Cue-Guided Search: A Computational Model of Selective Attention

KangWoo Lee, Hilary Buxton, and Jianfeng Feng

***Abstract*—Selective visual attention in a natural environment can be seen as the interaction between the external visual stimulus and task specific knowledge of the required behavior. This interaction between the bottom-up stimulus and the top-down, task-related knowledge is crucial for what is selected in the space and time within the scene. In this paper, we propose a computational model for selective attention for a visual search task. We go beyond simple saliency-based attention models to model selective attention guided by top-down visual cues, which are dynamically integrated with the bottom-up information. In this way, selection of a location is accomplished by interaction between bottom-up and top-down information. First, the general structure of our model is briefly introduced and followed by a description of the top-down processing of task-relevant cues. This is then followed by a description of the processing of the external images to give three feature maps that are combined to give an overall bottom-up map. Second, the development of the formalism for our novel interactive spiking neural network (ISNN) is given, with the interactive activation rule that calculates the integration map. The learning rule for both bottom-up and top-down weight parameters are given, together with some further analysis of the properties of the resulting ISNN. Third, the model is applied to a face detection task to search for the location of a specific face that is cued. The results show that the trajectories of attention are dramatically changed by interaction of information and variations of cues, giving an appropriate, task-relevant search pattern. Finally, we discuss ways in which these results can be seen as compatible with existing psychological evidence.**

***Index Terms*—Attention, bottom-up map, computer vision, cue-guided search, top-down map.**

I. Introduction

FINDING your friend in a crowded street is not an easy task because all faces in the scene will be similar to the face that you are looking for and will easily distract you. However, if you know where he/she is waiting or what color clothes he/she is wearing, it will be much faster to find this friend. That is, knowledge about a target such as its probable location, physical attributes, context, etc., guides our attention by providing a clue to indicate where a target may appear. This is called “top-down” or “cue-based” attention. In contrast, “bottom-up” or “saliency-based” attention is characterized by the intrinsic attraction of a stimulus. Just as the term ‘stimulus driven’ implies, attention control lies outside the organizm. That is, the strength of an external stimulus automatically draws attention to a particular location (see [13], [30], [37], [54], and [74] for a review of bottom-up versus top-down attention).

Bottom-up and top-down attention are also distinguished in terms of processing modes. Bottom-up attention is characterized by fast, massively parallel processing toward selecting stimuli-based on their saliency, while top-down attention is characterized slow, less efficient serial processing that is intentionally or volitionally controlled. The dichotomy between bottom-up and top-down driven attention often assumes that two different or independent neural processing paths are involved in the deployment of attention. Brain imaging studies provide some evidence for partially separated networks in brain areas that support these two different types of attention control [15], [55]. However, it is worthwhile to note that top-down-based attention is voluntarily controlled, but it is not necessarily required a cue to guide an attentional behavior. That is, we can intentionally assign our attention to the right or left side of a view without any cue. In this paper, our research is limited in a model behavior guided by top-down biasing, rather than a simple voluntary shift of attention.

Attentive processing in computer vision initially used saliency-based models, where the strength of response of feature detectors determined candidate locations by matching [14]. More elaborate models of attentive visual search, which combine multiscale image features into a single topological saliency map have been proposed more recently in the influential work of Itti *et al.* [34], [35]. In these models, competition among neurons in this map selects the winning location in a purely bottom-up manner using a dynamical neural network. In contrast, Tsotsos *et al.* [67] implemented a model of visual attention-based on the concept of selective tuning, which involves both bottom-up (feedforward) feature extraction and top-down (feedback) processing. In this model, attention acts to speed up search inherent in many visual tasks using a classical neural network. Learning of task-relevance for selective attention has also been implemented in a neural model [68]. Many other models have been proposed, e.g., recent extensions of the basic Itti and Koch model using task knowledge [50] or feedback for object selection [69]. However, useful predictions require that a specific visual task be performed so that models can be directly compared.

In our approach, visual attention as a human behavior is considered as an interactive process influenced by both bottom-up and top-down processing mechanisms. The interaction between these two mechanisms is critically important to understand current attentional behavior that is biased with respect to particular objects, spatial locations and time. We also use a dynamic neural network so the top-down mechanism may bias selective behavior in such a way that it speeds up (facilitates) or slows

Manuscript received February 26, 2003; revised September 23, 2004. The work of J. Feng was supported in part by UK EPSRC under Grants GR/R54569, GR/S20574, and GR/S30443.

K. Lee and H. Buxton are with the Department of Informatics, Sussex University, Brighton BN1 9QH, U.K.

J. Feng is with the Department of Mathematics, Hunan Normal University, Changsha 410006, P.R. China. He is also with the Center for Scientific Computing, Department of Computer Science, Warwick University, Coventry CV4 7AL, U.K. (e-mail: jianfeng.feng@warwick.ac.uk).

Digital Object Identifier 10.1109/TNN.2005.851787

down (interferes) with the processing of a given visual stimulus. Based on this perspective, where visual attention is defined as a three-way relationship among bottom-up input, top-down input and a response, we develop a computational model for a visual search task. The important aspects of the model are that the spatial allocation of attention "where we see" and the temporal order of attentional allocation of "when we see" are attained by the dynamic interaction between bottom-up input and top-down input in an interactive spiking neural network (ISNN). We also incorporate the learning of task-relevance in our model, using novel learning rules for the ISNN.

Using the ISNN model, we can manipulate the amount of bottom-up and top-down influence on a search task to investigate the dynamic and modulatory aspects of visual selective attention. In our experiments, we employed a face detection task to illustrate the performance of our model and contrasted it with a pure saliency-based model. Face detection is a challenging task in computer vision, often used in association with a face recognition system. Here we want only to extract the position of a face region without recognition of who it is. However, the task is still not easy because of variability in scale, translation, orientation and illumination, etc., of the faces. Various methods have been developed to solve this problem [28], [75]. In this paper, we combine simple face detection techniques and our selective attention mechanism, so that the model carries out a specific visual task to find a particular face when guided by a cue.

In Section II, the general structure of our model is briefly introduced and followed by a description of the top-down processing of task-relevant cues. This is then followed by a description of the processing of the external images to give three feature maps that are combined to give an overall bottom-up map. In Section III, the development of the formalism for our novel ISNN is given, with the interactive activation rule that calculates the integration map. The learning rules for both bottom-up and top-down weight parameters are given, together with some further analysis of the properties of the resulting ISNN. In Section IV, the results of the simulated model are given when applied to a face detection task to search for the location of a particular face that is cued. The results show that the trajectories of attention are dramatically changed by interaction of information and variations of cues, giving an appropriate, task-relevant search pattern. Next, in Section V, we discuss ways in which these results can be seen as compatible with existing psychological evidence and finally, in Section VI, draw some brief conclusions.

II. Model

A. General Structure

Fig. 1 shows the general structure of the model for cue-guided search. The model can be divided into three main processing modules—top-down module, bottom-up module and integration module—according to the direction of information processing as well as their role in the model. The original input is provided in a form of digitized images. From the original input image, some basic features including color, aspect ratio, symmetry, and ellipse are extracted in order to construct a bottom-up map, and color feature for a top-down map. First of all, the top-down processing module is to construct a top-down map and to assign its top-down input to the integration module on the basis of a cue that indicates a possible target might appear near the location. However, the location of a cue color region is not directly obtained by an explicit instruction (e.g., the right corner of an image or the center of an image), rather by a feature cue that is extracted from the original image. For instance, a particular color such as "a red colored cloth" that your friend may be wearing is used as a cue. So, the location of a cue color region is obtained from a series of color segmentation processes described in Section II-B. From the location of a cue color segmented region, a top-down input value is calculated with a Gaussian distance measure and is used for the subsequent integration processing.

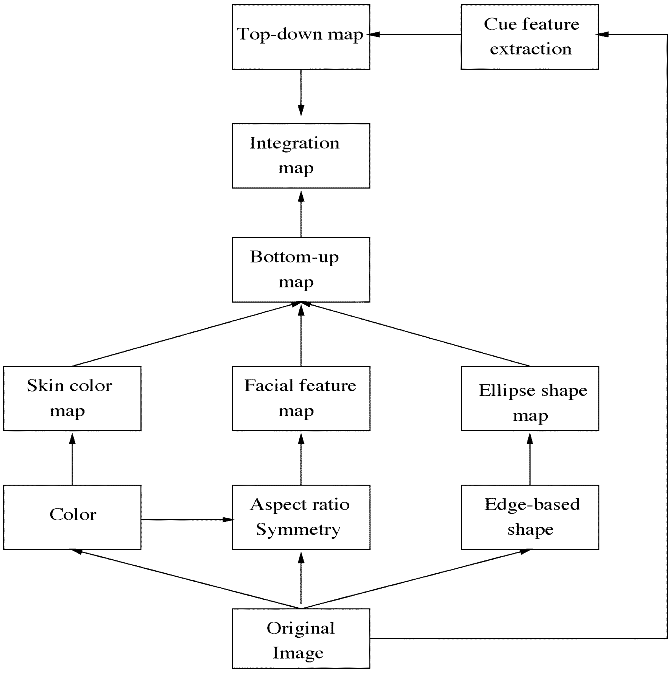


Fig. 1. General architecture of visual selective attention system. The system consists of three different submodules that form each different map—bottom-up, top-down and integration map. Within top-down and bottom-up modules, features are extracted in a series of segmentation processes and construct feature maps. In the bottom-up module, all the features are combined into a bottom-up map where a target candidate has a vector form of input. The top-down input is determined by the geometrical relationship and Gaussian distance between the location of a target and the location of a cue. Both inputs are integrated with ISNN that has a dynamic and modulatory property. The attentional allocation takes place in the order of the shortest interspike interval generated by ISNN.

In contrast, the bottom-up processing module extracts various features such as color, aspect ratio, symmetry, and shape that a target may share, and then constructs feature maps. In order to achieve a single and unified bottom-up map, we establish a set relationship between features depending on how many features are shared at a specific location of a target candidate. Each target candidate in the bottom-up map is represented in a vector form that consists of bottom-up feature values.

Finally, the integration processing module is to integrate both top-down input and bottom-up input from each map, and then produce an output in the form of interspike interval. A new neural network model called ISNN has been developed to obtain the output resulting from dynamic interaction between the two inputs. The model ISNN shows very interesting properties that dynamically correlate two input steams depending on whether

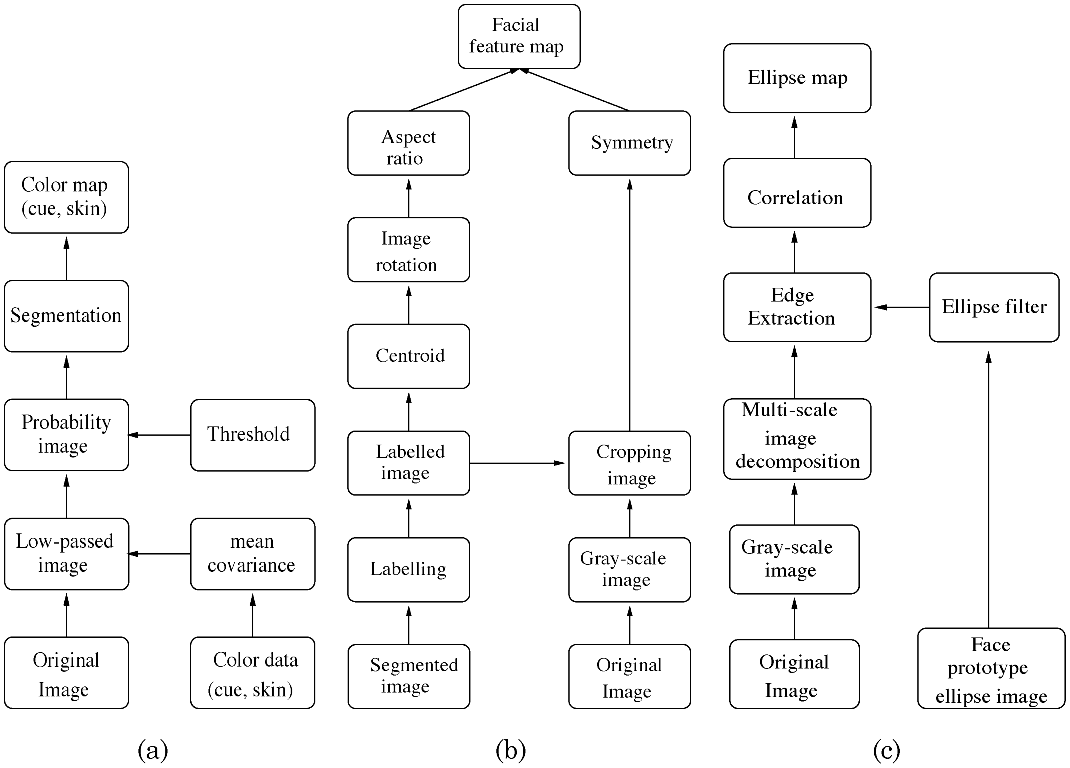


Fig. 2. Schematic processing for each feature map. (a) Cue or skin color map. (b) Facial feature map. (c) Ellipse shape map. The construction of each feature map is based on a kind of segmentation technique that segregates objects from their background and results in a binary image.

they are consistent or not. Using this output from ISNN, attentional windows are allocated in an ascending order. That is, the shorter the interval in spike trains is generated at a location, the earlier attention is allocated to that location.

B. Top-Down Processing Module

In the model, it is assumed that the properties of a cue are already known. In the psychological or neuroscience studies, many different kinds of signal or stimuli such as verbal instruction, color, arrow, shape, etc., have served as a cue. Throughout our experiment a single color cue is used. However, the use of a cue is not limited to color alone as here, but can generalize to various kinds of stimuli or stimulus combinations, etc.

Even though we can perceptually discriminate colors with very subtle differences, we do not actually have accurate knowledge of all these colors, so those used here were far fewer than we are able to discriminate. This also means that we just roughly know the color of a cloth a target person is wearing such as a red or yellow t-shirts, not color values in a chromatic-diagram. For this reason, we collected images to construct the four-color database (red, blue, green, and yellow) from various web sites and a training image set taken from street, campus, and so on. Since the common RGB representation of color images is susceptible to environmental illumination change, it is not suitable for constructing a color database. Thus, the color in the database was transformed into the chromatic representation in which illumination was eliminated

$$r = \frac{R}{(R+G+B)}$$
$$b = \frac{B}{(R+G+B)}$$

and then the mean of the r and b values and covariance are calculated from the sampled color pool. Fitting the mean and covariance matrix to a Gaussian distribution, we modeled the cue color probability that decides whether a pixel x belongs to the color

$$\begin{aligned} P(\text{pixel}) &= P\left(\text{pixel} \in \text{cuecolor} | r(\text{pixel}), b(\text{pixel})\right) \\ &= \exp\left(-0.5\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right) \end{aligned} \qquad (2.1)$$

where μ and Σ are the mean and covariance of the given color, respectively. Using the equation, the cue color region segmentation is carried out.

The schematic procedure for obtaining a top-down map is shown in Fig. 2(a). First of all, the original color image passes through a low-pass filter to get rid of noise, and then a cue color probability image is obtained by applying (1) to the low-pass image. Then, an adaptive thresholding method is introduced to determine an optimal threshold value that is used to segment the probability image into a binary image [58]. The coordinates of the each segmented region are obtained. It should be noted that the geometrical coordinates of the segmented region are not necessarily the center or centroid of the region, but can be varied according to the geometrical relationship between a cue and a target. For instance, the geometrical relationship between a face (target) and red t-shirts (cue) has "a-target-above-a-cue" relation not "a-target-below-a-cue." In our simulation, the coordinate of the segmented region extracted from cue color is defined the top-center of the segmented region. It should be noted that the cues, themselves, are nonspatial and consists of a given color. No direct spatial indication—e.g. right- or left-hand side—is given, but it is possible to use it as a cue.

The geometrical relationship also implies what the shape of the attentional window allocation should be determined with reference to a cue. Many attentional metaphors such as "spotlight" and "zoom-lens" suggest the attentional distribution be unitary or isotopic over a limited spatial extent [21], [57]. However, this view has been criticized with respect to object-based attention [19] and divided attention [44]. In this paper, we assume the coordinates of the cue position and the attentional distribution can be varied in terms of the geometrical relationship between a cue and a target. This means that the top-down input value for the interaction module is calculated by a geometrical relationship. For example, let us consider the case of "a-target-above-a-cue" geometrical relation. For any target candidates above a cue, top-down input value is computed as a Gaussian distance from the position of a cue to the center of a possible target candidate, whereas for any target candidates below a cue the top-down input value is set to 0. In an image, there could be more than one segmented of a cue color and more than one distance value can be obtained. In our simulation, the distance from a target to each segmented region of a color cue is defined as a shortest distance $D(\text{shortest})$ among calculated distances. To calculate the Gaussian distance from each segmented region of a color cue to a target candidates, then the distance was scaled into 0 to 10. Second, we applied the following equation to obtain the Gaussian distance:

$$D_x^y = \exp\Big(-\big(D(\text{shortest})_{x,y}\big)^2\Big)\Big) \tag{2.2}$$

where x and y are coordinates of a target candidate.

C. Bottom-Up Processing Module

As mentioned earlier, face detection techniques are used to extract a target candidate's features. Numerous methods have been proposed to detect a face in a given arbitrary image [28], [75]. Among them three different methods were employed for constructing the bottom-up map.

1) Skin Color Map: Skin color segmentation is a very simple, but very effective approach to face detection algorithms since it is well known that the hue of skin is roughly invariant across different ethnic groups and the skin color distribution lies a limited chromatic color space [8], [36], [48]. To build the skin color model that segments skin parts from nonskin parts, several approaches have been used to label a pixel as skin such as normalized RGB, HSI, YCrCb, etc. In our model, the procedure for skin color segmentation is the same as that of color cue segmentation, except using a skin color database. The skin color database was collected from images on internet web sites.

2) Facial Feature Map: Even though skin color segmentation provided a robust face detection algorithm, it also segments other body parts (e.g., naked hands and legs, etc.) and skin-color noise. Thus, two face features—aspect ratio and symmetry are introduced to distinguish among the skin color segmented regions. A face-like region may share these features but a less face-like region may not. A schematic procedure to extract these features is shown in Fig. 2(b).

- **Aspect ratio**

Aspect ratio between the width of face and height face is a unique feature of a face that discriminates it from other body parts. The golden ratio of an ideal face calculated by Govindaraju [25] is height/width $= (1+\sqrt{5})/2$. We cannot directly obtain the aspect ratio from the segmented region because the object in the region may incline toward a certain orientation, which means the elongation of the object is needed undo the translation by rotation.

First, to obtain the aspect ratio for each skin segmented region, the centroid of the region is calculated from the binary segmented image

$$\bar{x} = \frac{1}{A}\sum_{i=1}^{n}\sum_{j=1}^{m} jB(i,j)$$
$$\bar{y} = \frac{1}{A}\sum_{i=1}^{n}\sum_{j=1}^{m} iB(i,j) \tag{2.3}$$

where B is the n by m matrix of the skin color segmented region, and A is the area of the region from which a binary pixel is taken. Then, the area of the region from which the binary pixel is taken is rotated in the opposite direction from the angle of inclination. The angle is given by

$$\theta = \frac{1}{2}\text{a}\tan\frac{b}{a-c} \tag{2.4}$$

where $a = \sum_{i=1}^{n}\sum_{j=1}^{m}(x'_{ij})^2 B(i,j)$, $b = 2\sum_{i=1}^{n}\sum_{j=1}^{m}(x'_{ij}y'_{ij})B(i,j)$, and $c = \sum_{i=1}^{n}\sum_{j=1}^{m}(y'_{ij})^2 B(i,j)$ with $x' = x - \bar{x}$ and $y' = y - \bar{y}$.

After rotating the binary area, we determine the aspect ratio of the area. With respect to the variation of the aspect ratio of faces, if the value of aspect ratio is within the range from 0.8 to 2.0, we set the aspect ratio feature value to 1, otherwise set to 0.

- **Symmetry**

Symmetry is also an important feature to characterize a visual object and has been used to describe artificial and natural images [43] as well as to detect faces and parts of faces [59]. Kovesi [43] proposed an algorithm to quantify symmetry. The basic idea of his algorithm is that symmetry has a periodicity in the structure of the image. So, he calculated symmetry values at an orientation using even-symmetric and odd-symmetric Gabor filters. The procedure has been adopted in our model and presented as follows.

If I denotes the signal image and M_n^e and M_n^o denote even-symmetric and odd-symmetric wavelets at a scale n, we can obtain even and odd coefficients of each quadrature pair of filters as follows:

$$[e_n(x), o_n(x)] = [I(x)\otimes M_n^e, I(x)\otimes M_n^o] \tag{2.5}$$

where $\otimes$ is the convolution. The amplitude of the transform at a given wavelet scale is given by $A_n(x) = \sqrt{e_n(x)^2 + o_n(x)^2}$ and the phase is given by $\Phi_x(x) = \text{a}\tan 2(e_n(x), o_n(x))$. Then, the symmetry value is obtained by subtracting the absolute value of the even symmetric coefficient from the absolute value of the odd symmetric coefficient over all scales

$$Sym(x) = \frac{\sum_n ([|e_n(x)| - |o_n(x)|] - T)}{\sum_n A_n(x) + \varepsilon} \tag{2.6}$$

where ε is a small constant ($\varepsilon = 0.001$ in our simulations) to prevent division by zero and T is a noise compensation term.

To utilize the symmetry algorithm in our model, we first replace the skin segmented region with a gray scale image, and then apply the symmetry algorithm at four different orientations—0°, 45°, 90°, and 135°. The symmetry feature value is assigned 1 if a maximum symmetry value is at 90°, and 0.5 if the maximum symmetry value is at 45° or 135°. Otherwise, it is set to 0.

3) Ellipse Map: The other method to extract target features is based on the shape of the face contour [39], [62]. Typically, the frontal view of a face has an oval or round shape. Using a predefined standard contour pattern, the correlation values with a given image area are computed to locate a face part in the image. However, this method has been criticized because it is not efficient to deal with variations in shape, pose and so on. Nevertheless, it can provide partially useful information in the case where other features do not properly locate a target candidate (e.g., a background region surrounding faces shares the same color as face candidates and, thus, color-based feature extraction fails to separate a face from its background).

The procedure to extract an ellipse feature is shown in Fig. 2(c). First, the original color image is transformed into five gray level images which progressively interpolate the image, using a nearest neighbor interpolation method. In the nearest neighbor interpolation, the value of the new pixel is made the same as that of the closest existing pixel. When enlarging an image, the method duplicates pixels; when reducing the size of an image, it deletes pixels. The Sobel edge operator is used to obtain an edge image at each level. The edge images are convolved with an elliptical edge filter and then the ellipse-convolution images are obtained. Using a standard ellipse convolved image that is obtained from a face part only, the correlation value is computed at each pixel point of the ellipse convolved images. The correlation images at five different scales are rescaled to unify the image into a single image using the same interpolation method. Then, a threshold value is applied to segment possible face-like objects.

D. Bottom-Up Map

Up to now, the separate feature maps have been computed. However, it is necessary to combine the various feature maps into a single bottom-up map for a further processing. In a saliency-based model, feature fusion is accomplished with a scalar saliency value after normalization of each feature value [34]. In our model, on the other hand, a single and unified bottom-up map is accomplished by a set relation between features that describes whether they share a specific region or not. However, it should be noted that this Boolean operation is not biologically plausible since the operation simply combines features using logical operators (e.g., OR or AND). The relation between features in a segmented region can be expressed with a Venn diagram shown in Fig. 3. Let the problem space $\Re$ define the relation, then

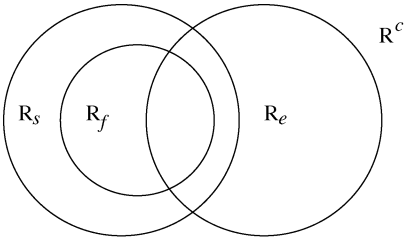


Fig. 3. Venn diagram that schematizes the relation between bottom-up features. The region R_s, R_f, R_e stands for the regions sharing skin color feature, facial feature, and ellipse feature. The region R^c denotes the complementary of the feature region R. Using a binary operation on the bottom-up feature map, five different subsets are obtained. The bottom-up input at a particular location is assigned based on these subsets.

$$\Re = R \cup R^c \tag{2.7}$$

where R is a region that shares at least one bottom-up feature and R^c is the complement of the set R. The region space R can be decomposed of three subregions that share each different feature and are denoted by R_s, R_f, and R_e:

$$R_s = \{\text{region} \mid \text{region that has a skin color feature}\}$$
$$R_f = \{\text{region} \mid \text{region that has a facial feature}\}$$
$$R_e = \{\text{region} \mid \text{region that has an ellipse shape feature}\}$$

that satisfy $R \supseteq (R_s \cup R_f \cup R_e)$. Since the region R_f is derived from R_s, R_f is a subset of R_s and has an inclusion relation $R_s \supset R_f$.

Using a binary operation on the subsets, we determine the relation between the subsets as shown in the Venn diagram—the region shared by only R_s ($R_s \cap \sim R_f \cap \sim R_e$), the region shared by only R_e ($R_e \cap \sim R_s$), the region shared by R_s and R_f not by R_e ($R_s \cap R_f \cap \sim R_e$), the region shared by R_s and R_e not by R_f ($R_s \cap R_e \cap \sim R_f$), and the region that shared by R_s, R_e and R_f ($R_s \cap R_f \cap R_e$). Also, R_f has two subsets corresponding to the regions that have two face features—aspect ratio and symmetry.

$$R_f1 = \{\text{region} \mid \text{region that has an aspect ratio feature}\}$$
$$R_f2 = \{\text{region} \mid \text{region that has a symmetry feature}\}$$

At a given region x, the element of the region R is given by

$$\begin{aligned} R(x) &= \{R_s(x), R_f(x), R_e(x), R^c(x)\} \\ &= \{E_s, E_{f1}, E_{f2}, E_e, E_c\} \end{aligned} \tag{2.8}$$

where E_s, E_{f1}, E_{f2}, E_e, and E_c are the elements of the set $R_s(x)$, $R_f(x)$, $R_e(x)$ and $R^c(x)$, respectively and $E_s \in \{0,1\}$, $E_{f1} \in \{0,1\}$, $E_{f2} \in \{0, 0.5, 1\}$, $E_e \in \{0,1\}$, and $E_c \in \{0,1\}$. So, after a binary operation on the feature maps, the segmented regions can be classified into one of these subsets that has bottom-up values in a vector form.

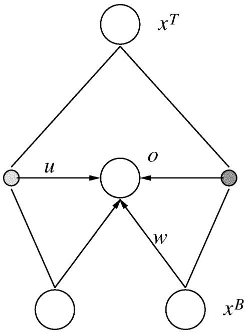


Fig. 4. Simple example of ISNN structure. The model has bottom-up (x^B), top-down input units (x^T) and output units (o). The bottom-up connection (w) links bottom-up units and output units, and multiplicative connection (u) links the two input units and output units. Therefore, an output unit receives two kinds of inputs—one driven by only bottom-up and the other driven by multiplication of both inputs. The output unit produces a spike if the membrane potential of the unit reaches a threshold. The interspike interval is used to measure the response of the unit.

III. Integration Module

The most important aspect of our model is that the top-down information obtained from a top-down cue and the bottom-up information obtained from various features are integrated through an ISNN that has been developed under consideration of the three-way relationship between bottom-up input, top-down input, and response. Similar consideration led to using a mutual information maximization approach for classical networks [3], [38]. The idea of the mutual information maximization is to maximally correlate two input streams and produce more gain when two input streams are coherent. Another approach is based on conditioning that associates two stimuli (classical conditioning) or response and feedback (instrumental conditioning) [2], [16]. In such a model, selection can be viewed as inference or expectation that predicts a reward (or unconditional stimulus) from an action or the conditional stimulus. In our model, we implemented the three-way relation with a leaky integrate-and-fire (IF) neural network to achieve a dynamic and modulatory property. In this section, we briefly describe the structure, learning rule and properties of the model.

A. Structure of ISNN

A simple example of the structure of ISNN is given in Fig. 4. The network consists of bottom-up input units x^B, top-down input units x^T and output unit o. The output unit o receives two kinds of weighted inputs—bottom-up input and multiplication between both inputs—at time t. The multiplication has an important nonlinear property that correlates the two inputs. The weighted sum for the jth output unit is given by

$$net_j(t) = \alpha \sum_{i=1}^{n} w_{ij}(t)x_i^B + \beta \sum_{i=1,r=1}^{n,m} u_{irj}(t)\left(x_i^B\right)^2 x_r^T \quad (3.1)$$

where B and T stand for the bottom-up and top-down inputs, n and m are the dimensions of the bottom-up and top-down inputs, and w and u are the bottom-up and multiplicative weights, respectively. A model that has a similar multiplicative operation is the sigma-pi network [41]. The constants α and β determine the amount of influence driven by bottom-up and top-down inputs on the net value. If $\alpha = 1$ and $\beta = 0$, the value net_j is determined by only bottom-up inputs. If $0 < \alpha < 1$ and $\beta = 1 - \alpha$, the value net_j is determined by both to a variable degree.

The second term in (9) can be considered as a correlation between two input sets because when two inputs are consistent, it produces a certain amount of gain, but when two inputs are inconsistent, it causes a cost to the network.

Another nonlinearity is implemented with a sigmoid function that may correspond to the processing at the level of soma. The sigmoid function has desirable properties; the output of the function will not be zero or one, and is located between these two values. The nonzero property is very useful with an interspike interval function, which we will explain later. So, the amount of activation driven by net_j is given by

$$y_j(t) = \frac{1}{1 + \exp\left(-net_j(t)\right)}. \quad (3.2)$$

In the IF model, a postsynaptic spike occurs if the summation of postsynaptic potential produced by the succession of input signals reaches a threshold [20], [24], [26], [40], [64] and the membrane is then reset to the resting potential (0 in our simulations). Conventionally, the model is described with a circuit that consists of a capacitor C is parallel with a resistance R driven by a current $I(t)$. The trajectory of the membrane potential can be expressed in the following form:

$$V(t+dt) = V(t) + RI(t)\frac{dt}{\tau_m} - V(t)\frac{dt}{\tau_m} \quad (3.3)$$

where τ_m is the membrane time constant of a neuron and we assume $R = 1$ in the following. The equation means the membrane potential V at time $t + dt$ is the sum of the potential V at the previous time t, the amount of ongoing current and the amount of decay.

If we limit our consideration to the special case of a cell firing a train of regularly spaced post synaptic potentials, we may write the voltage trajectory of the membrane potential in the following form by putting the leaky IF model with the sigma-pi like activation together:

$$\begin{aligned} V_j(t) &= V(t_0) + V(t_1) + \cdots + V(t_{n-1}) \\ &= y_j \left\{ 1 + \exp\left(-\frac{1}{k\tau_m}\right) + \exp\left(-\frac{2}{k\tau_m}\right) \right. \\ &\qquad \left. + \cdots + \exp\left(-\frac{n-1}{k\tau_m}\right) \right\} \\ &= y_j \sum_{i=0}^{n-1} \exp\left(-\frac{i}{k\tau_m}\right) \\ &= y_j \frac{1 - \exp\left(-\frac{n}{k\tau_m}\right)}{1 - \exp\left(-\frac{1}{k\tau_m}\right)} \end{aligned} \quad (3.4)$$

where the amplitude y_j decays exponentially with the membrane time constant τ_m and regularly spaced time $1/k$. A postsynaptic spike will be generated if the voltage of membrane potential V_j is equal to or larger than a threshold Vth

$$V\text{th} \leq y_j \frac{1 - \exp\left(-\frac{n}{k\tau_m}\right)}{1 - \exp\left(-\frac{1}{k\tau_m}\right)}. \quad (3.5)$$

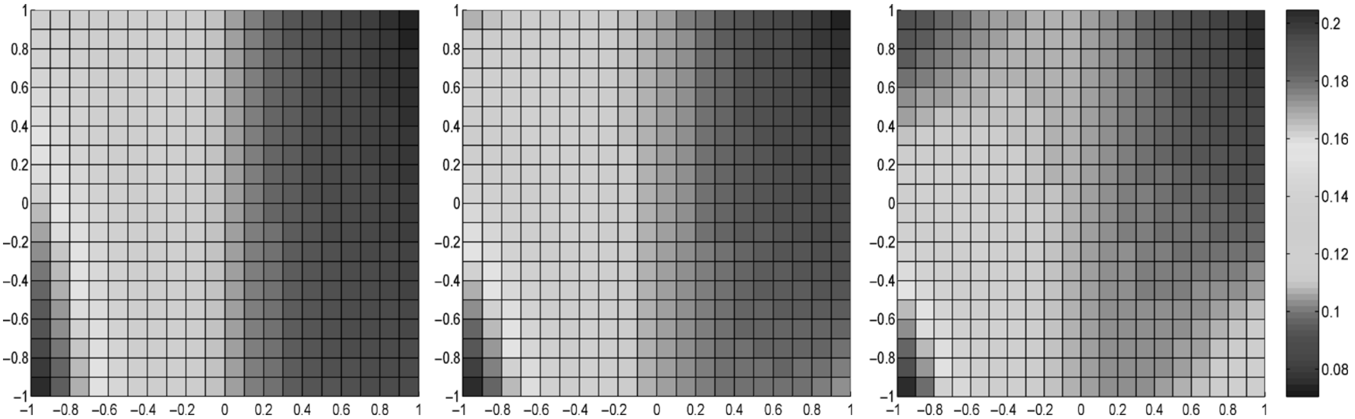


Fig. 5. Interspike interval plotted over bottom-up and top-down inputs. From left to right, the constant value α and β are set to (0.7, 0.3), (0.5, 0.5), and (0.3, 0.7), respectively. As the portion of α and β is changed, the interspike interval driven by bottom-up and top-down input is systematically changed. As the constant β is getting stronger, the influence of top-down input on the interspike interval is becoming stronger, and vice versa.

From the previous equation, the interspike interval n/k is determined

$$T_j = \frac{n}{k} = -\tau_m ln \left[1 - \frac{V\text{th}\left(1 - \exp\left(-\frac{1}{k\tau_m}\right)\right)}{y_j}\right]. \quad (3.6)$$

In our simulation, we use $V\text{th} = 5$, $k = 100$, and $\tau_m = 1$.

B. Learning Equation

In order to derive the learning equation here, we simply define an "error" as the difference between actual spike interval and desired spike interval. Thus

$$E = \frac{1}{2}\sum_{j=1}^{l}\left(T_j^d - T_j\right)^2 \quad (3.7)$$

where T_j^d is the jth desired spike interval and l is the number of output units. Since we want to find the weight values which minimize the error function, we can differentiate the error function w.r.t. the weight parameters

$$\frac{\partial E}{\partial w_{ij}} = \eta\alpha\left(T_j^d - T_j\right)\left[\frac{\tau_m}{y_j - V\text{th}\left(1 - \exp\left(-\frac{1}{k\tau_m}\right)\right)}\right] \times \left[\frac{V\text{th}\left(1 - \exp\left(-\frac{1}{k\tau_m}\right)\right)}{y_j}\right] y_j(1 - y_j)x_i^B. \quad (3.8)$$

Similarly, we can apply the learning rule for the secondary connection u_{rj}

$$\frac{\partial E}{\partial u_{irj}} = \eta\beta\left(T_j^d - T_j\right)\left[\frac{\tau_m}{y_j - V\text{th}\left(1 - \exp\left(-\frac{1}{k\tau_m}\right)\right)}\right] \times \left[\frac{V\text{th}\left(1 - \exp\left(-\frac{1}{k\tau_m}\right)\right)}{y_j}\right] y_j(1 - y_j)\left(x_i^B\right)^2 x_r^T. \quad (3.9)$$

In training, we assumed that a target is processed rapidly, whereas a nontarget is processed slowly. This assumption may reflect that we learn to respond to a cued object prior to noncued objects. This assumption could be supported by psychological findings in which search and object recognition are facilitated by implicit learning the relationship between context and target [10]–[12]. So, the desired interspike interval for a target and nontarget were set to 50 ms and 3000 ms respectably. We trained two cases (two learning examples)—a perfect target case (all bottom-up and top-down units corresponding to an ideal target were set to 1 and rest to 0) and a perfect nontarget case. The experiments then tested the generalization of this training.

C. Properties of ISNN

In order to provide an insight to the properties of an ISNN, some further analysis of (14) was carried out. First, the function in (14) for the interspike interval is very important in understanding the dynamics of error convergence. Note that it is a logarithmic function, which implies that if the inner value $1 - (V\text{th}(1 - \exp(-1/k\tau_m))/y_j)$ is equal to 0 or less than 0, the output will be infinite or imaginary. However, if $y_j > V\text{th}(1 - \exp(-1/k\tau_m)$, this always guarantees the network will converge to a certain stable state.

Second, the dynamic correlation between two inputs is analyzed by plotting the interspike interval over bottom-up input and top-down input in Fig. 5. The values of the constants α and β were systematically changed from left to right—(0.7, 0.3), (0.5, 0.5), and (0.3, 0.7), respectively. Note that, whatever the values of α and β, in the case that two inputs are correlated, the interspike interval will be shorter (positively correlated) or longer (negatively correlated) than in uncorrelated cases. Note also that, as the constant α gets larger, the influence of bottom-up input on the interspike interval is stronger. Similarly, the influence of top-down input on the interspike interval is stronger as the constant β gets larger. Finally, if two inputs are uncorrelated, the network pays a cost that causes an increase or decrease of the interspike interval in a manner that interferes with processing information (see Fig. 5).

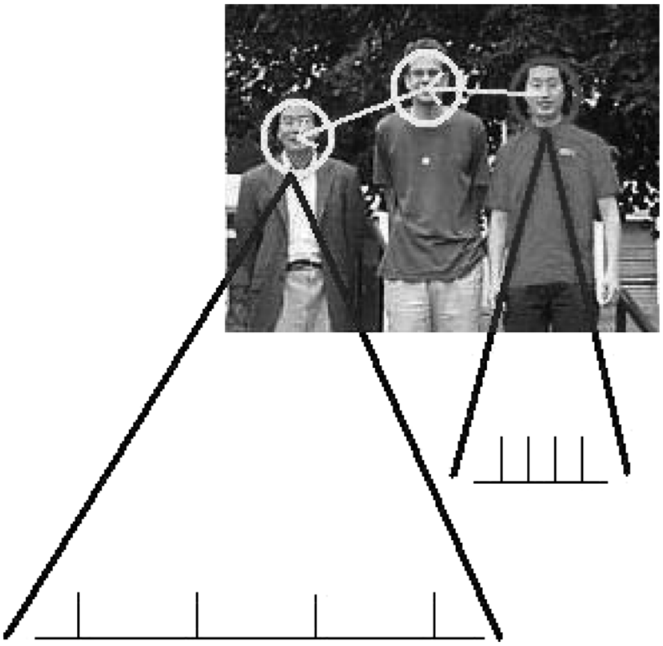


Fig. 6. Allocation of attention and interspike interval. The attentional window is allocated in ascending order from the location with the shortest interspike interval to the location with the longest interspike interval. This figure shows an example of how an attentional trajectory is created in our model.

D. Allocation of Attention

The attentional window is allocated at the possible target position in the original image after the interspike interval is calculated. The window is allocated in ascending order from the target with the shortest interspike interval to the target with the longest interspike interval (Fig. 6). In the following section, we show how trajectories of attention shifts are changed by different conditions of the constant α and β and different cue conditions.

IV. Simulation

We have investigated the performance of the system by manipulating various conditions. 100 images were obtained from natural environments such as a shopping center, university campus, garden and street that contain faces. Among them 95 images were used to test the model performance. Five of them were excluded because our model failed to detect a target face. Some of target face were not extracted from feature extraction procedure and a target face is occluded by other faces.

A. Bottom-Up and Top-Down Influence on a Face Search Task

As earlier, the influence of bottom-up and top-down input on a face search task can be manipulated by assigning different values of α and β. If $\alpha = 1$ and $\beta = 0$, the performance of the system is totally dependent on the bottom-up input. However, as the portion of β is getting increased, the influence of top-down input on the face search task is getting stronger. Accordingly, we changed the constant values α and β—(1.0, 0.0), (0.6, 0.4), (0.5, 0.5), and (0.4, 0.6).

100 images were used for this simulation, but five of them were excluded because the model fails to find a target face—e.g. a target face is occluded by another. The results of the system when finding a cued target were investigated. The results are presented in Figs. 7–10. The performance of the system is clearly improved by introducing top-down input and gets better as the constant β becomes larger. The attentional trajectories were also dramatically changed since the higher values of constant β force the system to attend the target candidate at the location of a cue color segmented region by increasing the correlation gain of the net value, whereas it detains attendance to target candidates in unindicated locations by a cue color as producing no gain or more interference, that reduces the level of the net value. These results are in accordance with much psychological evidence that shows facilitation and inhibitory effects of a cue on detecting a target [10], [47], [56]. In other words, the validity of a cue can be manipulated by changing the probability that a target appears in a cued location. If a target is given at a location regardless of a cue, the response should be based on only the target. However, if target is always given at

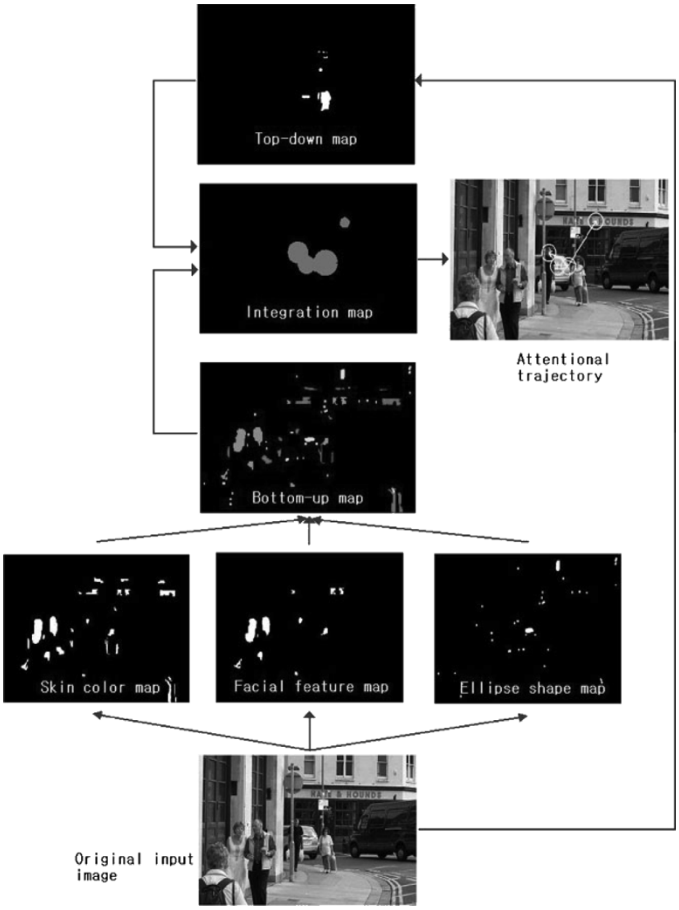


Fig. 7. Example of processing of the model with a natural image containing faces. In bottom-up map, different colors represent different overlaps between feature maps. The task given to the model is to find a lady who is wearing a yellow t-shirts. At the top of figure, the segmented regions of cue color are presented. The red color region indicates the overlap shared by three feature maps, the brown color region for the overlap two feature map, dark blue region for the skin color map, and bright blue for the ellipse map. The order of interspike interval is represented by the size of circle in the integration map. The red circle in attentional trajectory indicates the first location where attentional window is allocated. The model allocates the focus of attention to possible target locations in the order from the target location where the system generates the shortest interspike interval to the longest interspike interval. In other words, as the target location is more relevant to the indication of a cue that location is more likely to be selected at an early stage in the trajectory of attentional allocation.

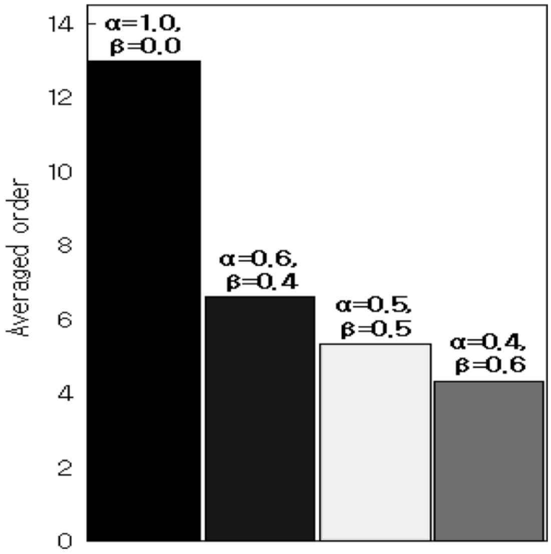


Fig. 8. Performance of the model when varying the constant values α and β. The averaged order ($=$ sum of order/number of images) in which a target is detected within the attentional trajectory was measured. Performance increases as the cue (top-down) weight is increased and reaches optimal value at $\beta = 0.6$.

a cued location, there is a strong correlation between a target and a cue. Therefore, a target that is highly correlated to a cue is detected faster, while a distraction (or possibly, a target) in an uncued location is inhibited [57], [60].

B. Comparison With Saliency-Based Model

Itti *et al.* [33], [34], [52] developed an elegant computational model of selective attention that traces back to Koch and Ullman's model [42]. The model is based on the saliency that is obtained from a center-surround operant mechanism working on many different levels of features including color, orientation and intensity. The attention of the system shifts from the strongest salient location to the next strongest salient location, and so on. The model provides a computational explanation about many bottom-up psychological results, so called "pop-out." However, it is difficult to say that the saliency is informative for a certain search task and it may seriously misguide the behavior of a system for a given task. It is worth while to show the performance of the system in comparison to our model because this provides an interesting contrast between a saliency-based model and more top-down model on our face detection task. The results obtained from two systems are presented in Fig. 11. The trajectories of attention in Itti's model were built by moving the focus of attention to the salient locations that give a strong difference from surrounding objects. However, these locations can be irrelevant, or even meaningless, since the salient locations obtained from the bottom-up features are task-independent and rely on only the center-surround operation. In contrast, our model has attention allocated to the locations of possible target candidates in the order of relevance to a cue.

C. Variation of Cues

In a real world situation there may be many objects that share a cue's feature—a red flag, a red postbox, and red t-shirts, etc. So, it is important for the system to deal with this situation and to show a robust response. An example of our results is presented in Fig. 12. The trajectories of attention were biased toward the positions of a cue color segmented region regardless how many cue color segmented regions were there. However, as the number of objects that share the same attribute of a cue increase, the order in which the system find a target is more likely delayed.

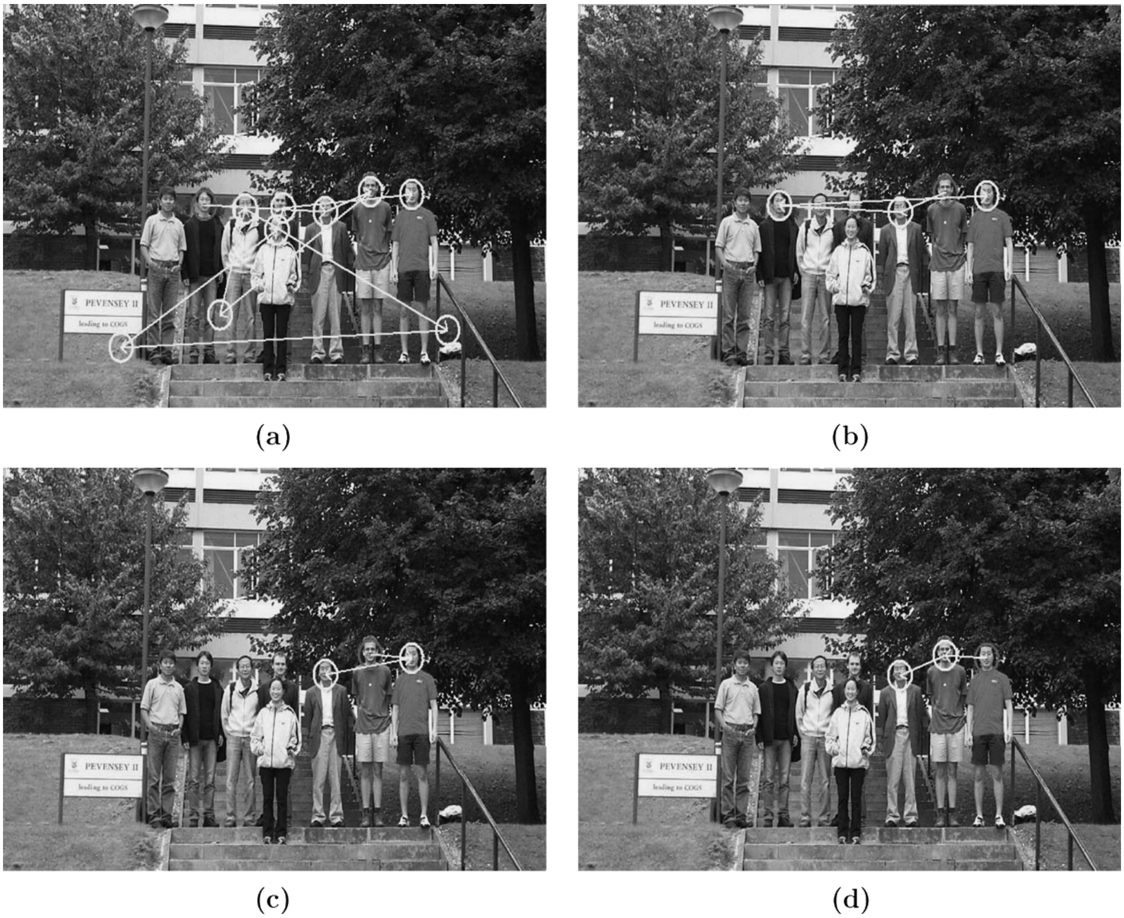
Fig. 13 illustrates how attentional allocation can be redirected by a different cue. In the upper image, attentional allocation was made by the red color—finding a man who has a red t-shirt, whereas in the lower image the attention allocation was done by blue color—finding a man who is carrying a blue plastic bag. The result shows the dramatic change of attentional trajectories that were guided by different cues in the same image.

V. Discussion

The focus of our work has been modeling of the interactive aspects of attention. Our computational model of selective attention integrates both bottom-up and top-down information, and dynamically allocates the focus of attention in space and time according to spike frequency in the ISNN. These properties provide a unique combination that distinguishes our approach from other computational models since the activation and learning rules here support intrinsic attention modulation. Even though we have not systematically tested performance against psychological and neuroscience evidence, our model is compatible with many known attentional behaviors. Here, we discuss the psychological and neuroscience plausibility of our model and its relationship to other computational models.

A. Top-Down Information as a Bias

Several theoretical frameworks for visual selective attention have been proposed to explain how the selection process can be accomplished, by focusing on different aspects of selective attention. Among them, Desimone and Duncan [17] suggested a persuasive theoretical account for visual attention which is called "biased competition." In their account, objects or locations in a visual scene compete to be selected for a further processing. However, objects or locations are not equally important, so that a certain object or location is preferred with respect to the current task. As mentioned earlier, two different mechanisms may be involved in selecting an object or location by imposing bias constraints on visual information processing. In bottom-up-based attention, the saliency of features that an object possesses plays an important role as a bias, so that our attention is easily drawn to the object that differs from the others or background. Many current computational models of visual selective attention are based on this saliency, constructing a "saliency map" from various features such as color, orientation, intensity, etc. [35], [42]. In this model, the saliency is computed as the difference between center and surround with different sizes of Gaussian filters. Therefore, the most salient location in the model stands for the maximum difference between center and surround features. However, saliency will be less effective if many other objects share the same feature properties as the target or if the target itself is not salient in the center-surround sense. Furthermore, Pashler *et al.* [54] challenged the classical view of bottom-up-based attention that stimuli with



(a) (b)

(c) (d)

Fig. 9. Trajectories of attention with different values α and β. By introducing the top-down influence ($\beta \neq 0$), the trajectories of attention are dramatically changed. Without any top-down influence ($\beta = 0$) (a), the attentional allocation is only based on the bottom-up information and is less constrained by a given stimulus. However, when top-down constraints are imposed on the task, the attentional allocation is guided by a cue (red color) (b), (c), and (d). (a) $\alpha = \mathbf{1.0}$, $\beta = \mathbf{0.0}$; (b) $\alpha = \mathbf{0.6}$, $\beta = \mathbf{0.4}$; (c) $\alpha = \mathbf{0.5}$, $\beta = \mathbf{0.5}$; (d) $\alpha = \mathbf{0.4}$, $\beta = \mathbf{0.6}$.

salient properties cause an involuntary or reflexive shift of attention. They claimed that the effect of top-down voluntary control is more pervasive, even in a case that stimuli easily pop-out, than that was previously thought. On the other hand, the selection of a target can also be biased by imposing constraints of top-down knowledge or cues on incoming bottom-up information. Classically, the role of cues has been investigated by Posner and his colleagues [56], [57] using a simple cue task. In a typical experiment, subjects were asked to detect a target as soon as possible. Prior to the target presentation, subjects were given a cue that indicated where a target would appear. The target would be presented either on the cued location (valid cue) or the opposite location (invalid cue). The results of this study showed that reaction time (RT) to the valid cue is faster than RT to no cue (a neutral condition), whereas RT to the invalid cue is slower than RT to no cue. In addition, Eriksen and Yeh [22] manipulated the degree of validity of a cue (i. e., the probability that a cue is valid) with different probabilities—40%, 70%, and 100%. They showed, for a valid case, RT decreased as the probability of a cue's validity increased, while for the invalid case, RT increased as the probability increased. Similarly, Downing and Pinker [18] showed that the RT increased rapidly as a target appeared *further away* from the cued location.

The interesting aspect of such cue experiments, as related to our model, is the manipulation of the influence of a cue by varying the probability of a cue's validity and distance between a cued location and a target's location. In our model, the top-down influence is achieved in two ways—manipulation of the portion of β toward α and the Gaussian distance measure between a location of a cue color segmented region and a target candidate's location. In some sense, the constants α and β can be considered as the strategic allocation of attention between bottom-up-based locations and cued locations. That is, if a cue is reliable (i.e., highly probable) indicating where a target will appear, a higher value β is useful to quickly detect a target. However, this is not helpful if a cue is unreliable (i.e., less probable) for the target, so the system should respond more on the basis of bottom-up information to be optimal.

Neuroimaging studies also provide insights that link brain areas and attentional function. Two segregated neural systems have been identified for goal-directed and stimulus-driven attention in the brain using various brain imaging techniques. In particular, Hopfinger *et al.* [29] showed that distinct networks were engaged by attention directing cues versus subsequent targets, using functional magnetic resonance imaging (fMRI). A network of cortical areas including superior frontal, inferior parietal and superior temporal brain regions were involved in top-down attentional control in response to an instructive cue. This control then biased activity in multiple visual cortical areas, resulting in selective sensory processing of relevant

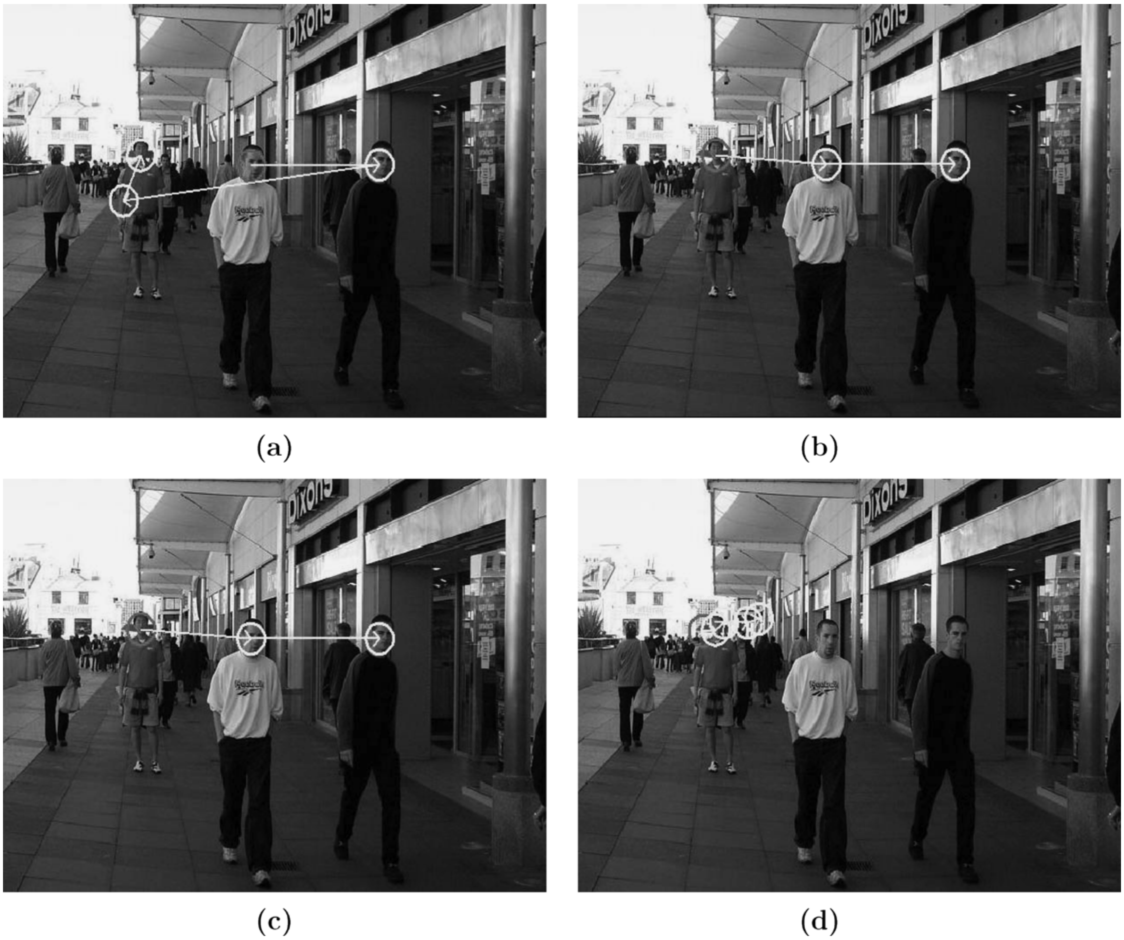

(a) (b) (c) (d)

Fig. 10. Another example of attention trajectories varied by different values α and β. Even a small amount of β is enough to change the trajectory of attention. Adding more to β produced a strong tendency that the trajectories of attention are attracted closely to the location of the region whose color matched the cued color. (a)$\alpha = \mathbf{1.0}$, $\beta = \mathbf{0.0}$. (b) $\alpha = \mathbf{0.6}$, $\beta = \mathbf{0.4}$. (c) $\alpha = \mathbf{0.5}$, $\beta = \mathbf{0.5}$. (d) $\alpha = \mathbf{0.4}$, $\beta = \mathbf{0.6}$.

visual targets. For example, with faces, Wojciulik *et al.* [72] demonstrated that face-specific neural response can be modulated by covert visual attention. In their experiment, subjects viewed brief displays each comprising two peripheral faces and two peripheral houses, and were asked to perform a matching task on either the two faces or the two houses without eye movement. They found that there was significantly less activity in a predefined face-selective region of interest (ROI) within human fusiform gyrus when faces were outside rather than inside the focus of attention.

B. Winner-Take-All Versus Attentional Modulation

The concept of competition is embedded in a winner-take-all (WTA) network in which units are mutually interconnected and are inhibited by each other [31], [52], [67]. Only one neuron corresponding to a location or an object in a given stimulus is selected at a time. Therefore, it can provide a unique solution at a given time by selecting the most salient location and suppressing all the others. To decide the next salient points, the suppressed units have to be revived and the previous winning point is excluded. Again, the competition occurs among the remaining neurons. This kill-and-revive scheme seems to be computationally expensive. Behind of the scheme, the limited processing capacity assumption takes an important rationale to explain why neurons inevitably compete with each other. That is, limited computational resources at a given processing stage are available for only a few neurons, so that neurons are forced to compete with each other in order to be selected. Ironically, this logic can be applied to the necessity of cooperation. That is, the limited computational resources may also require the cooperation of different brain areas or neural channels which may help to reduce the burden of processing in various ways. In Desimone and Duncans term, this cooperation means the "bias" that differentiates neural activities before or at least during competition [17]. This bias provides general criteria for what, or where, is selected in a task. That is, competition itself does not tell us what is relevant or not to a current behavior goal. Even though we do not directly implement WTA mechanism in our model for enhancing or suppressing neuronal activities, the second term in (9) correlates bottom-up and top-down inputs, and consequently it facilitates and suppresses neuronal activity according to the consistency between bottom-up and top-down information. Concerning the modulation effect itself, some issues have been controversial in the psychology and neuroscience communities. One question is whether the facilitation and inhibitory effects in the modulation of attention are caused by a single mechanism. The early explanation of these modulatory effects assumed that two different mechanisms may be involved because the amount of facilitation and inhibitory effect on RT or event related potential (ERP) is asymmetric [27], [56]. However, using a similar neural

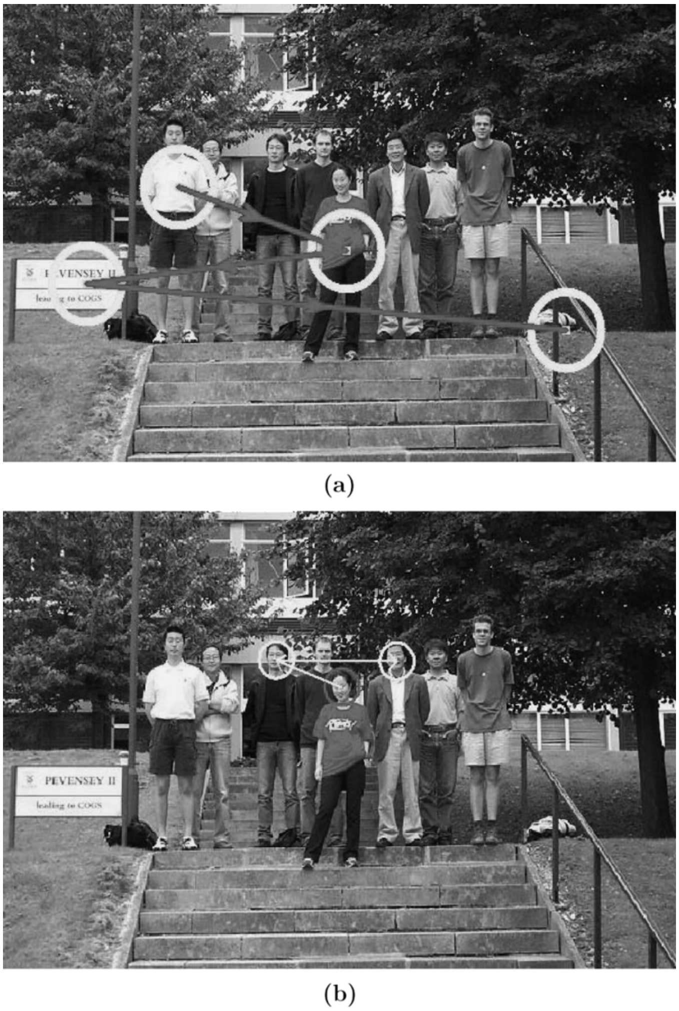


Fig. 11. Comparison with a saliency-based model. (a) The trajectory of attention obtained from Itti's model. (b) The trajectory of attention obtained from our model. The trajectory in (a) was biased on basis of saliency driven by only bottom-up features. In contrast, the trajectory in (b) was driven by the cooperation between bottom-up features as well as top-down task demand.

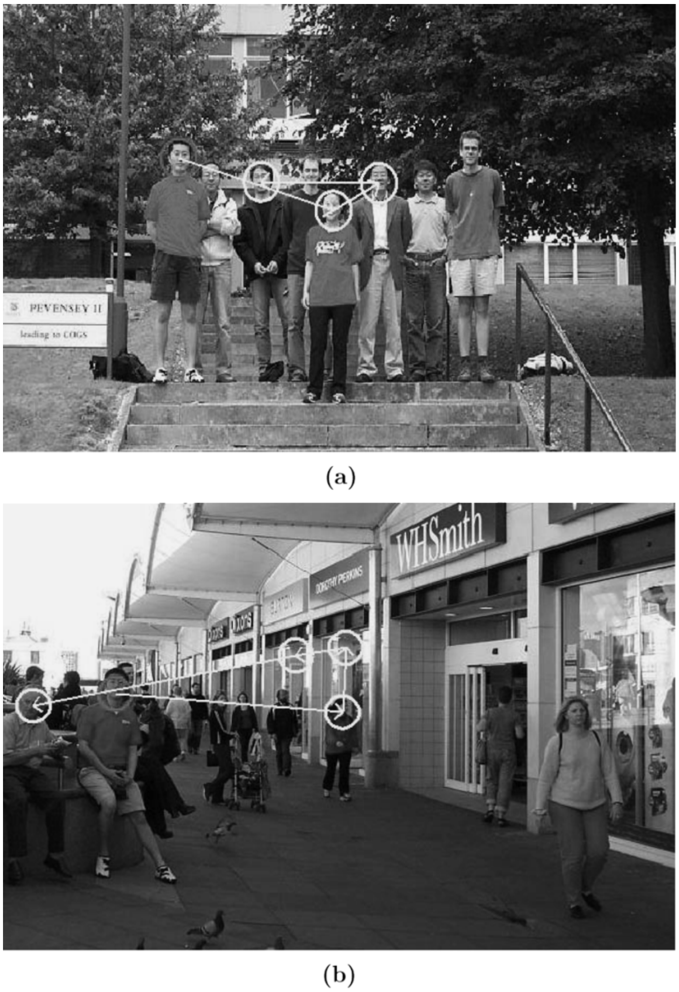


Fig. 12. Attention trajectories with different numbers of objects that share the same color attribute to a color cue. (a) Two objects share the same color attribute. (b) Many objects share the same color attribute. The search task for the system is to find a man who is wearing a red t-shirt.

network to ISNN, we recently showed the asymmetry may originate in the nonlinearity of neural activity [45]. We also suggested that this may be a common mechanism when more than two information sources are integrated such as in the Stroop effect [46], global-to-local interaction [51], as well as in the top-down cue effect [56]. In addition, the modulation effect suggests that the interaction between bottom-up and top-down information may be multiplicative, rather than a simple WTA mechanism. Neurophysiological and computational studies at a neuronal level showed that the response of a cell such as a pyramidal cell in cortex is dynamically modulated by AND-like or multiplicative operation that integrates incoming information from an early stage and feedback from a late stage, and that enhance neuronal responses to the attended location or relevant information, while suppressing those to unattended locations or irrelevant information, in a nonlinear manner [41], [61], [63], [66].

C. Early Versus Late Selection

The question "at which stage of visual information processing does attentional selection occur" has been debated in terms of supporting two different answers—early or late. According to the early selection view [6], the selection may occur early in the sequence of information processing stages, before identification of the stimulus and any semantic content, so that attention acts like a filter that works on low level physical properties such as color, orientation, etc. On the other hand, the late selection view asserts that selection may occur based on higher level content such as semantic information [53].

Even though the debate still has not been settled, some neuroscience studies shed light on the problem. Those studies indicate the brain regions of attentional modulation can extend from a very early stage (V1 and V2 [5], [7]) of ventral and dorsal pathways to a higher stage (parietal and inferior temporal area [9], [71]). However, more interestingly, the magnitude of attentional modulation increases as one moves up the cortical hierarchy.

In several computational models top-down influence on visual selection is accomplished in a manner that knowledge of a target leads to the level of feature map by applying some gains or gating networks [1], [4], [49], [73]. For example, in the Guided Search model, gain from the knowledge of target features such as red color, horizontal orientation, etc. is directly given to each

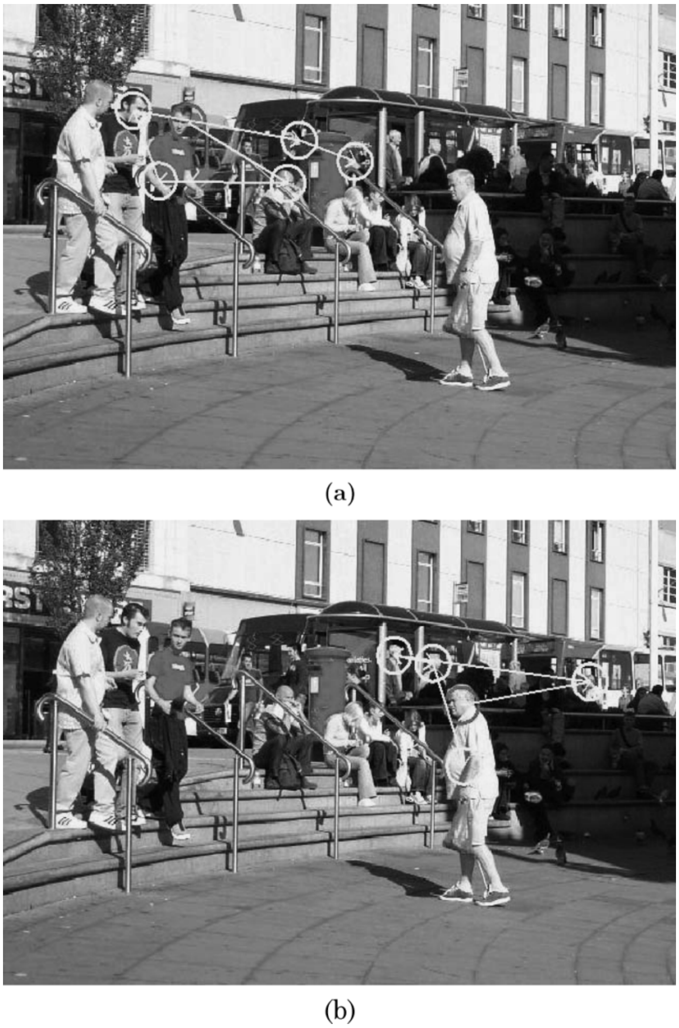


Fig. 13. Attention trajectories guided by different color cues. (a) A red color cue: the task is to find a man who is wearing a red t-shirt. (b) A blue color cue: the task is to find a man who is carrying a blue plastic bag.

corresponding feature map if bottom-up features are not sufficiently salient, and the activation driven from bottom-up processing is summed up with the top-down influence [73]. Therefore, in those models, the gain from top-down knowledge acts like a filter that works on the early stage of feature extraction. However, in our model, the top-down influence is not directly given to the feature level. The attentional modulation occurs at a late stage after feature extraction, where the form of information is more abstract rather than physical.

There is an advantage to top-down influence at an early stage, in that the influence increases feature contrast and, thus, eases the task of finding a target [33]. However, it is also possible to mislead a system in this way, by enhancing distractions that share similar feature properties. In contrast, for late selection, the form of the information is becoming more abstract in terms of size, rotation, translation, etc., and, thus, the top-down influence is less likely to affect such distractions, which share properties only at the feature level.

D. Spatial-Based Versus Object-Based Selection

Depending on "what is selected," the computational models of visual attention can be divided into two groups—spatial and object-based models. Most of the computational models (all of the current models except one (shown in the following), according to Sun and Fisher [65]) are based on space. For example, the saliency-based model proposed by Itti and Koch [32], [35] allocates an attentional window on the saliency map in which each coordinate is retinotopographically represented. That is, what is selected in their algorithm through a series of preprocessing stages is the "location" of saliency. The problem of this kind of scheme is that the location does not correspond to objects or meaningful regions in a visual scene. That is, in those space-based models, space does not mean the space that is occupied by an object, rather it is a salient "point" or "pixel," and exhaustively search a possible target location in a point-by-point (or pixel-by-pixel) way even though far fewer objects are contained in an image scene. In contrast, the object-based model proposed Sun and Fisher [65], the selection is accomplished a grouped features that may correspond to an object or its parts. They argued that object-based attention has advantage that space-based model does not have: 1) more efficient visual search; 2) less chance to select a nonsense or empty location; 3) naturally hierarchical selectivity. It showed more interesting and dynamic properties that have not been shown by space-based models. Attention in their model moves from the highest grouping salient point at a course scale to its salient parts at a finer scale according to a structured relation. However, parts of their algorithm were not completed—especially, they did not implement the segmentation process that is the core for the grouping process. Moreover, if we apply more strict criteria, that were shown in psychological tasks, in which objects are overlapped in the same place, attention is switched from one object to another, it is still questionable whether the model can carry out the task or not.

In comparison with both the point-by-point-based model and the object-based model, our model can be characterized as a region-based model. Attention is allocated at a segmented region where contains face-like features such as skin-color, symmetry, ellipse shape. The segmented region does not always contain a face and is not analyzed in a structured manner (e.g., from face to eyes). However, the attended regions are more meaningful in the statistical sense—a segmented region is more likely face-like than its background. Since far less number of target candidates were obtained through preprocessing stages, search in the model is more efficient than that in point-by-point search scheme.

In fact, spiking models developed by Wang [70] also belong to the group of object-based models. In his model, the networks can select several largest objects, which then alternate in time. Wang's model does not have the problems with the model proposed by Sun and Fisher [65]. The difference between our model and Wang's model lies in the fact that our model dynamically relies on the cue or top-down input rather than an object size to locate our attention. However our model and Wang's model share some commonalities. For example, synchronized oscillations are explicitly built in his model, we employ the positive correlation between the top-down and bottom-up inputs. As we all know, synchronized oscillations are special cases of positive correlations [23]. Of course, in the future developments, it would be very interesting to directly implement more biologically realistic mechanisms such as synchronized oscillations employed in Wang's model rather than correlations in our current model.

VI. Conclusion

In conclusion, we have developed a novel computational model of selective attention, based on an ISNN, for a visual search task in the context of face detection. Unlike a saliency-based model, attention in our model is guided by a top-down cue that is integrated with bottom-up information, and biases bottom-up processing toward the current task. The results from a series of simulations showed the dynamic change of attentional trajectories in accordance with strategic allocations of modulatory parameters α and β that affect the influence of the top-down cue, as well as dynamic effects due to variations of the cues themselves.

These results are compatible with psychological and neuroscience evidence as discussed previously. However, the model is limited to the "where task," and to the influence of top-down cues rather than language instruction, real world knowledge of goals, etc. These would require further extensions [50] but in principle could still use the same underlying ISNN model with modification of the top-down module. It is necessary to combine this with the "what task" for a recognition system in order to fully describe a system for selective attention that is useful to the computer vision community. This involves issues of object-based attention [33], [69] and will be a challenge for our further work.

References

[1] S. Ahmad and S. M. Omohundro, "Efficient visual search: a connectionist solution," in *Proc. 13th Annu. Meeting Cognitive Science Soc.*, Chicago, IL, 1991, pp. 52–53.

[2] C. Balkenius, "Attention, habituation and conditioning: toward a computational model," *Cogn. Sci. Quart.*, vol. 1, no. 2, 2000.

[3] S. Becker and G. E. Hinton, "Learning mixture models of spatial coherence," *Neural Computat.*, vol. 2, pp. 267–277, 1993.

[4] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1999, pp. 67–92.

[5] J. A. Brefczynski and E. A. DeYoe, "A physiological correlate of the 'spotlight' of visual attention," *Nature Neurosci.*, vol. 2, pp. 370–374, 1999.

[6] D. E. Broadbent, *Perception and Communication*. New York: Pergamon, 1958.

[7] C. Büchel, O. Josephs, G. Rees, R. Turner, C. D. Frith, and K. J. Friston, "The functional anatomy of attention to visual motion: A functional MRI study," *Brain*, vol. 121, pp. 1281–1294, 1998.

[8] H. Chang and U. Robles. (2000, May) Face Detection. [Online]http://ise0.stanford.edu/class/ee368a-proj00/project16/main.html

[9] L. Chelazzi, J. Duncan, E. K. Miller, and R. Desimone, "Responses of neurons in inferior temporal cortex during memory-guided visual search," *J. Neurophysiol.*, vol. 80, pp. 2918–2940, 1998.

[10] M. M. Chun, "Context cueing of visual attention," *Trends Cogn. Sci.*, vol. 4, pp. 170–178, 2000.

[11] M. M. Chun and Y. Jiang, "Contextual cueing: implicit learning and memory of visual context guides spatial attention," *Cogn. Psychol.*, vol. 36, pp. 28–71, 1998.

[12] ——, "Top-down attentional guidance based on implicit learning of visual covariation," *Psycholog. Sci.*, vol. 10, pp. 360–365, 1998.

[13] M. M. Chun and J. M. Wolfe, "Visual attention," in *Blackwell Handbook of Perception*, J. M. Goldstein, Ed. Oxford, U.K.: Blackwell, 2001, pp. 272–310.

[14] J. J. Clark and N. Ferrier, "Modal control of an attentive vision system," in *Proc. Int. Conf. Computer Vision*, Tarpon Springs, FL, 1988, pp. 514–523.

[15] M. Corbetta and G. L. Shulman, "Control of goal-direct and stimulus-driven attention in the brain," *Nature Neurosci. Rev.*, vol. 3, pp. 201–215, 2002.

[16] P. Dayan, S. Kakade, and P. R. Montague, "Learning and selective attention," *Nature Neurosci. Rev.*, vol. 3, pp. 1218–1223, 2000.

[17] R. Desimone and J. Duncan, "Neural mechanism of selective attention," *Annu. Rev. Neurosci.*, vol. 18, pp. 193–222, 1998.

[18] C. J. Downing and S. Pinker, "The spatial structure of visual attention," in *Attention and Performance*, M. I. Posner and O. S. M. Marin, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1985, pp. 160–174.

[19] J. Duncan, "Selective attention and the organization of visual information," *J. Exp. Psychol.: General*, vol. 113, pp. 501–517, 1984.

[20] J. F. Feng, Ed., *Computational Neuroscience: A Comprehensive Approach*. Boca Raton: Chapman & Hall/CRC, 2003.

[21] C. W. Eriksen and J. D. James, "Visual attention within and around the field of local attention: a zoom lens model," *Perception Psychophys.*, vol. 40, no. 4, pp. 225–240, 1986.

[22] C. W. Eriksen and Y.-Y. Yeh, "Allocation of attention in the visual field," *J. Exp. Psychol.: Human Perception and Performance*, vol. 11, pp. 583–597, 1985.

[23] J. F. Feng and D. Brown, "Impact of correlated inputs on the output of the integrate-and-fire models," *Neural Computat.*, vol. 12, pp. 671–692, 2000.

[24] J. F. Feng, Y. L. Sun, G. Wei, and H. Buxton, "Training the integrate-and-fire model with the informax principle II," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 326–336, Mar. 2003.

[25] V. Govindaraju, "Locating human faces in photographs," *Proc. Int. J. Computer Vision*, vol. 19, no. 2, pp. 129–146, 1996.

[26] H. Vasilaki, J. F. Feng, and H. Buxton, "Temporal album," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 439–443, Mar. 2003.

[27] S. A. Hillyard and L. Anllo-Vento, "Event-related brain potentials in the study of visual selective attention," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 781–787, 1998.

[28] E. Hjelmas and B. K. Low, "Face detection: a survey," *Comput. Vis. Image Understanding*, vol. 83, pp. 236–274, 2001.

[29] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The neural mechanisms of top-down attentional control," *Nature Neurosci.*, vol. 3, pp. 284–291, 2000.

[30] G. W. Humphreys and V. Bruce, *The Psychology of Attention*. Hillsdale, NJ: Lawrence Erlbaum, 1989.

[31] G. Indiveri, "A neuromorphic VLSI device for implementing 2-D selective attention," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1455–1463, Nov. 2001.

[32] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, pp. 1489–1506, 2000.

[33] ——, "Computational modeling of visual attention," *Nature Neurosci. Rev.*, vol. 21, pp. 314–329, 2001.

[34] ——, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 21, pp. 314–329, 2001.

[35] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[36] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE Press, 1999, pp. 274–280.

[37] N. Kanwisher and E. Wojciulik, "Visual attention: insights from brain imaging," *Nature Neurosci. Rev.*, vol. 1, pp. 91–100, 2000.

[38] J. Kay and W. A. Phillips, "Activation functions, computational goals and learning rules for local processors with contextual guidance," *Neural Computat.*, vol. 8, pp. 895–910, 1997.

[39] H.-S. Kim, W.-S. Kang, J.-I. Shin, and S.-H. Park, "Face detection using template matching and ellipse fitting," *IEICE Trans. Inf. Syst.*, vol. 11, pp. 2008–2011, 2000.

[40] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford Univ. Press, 1999.

[41] C. Koch and T. Poggio, "Multiplying with synapses and neurons," in *Single Neuron Computation*, T. McKenna, J. Davis, and S. F. Zornetzer, Eds. New York: Academic, 1992, pp. 315–345.

[42] C. Koch and S. Ullman, "Shifts in selective visual attention: toward the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, 1985.

[43] P. Kovesi, "Symmetry and asymmetry from local phase," in *Proc. 10th Australian Joint Conf. Artificial Intelligence*, 1997, pp. 15–20.

[44] A. F. Kramer and S. Hahn, "Splitting the beam: distribution of attention over noncontiguous regions of the visual field," *Psycholog. Sci.*, vol. 6, pp. 381–386, 1995.

[45] K. W. Lee, J. F. Feng, and H. Buxton, "A dynamic neural network model on global-to-local interaction over time course," in *Proc. 9th Int. Conf. Neural Information Processing*, Singapore, 2002, pp. 17–21.

[46] C. M. MacLeod and P. A. MacDonald, "Inter-dimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention," *Theor. Comput. Sci.*, vol. 4, pp. 383–391, 2000.

[47] E. A. Maylor, "Facilitatory and inhibitory components of orienting in visual space," in *Attention and Performance 11*. Hillsdale, NJ: Lawrence Erlbaum, 2002, pp. 189–204.

[48] B. Menser and M. Brunig, "Segmentation of human face in color images using connected operators," in *Proc. Int. Conf. Image Processing*, 1999, pp. 533–536.

[49] M. C. Mozer and M. Sitton, "Computational modeling of spatial attention," in *Attention*, H. Pashler, Ed. London, U.K.: Univ. College London Press, 1998, pp. 341–393.

[50] V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," in *Biologically Motivated Computer Vision*, H. H. Bulthoff, S.-W. Lee, T. A. Poggio, and C. Wallraven, Eds. New York: Springer-Verlag, 2002, pp. 453–461.

[51] D. Navon, "Forest before trees: the precedence of global features in visual perception," *Cogn. Psychol.*, vol. 9, pp. 353–383, 1977.

[52] E. Niebur, L. Itti, and C. Koch, "Controlling the focus of visual selective attention," in *Models of Neural Networks 4*. New York: Springer-Verlag, 2002, pp. 247–276.

[53] D. A. Norman, "Toward a theory of memory and attention," *Psycholog. Rev.*, vol. 75, pp. 522–536, 1968.

[54] H. Pashler, J. C. Johnston, and E. Ruthruff, "Attention and performance," *Annu. Rev. Psychol.*, vol. 52, pp. 629–651, 2001.

[55] G. A. Patel and K. Sathian, "Visual search: bottom-up or top-down?," *Frontiers in Biosci.*, vol. 5, pp. 169–193, 2000.

[56] M. I. Posner, M. J. Nissen, and W. C. Ogden, "Attended and unattended processing modes: the role of set for spatial location," in *Modes of Perceiving and Processing Information*, H. L. Pick and I. J. Saltzman, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1977, pp. 160–174.

[57] M. I. Posner, C. R. R. Snyder, and B. J. Davidson, "Attention and the detection of signals," *J. Exp. Psychol.: General*, vol. 109, pp. 160–174, 1980.

[58] D. Rademacher. (2001, Dec.) Face detection. [Online]http://egweb.mines.edu/eges512/projects/face/Rademacher.PDF

[59] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context free attentional operator: the generalized symmetry transform," *Int. J. Comput. Vision*, vol. 14, pp. 119–130, 1995.

[60] A.-M. Schuller and B. Rossion, "Spatial attention triggered by eye gaze increases and speed up early visual activity," *Neuroreport*, vol. 12, pp. 2381–2386, 2001.

[61] M. Siegel, K. P. Körding, and P. König, "Integrating top-down and bottom-up sensory processing by somato-dendritic interactions," *J. Computat. Neurosci.*, vol. 8, pp. 161–173, 2000.

[62] S. A. Sirohey, "Human Face Segmentation and Identification," AU: WHAT DEPARTMENT?, Univ. Maryland, College Park, MD, Tech. Rep. CS-TR-3176, 1993.

[63] M. W. Spratling, "Cortical region interactions and the functional role of apical dendrites," *Behav. Cogn. Neurosci. Rev.*, vol. 1, no. 3, pp. 219–228, 2002.

[64] R. B. Stein, "Some model of neural variability," *Biophys. J.*, vol. 7, pp. 37–68, 1967.

[65] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artif. Intell.*, vol. 1, pp. 77–123, 2003.

[66] S. Treue, "Neural correlates of attention in primate visual cortex," *Trends Neurosci.*, vol. 24, no. 5, pp. 295–300, 2001.

[67] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nufl, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, pp. 507–545, 1995.

[68] P. van de Laar, T. Heskes, and S. Gielen, "Task-dependent learning of attention," *Neural Netw.*, vol. 10, pp. 981–992, 1997.

[69] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch *et al.*, "Attentional selection for object recognition-a gentle way," in *Biologically Motivated Computer Vision*, C. Bulthoff *et al.*, Eds. New York: Springer-Verlag, 2002, pp. 472–479.

[70] D. L. Wang, "Object selection based on oscillatory correlation," *Neural Netw.*, vol. 12, pp. 579–592, 1999.

[71] E. Wojciulik and N. Kanwisher, "The generality of parietal involvement in visual attention," *Neuron*, vol. 23, pp. 747–764, 1999.

[72] E. Wojciulik, N. Kanwisher, and J. Driver, "Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study," *J. Neurophysiol.*, vol. 79, no. 3, pp. 1574–1578, 1998.

[73] J. M. Wolfe, "Guided search 2.0: a revised model of visual search," *Psychonom. Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[74] ——, "Visual attention," in *Seeing*, K. K. De Valois, Ed. San Diego, CA: Academic, 2000, pp. 335–386.

[75] M.-H. Yang, D. Kriegman, and N. Ahuja, "Face detection in images: a survey," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

KangWoo Lee received the Ph.D. degree from the Department of Informatics, Sussex University, Brighton, U.K.

He is currently a Postdoc in Korea.

Hilary Buxton is currently a Professor of informatics at Sussex University, Brighton, U.K. Her research interest is in cognitive vision.

Jianfeng Feng is a Professor in the Center for Scientific Computing and Computer Science, Warwick University, Coventry, U.K. His research interest is spiking neural networks and their applications.