An Improved Neural Architecture for Gaze Movement Control in Target Searching

Jun Miao, Lijuan Duan, Laiyun Qing, and Yuanhua Qiao

Abstract—This paper presents an improved neural architecture for gaze movement control in target searching. Compared with the four-layer neural structure proposed in [14], a new movement coding neuron layer is inserted between the third layer and the fourth layer in previous structure for finer gaze motion estimation and control. The disadvantage of the previous structure is that all the large responding neurons in the third layer were involved in gaze motion synthesis by transmitting weighted responses to the movement control neurons in the fourth layer. However, these large responding neurons may produce different groups of movement estimation. To discriminate and group these neurons’ movement estimation in terms of grouped connection weights form them to the movement control neurons in the fourth layer is necessary. Adding a new neuron layer between the third layer and the fourth layer is the measure that we solve this problem. Comparing experiments on target locating showed that the new architecture made the significant improvement.

I. INTRODUCTION

Here are generally two kinds of top-down cues used for gaze movement control in target search: the cues about targets such as color, shape or scale [1-4] and the cues about the visual context that contains the target and the relevant objects or environmental features with their spatial relationship [5-7]. The second kind of top-down cues, i.e., the visual context, was reported to play a very helpful role in humans’ top-down gaze movement control for target searching. Through examining the response time (RT), psychological experiments show that the response time can be decreased dramatically when the relationship between the background and the location of an object in a trail (image) is known. Chun [6] stated this reduction on RTs was influenced by “Contextual Cueing”. Henderson and his colleges examined many other indexes besides RTs, such as fixation location, saccade length and the relationship among the sequenced fixation locations [7]. They found that fixation location can be predicted based on the combination of current location and context. In other words, the local feature of current fixation location and its peripheral areas can influence the next fixation position.

A large proportion of computational models rarely took advantage of context cues in target search but used the object-centered matching techniques. It means they did not predict where the targets are but compared the object features with each image window to verify if that window is the location of the target to be searched. Especially for the classical object detection methods, an original image are usually rotated $m$ times and rescaled $n$ times and then an object detector moves pixel by pixel on the transformed images $l$ times to compare each image window with the target features. Thus the detector will spend totally $mnl$ times to locate targets in the original image. This technique generally considers each object is independent and neglects the relevance between the target and the relevant objects or environmental features.

In literature, there is a small amount of research work have been done by using visual context on object searching. Torralba [8] used global features or global context to predict a horizontally long narrow region where the target is more likely to appear. Since it does not provide an accurate estimation on the horizontal coordinate, Torralba suggested using an object detector to search the target in that predicted region for accurate localization, which was implemented in the literature [9, 10]. Kruppa, Santana and Schiele [11] used an extended object template that contains local context to detect extended targets and infer the location of the target via the ratio between the size of the target and the size of the extend template. Bergboer, Postma and Herik [12] introduced local-contextual information to verify the candidates provided by an object detector, in order to reduce the false detection rate.

Different from the above methods which adopted either global context or local context cues, Miao and et al. proposed a serial of neural coding networks [13-15] using both global context and local context cues for target searching (with reference to Fig. 1). In [13], they proposed a visual perceiving and eyeball-motion controlling neural network to search target by reasoning with visual context encoded with a singe-
cell-coding mechanism. This representation mechanism led to a relatively large encoding quantity for memorizing the prior knowledge about the targets’ spatial relationship contained in the visual context. In [14], they improved the network by using the population-cell-coding mechanism. They decreased the encoding quantity for representing visual context largely. However, the disadvantage of this structure is that all the larger responding neurons in the third layer were involved in gaze motion synthesis by transmitting weighted responses to the movement control neurons in the fourth layer. These larger responding neurons may produce different groups of movement estimation. Due to a relatively big variation of the final gaze points that represent the target centers located, the system has to run multiple times to obtain a fine target locating result by computing the maximum density of multiple search results.

It is necessary to discriminate and group the movement estimations from the population coding neurons in the third layer of the system [14]. In terms of grouped connection weights form the population coding neurons to the movement control neurons in the fourth layer, a new movement coding neuron layer is inserted between the third layer and the fourth lay in the previous structure, which is the key measure that we solved this problem and presented in this paper. Comparing experiments on target locating showed that the new architecture made the significant improvement.

This paper is organized as follows: Section II describes the improved neural structure using the grouped population-cell-coding mechanism and the relational principles on encoding visual context and controlling gaze movement in target search. Bayesian learning properties of the grouped population-cell-coding are discussed in Section III. Comparative experiments for original and improved coding mechanisms on a real image database are reported and analyzed in Section IV. Conclusions are given in the last section.
features and encodes the current visual field image in terms of connection weights between the second layer and the third layer. The second part is “movement encoding and decoding”, which includes the last three layers: the third layer - VF-image coding neurons, the fourth layer- movement coding neurons and the fifth layer- movement control neurons. It encodes the spatial relationship either between two object centers or between the center of the target and the center of the current visual field image into the connection weights between the third layer and the fourth layer, and between the fourth layer and the fifth layer, which correspond to the horizontal and vertical shift distances (Δx, Δy) from the center position (x, y) of the current visual field to the center of the target.

A. Receptive Field Image Encoding

The second layer of the system consists of feature neurons which extracted different features for encoding each receptive field image received by the input neurons in the first layer. The improved system adopts the same features as used in [14], which are a set of extended LBP (local binary pattern) features. LBP is an 8-bit binary code for representing one of 256 patterns for each image block of 3×3 pixels [16]. They are widely used recently because of their powerful representation and fast computation characteristics. The original LBP coding features are discrete codes from 0~255 which extracted different features for encoding each receptive output extended the LBP feature is computed with a continuous representation and fast computation characteristics. The second layer of the system consists of feature neurons with the largest responses which extracted different features for encoding each receptive output block of 3×3 pixels or the center of the target.

\[ R_{ij} = f_j(x_{ij}) = \sum_{l=0}^{7} w_{ij}(x_{il} - x_{in}) \]  
\[ w_{ij} = (-1)^l \]  
\[ b_j = \begin{cases} 0 & \text{if } (x_{il} - x_{in}) < 0, \text{ (l=0~7)} \\ 1 & \text{otherwise} \end{cases} \]

where the vector \( X_i = (x_{i0}, x_{i1}, ..., x_{in})^T \) represents the \( i \)th image block of 3×3 pixels or the \( i \)th receptive field image of 3×3 input neurons; The term \( R_{ij} \) represents the response of the \( ij \)th feature neuron which extracts the \( j \)th feature from the \( i \)th image block or receptive field, and \( j \) is the discrete code among 0~255, which corresponds to a 8-bit binary code: \( b_0, b_1, ..., b_3, b_2 \), where \( b_j \) is expressed in Equation (2).

In our coding system illustrated in Fig. 3, for each receptive field image \( X_j \), there are 256 feature neurons in the second layer extracting the above extended LBP features \( R_{ij} = f_j(x_{ij}) \) \((j=0~255)\) and only the first \( m \) \((1 \leq m \leq 256)\) neurons with the largest responses \( \{ R_{ij} = f_j(x_{ij}) \} \in \{ R_{ij} \}, j=1\sim m, j=0\sim 255 \) win through the competition to build a weighted connection to the neurons in the third layer. To maximally decrease the encoding quantity of connection weights, \( m \) may be set to 1 for enough sparsity.

\begin{figure}[h]
\centering
\includegraphics[width=0.7\textwidth]{fig4.png}
\caption{Extend LBP features extracted by 256 feature neurons, each of which is computed by a sum of eight pairs of differences between surrounding pixels (labels=0~7) and the central pixel (label=8) in its receptive field (RF)=3×3 input neurons (pixels). They are illustrated in the 256 feature templates above, in which the gray box represents weight 1 while the black box represents weight -1.}
\end{figure}

B. Movement Encoding

With reference to Fig. 3, the fourth layer of the improved system consists of movement coding neurons which encode the movement (Δx, Δy) represented by the responses of two movement control neurons in the fifth layer and group the movement estimations from the population coding neurons in the third layer. In our system, the visual field of each scale is composed of 16×16 input neurons. Then the horizontal movement Δx and the vertical movement Δy are quantified into the integers \( u \) and \( v \) respectively, both of which range from -7 to 8, i.e., 16 numbers: -7, -6, ..., -1, 0, 1, ..., 7 and 8. The negative numbers represent the leftward or downward moving distances, \( u \) represents no movement, and the positive numbers represent the rightward or upward moving distances. Therefore there are totally 16×16=256 movements to be encoded to \( \{(u, v) \mid u=-7~8, v=-7~8 \} \), i.e., \((-7, -7), (-7, -6), \ldots, (0, 0), \ldots, (8, 7) \) and \( (8, 8) \). They are represented by the 256 movement coding neurons in the fourth layer. The connection weights between the \( uv \)th movement coding neuron and the two movement control neurons can be computed with Hebbian rule:

\[ w_{uv,\Delta x}(0) = 0, \quad \Delta w_{uv,\Delta x}(0) = \gamma R_{uv,\Delta x} \]  
\[ w_{uv,\Delta x}(1) = w_{uv,\Delta x}(0) + \Delta w_{uv,\Delta x}(0) = \gamma R_{uv,\Delta x} \]  
\[ w_{uv,\Delta y}(0) = 0, \quad \Delta w_{uv,\Delta y}(0) = \gamma R_{uv,\Delta y} \]  
\[ w_{uv,\Delta y}(1) = w_{uv,\Delta y}(0) + \Delta w_{uv,\Delta y}(0) = \gamma R_{uv,\Delta y} \]

where \( R_{\Delta x} \) and \( R_{\Delta y} \) are the responses of the two movement control neurons which output the spatial relationship or distance (Δx, Δy) to control the movement; \( \gamma \) and \( R_{uv} \) are the learning rate and the response of the \( uv \)th movement coding neuron respectively. Both of \( \gamma \) and \( R_{uv} \) are set to 1 for simplifying computation, and then Equations (3) and (4) are simplified to Equations (5):
\begin{align}
\begin{cases}
w_{iv,\Delta x}(1) = \Delta x = u \\
w_{iv,\Delta y}(1) = \Delta y = v
\end{cases}
\end{align} \tag{5}

C. Visual Context Encoding

In our paper, the visual context refers to the visual field image and the spatial relationship \((\Delta x, \Delta y)\) from the centers of the visual field to the center of the target. Thus encoding such context needs the calculation and storage of the representation coefficients of the spatial relationship and the visual field images which are centered at all the possible positions surrounding the target center and are at all the possible scales. The algorithm is described as follows:

BEGIN LOOP1
Select a scale \(s\) from the set \(\{s\}\) for the current visual field;
BEGIN LOOP2
Select a starting gaze point \((x_J, y_J)\) as the center of the visual field from an initial point set \(\{(x_J, y_J)\}\) distributed in the context area of the target;
1. Input an image from the current visual field, and output a relative position prediction for the real relative position of target center \((\Delta x, \Delta y)\) in terms of gaze movement distances \((\Delta x, \Delta y)\);
2. If the prediction error \(Er = \sqrt{(\Delta x - \Delta x)^2 + (\Delta y - \Delta y)^2}\) is larger than the maximum error limit \(ER(s)\) for the scale \(s\) of the current visual field, move the center of the visual field to the new gaze point \((x + \Delta x, y + \Delta y)\); go to 1 until \(Er \leq ER(s)\) or the iteration number is larger than a maximum limit;
3. If \(Er > ER(s)\), generate a new VF-image coding neuron (let its response \(R_i = 1\)); encode the visual context by computing and storing the connection weights \(\{w_{ij,k}\}\) (initialized to zeros) between the new VF-image coding neuron and the feature neurons (their responses \(R_j = f_1(x_j)\)) and the connection weights \(w_{uv}\) (initialized to zeros) between the new VF-image coding neuron and the movement coding neurons (let their response \(R_{uv} = 1\), \((u,v)\) encoding the corresponding \((\Delta x, \Delta y)\) respectively using the Hebbian rule \(\Delta w_{a,b} = \alpha R_k R_{a,b}\);
END LOOP2

The key part of the algorithm is the dynamical generation of VF-image coding neurons. The image coding neurons are connected by the feature neurons in the second layer and the movement coding neurons in the fourth layer with two group connection weights \(\{w_{ij,k}\}\) and \(\{w_{uv}\}\). They are generated when the coding system can not search the target in the given characteristic. The encoding of visual field images and the spatial relationship are formulated in the following two sections.

1) Encoding of Visual Field Images

The \(k\)th VF-image coding neuron in the third layer represents or encodes a visual field image pattern \(X(k)\) with a group of connection weights \(\{w_{ij,k}\}\) between the RF-feature extracting neurons in the second layer and the \(k\)th image coding neuron. The \(i\)th RF-feature neuron extracts the \(i\)th feature \(R_j = f_j(x_j)\) \((1 \leq i \leq n, 0 \leq j \leq 255)\) from the \(i\)th receptive field image \(X(j)\). All the receptive field images \(\{X(j)\}\) compose the visual field image \(X(k)\). The connection weights \(\{w_{ij,k}\}\) are computed with Hebbian rule:

\[
\begin{align*}
\Delta w_{a,b}(t) &= \alpha R_k R_{a,b} \\
 w_{a,b}(t+1) &= w_{a,b}(t) + \Delta w_{a,b}(t)
\end{align*} \tag{6}
\]

where \(\alpha\) is the learning rate; \(t\) is the iteration number; \(R_k\) and \(R_{a,b}\) are responses of two neurons which are connected by a synapse with a connection weight \(w_{a,b}\). Thus each weight \(w_{ij,k}\) between the \(i\)th RF-feature extracting neuron and the \(k\)th VF-image coding neuron is formalized in Equation (7):

\[
\begin{align*}
\Delta w_{a,b}(0) &= 0, \\
 \Delta w_{a,b}(0) &= \alpha R_k R_{a,b} \\
 w_{a,b}(0) &= w_{a,b}(0) + \Delta w_{a,b}(0) \\
 w_{a,b}(1) &= w_{a,b}(0) + \Delta w_{a,b}(1) \\
 w_{a,b}(t) &= w_{a,b}(t) + \Delta w_{a,b}(t)
\end{align*} \tag{7}
\]

where \(\alpha\) and \(R_k\) are the learning rate and the response of the \(k\)th VF-image coding neuron respectively. Both of them are set to be 1 for simplifying computation, and then Equation (7) is changed to Equation (7a):

\[
\begin{align*}
w_{ij,k}(1) &= f_j(x_j) \tag{7a}
\end{align*}
\]

According to our experiments, the performance of the system using multi-step learning which updates the weights \(\{w_{ij,k}\}\) with more than one steps is almost the same to the performance of the system updating the weights \(\{w_{ij,k}\}\) with only one step. This was proved through a theoretical analysis in [15]. Therefore, we used the simple, fast and efficient way to encode visual images described by Equations (7) or (7a).

The lengths of all the weights \(\{w_{ij,k}\}\) are finally normalized to one for unified similarity computation and comparison.

2) Encoding of Spatial Relationship

The spatial relationship \((\Delta x, \Delta y)\) between the center of the \(k\)th visual field image and the center of the target can be encoded into its integer form \((u,v)\). And then the connection weight \(w_{k,uv}\) between the \(k\)th VF-image coding neuron and the \(uv\)th movement coding neuron can be computed with Hebbian rule:

\[
\begin{align*}
w_{k,uv}(0) &= 0, \\
 \Delta w_{k,uv}(0) &= \beta R_k R_{uv} \\
 w_{k,uv}(1) &= w_{k,uv}(0) + \Delta w_{k,uv}(0) = \beta R_k R_{uv}
\end{align*} \tag{8}
\]

where \(\beta, R_k\) and \(R_{uv}\) are the learning rate, the responses of the \(k\)th VF-image coding neuron and the \(uv\)th movement coding neuron respectively. Similarly, all of them are set to 1 for
simplifying computation, and then Equations (8) are simplified to Equations (8a).

\[
\begin{align*}
W_{k,uv}(1) & = 1 \\
W_{k,uv}(1) & = 1
\end{align*}
\]  
(8a)

D. Visual Context Decoding for Gaze Movement Control

Visual context decoding includes the responding of population coding neurons and the decoding of spatial relationship. The decoded spatial relationship has a direct relation to the control of gaze movement for target search. They are formulated in following sections.

1) Responding of Population Coding Neurons

When the coding system inputs a visual field image \( Y \) for test, population neurons in the third layers may respond through competition among the total \( N \) coding neurons to represent a visual field image pattern. With reference to Fig. 3, for the \( i \)th receptive field image \( Y_i \), the \( k \)th coding neuron inputs \( m \) responses \{ \( R_{ij} \) \} \((1 \leq j' \leq m \leq 256)\) weighted by \{ \( w_{ij'} \) \} from \( m \) feature neurons which extract features \{ \( R_{ij} = f_{ij}(Y_i) \) \} from \( Y_i \). Therefore for the visual field image \( Y \) which is composed of the receptive field images \{ \( Y_i \) \} \((1 \leq i \leq n)\), the response of the \( k \)th coding neuron in the third layer is:

\[
R_k = F_k(Y) = F_k(Y_1, Y_2, ..., Y_n)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} R_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} f_{ij}(Y_i)
\]  
(9)

where \( w_{ij} \in \{ w_{ij'} \} \), \( R_{ij} \in \{ R_{ij} \} \), \( f_{ij}(Y_i) \in \{ f_{ij}(Y_i) \} \), \( j'=1-m \) and \( j=0-255 \); The weights \{ \( w_{ij} \) \} are obtained at the encoding or training stage discussed in Section IIC-1; \( R_{ij} \) is the response of the \( j \)th feature neuron for the receptive field image \( Y_i \), belonging to the first \( m \) largest responses among the total feature responses \{ \( R_{ij} \) \}. Substituting Equation (7a) into (9), we get Equation (9a).

\[
R_k = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} f_{ij}(Y_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}(X^{(k)}_i) f_{ij}(Y_i)
\]  
(9a)

Let

\[
W_{X^{(k)}} = (w_{j=1,j'=1} w_{j=1,j'=2} ... w_{j=n,j'=m,k})^T, \]

\[
f_{X^{(k)}} = (f_{j=1}(X^{(k)}_1)f_{j=2}(X^{(k)}_2)...f_{j=m}(X^{(k)}_m))^T,
\]

and \( f_Y = (f_{j=1}(Y_1)f_{j=2}(Y_2)...f_{j=n}(Y_n))^T \), then Equation (9a) is changed to its inner product form between two groups of features shown in Equation (9b):

\[
R_k = W_{X^{(k)}} f_Y = f_Y^T f_{X^{(k)}}
\]  
(9b)

Equation (9b) indicates that the response of the \( k \)th coding neuron in the third layer is a similarity measurement between the new image \( Y \) and the \( k \)th visual field image pattern \( X^{(k)} \) memorized in the coding system.

2) Decoding of Spatial Relationship for Gaze Movement Control

Gaze movement control is directly responsible for visual object research. This has been implemented in a structure that consists of three layers of neurons: VF-image coding neurons, movement coding neurons and movement control neurons (see Fig. 3). The movement control neurons, divided into \( \Delta x \) and \( \Delta y \) neurons, whose responses \( R_{\Delta x}, R_{\Delta y} \) represent the relative position \( (\Delta x, \Delta y) \) of the target to the current gaze point \((x, y)\) or the center of the current visual field image. For the current visual field image input, the first \( M \) image coding neurons which have the largest responses among the totally generated \( N \) coding neurons \((1 \leq M \leq N)\) play the main role in activating the movement coding and control neurons. If \( M=1 \), it is the single-cell-coding controlling mechanism; otherwise it is the population-cell-coding mechanism [14, 15].

Our experiments showed that \( M \) is not a stable parameter to be selected directly for the system’s best generalization performance. Instead [15], we use a similarity factor \( P = R_{\Delta x}/R_{\Delta y} \) to control \( M \), where \( R_{\Delta x} \) and \( R_{\Delta y} \) are the \( M \)th largest and the first largest responses of the VF-image coding neurons respectively.

If the first \( M \) image coding neurons may produce \( L \) different movement estimations, then they are grouped to \( L \) sets \((1 \leq L \leq 16 \times 16 = 256 \) (movement coding neurons), see Section IIB) and each image coding neuron in the \( l \)th set \((l=1-L)\) set is connected to the \( u l \)th movement coding neuron whose movement code is \((u_l, v_l)\). Let \( M_l \) represent the population number of image coding neurons in the \( l \)th set, then we have:

\[
M = \sum_{l=1}^{L} M_l
\]  
(10)

and the response of the \( u l \)th movement coding neuron is:

\[
R_{u l} = \sum_{k'=1}^{M_l} w_{k'u l} R_{k'},
\]  
(11)

where \( R_{k'} \) is the response of the \( k \)th image coding neuron, i.e., the \( k \)th image coding neuron of the \( l \)th set in the total \( L \) sets; \( w_{k'u l} \) is the connection weight from the \( k \)th image coding neuron to the \( u l \)th movement coding neuron in the fourth layer. Substituting Equation (8a) into Equation (11), we get:

\[
R_{u l} = \sum_{k'=1}^{M_l} R_{k'}
\]  
(11a)

With reference to Fig. 3, each \( u l \)th movement coding neuron in the fourth layer is designed to receive the local lateral excitation from the \((5 \times 5 \times 1)=24\) surrounding neurons and the global inhibition from all other neurons in the same layer. The role of local lateral excitation is to fuse the responses of the movement coding neurons in a neighborhood and the global inhibition’s function is to select a maximum
through winner-take-all (WTA) mechanism. The fused response is:

$$R_{uv} = \sum_{(u,v) \in Nh(u,v)} w_{uv} R_{uv}$$  \quad (12a)$$

where $Nh(u,v)$ represents the neighborhood of 5x5=25 movement coding neurons centered at $(u,v)$ and the $w_{uv}$ is the weight between the $uv$th movement coding neuron and the $u/v$ th movement coding neuron, which is designed as:

$$w_{uv} = \frac{1}{1 + \sqrt{\frac{(u-u_j)^2 + (v-v_j)^2}{5^2 + 5^2}}}$$ \quad (13)$$

Substituting Equations (9b), (11a) and (13) into Equation (12), the response of the $u/v$ th movement coding neuron is represented by Equation (12a).

$$R_{uv} = \sum_{(u,v) \in Nh(u,v)} \frac{[f_Y^T (\sum_{k=1}^{M} f_{X(k_i,i)}^T)]}{1 + \sqrt{\frac{(u-u_j)^2 + (v-v_j)^2}{5^2 + 5^2}}}$$ \quad (12a)$$

After the global inhibition or WTA competition, the response of the final winner- the $u/v$ th movement coding neuron, is set to 1 for simplifying computation:

$$R_{u,v} = 1$$ \quad (14)$$

where $l^* = \text{arg Max}\{R_{u,v} \mid R_{u,v} > Th\}$ and $Th$ is a threshold.

Then the responses of two movement control neurons can be formulated as:

$$\begin{align*}
R_{\Delta x} &= w_{u,v} R_{u,v}, \\
R_{\Delta y} &= w_{u,v} R_{u,v}
\end{align*}$$ \quad (15a)$$

Substituting Equations (5) and (14) into Equation (15), the decoded spatial relationship or movement control is represented by Equation (15a):

$$\begin{align*}
R_{\Delta x} &= \Delta x = u^*_l, \\
R_{\Delta y} &= \Delta y = v^*_l
\end{align*}$$ \quad (15a)$$

where $l^*$ are decided by Equation (16):

$$l^* = \text{arg Max}\{\sum_{(u,v) \in Nh(u,v)} \frac{[f_Y^T (\sum_{k=1}^{M} f_{X(k_i,i)}^T)]}{1 + \sqrt{\frac{(u-u_j)^2 + (v-v_j)^2}{5^2 + 5^2}}} > Th \mid l = 1, \ldots, L\}$$ \quad (16)$$

Equation (16) means if the maximum of fused response of the $u/v$ th movement coding neuron is larger than a threshold $Th$, then the $l$ th movement estimation is selected as the optimal.

Formulae (15a) and (16) describe the entire spatial relationship decoding or gaze movement control mechanism. It indicates that the gaze movement distance $(\Delta x, \Delta y)$ induced by an input image $Y$ is decided by the movement code $(u^*_l, v^*_l)$ of the $u/v$ th movement coding neuron. This coding neuron is activated by the $l$ th group of image coding neurons. It fuses the responses of neighbor neurons and produces the maximum response or movement estimation. The response of the neuron in the neighborhood of the $u/v$ th movement coding neuron is a similarity measurement between the extracted features $f_Y^T$ from the new VF-image $Y$ and the sum of features $\{f_{X(k_i,i)}^T \mid k_i = 1, \ldots, M_i\}$ extracted from $M_i$ training VF-images $\{X(k_i) \mid k_i = 1, \ldots, M_i\}$ which are encoded or represented by the $l$ th group of image coding neurons.

An entire algorithm for gaze movement control for target search is given as follows:

BEGIN LOOP1 Select a starting gaze point $(x_J, y_J)$ as the center of the visual field from a random initial point set $\{(x_s, y_s)\}$ distributed in the image area;

BEGIN LOOP2 Select a scale $s$ from the set $\{s_i\}$ for the current visual field in the order of from the maximum to the minimum;

Input an image from the current visual field, and output a estimated relative position in terms of gaze movement $(\Delta x^*, \Delta y^*)$ for the real relative position of the target center $(\Delta x, \Delta y)$;

END LOOP2

The position of the target center $(x, y)$ starting from the initial gaze point $(x_J, y_J)$ is predicted by

$$\begin{align*}
\hat{x}_j &= x_J + \sum_i \Delta \hat{x}_j \\
\hat{y}_j &= y_J + \sum_i \Delta \hat{y}_j
\end{align*}$$

END LOOP1

The algorithm uses a gradual search strategy that move an initial gaze point to the center of target from the largest visual field to the smallest visual field by decoding global and local context.

III. LEARNING PROPERTIES OF GROUPED POPULATION CODING

In our proposed coding system, the visual context is encoded into the connection weights of the neural coding structure. The Hebbian rule is the fundamental learning or encoding rule. The system dynamically generates $N$ coding neurons to represent $N$ visual context patterns $\{X^{(k)}, (\Delta x^k, \Delta y^k)\}$ $(1 \leq k \leq N)$ at the training stage and estimates the relative position $(\Delta x, \Delta y)$ of the target from a new visual field image $Y$ at the test stage.

The probability of the relative position $(\Delta x, \Delta y)$ of the target or the corresponding movement code $(u_l, v_l)$
estimated from the new image \( Y \) with the encoded visual context \( \{ (X_k, (\Delta x_k, \Delta y_k)) | 1 \leq k \leq N \} \) can be described by the Bayesian statistical learning in Equations (17).

\[
P(\Delta x, \Delta y | Y) = \frac{P(Y | u, v)P(u, v)}{P(Y)} = \frac{\alpha \sum_{k=1}^{M} f_k(Y | X)}{P(Y)} \tag{17}
\]

Then the relative position \((\Delta x, \Delta y)\) of the target or the movement code can be estimated with the Maximum a Posterior Probability (MAP) rule:

\[
(\Delta x, \Delta y) = (u_k, v_k) = \arg \max \{P(u_k, v_k | Y)\}
\]

\[
= \arg \max \{\{f_k(Y | X)\} | l = 1, ..., L\} \tag{18}
\]

It is similar to the equations (15a) and (16). The difference is that Equation (16) does an additional manipulation of fusing the adjacent movement estimations and thresholding.

IV. EXPERIMENTS ON CONTEXT CODING FOR GAZE MOVEMENT CONTROL IN TARGET SEARCH

We implemented the improved visual context coding system using grouped population-cell-coding mechanism and compare it with the system using non-grouped population-cell-coding in [14] for target searching experiments. The head-shoulder image database from the University of Bern [17] has been used. In this database, there are a total of 300 images with 30 people in ten different poses (ten images each person). The image size is 320×214 pixels. The average radius of the eyeballs of these 30 persons is 4.02 pixels. Fig.5 illustrates the first ten images. The two coding systems are compared by applying them to search the left eye centers.

![Fig. 5. Face database of the University of Bern (320x214 pixels)](image)

A. Structures of the Coding Systems

All the implemented coding systems have a group of visual fields at five scales (256×256, 128×128, 64×64, 32×32 and 16×16 pixels) which are used to input global or local images by sampling the training and test images (320×214 pixels). The \( ER(s) \) in Section IIC for five scales of the visual fields is set to 16, 8, 4, 2 and 1 respectively. For each scale or resolution, there are same 16×16 input neuron arrays with different intervals (16, 8, 4, 2 and 1 pixels) in the first layer of a coding system. These neurons simulate the distribution of visual sensing neurons in primate’s retinas. Each input neuron samples a pixel or a small region (16×16, 8×8, 4×4, 2×2 and 1×1 pixels) at the corresponding position in images. There are totally 5×16×16=1280 input neurons in the first layer of the coding structure.

Fig.2 shows that there are 256 kinds of extended LBP features for each receptive field. In the second layer of the coding system, each feature neuron extracts an extended LBP feature from its receptive field of 3×3 input neurons. Each receptive field has 1/2 overlap with its neighboring receptive fields in five visual fields. Thus there are totally \( [16-(3-1)]^2×256×5=250880 \) feature extracting neurons. Among these feature neurons at most 250880×(1/256)=980 neurons (the first \( m \) feature neurons with largest responses, \( m=1 \) for sparsity, see Section IIA) contribute to activate the population coding neurons in the third layer. The number of coding neurons in the third layer is dependent on the natural categories of the visual context encoded by the system. The number of movement coding neurons in the fourth layer is 16×16=256. The number of gaze movement control neurons in the fifth layer is two.

B. Experiments on Gaze Movement Control for Target Search

![Fig. 6. Visual context learning and testing. (a) Encoding or learning visual context between the eye center and a group of initial gaze points placed in a uniform distribution; (b) Decoding or testing for gaze movement control for the eye center search from a group of initial gaze points placed in a random distribution.](image)
of 210 images (nine images in other poses each person). The second experiment (Exp.2) used the training set of 90 images (9 people, 10 images each person) and the test set of 210 images (21 people, 10 images each person) respectively.

We compare the improved coding system with the coding system [14] on the database. Table I listed the details of the number of coding neurons generated in layer 3, the mean and the standard deviation of locating errors and the comprehensive test error.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Coding System (P, Th)*</th>
<th>Number of Coding Neurons in Layer 3</th>
<th>Locating Error (Unit: Pixel)</th>
<th>Mean (m)</th>
<th>Standard Deviation (σd)</th>
<th>Comprehensive Error $\sqrt{\text{m}^2 + \text{σd}^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp.1 (30 vs. 270)</td>
<td>the system [14] (P = 0.6)</td>
<td>2519</td>
<td>5.05*</td>
<td>9.70*</td>
<td>10.94*</td>
<td></td>
</tr>
<tr>
<td>this paper (P = 0.7, Th = 0.06)</td>
<td>4579</td>
<td>2.10</td>
<td>4.93</td>
<td>5.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp.2 (90 vs. 210)</td>
<td>the system [14] (P = 0.8)</td>
<td>5405</td>
<td>3.66*</td>
<td>6.98*</td>
<td>7.88*</td>
<td></td>
</tr>
<tr>
<td>this paper (P = 0.7, Th = 0.08)</td>
<td>16432</td>
<td>1.50</td>
<td>2.83</td>
<td>3.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: Here are test results without the post-processing adopted in [14] which runs multiple searches and uses the maximum density of multiple searching results as the center of the located object. With reference to Section IID-2, P is the parameter to control the number of population neurons involved in coding and Th is the experience threshold for movement estimation.

From Table I, it can be learned that the improved system achieved an accuracy of target localization which is about two times higher than the accuracy of the system [14]. At the same time, we noticed the new system also generated 1~2 time(s) larger encoding quantity in terms of the number of coding neurons than the system [14]. It is a cost to avoid running multiple searches and using the maximum density of multiple searching results as the center of the located object as the system [14] did.

V. CONCLUSION AND DISCUSSION

Compared with the four-layer neural structure proposed in [14], a new movement coding neuron layer is inserted between the third layer and the fourth layer for finer gaze motion estimation and control. The new structure preliminarily overcame the disadvantage of non-grouped population neurons involved in motion estimation. The measure discriminating and grouping these neurons' movement estimation produced the significant improvement. Experimental results showed that the system achieved two times higher the accuracy than that of the system [14]. The new system avoided running multiple searches for a fine searching results with the cost of generating 1~2 time(s) higher encoding quantity than the system [14]. In the future, the research will be focused on how to integrate the movement estimations from global and local context with a lower encoding quantity for better searching results.

REFERENCES