribed regions from dynamic scenes that contain relevant objects and events for high-level scene recognition. This mechanism is indispensable for not only human but also robots and computer vision systems to act effectively in a real world under constraints of temporal and computational resources. Visual attention is controlled by both stimulus-driven bottom-up factors and volition-driven top-down factors [1]. The former attention is rapid and task-independent but the latter attention is slow and task-dependent. There have been a lot of psychological and computational models of visual attention, in which the feature integration theory of Treisman [2], the guided search model of Wolfe [3] and the saliency map model of Koch & Ullman [4, 5] are well-known and have a great influence on later studies.

In this paper, we propose a computational model of visual attention which performs stochastic attentional selection and shift on the visual attention pyramid that is computed for each frame of a video sequence. The characteristics of our model are outlined as follows. First, the model computes a multi-level saliency map, which we call the visual attention pyramid, according to a rareness criteria [6] by using intensity contrast, saturation contrast, hue contrast, orientation and motion energy on a Gaussian resolution pyramid. Second, stochastic

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

attentional selection and shift on the visual attention pyramid is performed on mechanisms of the dynamic maintenance of IOR(Inhibition Of Return) [7], the bottom-up spatial modulation of attention and the adaptive competitive filtering of attention, which enable the model to control convergence and divergence of attention and the whole-part shift of attention. Performance of the model is evaluated for stochastic visual pop-out to artificial pop-out targets and stochastic attentional selection and shift, especially the whole-part attention shift and the motion-follow attention, in daily scenes.

This paper is organized as follows. Section 2 describes an organization of the visual attention pyramid and section 3 presents mechanisms of stochastic attentional selection and shift. Experimental results are shown in section 4 and we conclude our work in section 5.

2 Visual Attention Pyramid

2.1 Visual Features on a Gaussian Resolution Pyramid

In the human vision system, it is known that there exist parallel perceptual channels for different spatial frequencies and temporal frequencies. In computer processing of images, these channels can be represented on a Gaussian resolution pyramid [8, 9]. In our model, a Gaussian resolution pyramid of intensity, saturation and hue is firstly obtained for each image of a video sequence as a basis for computing early visual features. Then for each point of each level of the Gaussian resolution pyramid, intensity contrast, saturation contrast and hue contrast and also orientation of intensity, saturation and hue are computed as static visual features and motion energy [10] is computed as a dynamic visual feature for a sequence of intensity images.

Intensity contrast C_I and saturation contrast C_S are respectively computed by convolving intensity and saturation with a LoG(Laplacian of a Gaussian) kernel at each point. Since a hue value $h(h \in [0, 2\pi))$ represents a spectrum of color on a continuous color circle, hue contrast C_H is obtained as follows. Let $P(p)$ be a set of points around a point p which is used for convolution with a LoG kernel. First, for each point $p' \in P(p)$, the difference $d(p', p) = |h(p) - h(p')|$ is calculated, where $h(p)$ and $h(p')$ are hue of the point p and p' respectively. Next, a relative hue value $h'(p')$ of each point $p' \in P(p)$ to a hue value of the point p is computed by

$$h'(p') = \begin{cases} \frac{\pi - d(p',p)}{\pi} \cdots d(p', p) \leq \pi \\ \frac{d(p',p) - \pi}{\pi} \cdots d(p', p) > \pi \ . \end{cases} \quad (1)$$

Then, hue contrast is obtained by convolving these relative hue values with a LoG kernel. However, if hue of a point p is indefinite, hue contrast of the point p is set to 0. Also, if hue of a point $p' \in P(p)$ is indefinite, $d(p', p)$ is supposed to be 0.

Intensity orientation O_I, saturation orientation O_S and hue orientation O_H of 0°, 45°, 90° and 135° are respectively computed by convolving intensity, saturation and hue with a Gabor kernel of 0°, 45°, 90° and 135° at each point,

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

where convolution for hue orientation is performed by using relative hue values in a similar way as hue contrast. The maximum value of intensity, saturation and hue orientation defines a value of orientation O for each point.

For a temporal sequence of Gaussian resolution pyramids, motion energy is computed by convolving intensity with a spatio-temporal Gabor kernel at each point of each level of them [9–11]. In our model, full motion energies of 8 directions of 0°, 45°, 90°, 135°, 180°, 225°, 275° and 315° and 2 speeds of low and high are computed as a dynamic visual feature.

2.2 Multi-level Saliency Map

The multi-level saliency map SP is obtained by using intensity contrast, saturation contrast, hue contrast, orientation and full motion energy of each point of each level of a Gaussian resolution pyramid. We call this map and also a modulated map of it by spatial attention and competitive filtering the visual attention pyramid. Each feature is composed of plural dimensions to represent conspicuity competitively. Feature of intensity contrast is composed of two dimensions of on-center off-surround response and on-surround off-center response. Feature of saturation is also composed of two dimensions of on-center off-surround response and on-surround off-center response. Feature of hue contrast is composed of six dimensions which correspond to red, yellow, green, cyan, blue and magenta. Feature of orientation is composed of four dimensions which correspond to orientation of 0°, 45°, 90° and 135°. Feature of motion energy is composed of sixteen dimensions for eight directions and two speeds. Conspicuity of feature is obtained according to the rareness criteria among dimensions which ensures that the fewer a dimension of feature appears in an image the more conspicuous the region of the dimension is in the image.

On-center off-surround response and on-surround off-center response of intensity contrast are respectively represented by the positive and the negative signs of LoG convolution. Strength of contrast is represented by the absolute value of LoG convolution. For each level, let maximum strength of on-center off-surround contrast and on-surround off-center contrast be C^*_{Ion} and C^*_{Ioff} respectively. By defining the threshold rate of rareness $\theta_R (0 \leq \theta_R \leq 1)$, rareness of on-center off-surround contrast and on-surround off-center contrast are given by $r(n_{Ion})$ and $r(n_{Ioff})$ for a monotonic decreasing function $r(n)$ and the number n_{Ion} and n_{Ioff} of points in the level whose strength of contrast are not less than $\theta_R \times C^*_{Ion}$ and $\theta_R \times C^*_{Ioff}$ respectively. We call $r(n)$ the rareness function and use $r(n) = 1/\sqrt{n}$. Then, intensity conspicuity $S_{IC}(p)$ is obtained for strength of on-center off-surround contrast $C_{Ion}(p)$ and strength of on-surround off-center contrast $C_{Ioff}(p)$ of each point p by

$$S_{IC}(p) = \frac{(r(n_{Ion}) \times C_{Ion}(p) + r(n_{Ioff}) \times C_{Ioff}(p)) \times max(C^*_{Ion}, C^*_{Ioff})}{C^*_{IR}} \quad (2)$$

where C^*_{IR} is the maximum value of rareness-weighted intensity contrast, that is given in the first term of the numerator, for all points in the level. Intensity conspicuity is normalized to take a maximum value of $max(C^*_{Ion}, C^*_{Ioff})$.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Saturation conspicuity S_{SC} is obtained in the same way as intensity conspicuity based on on-center off-surround contrast and on-surround off-center contrast of saturation. Orientation conspicuity S_{OR} and motion energy conspicuity S_{ME} are computed in the similar way for four dimensions of orientation and sixteen dimensions of motion energy respectively.

Hue conspicuity S_{HC} is obtained from hue and hue contrast in the same idea but somewhat complicated way. Firstly, for each dimensional hue of red($R_e = 0 or 2\pi$), yellow($Y_e = \pi/3$), green($G_r = 2\pi/3$), cyan($C_y = \pi$), blue($B_l = 4\pi/3$) and magenta($M_a = 5\pi/3$), maximum hue contrast C^*_{Re}, C^*_{Ye}, C^*_{Gr}, C^*_{Cy}, C^*_{Bl} and C^*_{Ma} of the dimension is calculated as follows. Let $A(h^*)(h^* = R_e, Y_e, G_r, C_y, B_l, M_a)$ be a set of points whose hue $h(p)$ satisfies $|h(p) - h^*| \leq \pi/6$. Then $C^*_{h^*}$ is given as the maximum hue contrast among any hue contrast of points in $A(h^*)$ for $h^* = R_e, Y_e, G_r, C_y, B_l, M_a$. Next, rareness $r(n_{h^*})$ of each dimension $h^*(h^* = R_e, Y_e, G_r, C_y, B_l, M_a)$ is computed by applying the rareness function $r(n)$ to the number n_{h^*} of points in $A(h^*)$ whose hue contrast are not less than $\theta_R \times C^*_{h^*}$, where θ_R is the threshold rate of rareness. By using these rareness values, hue contrast $h(p)$ for each point p is weighted in the following way. If the point p has one of dimensional hue $h^*(h^* = R_e, Y_e, G_r, C_y, B_l, M_a)$, weighted hue contrast is given as $C^R_H(p) = r(n_{h^*}) \times C_H(p)$. Otherwise, weighted hue contrast $C^R_H(p)$ of the point p is obtained by using rareness values $r(n_{h_1})$ and $r(n_{h_2})$ of two neighbor dimensional hue h_1 and h_2 on the color circle and hue difference d_{p,h_1} and d_{p,h_2} between hue of the point p and these two neighbor dimensional hue h_1 and h_2 as

$$C^R_H(p) = \begin{cases} r(n_{h_1}) \times C_H(p) & \cdots d_{p,h_1} < d_{p,h_2} \\ r(n_{h_2}) \times C_H(p) & \cdots d_{p,h_1} > d_{p,h_2} \\ \frac{r(n_{h_1}) + r(n_{h_2})}{2} \times C_H(p) & \cdots d_{p,h_1} = d_{p,h_2} \ . \end{cases} \tag{3}$$

However, in case hue of a point p is indefinite, $C^R_H(p) = 0$. Hue conspicuity $S_{HC}(p)$ is obtained by $S_{HC}(p) = C^R_H(p) \times (C^*_H / C^*_{HR})$, where $C^*_H = max(C^*_{R_e}, C^*_{Y_e}, C^*_{G_r}, C^*_{C_y}, C^*_{B_l}, C^*_{M_a})$ and C^*_{HR} is the maximum value of rareness-weighted hue contrast.

The static saliency $S_S(p)$ is obtained for intensity conspicuity $S_{IC}(p)$, saturation conspicuity $S_{SC}(p)$, hue conspicuity $S_{HC}(p)$ and orientation conspicuity $S_{OR}(p)$ by

$$S_S(p) = \sum_{i=IC,SC,HC,OR} (w_i \times S_i(p)) \tag{4}$$

where $w_i (i = IC, SC, HC, OR)$, whose sum is 1, are weights for combination. The dynamic saliency $S_D(p)$ is obtained only from motion energy conspicuity as $S_D(p) = S_{ME}(p)$. Finally, the saliency $S(p)$ is obtained as weighted sum of the static saliency and the dynamic saliency by

$$S(p) = w_S \times S_S(p) + w_D \times S_D(p) \tag{5}$$

where w_S and w_D, sum of which is 1, are weights for composition. The multi-level saliency map is generated by computing this saliency value at every point of every level of a Gaussian resolution pyramid.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

3 Stochastic Attentional Selection and Shift

3.1 Attention Probability Space

In our model, an attention probability space is introduced on the visual attention pyramid. The attention probability on the visual attention pyramid is adaptively determined based on modulating saliency on the multi-level saliency map with attentional IOR, spatial attention and global competitive inhibition.

The attention probability space on the visual attention pyramid is defined as follows. First of all, it is assumed that attention to each point at each level of the visual attention pyramid, which we call an attention event, is mutually independent. That is, the model can discriminatively direct attention to two points in different levels one of which spatially includes another. Let $A(n, x, y)$ and $a(n, x, y)$ be an attention event and an attention value to a point $p(n, x, y)$ at a level $n(n = 0, \ldots, N-1$, where the original image level is 0) of the visual attention pyramid. Then, the attention probability $prob(n, x, y)$ of $A(n, x, y)$ is defined by

$$prob(n, x, y) = \frac{a(n, x, y)}{\sum_{n=0}^{N-1} \sum_{x,y} a(n, x, y)} \ . \tag{6}$$

This probability satisfies following conditions:

1. for each attention event A, $0 \leq prob(A) \leq 1$,
2. for the whole attention event Ω, $prob(\Omega) = 1$,
3. for mutually exclusive attention events $A_i(i = 1, 2, \ldots)$, $prob(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} prob(A_i)$.

By the way, since IOR inhibits attention from returning to the previously attended location, attention is set to 0 for IOR-imposed points on the visual attention pyramid. Moreover, in our model, it is assumed that saliency probabilistically receives spatial modulation around a focus of attention immediately before and also is filtered by global competitive inhibition. We call the former the bottom-up spatial attention and the latter the competitive filtering.

Let $p(l, x_l, y_l)$ be a center (x_l, y_l) of spatial attention at a level l and σ_l be an extent of the spatial attention at the level l. Then, the bottom-up spatial attention $\beta_n(x, y)$ at each level $n(n = 0, \ldots, N-1)$ is defined by a Gaussian function as

$$\beta_n(x, y) = \lambda^n exp\left(-\frac{(x - x_n)^2 + (y - y_n)^2}{2 \times (\frac{\sigma_l}{2^{n-l}})^2}\right) \tag{7}$$

where $p(n, x_n, y_n)$ is a center (x_n, y_n) of spatial attention at a level n which corresponds to $p(l, x_l, y_l)$ and λ is the level multiplier of spatial attention. In case $\lambda > 1$, it represents that attention has tendency to be paid to the whole rather than the part. Saliency $S_n(x, y)$ of a level n is modulated by the bottom-up spatial attention as $\beta_n(x, y) * S_n(x, y)$, where $*$ represents the pointwise product of $\beta_n(x, y)$ and $S_n(x, y)$.

The competitive filtering is performed for the visual attention pyramid constrained by IOR and the bottom-up spatial attention. Let γ be the competitive

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

filtering rate which is specified as a rate for the maximum value a^* of attention in all levels of the pyramid. Then attention of any point whose attention value is less than $\gamma \times a^*$ is set to 0 by the competitive filtering. The competitive filter is denoted by $\psi(\gamma)$.

Consequently, the attention probability space is expressed by

$$APS := (AP(SP, Q_{IOR}, \beta(l, x_l, y_l, \sigma_l), \psi(\gamma)), prob) \tag{8}$$

where $AP(SP, Q_{IOR}, \beta(l, x_l, y_l, \sigma_l), \psi(\gamma))$ is a visual attention pyramid in which a multi-level saliency map SP is modulated by IOR, the bottom-up spatial attention β and the competitive filter ψ, Q_{IOR} is a set(queue) of IOR-imposed points and $prob$ is attention probability defined by the expression (6) on AP.

3.2 Mechanism of Stochastic Attentional Selection and Shift

In attention for a continuous scene, it is necessary to perform attentional selection and shift on a sequence of image frames over time. In our model, stochastic attentional selection and shift is performed for a sequence of visual attention pyramids through mechanisms of the dynamic maintenance of IOR, the bottom-up spatial attention probabilistically imposed according to the share of attention immediately before, and the adaptive competitive filtering of attention.

In the dynamic maintenance of IOR, IOR continues over a certain duration of frames so long as saliency of IOR-imposed points does not change, but it is released when more than some degree of saliency change occurs at the points. That is, for each point p in Q_{IOR}, let $S^b(p)$ and $S^c(p)$ be saliency at the time when it added to Q_{IOR} and the current time respectively. Then if the difference $|S^c(p) - S^b(p)|$ is not less than a threshold δ, that is called the IOR-release threshold, the point is removed from Q_{IOR}.

The bottom-up spatial attention is probabilistically imposed according to the share of attention immediately before. Let $a(l, x, y)$ be an attention value of a point $p(l, x, y)$ in a level l which is selected as a focus of attention. Then, the share of attention η is given as $\eta = a(l, x, y)/a(l)$ for the sum of attention $a(l)$ of the level l. According to the share of attention, the probability with which bottom-up spatial attention acts on is obtained as

$$p = \frac{1}{1 + exp\left(-\frac{\eta - \theta_S}{T_S}\right)} \tag{9}$$

where θ_S is a transition threshold that gives the probability 0.5 with which spatial attention acts on and T_S is a temperature which provides dependency between probability of spatial attention and share of attention. The bottom-up spatial attention is computed by the expression (7). A standard deviation σ_0 of spatial attention at level 0 is given by $\sigma_0 = 2^l \times \sigma_l$ for a standard deviation σ_l of spatial attention at the level l. The bottom-up spatial attention affects only for the next attentional selection. When no bottom-up spatial attention is generated, it is supposed that default spatial attention with a large standard deviation is

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

imposed on the center of an image. This causes a tendency that attention is directed to the center of an image and also the whole-part attentional bias that is provided by the level multiplier of spatial attention.

The competitive filtering narrows the range of attention down to relatively high attention area, which is adaptively performed according to change of the maximum saliency in response to change of IOR and modulation of saliency by the bottom-up spatial attention. Low attention area gradually becomes a target of stochastic attentional selection as high attention points inhibited by IOR but this attention shift can be probabilistically cyclic in accordance with the continuation time of IOR.

Thus, in our model, the dynamic maintenance of IOR, the bottom-up spatial attention probabilistically imposed according to the share of attention immediately before, and the adaptive competitive filtering of attention can achieve stochastic shift from high attentional points to low attentional points, convergence and divergence of attention, and the whole-part shift of attention.

4 Experimental Results

4.1 Experimental Framework

To evaluate dynamics and performance of stochastic attentional selection and shift, several experiments were conducted by using animated still video clips for stochastic visual pop-out test, still video clips which contains objects that are designed to attract attention in the real-world, and motion video clips which contains moving objects and persons in daily scenes.

Main parameter values used in experiments are as follows. The number of levels of a Gaussian resolution pyramid is 5. Among these levels, the visual attention pyramid is generated for reduced images from level 1 to level 4. In saliency computation, $\theta_R = 0.75$, $(w_{IC}, w_{SC}, w_{HC}, w_{OR}) = (0.25, 0.1, 0.4, 0.25)$ and $(w_S, w_D) = (0.05, 0.95)$. As for IOR, IOR continues 2 frames by default and $\delta = 0.1$. As for bottom-up spatial attention, $\lambda = 1.25$, $\theta_S = 0.25$, $T_S = 0.1$ and σ_l is set to 8 for any level l. As for competitive filtering, $\gamma = 0.8$. Four foci of attention are sequentially selected on each visual attention pyramid, that is, in each image frame.

4.2 Stochastic Visual Pop-out

Animated still video clips of 25 frames were used for experiments of the intensity, hue and orientation pop-out. Fig. 1 shows images for pop-out and foci of attention on them. The number of foci of attention is 100 in total since 4 foci of attention is selected in 1 frame. The left image of Fig. 1 shows intensity pop-out of a white circle in a certain frame. The result of the stochastic intensity pop-out is summarized as follows. First of all, 47 foci (47%) of attention were directed to a white circle. In frame by frame, at least one focus was directed to a white circle at 22(88%) frames of 25 frames. For each of 9 black circles, 5.89 foci (5.89%) of

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

attention were directed in 2.44(9.8%) frames on average though distribution was uneven. The whole-part attention shift was occurred 3 times from level 1 to level 2 and also 3 times from level 2 to level 1 around the white circle. From these results, it was confirmed that attention was almost always directed to a white circle and sometimes to black circles, that is, the white pop-out occurred clearly but stochastically. The hue pop-out and the orientation pop-out were similarly observed as shown in middle and right images in Fig. 1.

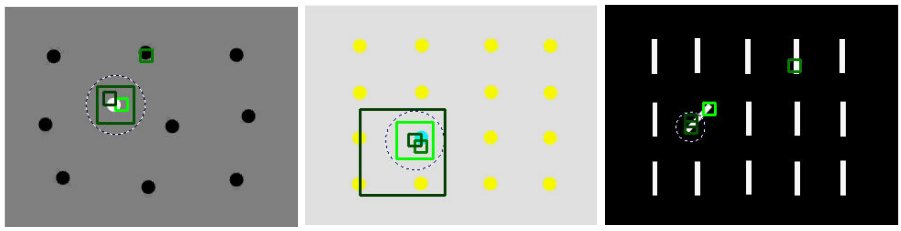

Fig. 1. Stochastic visual pop-out on images for the intensity pop-out (left), the hue pop-out (middle) and the orientation pop-out (right). (Squares represent foci of attention and their size represents their levels on the visual attention pyramid. That is, small, middle and large squares represent level 1, level 2 and level 3 respectively. Brightness represents the order of attention in an image. Dotted circles represent bottom-up spatial attention.)

4.3 Stochastic Attentional Selection and Shift for Static Scenes

A still video clip which contains traffic signs as objects designed to attract attention was used for this experiment. This video clip consists of 25 frames and the number of foci of attention is 100 in total. Fig. 2 shows snapshots and foci of attention. In this experiment, 84 foci (84%) of attention were directed to two traffic signs. This included attention to the upper traffic sign, the lower traffic sign and an area which contained both traffic signs, one which is shown in the left image of Fig. 2. In frame by frame, 23(92%) frames of 25 frames were directed to one or both of traffic signs. Other attended objects or areas were a car and an area ahead which contains a sky, a mountain and a road, which are shown in middle and right images of Fig. 2. The 16 foci of attention were directed to these objects or areas in 7 frames. The whole-part shift of attention were observed 15 times (15%), which included 7 times from the part to the whole and 8 times from the whole to the part. An attention shift from the lower traffic sign at level 1 to an area which contains two traffic signs at level 3 then to the upper traffic sign at level 1 was a typical whole-part attention shift. Also from car at level 2 to its part at level1 was a typical example of the whole-part attentional shift. Thus, it was confirmed that attention was directed to traffic signs designed to attract attention with high probability and the whole-part attentional shift was occurred with a certain probability.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

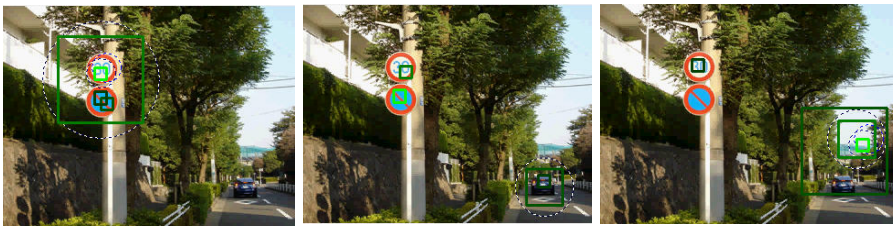

Fig. 2. Stochastic attentional selection and shift on a scene which contains traffic signs.

4.4 Stochastic Attentional Selection and Shift for Dynamic Scenes

Motion video clips of a person performing a sign language and a dynamic scene at a crossing were used for experiments.

A sign language video clip[1] consists of 30 frames which contains 120 foci of attention. In the experiment, attention was directed to a face, a moving right hand and a moving right arm. The left image of Fig. 3 shows an example of a person performing a sign language and foci of attention. The 78(65%), 24(20%), and 18(15%) foci of attention were directed to the right hand, the face and the right arm respectively. In frame by frame, 24, 10 and 10 frames were directed to the right hand, the face and the right arm respectively and attention followed the moving right hand and arm. The whole-part shift of attention were observed 31 times (26%) around the face and the right hand, which included 13 times from the part to the whole and 18 times from the whole to the part. From these observations, it was confirmed that attention was directed to and followed moving parts of a person with high probability, and that attention was directed to the face with a certain probability because it was salient in hue. Also it was confirmed that the whole-part attentional shift was occurred around these salient parts of a person.

A crossing video clip consists of 70 frames which contains 280 foci of attention. The middle and right images of Fig. 3 show snapshots and foci of attention. In this video clip, a lot of attention were paid to a person walking from left to right, two bicycles toward the left and one bicycle toward the right. In total, 244 foci (87%) of attention were directed intermittently to these person and bicycles. In frame by frame, attention was directed to at least one of them in 68 (97%) frames. Other attended objects or areas were a traffic signal, a road ahead and a area in the right side which contained buildings and roadside trees. The 36 foci of attention were directed to these objects or areas in 21 frames. The whole-part shift of attention did not occur so much in this video clip. Only 7 shifts from the part to the whole and 8 shifts from the whole to the part were observed. From these analyses, it was confirmed that in case plural objects were moving such as at a crossing, attention was directed intermittently to those with high probability and also static salient objects were attended to with low probability.

[1] This clip is part of a video open to the public on the web by NTT Data corporation.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

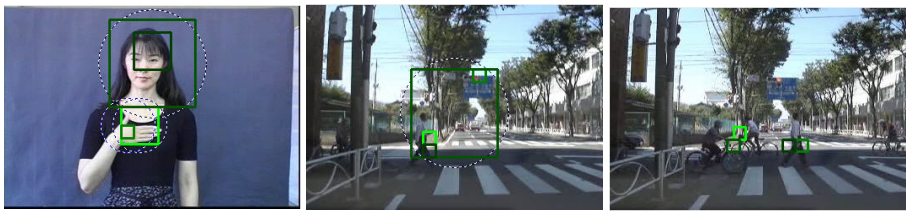

Fig. 3. Stochastic attentional selection and shift for a person performing a sign language (left) and a scene at a crossing (middle and right).

5 Conclusions

We have proposed a computational model of visual attention which performs stochastic attentional selection and shift on the visual attention pyramid that is computed for each frame of a video sequence. Through experiments, it was confirmed that this model achieved stochastic visual pop-out to artificial pop-out targets and stochastic attentional selection and shift, especially the whole-part attention shift and the motion-follow attention, in daily scenes.

Acknowledgments. This work was supported in part by Grant-in-Aid for Scientific Research No.18500121 from Japan Society for Promotion of Science.

References

1. Corbetta, M., Shulman, G.L.: Control of Goal-directed and Stimulus-driven Attention in the Brain. Nature Reviews Neuroscience **3** (2002) 201–215
2. Treisman, A.M., Gelade, G.: A Feature Integration Theory of Attention. Cognitive Psychology **12** (1980) 97–136
3. Wolfe, J.M., Cave, K., Franzel, S.: Guided Search: An Alternative to the Feature Integration Model for Visual Search. Journal of Experimental Psychology: Human Perception and Performance **15** (1989) 419–433
4. Koch, C., Ullman, S.: Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. Human Neurobiology **4** (4) (1985) 219–227
5. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-based Visual Attention for rapid Scene Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (11) (1998) 1254–1259
6. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. LNAI, Vol. 3899. Springer-Verlag, Berlin Heidelberg (2006)
7. Klein, R.M.: Inhibition of Return. Trends in Cognitive Sciences **4** (4) (2000) 138–147
8. Burt, P.E. and Adelson, E.H.: The Laplacian Pyramid as a Compact Image Code. IEEE Trans. on Communications **31** (4) (1983) 532–541
9. Mallot,H.A.: Computational Vision. The MIT Press (2000)
10. Adelson, E.H., Bergen, J.R.: Spatiotemporal Energy Models for the Perception of Motion. J. Opt. Soc. Am. A **2** (2) (1985) 284–299
11. Burr, D.C.: Motion Perception, Elementary Mechanisms. In: Arbib, M.A.(ed.): The Handbook of Brain Theory and Neural Network. The MIT Press (2003) 672–676