Lee KW. Guiding attention by cooperative cues. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 23(5): 874–884 Sept. 2008

Guiding Attention by Cooperative Cues

KangWoo Lee

School of Media, Soongsil University, 1-1, Sangdo-5 Dong, Dong-Jak Ku, Seoul, 156-743, Korea

E-mail: kangwooster@gmail.com

Received February 14, 2007; revised May 10, 2008.

Abstract A common assumption in visual attention is based on the rationale of "limited capacity of information processing". From this view point there is little consideration of how different information channels or modules are cooperating because cells in processing stages are forced to compete for the limited resource. To examine the mechanism behind the cooperative behavior of information channels, a computational model of selective attention is implemented based on two hypotheses. Unlike the traditional view of visual attention, the cooperative behavior is assumed to be a dynamic integration process between the bottom-up and top-down information. Furthermore, top-down information is assumed to provide a contextual cue during selection process and to guide the attentional allocation among many bottom-up candidates. The result from a series of simulation with still and video images showed some interesting properties that could not be explained by the competitive aspect of selective attention alone.

Keywords selective attention, cooperation, competition, cooperative cues, guidance

1 Introduction

Due to its intricate and manifold nature, visual attention has been investigated by a wide spectrum of approaches that lead to controversial issues. There is still much debate on where the selection process occurs in information processing stages (early vs. late selection), the location of attention (spatial vs. object), the direction of information flow (bottom-up vs. top-down), etc[1,2].

Regardless of the controversies surrounding each issue, the above problems are all linked by a common assumption of why attention is needed. The common assumption of the necessity of attention is the limited amount of computational resource that is available for a given task or process. The basic purpose of attention is to avoid possible information overload in order to protect a mechanism of limited capacity[3]. The necessity of attention to overcome resource limitation becomes clear if we consider the analogy between a computer with limited capacity and its use of resources in presence of a huge amount of input data, not all of which is relevant to a current task. If the system can selectively process a small portion of information that is relevant to a current task, it can increase the efficiency of processing and prevent a breakdown caused by an overload in memory or long processing time.

Computational models for selective attention have grown on the soil of this diversity during the last decade. Many of them are inspired by the idea "process only a small portion of selected information". This idea is very attractive for computational modelers because it may ultimately improve computational efficiency and thus, allow a system to deal with real world problems. In spite of some success of current computational models, they all have fundamental limitations when it comes to explaining the richness of selective visual attention. In particular, most computational models fail to describe how functional modules are coordinated, how top-down information is utilized to select incoming visual information, how attention modulates a visual stimulus from the external world, and how a system shows differentially biased behavior to the current information in the context of different tasks.

This paper is devoted to a computational account of how knowledge of a cue works with incoming information. In this approach, visual attention is considered as an integrative process influenced by both bottom-up and top-down processing mechanisms. The interaction between these two mechanisms is critically important to understand attentional behaviors that are biased with respect to a particular object or spatial location. In the following section, we provide an introductory review of some theoretical and computational issues on selective attention. Second, a spiking neural network called Interactive Spiking Neural Network (ISNN) is explained

Regular Paper

and the process of integrating bottom-up and top-down information using interactive activation rule is presented along with the process of constructing the integration map. This is followed by a computational model of selective attention guided by contextual cues. Fourth, simulation results with various cue conditions including point, color, and motion are presented. Finally, a short summary is given at the end of this paper.

2 Reviews on Related Work

2.1 Limited Capacity and Competition

The assumption of limited capacity was originally conceptualized by Broadbent[3]. In his theory, which is known as filter theory, only a small portion of the incoming information is passed through selective filter and the rest of information is shut out from further analysis. From the limited resource assumption, Desimone and Duncan[4] suggested an influential theory of visual attention on the basis of behavioral and neural studies. According to their study, the receptive field (RF) can be viewed as a critical visual processing resource for which objects in the visual filed must compete because the information available about any given object will decline as more objects are added to RFs. Therefore, a cell's activity is degraded if more than one stimulus falls into RF in comparison with the activity evoked by a single stimulus presented within RF. Furthermore, Kastner *et al.*[5] argued that multiple objects in a restricted RF interact in a mutually suppressive way such that the pertinent neurons try to get hold of required processing resource. This competition in the presence of multiple objects results in the degrading responses of the cell.

A possible solution to the degraded neural activity is to selectively enhance an object and suppress others through the competition. However, cells in a competition are not equally selectable; some cells are more likely to win while others are not. The competition is biased towards information that is currently relevant to behaviors.

2.2 Competition in Computational Models

This subsection introduces computational models of selective attention in terms of the flow of information processing and selection process. These two dimensions shed light on how competitive mechanism is implemented in computational models.

2.2.1 Processing Stages in Selective Attention

Roughly speaking, any computational model for visual attention has two distinctive processing stages: preattentive and attentive stages. This distinction is due to the assumption of resource limitation.

In preattentive processing, 3 assumptions are commonly made in many computational models: 1) preattentive processing is unlimited in capacity; 2) information is processed in a bottom-up and massively parallel manner; and 3) information processing is independent. Therefore, for a given stimulus, different features such as color, intensity, orientation, and movement, are extracted by different processing channels in a parallel way as in, for example, Itti's saliency-based model[6,7].

At this stage, two different mechanisms are widely used for processing a given features. First, a multi-resolution mechanism is used for obtaining an image representation from a coarse spatial scale to a finer spatial scale, with the zoom lens metaphor embedded in the mechanism[8]. The information carried by different spatial scales can be used for different purposes. In Deco's model[9], the coarsest level of spatial resolution is utilized to find the location of an interesting object in a priority map, whereas detailed spatial resolution is used for identifying what object is. Second, a center-surround mechanism is used for achieving the contrast within a channel. In computer vision, this mechanism is widely used for detecting local edge in an image. In general, it is assumed that there is homogeneity within an object or a part of an object and discontinuity between objects or parts when detecting a local edge. The homogeneous parts of the image nullify the response of a center-surround filter. Conceptually, the center-surround mechanism for edge-detection is the same for bottom-up saliency detection in which attention is directed to a unique object among similar objects.

The preattentive stage is followed by an attentive process that can be characterized by a serial process in which only one item is processed at a time. In this stage, features obtained from different channels are combined to construct a saliency map. Even though saliency can be defined at many different levels from a feature to a semantic level, saliency in most current models is defined at the feature level. The important factor in guiding bottom-up attention is feature contrast rather than absolute feature values.

Once a saliency map has been constructed, a location has to be selected for the deployment of an attentional window. A "winner-take-all" (WTA) network is commonly used to allocate attention[6,7,10,11]. In the network that receives input from a saliency map, only one unit is allowed to be active at a given time for serial processing, and the others are suppressed. In other words, biased competition is accomplished through the WTA network. In order to prevent reallocation of attention to this winning location, it is excluded from the

saliency map after processing.

2.2.2 Selection Process in Attention

Depending on where selection process occurs, and which level of information is selected, computational model can be discriminated into two classes of models — early and late selection models. Most current computational models are based on early selection. Since selection is accomplished by the saliency calculated from the center-surround feature contrast, the selected location does not meaningfully correspond to the location of an object. Rather, it simply corresponds to the location where it gives the strongest contrast. Furthermore, top-down knowledge is directed to the early stage of information processing, immediately, before or after the feature extraction process[12]. In contrast, a few models have implemented a late selection. For instance, in Sun and Fisher's model, the feature elements such as color, intensity, and orientation are grouped into more meaningful perceptual units (objects) before attentive process operates[8,13].

Regarding the competition mechanism, the early vs. late selection has implications for important issues. If the RF were viewed as a critical visual processing resource for which objects in the visual field must compete, the RF property of increasing size along the visual pathways implies: 1) a cell in the higher processing stage has a relatively larger RF size with a greater chance of having more objects than the cell in a lower layer; 2) competition for processing resource becomes more intense in the upper ladder of the hierarchy; and thus 3) stronger attentional modulation effects will be found at the higher stages. From the above reasons, it can be induced that the attentional effect increases from a lower stage to higher stage, rather than being attenuated from a lower stage to a higher stage. Also, this means that the attention is object-based rather than feature-based.

2.3 Competition in Computational Models

The biased competition hypothesis implies that neurons at a given processing stage take part in an inevitable war for resources. The relationship among neurons is considered as mutually exclusive and there seems to be little chance for cooperation to solve the limited resource problem, since competition is the main mechanism of the selection process. As noted previously, the concept of competition is embedded in the WTA network in which units are mutually interconnected and are inhibited by each other. In those models, only one neuron corresponding to a location or an object in a given visual stimulus is selected at a time.

In spite of the fact that the limited resource assumption provides a logical basis for the inevitability of competition, the same logic can be equally applied to the necessity of cooperation. That is, the limited resource assumption may also require the cooperation of different brain areas or neural channels which may help to reduce the burden of processing in various ways. The cooperative information from other brain areas does not simply contribute to the enhancement or suppression of neural activities at a given processing stage. It provides general criteria for what or where are selected in a task. Top-down knowledge and contextual information provide critical criteria that allow a system to selectively process current information.

Neurophysiological evidence is given by Rainer *et al.*[14], who recorded cell activity in the prefrontal cortex of a monkey during a "delayed-matching-task". In the task, a monkey was required to find a target object in a stimulus scene containing many objects, and to remember its location until a test stimulus was given. They found that the activity of the neurons in the cortex reflected the target location alone and was maintained during the delay. This result suggested that the relevant neural activity corresponding to a target was maintained during the delay and was involved in the selection of the location where a matching object would be given. That is, remembering only target location (or cued location) overcomes the severe limitation of the capacity of working memory.

2.4 Integration of Cooperative Information

The argument "*cooperative mechanism of visual attention is critical for the selection process on current information*" raises another question — how does the cooperative mechanism work for selection of a location or object? Basically, we argue that the information from other processing channels provides a context or bias to the network that receives incoming neural activities. This context or bias helps the network to interpret the incoming neural signal by setting a criterion for determining whether the signal is relevant or irrelevant to the current behavior or information processing.

Recently, Spratling[15] investigated differentiated roles of apical and basal dendrites of a pyramidal cell. A typical pyramidal cell has two separate dendritic arbors that receive different information sources. This anatomical segregation of dendrites of a pyramidal cell may suggest that the dendrites receive information from two distinctive sources — feedforward information to the basal dendrite and feedback information to the

apical dendrite. Spratling[15] speculated that the distal and proximal dendrites of pyramidal cells act as separate compartments and contribute to different functional roles for information processing. Since apical inputs have weaker effects on the output activity than basal inputs, the apical dendrite is considered to take a role in modulating the responses of the cell. That is, for such neurons, sensory-driven, feedforward information is applied to the basal dendrite while top-down and feedback information arrives at the apical dendrites.

Interestingly, Treue *et al.*[16] showed that the attentional modulation effect on sensory selectivity is multiplicative. They measured the tuning curve of direction-selective neurons in the middle temporal (MT) visual areas of a monkey while the animal was attending to moving random dot patterns guided by a spatial cue. The result showed that attention increased the response to all attended stimuli by the same proportion ("multiplicative modulation") along the different degrees of the orientation, without narrowing the width of the tuning curve. Reynolds *et al.*[17] systemically investigated this relationship between the amount of attentional modulation and stimulus strength as they manipulated a range of luminance contrast. They showed that the attentional modulation effect on V4 neurons to a low contrast stimulus is larger than that to a high contrast stimulus. These results are compatible with our models in which there exists a multiplicative relationship between the two inputs from different information sources, and the attentional modulation effect varies with respect to the strength of the inputs.

3 Interactive Spiking Neural Network

3.1 3-Way Interaction Between Bottom-Up, Top-Down Input and Output Units

In attentional tasks developed by Posner and his colleagues[18,19], subjects in their experiments were asked to respond to the onset of a target light as soon as they detected it. A cue can be presented at one of several possible locations prior to a target appearing. This cue was given to inform where a target would appear. Experimenters manipulated the validity of the cue by intermittently showing the target at different location from the cue. That is, the target might appear at the location where a cue indicated. These studies showed that attention to a location guided by a top-down cue benefited the response for a target if it appeared at the same location indicated by the cue but induced the subjects to make an incorrect response if the cue was invalid.

From the experiments 3-way relationship between a target (bottom-up input in this model), a cue (top-down input) and a response (output) can be established. The response in the attentional task can be represented as a conditional probability of output y jointly given by the bottom-up stimulus B and the top-down stimulus T. Now we may ask "how can these two types of stimulus influence the responses in the attentional task?" Consider a typical "cueing task" in which a cue stimulus precedes a target stimulus and it provides the information where the target stimulus will appear. The target stimulus may appear at the location indicated by the cue or the opposite location. This simple task provides interesting insights on the question. First, the response is basically determined by a target stimulus because observers in the task should determine where the target stimulus is, not where a cue stimulus indicates. Second, we may say from the first fact that no response can be determined by a cue stimulus alone. Third, there is an interaction effect on the response from the cue and target stimulus. A response to a test stimulus is facilitated, when a cue consistently indicates where a target stimulus appear, whereas it is interfered when a cue inconsistently indicated in comparison with a neutral condition (no cue is given). That is, the two stimuli are correlated.

Based on the 3-way relationship between target, cue stimulus and response, a model of attentional output can be described from (1) as follows.

$$P(y|B,T) = f(B, g(B,T)). \tag{1}$$

The attentional response y is a function of the two terms driven by bottom-up stimulus B and the interaction between bottom-up and top-down stimulus $g(B,T)$. The interaction term can be thought as the modulatory effect that is added to the bottom-up driven output. The multiplicative operation was used to model the interaction between the bottom-up and top-down inputs. This correlates the two inputs and establishes a cooperative relationship between them.

3.2 Structure of ISNN

Based on this conceptual framework, we have developed an Interactive Spiking Neural Network (ISNN) using a leaky Integrate-and-Fire (IF) neural network[20−22]. A simple example of the structure of the ISNN is given in Fig.1. 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 two inputs at time t. The multiplication has an important non-linear property that correlates the two inputs. The net value for the j-th output unit is given

by:

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

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. The constants α and β determine the amount of influence driven by the bottom-up and top-down inputs on the net value. If $\alpha = 1$ and $\beta = 0$, the value net_j is determined by only the bottom-up inputs. If $0 < \alpha < 1$ and $\beta = 1 - \alpha$, the value net_j is determined by both inputs in varying proportion.

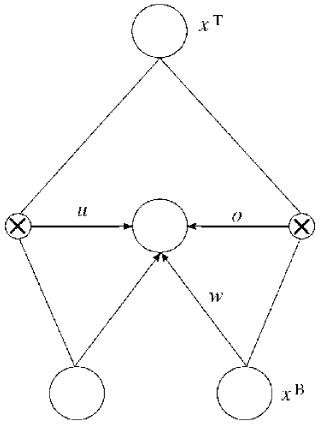

Fig.1. Simple example of ISNN structure. The model has bottom-up x^{B} and top-down input units x^{T}, and output units o. The bottom-up connection w links bottom-up units and output units, and the multiplicative connection u links the two input units and output units. Therefore, an output unit receives 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.

The second term in (2) 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 the soma. The sigmoid function has desirable properties; the output of the function will not be zero or one. So, the amount of activation driven by net_j is given by:

$$y_j(t) = \frac{1}{1 + \exp(-net_j(t))}. \tag{3}$$

In the IF model, a postsynaptic spike occurs if the summation of postsynaptic potential produced by the succession of input signals reaches a threshold[23]. Conventionally, the model is described with a circuit that consists of a capacitor C in parallel with a resistance R driven by a current $I(t)$. The trajectory of the membrane potential can be expressed in the following form.

$$V(t + \mathrm{d}t) = V(t) + RI(t)\frac{\mathrm{d}t}{\tau_{\mathrm{m}}} - V(t)\frac{\mathrm{d}t}{\tau_{\mathrm{m}}}, \tag{4}$$

where τ_{m} is the membrane time constant of a neuron. The equation means the membrane potential V at time $t + \mathrm{d}t$ 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 sigmoidal activation together:

$$V_j(t) = y_j \frac{1 - \exp(-n/k\tau_{\mathrm{m}})}{1 - \exp(-1/k\tau_{\mathrm{m}})}, \tag{5}$$

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 V_{th}.

$$V_j(t) \leqslant y_j \frac{1 - \exp(-n/k\tau_{\mathrm{m}})}{1 - \exp(-1/k\tau_{\mathrm{m}})}. \tag{6}$$

From the equation above, the interspike interval n/k is determined:

$$T_j = \frac{n}{k} = -\tau_{\mathrm{m}} \ln\Big[1 - \frac{V_{\mathrm{th}}(1 - \exp(-1/k\tau_{\mathrm{m}}))}{y_j}\Big]. \tag{7}$$

3.3 Learning Equation

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

$$E = \frac{1}{2}\Big(\sum_{j=1}^{l} T_j^{\mathrm{d}} - T_j\Big)^2, \tag{8}$$

where T_j^{d} is the j-th desired spike interval and 1 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.

$$\frac{\partial E}{\partial w_{ij}} = \eta\alpha(T_j^{\mathrm{d}} - T_j)\Big[\frac{\tau_{\mathrm{m}}}{y_j - V_{\mathrm{th}}(1-\exp(-1/k\tau_{\mathrm{m}}))}\Big] \Big[\frac{V_{\mathrm{th}}(1-\exp(-1/k\tau_{\mathrm{m}}))}{y_j}\Big] y_j(1-y_j)x_i^{\mathrm{B}}. \quad (9)$$

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

$$\frac{\partial E}{\partial u_{irj}} = \eta\beta(T_j^{\mathrm{d}} - T_j)\Big[\frac{\tau_{\mathrm{m}}}{y_j - V_{\mathrm{th}}(1-\exp(-1/k\tau_{\mathrm{m}}))}\Big] \Big[\frac{V_{\mathrm{th}}(1-\exp(-1/k\tau_{\mathrm{m}}))}{y_j}\Big] y_j(1-y_j)(x_i^{\mathrm{B}})^2 x_r^{\mathrm{T}}. \quad (10)$$

4 Model of Selective Attention

4.1 General Structure of Selective Attention Model

We introduce some assumptions that are used as a blueprint for constructing a computational model of selective attention. First, preattentive features such as color, shape or facial feature are treated as bottom-up information. Second, top-down information is utilized by establishing the relationship between a cue and preattentive features. Third, both bottom-up and top-down inputs are integrated to construct an integration map via ISNN. Fourth, the attentional allocation is ordered by the consistency of the two inputs. This means that the location that is the most activated by bottom-up input, and that is associated by a top-down cue, is the first to be visited with an attentional window.

Fig.2 shows the general structure of the selective model. The model can be divided into 3 main processing submodules — bottom-up processing module, bottom-up processing module and integration module. The original input is given in the form of digitized images. From the original input image some bottom-up features including skin color, facial features (aspect ratio and symmetry), and ellipse (or eigenface) shape are extracted to construct a bottom-up map. Combination of features into a single bottom-up map is accomplished by a set relation between feature regions defining whether features share a specific region or not. To

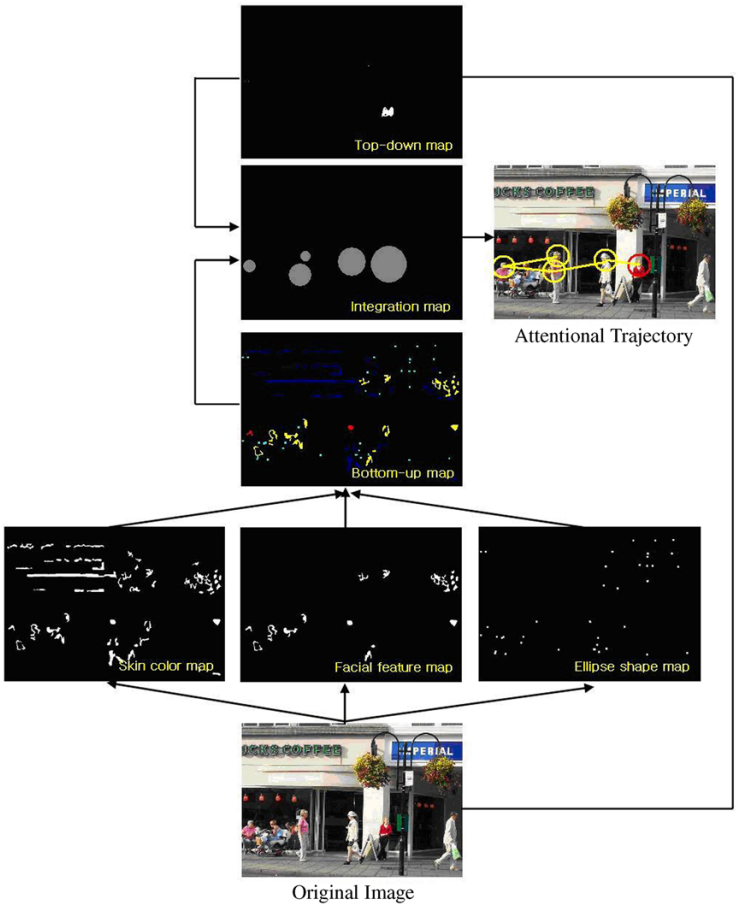

Fig.2. General model architecture of selective attention. The model consists of 3 different sub-modules that each forms a different map — bottom-up, top-down, and integration map.

extract top-down features, the location of a cue and its region are obtained from a series of segmentation processes. From the location of a cue segmented region, a top-down input value is calculated by measuring the distance between the locations of a cue segmented region and a target candidate, and is used for the subsequent integration processing. The integration processing module integrates both top-down input and bottom-up input, and produces an output in the form of an interspike interval. The attentional window is allocated in ascending order from the location with the shortest interspike interval to the location to the longest interspike interval calculated from ISNN. In a sense, the first location allocated by the attentional windows will have the highest consistency between the cue and the top-down features.

4.2 Bottom-Up Processing Module

For the first step of bottom-up processing, an input image is decomposed into 3 different sub-modules that extract the features of a target's features such as skin color, facial features, ellipse shape (or eigenface) (for a face searching task) or round shape (for ball searching task). Roughly speaking, these submodules can be considered as preattentive processing that occurs before the engagement of attention. As found in many computational models, we assume this bottom-up processing to take place across all submodules in parallel. In this subsection we only describe the submodules for searching a face, not searching a ball. However, the same algorithm can be modified and used for searing a ball.

4.2.1 Skin Color Map

To build a skin color model, we used normalized RGB since the common RGB representation of color images is susceptible to environmental illumination change. The RGB color coding is transformed as follows:

$$\begin{cases} r = \dfrac{R}{(R+G+B)}, \\ g = \dfrac{G}{(R+G+B)}, \\ b = \dfrac{B}{(R+G+B)}. \end{cases} \tag{11}$$

Since b does not contain significant information of skin color, it can be eliminated[24]. The probability of skin color for a pixel x was obtained by applying the normal distribution to the pixel using the following equation.

$$P(x) = \exp(-0.5((x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu))), \tag{12}$$

where μ and Σ are the mean and covariance of the given colors, respectively. Using this equation, the skin color segmentation was carried out by imposing threshold value to the probability. Then, we assigned the skin color feature value 1 to each of the segmented region.

4.2.2 Facial Feature Map

Skin color-based method will segment many parts of a body; segmentation was constrained to facial part by using two facial features, i.e., aspect ratio and symmetry. First, the aspect ratio between the width and height of a face is a unique feature that can be distinguished from other body parts. The golden ratio of a face has been calculated as $(1+\sqrt{5})/2$[25]. The aspect ratio of each segmented region was obtained after elongating and rotating the object to 90°. If the variation of the aspect ratio is within 1.2∼2.0 range, the feature value was set to 1; otherwise, it was set to 0.

Second, symmetry is an important feature that has been used to detect faces or parts of faces[26]. The symmetry feature extraction is based on the algorithm proposed by Kovesi[27]. The basic idea of his algorithm is that images contain symmetry whose periodic nature can be reflected as the maximum or minimum point in the frequency domain. The algorithm was applied at four different orientation angles: 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°. For all other cases, the symmetry feature value is set to 0.

4.2.3 Ellipse Shape Map

Another method to extract target features is based on the shape of the face contour[28,29]. Typically, the frontal view of a face has an oval or round shape. Using a predefined standard contour pattern, the correlation values within 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 orientation, etc. Nevertheless, it can provide partially useful information in the case where other features do not properly locate a target candidate. To do this, the original image is transformed into a gray scale image, and the gray image is down-sampled into 5 different levels of different sizes. The images are (vertically and horizontally) convolved by ellipse-shaped filters. The convolved images are combined into a single ellipse convolved image. Then, a threshold value is applied to segmenting possible face like objects. The separate feature maps are combined into a single bottom-up map in which each region contains the feature values described above. This

is accomplished by the set relationship among the segmented regions from the bottom-up feature maps. For simplicity, the bottom-up feature maps in binary forms are added up and the corresponding feature values are assigned to the center of the each overlapping region.

4.3 Top-Down Processing Module

In psychological or neuroscientific studies, many different kinds of signal or stimuli such as verbal instruction, color, arrow, shape etc. have served as a cue. In our simulation, color, motion and pointing cues are used for showing the cooperative aspects of selective attention. The prior knowledge of cue stimulus is assumed since the cue may provide the context of where attention should be allocated. That is, we strategically allocate our attention when we try to find a person in a crowded scene according to the prior knowledge such as meeting point, color of clothes, etc. This may lead to a further assumption that a target object is closely located with a given cue.

Based on the assumptions, the geometrical relationship between a target and cue is established after extracting cue features such as color, motion and pointing direction. For instance, the geometrical relationship between a face (target) and the color of clothes (cue) has "a-target-above-a-cue" not "a-target-below-a-cue".

To construct top-down map, the region of cue features is obtained from feature extraction algorithms such as color segmentation for color cue, temporal difference of frames for motion cue etc. The segmented regions of a top-down cue are located to the top-down map. After the location of cue features is obtained, the distances between the regions of a cue feature and a target candidate (obtained from bottom-up map) are calculated. The distances are normalized to range from 1.0 to 0.0, representing the shortest and the longest distance respectively, if the location of a target candidate satisfies the constraint of the geometrical relationship. Otherwise, the distance is set to 0. However, it should be noted that this geometrical relationship cannot be applied to the cases where motion or pointing cues are used. In these cases geometrical distances between the locations of cue features and target candidates are calculated without the constraint. The obtained distance values are used for top-down input of ISNN.

4.4 Constructing Integration Map

The inputs from both bottom-up and top-down maps are integrated through an ISNN that has been developed by consideration of the 3-way relationship between bottom-up, top-down inputs and output. At the location of each target candidate, the output can be obtained as two inputs are fed into the ISNN. Since an image may contain more than two locations of a cue feature (i.e., there can be two persons who are wearing yellow t-shirts in finding "a person who is wearing a yellow t-shirts"), the location of a target candidate may have more than two top-down input values calculated to each location of the feature. This may cause for an attention window to revisit the same location again. To prevent this, we choose only the biggest top-down input value, and use it as the top-down input for ISNN. After obtaining the output of ISNN, the integration map can be constructed using interspike interval at the locations of target candidates. The attentional window moves from the location with the shortest interspike interval to the location with the longest interspike interval in the ascending manner.

5 Simulations

A series of simulations have been done to investigate the performance of the model. Still and video images obtained from natural environments, such as a street, a campus and a laboratory, were used as input images. The first simulation was designed to compare the allocation of attention resulting from the WTA network with the allocation of attention from our model. The second simulation showed how a contextual cue guides the allocation of attentional windows. The third simulation illustrates that attentional trajectories are modulated by the interaction between bottom-up and top-down inputs. The final simulation concerned the attentional shift from one face to another and maintenance of attention on a target face guided by a motion cue.

5.1 Constructing Integration Map

As noted earlier, a WTA network is commonly used to determine the allocation of attention in many models. In the scheme of the allocation of attention used in a WTA network, it receives inputs from a saliency map driven by bottom-up features, and the location with the biggest output from a WTA network is selected. Then, the location with the second biggest output from the WTA network is selected. Therefore, the allocation of attentional window is totally determined by bottom-up features as shown in Fig.3. That is, the more face-like features the location of a target candidate has, the more possibility it has to be attended to. In contrast, our model utilized not only inputs driven bottom-up features, but also inputs driven from by top-down cue. Both inputs are cooperatively interacted, and thus the allocation of attention is guided by the top-down cue. Thus, the attention is allocated to the target candidates

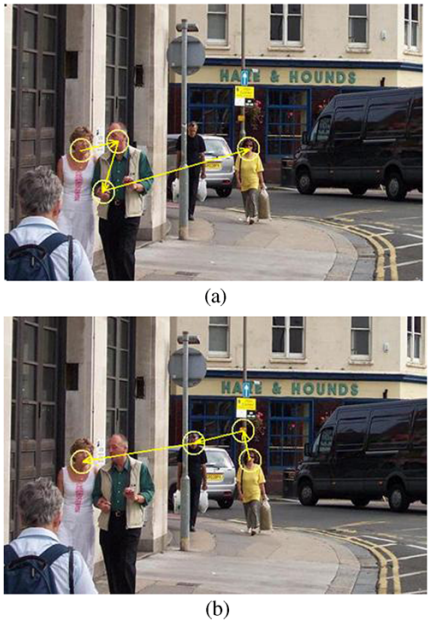

Fig.3. Comparison between attentional trajectories resulted from WTA network (a) and ISNN (b). The task here is to find a person who is wearing a yellow t-shirt. For this task, a yellow cue color was given to our model, but no cue color was utilized for a WTA network.

near the location of a cue feature even though the candidate has less face-like features as shown below.

5.2 Resolving Ambiguity by a Cooperative Cue

In this simulation we used two objects — a red ball and blue ball that are only different in color dimension, but not in shape. The round shape of the balls is extracted by Hough circle detection and simply represented by a binary value 1[30], whereas the color of the ball is extracted by a simple color segmentation algorithm and then assigned (1, 0) for the red color and (0, 1) for the blue color. After we trained the network using both shape and color, we tested the attentional behaviors of the model guided by the pointing gesture. If no pointing cue is given, both balls have almost equal chance to be a target depending on their weighting parameters. However, if the pointing cue is given, the ball indicated by the cue is more likely to be a target. In Fig.4 the balls pointed at by a finger obtained the attention.

For the computational models, in which the concept of competition is embedded in winner-take-all networks, units are mutually interconnected and are inhibited by each other. In these models, only one neuronal unit is allowed to be active at a time, whereas the others are suppressed. So, if an input equally activates the units

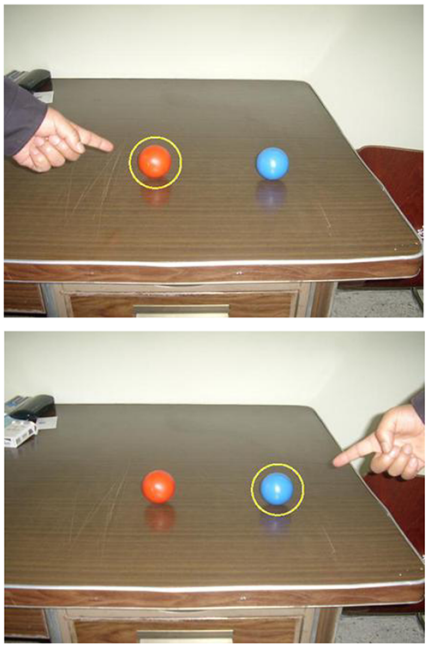

Fig.4. Allocation of attention guided by a pointing cue. The red and blue balls are equally to be a target if no cue is given. However, if the point cue is presented, the balance between the two balls is broken and the ball indicated by the finger pointing has more activation. The attentional window is allocated on the object according to the level of activation.

that correspond with blue and red balls, no competition between them can resolve this ambiguity at all and attention cannot be allocated. However, this ambiguity can be resolved when the contextual cue is introduced.

5.3 Top-Down Influence on the Modulation of Attentional Trajectories

The top-down influence on the modulation of attentional trajectories was investigated by assigning different values of α and β. If $\alpha = 1.0$ and $\beta = 0.0$, then the attentional trajectory is totally dependent on the bottom-up input. However, as the value of β is increased, the influence of top-down input on determination of the trajectory is stronger. Accordingly, the constant values of α and β were changed to $(1.0, 0.0)$ $(0.6, 0.4)$, $(0.5, 0.5)$, $(0.4, 0.6)$.

One of the results was presented in Fig.5. The task for the model is to find the person who is wearing a blue t-shirt. The attentional trajectories were dynamically changed with different values of α and β since the higher values of constant β force the model to attend to the target candidate at the location of cued color by increasing the correlation gain of the net value, whereas it detains attendance to the location where no cue color indicates by little gain or more interference.

Fig.5. Attention trajectories varied by different values of α and β. (a) $\alpha = 1.0$, $\beta = 0.0$. (b) $\alpha = 0.6$, $\beta = 0.4$. (c) $\alpha = 0.5$, $\beta = 0.5$. (d) $\alpha = 0.4$, $\beta = 0.6$. 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 a cue color segmented region.

5.4 From One Face to Another

In this simulation, video images (5 images per a second) were used for testing the performance of the model. The task for this simulation is to find the person who is waving his hand. The motion information obtained from the difference between images at two time frames ($t-1$ and t) was utilized as a cue. The first locations where an attentional window was allocated were marked with the yellow circle. As shown in Fig.6, the attention of the model was maintained at the person's face while he was waving his hand, and then moved again to the face of the other person who was waving his hand as well. Interestingly, the locus of attention stays on the location of a particular face by the motion cue, then shifts to the other face indicated by it. The results implied that the knowledge of a cue actively involves in the attentional control to engage attention to a particular location, and to shift to other locations.

Fig.6. Sustaining and shifting attention guided by a motion cue. The task for the model is to find the person who is waving his hand. The model's attention is focused on the face of the person waving his hand as shown in each column, and when that person stops waving his hand and a second person starts, the model again shifts its focus to the second person's face.

6 Conclusion

In summary, we demonstrated cooperative aspects of selective attention using a computational model in which two input streams (bottom-up and top-down) cooperate and integrate. The cooperative and integrative aspect of the model not only provides selection criteria for which current incoming information is relevant or not, but also dynamically modulates the information through a multiplicative correlation mechanism by enhancing relevant information or suppressing irrelevant

information. In this context, the limited capacity assumption was criticized because of a logical deficiency in supporting the necessity of selective attention as well as in implementing a selection mechanism.

References

[1] Humphreys G W, Bruce V. Visual Cognition: Computational, Experimental and Neuropsychological Perspectives. East Sussex: Lawrence Erlbaum Associates Ltd., UK, 1987.

[2] Pashler H E. The Psychology of Attention. Cambridge, Mass.: MIT Press, 1998.

[3] Broadbent D E. Perception and Communication. London: Pergamon, 1958.

[4] Desimone R, Duncan J. Neural mechanism of selective attention. *Annual Review of Neuroscience*, 1995, 18: 193–222.

[5] Kastner S, DeWeerd P, Desimone R *et al.* Mechanisms of directed attention in ventral extrastriate cortex as revealed by functional MRI. *Science*, 1998, 282(5386): 108–111.

[6] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254–1259.

[7] Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 2000, 40(10-12): 1489–1506.

[8] Sun Y, Fisher R. Hierarchical selectivity for object-based visual attention. In *Proc. 2nd Workshop on Biologically Motivated Computer Vision* (*BMCV'02*), Tuebingen, Germany, 2002, pp.427–438.

[9] Deco G, Schurmann B. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Research*, 2000, 40(20): 2845–2859.

[10] Carota L, Indiverib A G, Dantec V. A software-hardware selective attention system. *Neurocomputing,* 2004, 58-60: 647–653.

[11] Standage D I, Trappenberg T P, Klein R M. Modelling divided visual attention with a winner-take-all network. *Neural Networks*, 2005, 18(5/6): 620–627.

[12] Wolfe J M, Gancarz G. Guided Search 3.0: A Model of Visual Search Catches Up with Jay Enoch 40 Years Later. Basic and Clinical Applications of Vision Science, Lakshminarayanan V (ed.), Dordrecht: Kluwer Academic Press, Netherlands, 1996, pp.1989–1992.

[13] Sun Y, Fisher R. Object-based visual attention for computer vision. *Artificial Intelligence*, 2003, 146(1): 77–123.

[14] Rainer G, Assad W F, Miller E K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, 1999, 393: 577–579.

[15] Spratling M W. Cortical region interactions and the functional role of apical dendrites. *Behavioral and Cognitive Neuroscience Reviews*, 2002, 1(3): 219–228.

[16] Treue S, Martinez-Trujillo J C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 1999, (6736): 575–579.

[17] Reynolds J H, Pasternak T, Desimone R. Attention increases sensitivity of v4 neurons. *Neuron*, 2000, 26(3): 703–714.

[18] Posner M I. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 1980, 32: 3–25.

[19] Posner M I, Snyder C R R, Davidson B J. Attention and the detection of signals. *Journal of Experimental Psychology: General,* 1980, 109: 160–174.

[20] Lee KW, Feng J, Buxton H. Selective attention for cue-guided search using a spiking network. In *Proc. International Workshop on Attention and Performance in Computer Vision* (*WAPCV'03*), Graz, Austria, 2003, pp.27–33.

[21] Lee KW. Computational model of selective attention: Integrative approach [Dissertation]. University of Sussex, 2004.

[22] Lee KW, Feng J, Buxton H. Cue-guided search: A computational model of selective attention. *IEEE Transactions on Neural Networks*, 2005, 16(4): 910–924.

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

[24] Vezhnevets V, Sazonov V, Andreeva A. A survey on pixel-based skin color detection techniques. In *Proc. Graphicon-2003*, Moscow, Russia, 2003, pp.85–92.

[25] Govindaraju V. Locating human faces in photographs. *International Journal of Computer Vision*, 1996, 19(2): 129–146.

[26] Reisfeld D, Wolfson H, Yeshurun Y. Context free attentional operator: The generalized symmetry transform. *International Journal of Computer Vision*, 1995, 14: 119–130.

[27] Kovesi P. Symmetry and asymmetry from local phase. In *Proc. Tenth Australian Joint Conference on Artificial Intelligence* (*AI'97*), Perth, Australia, 1997, pp.185–190.

[28] Kim H-S, Kang W-S, Shin J-I, Park S-H. Face detection using template matching and ellipse fitting. *IEICE Transaction on Information and System*, 2000, E38-D(11): 2008–2011.

[29] Sirohey S A. Human face segmentation and identification. Technical Report CS–TR–3176, University of Maryland, 1993.

[30] Borovicka J. Circle detection using hough transforms documentation. http://linux.fjfi. cvut.cz/∼pinus/bristol/image-proc/hw1/report. pdf, 2003.

KangWoo Lee received the Ph.D. degree in computer science and artificial intelligence from the University of Sussex, Brighton, UK in 2004. From 2004 to 2005, he was a Brain Korean 21 Postdoctoral Fellow in the Center for Human-Robot Interaction, Korea Advanced Institute of Science and Technology (KAIST). Since 2006, he has been with Soongsil University, Seoul, Korea, where he is currently a research professor with the School of Media. He has authored 2 research book chapters and over 20 refereed papers in his research areas. His research interests include human robot interaction, computer vision, neural network, and cognitive modeling.