Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2008) 1813–1823
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Spatial relationship representation for visual object searching

Jun Miao[a,*], Lijuan Duan[b], Laiyun Qing[c], Wen Gao[a,e], Xilin Chen[a], Yuan Yuan[d]

[a]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[b]College of Computer Science and Technology, Beijing University of Technology, Beijing 100022, China
[c]School of Information Science and Engineering, Graduate University of the Chinese Academy of Sciences, Beijing 100049, China
[d]School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, UK
[e]Institute for Digital Media, Peking University, Beijing 100080, China

Available online 10 March 2008

Abstract

Image representation has been a key issue in vision research for many years. In order to represent various local image patterns or objects effectively, it is important to study the spatial relationship among these objects, especially for the purpose of searching the specific object among them. Psychological experiments have supported the hypothesis that humans cognize the world using visual context or object spatial relationship. How to efficiently learn and memorize such knowledge is a key issue that should be studied. This paper proposes a new type of neural network for learning and memorizing object spatial relationship by means of sparse coding. A group of comparison experiments for visual object searching between several sparse features are carried out to examine the proposed approach. The efficiency of sparse coding of the spatial relationship is analyzed and discussed. Theoretical and experimental results indicate that the newly developed neural network can well learn and memorize object spatial relationship and simultaneously the visual context learning and memorizing have certainly become a grand challenge in simulating the human vision system.

Keywords: Sparse coding; Spatial relationship; Visual context; Neural network; Object searching

1. Introduction

Quite a few psychological experiments [7,10,11,25,27] have supported the hypothesis that humans cognize the world using visual context or spatial relationship. Research on improving efficiency in learning and memorizing visual context has equally great significance for both theoretic exploration and practical application. For example, one of the main applications is object searching or detection, which is also an important function of the human vision system. However, in the study of object searching, most work [9,14,23,29,39] has focused on object-centered detection, which means, given the task of locating an object in images, the system is designed as a machine to compare each image window with the object template. These methods have dominated the research of object detection. For example, Osuna et al. [23] used face and non-face samples to train a face template for face detection. Schneiderman and Kanade [29] designed a statistical histogram method for 3D object detection and applied it to faces and cars. Viola and Jones [39] developed Boosting-based template method for real-time face detection. Chen and Yuille [9] implemented a system for text detecting in natural scenes with AdaBoost. Garcia and Delakis [14] proposed a face detector using convolutional neural network. In Refs. [2,26], the authors proposed general frameworks for face detection and object detection, respectively. In the first framework of Ai et al., the system first processed color images through skin color region segmentation, then used an average face template for those skin regions to choose face candidates; and finally reduced false faces from the candidates by a face and non-face classifier for true face verification. In the second framework of Papageorgiou et al., the system learned a subset of overcomplete wavelet base functions from a coarse scale to a fine scale for object presentation, and search objects from

*Corresponding author at: Digital Media Research Center, Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Haidian District, Beijing 100080, China. Tel.: +86 10 62600747; fax: +86 10 58858301.
E-mail address: jmiao@ict.ac.cn (J. Miao).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.11.030

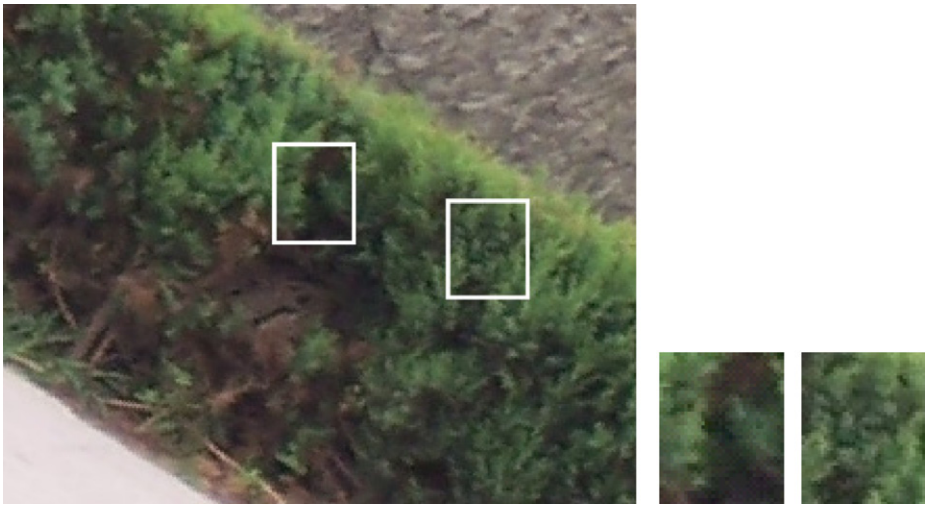


Fig. 1. Face-like image patterns in the context of environment and in isolation, respectively.

shifting windows on images by template matching with SVM and wavelet features input. It can be learned that the two general frameworks for face/object detection have a common characteristic: using the top-down or coarse-to-fine template matching method which does not utilize the context between faces and bodies or the spatial relationship between objects or between object and environmental features. Isolating an object from its surrounding features may risk false detection when the object itself cannot supply enough information for template matching. For example, humans may have different perceptions of a face-like image patterns in the context of environment and in isolation, respectively, which was discussed by Sung and Poggio [31] in 1998. Fig. 1 illustrates such face-like image patterns.

Most of the methods above, when used for visual object search, seldom consider the context between visual features, which are rather critical for the visual function of human beings. Only a few researchers have utilized context for the study of object searching and locating. Among them, Kruppa et al. [18] made use of local context to find faces. Paletta and Greindl [24] designed a method to detect objects with context in video. Strat and Fischler [30] implemented a context-based vision system, which recognized objects using information from both 2D and 3D imagery. Torralba et al. [36–38] introduced probability and statistical framework to detect objects with context. Besides, visual context is also used in visual perception and recognition [28] as well as image indexing or retrieval [6,8,12,16,20,32,33].

One possible reason for most work seldom using context is that learning context usually requires large memory that a practical system generally cannot afford. However, as we mentioned, quite a few of psychological experiments support the theory that context is the way that humans cognize the world. Furthermore, when humans perceive an image, e.g. a human face image, only a few neurons respond in their visual cortex [40]. This is the strategy of sparse coding for human visual neural system. Mathematically, sparse coding means that an image can be approximately represented by a group of sparse coefficients corresponding to a number of bases or features. Many approaches have been proposed to find such sparse bases. For example, Olshausen and Field [22] suggested a strategy of overcomplete basis set for visual cortex area V1. Hyvarinen and Hoyer [17] proposed a two-layer neural network to learn sparse coding for simple and complex cell receptive fields. Lee and Seung [19] used the method of non-negative matrix factorization for sparsely representing parts of objects.

Seldom has research work been found to implement learning and memorizing spatial relationship in the form of neural network with sparse coding features. As the human visual system is one kind of neuronal network, this paper study this significant question in theory and through experiments. Three aspects are discussed: (1) design of sparse coding features, (2) structure of sparse coding neural network, and (3) efficiency of learning and memorizing spatial relationship between initial positions and object positions.

The main contribution of this paper is: (1) a new type of neural network proposed for sparse coding of visual context or spatial relationship, (2) a significant simulation for human visual object searching mechanism, and (3) a group of experiment results which indicate that the sparse coding of spatial relationship becomes a grand challenge in simulating human visual information coding system.

In the following paragraphs, the designed sparse features are introduced in Section 2; the entire structure of the sparse coding neural network is given in Section 3; Section 4 discusses the detailed mechanism of spatial relationship learning and memorizing; in Section 5, experiments on a practical image database are analyzed; and discussion is given in Section 6.

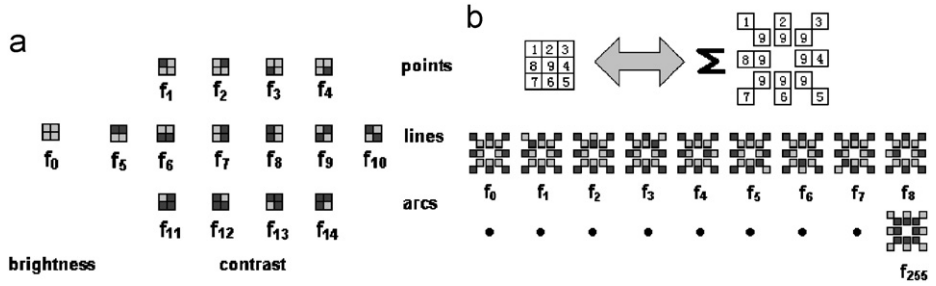


Fig. 2. Two sets of features that are extended from the widely used ones: (a) extended Haar-like features (receptive field = 2 × 2 pixels) and (b) extended LBP features (receptive field = 3 × 3 pixels).

2. Sparse coding features

The reason why most researchers or engineers would rather not use visual context is perhaps that it requires too much memory. Studies on how to use concise and efficient features, such as sparse coding features, seem quite important. Bell and Sejnowski proved that independent components of natural scenes are edge filters [5], which can be viewed as one kind of sparse bases for images. Gabor functions are used for modeling the receptive field of simple cells of the primary visual cortex V1, whose sparse coding mechanism has been used in image representation or early visual information coding [13,15,34]. A set of features called local binary patterns (LBP) used by Ahonen et al. [1] can be also adapted for image sparse representation.

We designed two groups of features that are extended from the widely used features: Haar-like features and LBP features (Fig. 2a and b), and we extend the former into several single or integrated scale forms (Fig. 4a–c).

Fig. 2a shows a set of extended Haar-like features for receptive field = 2 × 2 pixels. Two types of features are given: brightness $\vec{f}_0$ and contrasts $\vec{f}_1 \sim \vec{f}_{14}$. Among them, the 14 contrast features are actually representing three kinds of geometrical features, which are points, line segments and arcs with different positions or orientations. A gray small box in the feature patterns in Fig. 2 represents one excitatory input with a positive weight and a black box represents one inhibitive input with a negative weight.

A set of extended LBP features are illustrated in Fig. 2b. Basic LBP [1] is a kind of binary code for representing one of 256 patterns for image blocks of 3 × 3 pixels. Original LBP only output a discrete number from 0 to 255 to encode a local image pattern instead of producing a continuous comparable value. We extend LBP features by assigning continuous values to them with the following definition:

$$f_k(\vec{X}_i) = \frac{1}{8}\sum_{j=1}^{8} |x_{ij} - x_{i9}|,$$

where vector $\vec{X}_i = (x_{i1} \quad x_{i2} \quad \dots \quad x_{i9})^T$ represents the ith image block of 3 × 3 pixels and k is a discrete number among 0–255, which responds to a 8-bit binary code

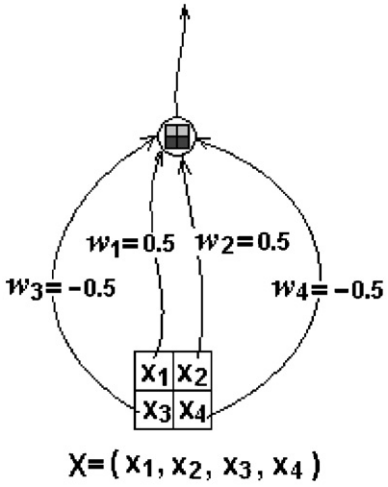


Fig. 3. An example for feature neurons on scale = 1, extracting features from a receptive field = 2 × 2 pixels.

$\mathrm{LBP}_k(\vec{X}_i) = (b_{i1} \quad b_{i2} \quad \dots \quad b_{ij} \quad \dots \quad b_{i8})$, where

$$b_{ij} = \begin{cases} 1 & \text{if } (x_{ij} - x_{i9}) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Fig. 3 shows an example of how a feature neuron on scale 1 extracts features from a receptive field of 2 × 2 pixels. Thus, a feature pattern could be represented by a vector with a group of weights (here are four weights). Generally, all weights in each feature vector are normalized to length 1 for unified feature response or similarity computation and comparison.

Fig. 4a–c shows three extended Haar-like feature forms on scales 1–3, which extract average brightness or contrasts from the output of 1 × 1 = 1 (in a receptive field of 2 × 2 pixels), 3 × 3 = 9 (in a receptive field of 4 × 4 pixels) and 7 × 7 = 49 (in a receptive field of 8 × 8 pixels) feature neurons on scale = 1, respectively. For the purpose of sparseness, the first m largest responses from these 15 features are reserved for the next step of information processing. For the example, m could be set to 2, 10 and 10 for the scales 1–3, respectively. For the extended LBP features above, m is 1 for its encoding properties.

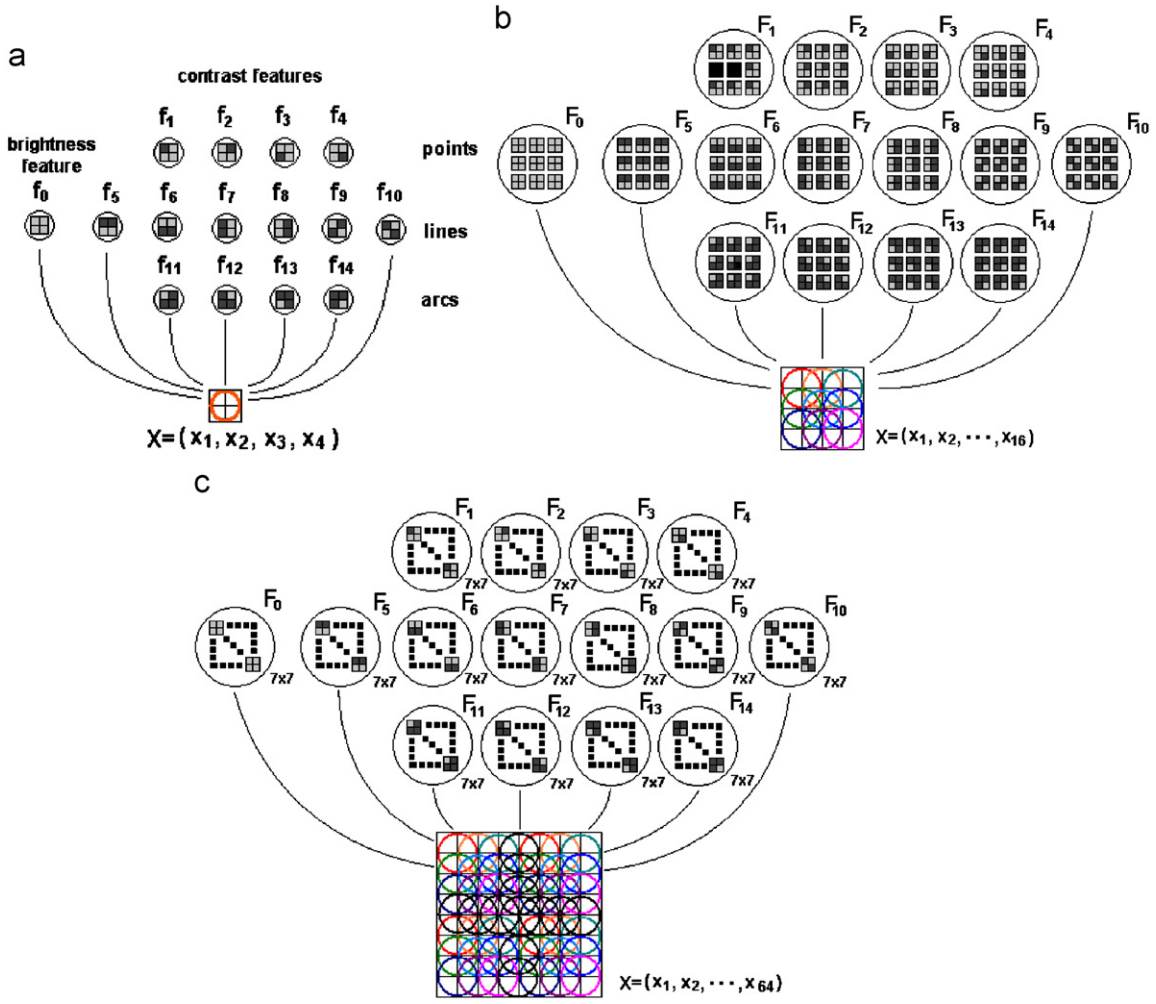


Fig. 4. Extended Haar-like features on different scales. (a) Feature neurons on scale = 1, receptive field = 2×2 pixels, extracting brightness or contrast features. (b) Feature neurons on scale = 2, receptive field = 4×4 pixels, extracting average brightness or contrast features from $3 \times 3 = 9$ feature neurons on scale = 1. (c) Feature neurons on scale = 3, receptive field = 8×8 pixels, extracting average brightness or contrast features from $7 \times 7 = 49$ feature neurons on scale = 1.

The proposed sparse features on different scales (1–3) compute mean brightness and contrasts from their various receptive fields (2×2, 4×4, and 8×8 pixels), which have certain properties of shift invariance. In other words, the features on scales 1–3 permit a certain degree of shift in the ranges of 2×2, 4×4, and 8×8 pixels.

3. Sparse coding neural network

As introduced in Section 1, in the field of object detection, most work focuses on object-centered detection and only a few researchers use context for object detection. Here the context means the spatial relationship between adjacent parts or local objects. For example, the head is always above the shoulders and the eyes are always on the face. Moreover, seldom has research work been found to learn the spatial relationship in the form of neuronal networks that the human visual system seems to use. We propose a neural network for learning and memorizing visual context, which is illustrated in Fig. 5. A related object search algorithm is given in Fig. 6.

The proposed neural network consists of two parts. One is a local image content coding structure, which inputs the local images from a group of visual fields at corresponding resolutions and memorizes the current local image pattern with sparse features. The second part is a coding structure which memorizes the spatial relationship in terms of horizontal and vertical shift distances $(\Delta x, \Delta y)$ from the center (x, y) of the current visual field to the object position $(x + \Delta x, y + \Delta y)$. The two structures naturally incorporate into an entire one and cooperate to code local image patterns and memorize their spatial relationship in a repeated mode from a global low resolution to a local high resolution.

Fig. 6 describes the object locating procedure using the visual context memory. Starting from any given initial position (x, y), the two parts of the system work together to perceive and move the local image center or gaze point in a

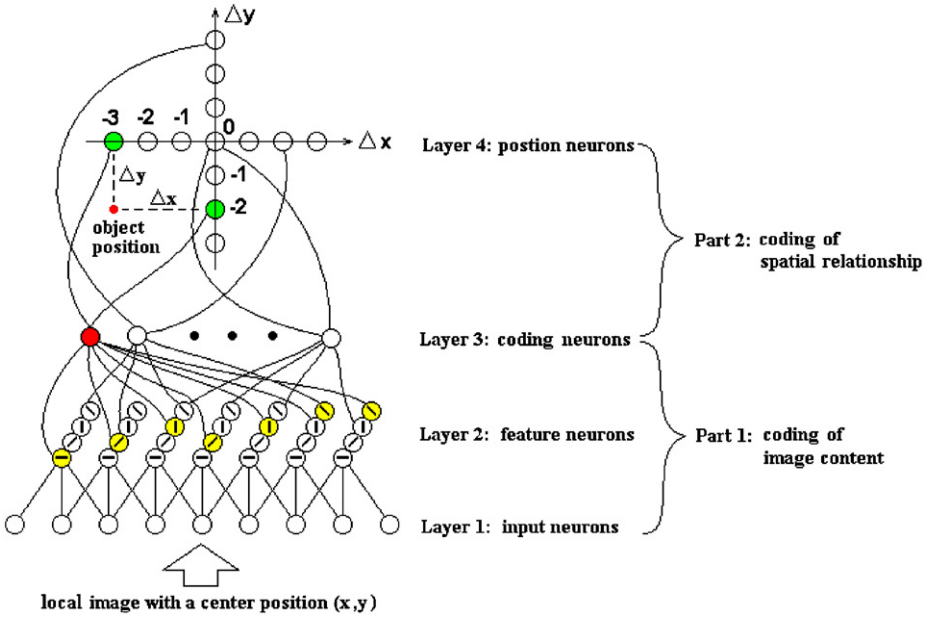


Fig. 5. Sparse coding neural network for learning and memorizing spatial relationship $(\Delta x, \Delta y)$ from any given initial position (x, y) to the object position $(x+\Delta x, y+\Delta y)$.

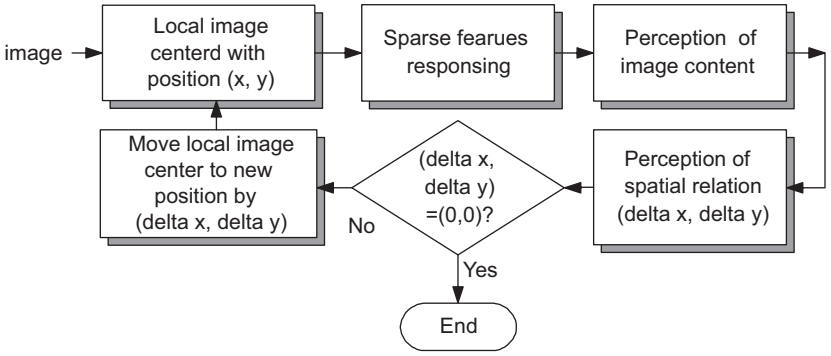


Fig. 6. Object search using visual context memory in the current visual field.

repeated mode from a global largest visual field to a local smallest visual field until the system does not move its gaze point anymore $(\Delta x = 0, \Delta y = 0)$.

3.1. Local image content coding part

From the strict point of view, visual context actually includes two components: local image patterns or objects and the spatial relationship between them. Local image or object coding is the indispensable part for visual context learning. With reference to Fig. 5, this part consists of three layers of neurons, the first layer—input neurons, the second layer—feature neurons, and the third layer—coding neurons. With reference to Fig. 7, the first layer inputs local images centered with any given initial position (x, y) from the current visual field. The second layer extracts features such as brightness and edges. These features are involved in competition and only sparse winners contribute to the responses of the neurons in the next layer. The third layer is composed of coding neurons which memorize different local image patterns or objects.

For example, for the feature neurons on scale = 1 (Fig. 4a), let vector $\vec{X}_i = (x_{i1} \quad x_{i2} \quad x_{i3} \quad x_{i4})^T$ represent the ith image window of 2×2 pixels and vector $\vec{f}_{ij} = (a_{j1} \quad a_{j2} \quad a_{j3} \quad a_{j4})^T$ represent the jth feature extracting pattern for $\vec{X}_i$, then the feature response $r_{ij} = f_{ij}(\vec{X}_i)$ can be obtained by the inner product computation:

$$r_{ij} = f_{ij}(\vec{x}_i) = <\vec{f}_{ij}, \vec{x}_i> = \sum_{k=1}^{4} a_{jk} x_{ik}.$$

Generally, a neuron is firing only if its response is higher than a threshold, for example, threshold = 0. Thus, the actual response of a neuron is

$$r_{ij} = \begin{cases} \langle \vec{f}_{ij}, \vec{x}_i \rangle & \text{if } (\langle \vec{f}_{ij}, \vec{x}_i \rangle) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Mathematically, these features constitute a set of non-orthogonal bases in the local feature vector space for describing image window patterns. For example, with reference to Fig. 4a, $\vec{f}_{i0}=(1,1,1,1)/\sqrt{4}$, $\vec{f}_{i1}=(-3,1,1,1)/\sqrt{12}$, $\vec{f}_{i5}=(-1,-1,1,1)/\sqrt{4}$, $\vec{f}_{i11}=(3,-1,-1,-1)/\sqrt{12}$, in which the brightness feature vector $\vec{f}_{i0}$ is orthogonal to any one of the contrast feature vectors $\vec{f}_{i1}\sim\vec{f}_{i14}$. Generally, the brightness feature $\vec{f}_{i0}$ has the largest response to any image window input $\vec{X}_i$ except that in a few of cases the contrast feature of "point" or "arc" has the largest responses. If we select the first two largest responding features $\vec{f}_{i0}$ and $\vec{f}_{ik}$ ($k=1$–14), then the image window pattern $\vec{X}_i$ can be approximately reconstructed by a sum of sparse weighted $\vec{f}_{i0}$ and weighted $\vec{f}_{ik}$ ($k=1$–14), i.e.:

$$\vec{X}_i \approx b_{i0}\vec{f}_{i0}+b_{ik}\vec{f}_{ik},$$

where $b_{i0}=r_{i0}=f_{i0}(\vec{X}_i)$ and $b_{ik}=r_{ik}=f_{ik}(\vec{X}_i)$ ($k=1$–14). In other words, image window pattern $\vec{X}_i$ can be represented by two reconstructed coefficients b_{i0} and b_{ik} or two feature neurons' responses $f_{i0}(\vec{X}_i)$ and $f_{ik}(\vec{X}_i)$ ($k=1$–14).

From the point of view of feature reduction in pattern recognition, the first m sparse features $(\vec{f}'_{i1},\vec{f}'_{i2},\ldots,\vec{f}'_{im})$ that have the largest responses $(r'_{i1}=f'_{i1}(\vec{X}_i), r'_{i2}=f'_{i2}(\vec{X}_i),\ldots, r'_{im}=f'_{im}(\vec{X}_i))$ to the image window pattern $\vec{X}_i$ could approximately describe or represent the $\vec{X}_i$ at the cost of minimum reconstruction error. Generally, m is less than the pixel number or dimension of the image window input $\vec{X}_i$. As illustrated in Fig. 4a, the size of the input image window or receptive field of feature neurons is $2\times 2=4$ pixels. Thus, the dimension of image window pattern $\vec{X}_i$ is 4. For the purpose of producing sparse features, m is set as 2, which is less than the number of pixels of the image window input $\vec{X}_i$, i.e.:

$$\vec{X}_i \approx \sum_{j=0}^{m} b'_{ij}\vec{f}'_{ij}.$$

Fig. 7 shows the local image coding structure in which the kth coding neuron receives inputs weighted with $w_{k,ij}$ from the ijth feature neuron on scale 1 with response r'_{ij} for their ith image window input $\vec{X}_i$. So the coding neuron's response $R_k=F(\vec{X})$, for the local image $\vec{X}=(\vec{X}_1,\vec{X}_2,\ldots,\vec{X}_N)$ which is composed of the image window input $\vec{X}_i$, is

$$R_k=F(\vec{X})=F(\vec{X}_1,\vec{X}_2,\ldots,\vec{X}_N)=\sum_{i=1}^{N}\sum_{j=1}^{m} w_{k,ij}f'_{ij}(\vec{X}_i)=\sum_{i=1}^{N}\sum_{j=1}^{m} w_{k,ij}r'_{ij},$$

where the weights $w_{k,ij}$ are acquired in the learning stage according to Hebbian rule $w_{k,ij}=\alpha R_k r_{ij}$, in which R_k is set to 1 to represent the response of the kth coding neuron who is generated for memorizing a new local image pattern, and α is also set to 1 for simplification. All the weights $w_{k,ij}$ are normalized to length 1 for unified similarity computation and comparison.

Similarly, replacing the feature on scale 1 in Fig. 7 with other ones in Figs. 4b and 4c will produce the same results.

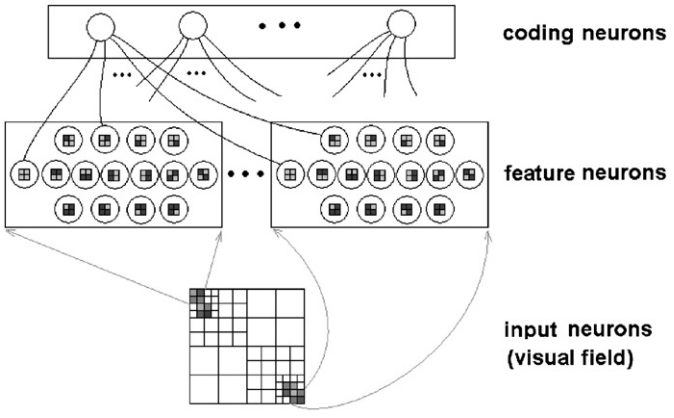


Fig. 7. Local image coding structure.

3.2. Spatial relationship coding part

As Fig. 8 shows, the spatial relationship coding structure consists of two layers of neurons: coding neurons and position neurons. The coding neurons, as discussed in Section 3.1, memorize different local image patterns. The position neurons, divided into Δx- and Δy-position neurons, represent the object position $(x+\Delta x, y+\Delta y)$ away from the center of the current visual field at a corresponding resolution.

For the local image input from the current visual field, there must be a coding neuron with a maximum response, which win the competition from all the coding neurons and represent the current local image pattern. In the learning stage, if the kth coding neuron has the maximum response and the object position is $(x+\Delta x, y+\Delta y)$ from the center of the current local image, two connections will be generated between the kth coding neuron and two position neurons: Δx- and Δy-position neurons (see Fig. 8). The weights $w_{k,\Delta x}$ and $w_{k,\Delta y}$ on the two connections could be learned by the Hebbian rule.

4. Mechanism of perceiving, learning and memorizing spatial relationship

The system perception of object positions includes a series of procedures of local image pattern recognition and object position prediction according to the learned visual context, which begin with an initial center position (x,y) and stop at a finally predicted end position ($\Delta x=0$, $\Delta y=0$). The procedure for local image pattern recognition is achieved by the coding neuron producing the largest response among all the coding neurons and becoming a winner through competitive interaction. The procedure for object position prediction is achieved by a winner coding neuron which activates Δx- and Δy-position neurons

according to the learned context. The two procedures cooperate to recognize and predict in a repeated mode from a global low resolution to a local high resolution until the system's position prediction stays unchanged ($\Delta x = 0$, $\Delta y = 0$).

The learned visual context is preserved in the connection weights of the neural networks. Hebbian rule is the fundamental learning rule, i.e., $w_{ij} = \alpha R_i R_j$, where w_{ij} is connecting weight; α is the learning rate; R_i and R_j are responses of two neurons that are connected mutually. The learning algorithm is as shown in Fig. 9.

From the above-mentioned learning algorithm, it can be learned that the neural network memorizes "support weight vectors" that are distributed on the borders between different local object classes. With reference to Fig. 10, there are four local object classes learned with weight vectors $\vec{W}_1$, $\vec{W}_2$, $\vec{W}_3$, and $\vec{W}_4$. For a local image $\vec{X}$, it is first classified to one object class according to its projections to the four class vectors, and then is mapped to object positions according to the connections from coding neurons to position neurons. The number of such "support weight vectors" is dependent on the local object categories in images. Since this number may be rather large, sparse coding should be considered. Quite a few physiological and simulation experiments [17,19,22,40] also indicate that the human visual system uses sparse coding for visual perception and cognition.

5. Experiments

The neural network is applied to coding the spatial relationship between human facial features and environmental features, for example, between the given initial positions and the eye center. It is carried out on the still face image database of the University of Bern [35], which includes totally 300 images (320×214 pixels) with 30 people (10 images each person) in 10 poses. Fig. 11 illustrates the first ten 10 images.

5.1. Coding structures

We designed five coding systems using extended LBP features, extended Haar-like features on three scales and a multi-scale features which integrated three scales of extended Haar-like features, respectively. A group of visual fields on five different scales (16×16, 32×32, 64×64,

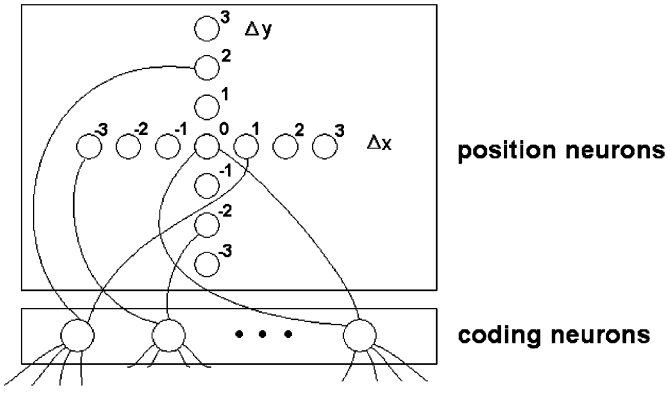


Fig. 8. Spatial relationship coding structure.

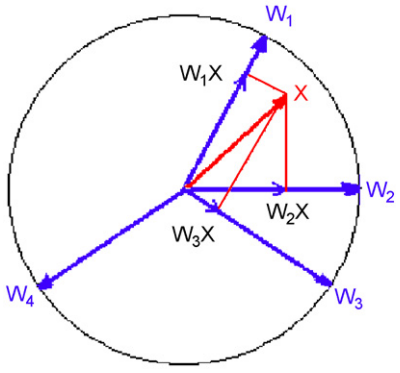


Fig. 10. Visual context learning in terms of support weight vectors memorizing.

LOOP1 Take next gaze point from all given initial gaze points

LOOP2 Take next visual field scale from Maximum to Minimum

1. Predict the object position (horizontal distance and vertical distance);
2. If prediction result is not correct, generate a new coding neuron (let response = 1); else go to 4;
3. Compute connection weights between the new generated coding neuron and lower feature neurons and that between the new generated coding neuron and position neurons with Hebbian rule $w_{ij} = \alpha R_i R_j$;
4. Move the current gaze point (visual field center) to the position of the object in the current visual field

END LOOP2

END LOOP1

Fig. 9. The learning algorithm for memorizing spatial relationship.

Fig. 11. Examples from face database of the University of Bern (320 × 214 pixels).

128 × 128 and 256 × 256 pixels) are used to input local images from the training and test images (320 × 214 pixels). For each scale or resolution, with reference to Fig. 7, there is a corresponding 16 × 16 input neuron array in common but with different intervals of 1, 2, 4, 8 and 16 pixel(s). So there are totally $5 \times 16 \times 16 = 1280$ input neurons in the first layer.

With reference to Fig. 4, the sizes of the receptive fields for extended LBP features are 3 × 3 pixels. There are 256 types of such features for input neurons at five resolutions, thus there are totally $256 \times 5 \times (16-2)^2 = 50{,}175$ feature neurons for the first system, in which only $50{,}175 \times (1/256) = 980$ neurons (the first m largest responding neurons, $m = 1$, see Section 2) win the competition. The sizes of the receptive fields for extended Haar-like features are 2 × 2, 4 × 4 and 8 × 8 pixels, respectively, each of which has 1/2 overlap between neighboring receptive fields. There are 15 types of such features for input neurons at five resolutions, thus there are totally $15 \times 5 \times [((16/2) \times 2-1)]^2 = 16{,}875$, $15 \times 5 \times [((16/4) \times 2-1)]^2 = 3675$, $15 \times 5 \times [((16/8) \times 2-1)]^2 = 675$, and $16{,}875 + 3675 + 675 = 21{,}225$ feature neurons, respectively, for the rest of four systems, in which only $16{,}875 \times (2/15) = 2250$, $3675 \times (10/15) = 2450$, $675 \times (10/15) = 450$ and $2250 + 2450 + 450 = 5150$ neurons (the first m largest responding neurons, $m = 2$, 10 and 10, see Section 2) win the competition and contribute to activate the coding neurons in the third layer.

The number of the coding neurons in the third layer is dependent on natural categories of local image patterns that the system learned. The number of position neurons in the fourth layer is $2 \times 16 = 32$, which represents 16 positions in x- and y-directions, respectively, and corresponds to 16 × 16 input neuron arrays for all the five visual fields in the first layer.

5.2. Training and testing

Two experiments for each system, totally 10 experiments were done on the face database of the University of Bern.

As illustrated in Figs. 12 and 13, training was with a group of initial positions in even distribution while testing was with a group of initial positions in random distribution. Given an initial position, the system was trained or tested to memorize or search the eye centers.

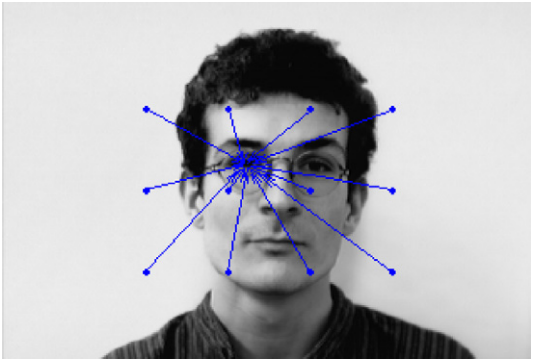


Fig. 12. Training for memorizing eye center positions from a group ofinitial positions in even distribution.

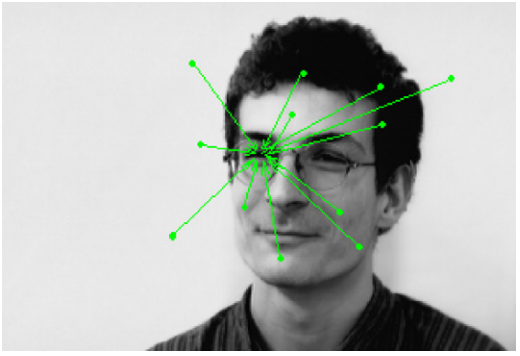


Fig. 13. Testing for predicting eye center positions from a group of initial positions in random distribution.

In the first experiment (#1) for each system, 30 images of 30 people (one frontal image each person) were trained with 368 initial gaze point positions on each image, and the rest of 270 images were tested at 48 random initial gaze point positions on each image. In the second experiment (#2) for each system, 90 images of nine people (10 images each one) were trained with 368 initial gaze point positions on each image, and the rest of 210 images were tested at 48 random initial gaze point positions on each image. The average locating error, the number of local object categories coded, and the number of connections between feature neurons and coding neurons are listed in Table 1.

Table 1 shows that the performance of the system with extended Haar-like features is better than that with the

Table 1
Performances of the first two systems (M: million)

System (sparse feature used)	No. of feature neurons	No. of coding neurons		No. of connections between feature neurons and coding neurons (M)		Average locating error (pixel)	
		#1	#2	#1	#2	#1	#2
Extended Haar-like on scale 1	2250	6026	18,988	13.5585	42.723	5.47	8.25
Extended LBP	980	3397	10,511	3.32906	10.30078	8.23	10.46

Table 2
Performances of the last four systems using extended Haar-like features (M: million)

System (extended Haar-like feature scale)	No. of feature neurons	No. of coding neurons		No. of connections between feature neurons and coding neurons (M)		Average locating error (pixel)	
		#1	#2	#1	#2	#1	#2
1	2250	6026	18,988	13.5585	42.723	5.47	8.25
2	2450	5724	16,532	14.0238	40.503	5.19	8.84
3	450	6574	20,374	2.9583	9.1683	7.97	14.89
Multi	5150	5961	18,879	30.6992	97.227	5.52	8.30

extended LBP features. The reason is that the latter produces too few coding neurons for visual context. From Table 2, it can be learned that the systems with extended Haar-like features on scales 1, 2 and multi-scale have almost the same average locating error. However, the system with multi-scale features have generated 30.6992 million and 97.227 million connections between feature neurons and coding neurons, respectively, in experiments 1 and 2, which are far more than that of the systems with the features on scales 1 and 2. Although the system with the feature on scale 3 produced much fewer connections, its average locating error is about 2 and 6 pixels higher than that of the other three systems, respectively. From the comparison, it can be concluded the extended Haar-like features whose receptive field is larger than 4×4 pixels contribute little for efficiently coding of visual context, which indicates that the features with larger receptive fields have poorer coding ability for the spatial relationship.

Figs. 14 and 15 show the statistical results for eyeball center searching by the last four systems labeled respectively with the feature scales they used, in which the horizontal axis represents the percentage of the distance between the searching results and the ground truth over the distance between the actual left and right eye centers. The vertical axis represents the accumulative correct locating rate. From the two figures, besides the similar conclusion from Table 1, it shows that experiment 1, where training and testing faces are in different poses from same persons, is better at generalization than experiment 2, in which training and testing faces are from different persons.

To simulate human retina, five overlapped visual fields (16×16, 32×32, 64×64, 128×128 and 256×256 pixels)

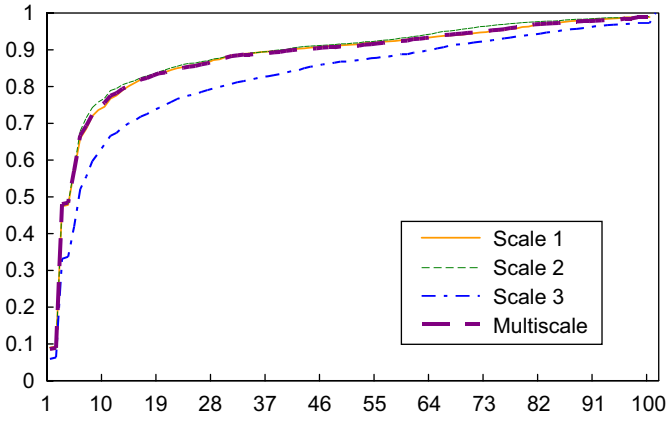


Fig. 14. Test results of experiment 1 using extended Haar-like features.

with a common center are designed in the object searching system. The system locates the object (eye center) in a sequence of five steps that are from the largest visual field to the smallest visual field. This dependence leads that the location error in the last visual field could be transferred or accumulated in the successive locations in next visual fields. As a result of limited training images (30 and 90 images for experiments 1and 2, respectively), the system's search performance sometimes cannot be avoid falling to local minimum. This problem could be solved by increasing the number of training images that has the similar distribution with the test images. Another possible solving method could be tried by integrating five visual fields into an entire visual field and searching object on the basis of this integrated visual field.

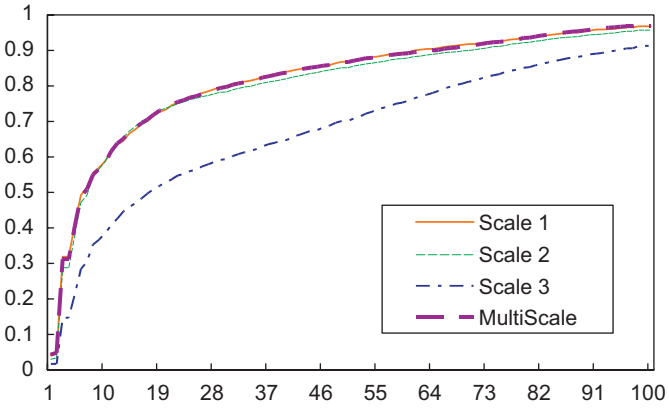


Fig. 15. Test results of experiment 2 using extended Haar-like features.

6. Discussion

Visual context is relative to the spatial relationship between local image patterns or objects. This paper proposed a novel and feasible type of neural network for coding and representing such context. Because learning and memorizing visual context indirectly become the representation for various local objects in images, the coding quantity is positively proportional to the number of these local objects. In the experiments in Section 5, the first system generated 6026 and 18,988 local object categories with about 13 and 42 million connections for memorizing spatial relationship from the initial positions to the object position, which seems rather large for the images with some human heads in the blank background. Similar results can be found in other systems. For the arbitrary object images with unlimited backgrounds, it can be inferred there will be larger number of local image coding neurons generated. Dose the human visual system produce so many coding units for visual perception and cognition? What sparse features are used by the human visual system? Does the human brain work as the sparse coding neural network we described here? It indicates that visual context or spatial relationship coding has certainly become a grand challenge in simulating the human vision system.

To face this challenge, in the future studies, the selective attention mechanism [3,4,21] can be introduced as a preceding procedure for the spatial relationship coding. It is a data-driven method for initial object or gaze point selection, which is different with the task-driven method in this paper and maybe helps to decrease the coding quantity by merging a large number of initial evenly distributed object positions to much less initial attentive object positions.

Acknowledgments

The authors would like to thank Shengye Yan for providing example figures. The authors are grateful to the anonymous referees for constructive comments. This research is partially sponsored by NSFC (Nos. 60673091, 60533030, 60702031 and 60772071), Hi-Tech R&D Program of China (Nos. 2006AA01Z122 and 2007AA01Z163), Natural Science Foundation of Beijing (Nos. 4072023 and 4061001), Beijing Municipal Education Committee (No. KM200610005012), "100 Talents Program" of CAS, and ISVISION Technologies Co., Ltd.

References

[1] T. Ahonen, A. Hadid, M. Pietikainen, Face recognition with local binary patterns, Lecture Notes in Computer Science-Proceedings ECCV2004, vol. 3021, 2004, pp. 469–481.
[2] H. Ai, L. Liang, G. Xu, A general framework for face detection, Lecture Notes in Computer Science-Proceedings ICMI2000, vol. 1948, 2000, pp. 119–126.
[3] S.W. Ban, M. Lee, Selective attention-based novelty scene detection in dynamic environments, Neurocomputing 69 (2006) 1723–1727.
[4] C. Bartolozzi, G. Indiveri, Selective attention implemented with dynamic synapses and integrate-and-fire neurons, Neurocomputing 69 (2006) 1971–1976.
[5] A.J. Bell, T.J. Sejnowski, The independent components of natural scenes are edge filters, Vision Res. 37 (23) (1997) 3327–3338.
[6] S. Berretti, A. Del Bimbo, E. Vicario, Weighted walkthroughs between extended entities for retrieval by spatial arrangement, IEEE Trans. Multimedia 5 (1) (2003) 52–70.
[7] I. Biederman, R.J. Mezzanotte, J.C. Rabinowitz, Scene perception: detecting and judging objects undergoing relational violations, Cogn. Psychol. 14 (1982) 143–177.
[8] S.K. Chang, Q.Y. Shi, C.W. Yan, Iconic indexing by 2-D strings, IEEE Trans. Pattern Anal. Machine Intell. 9 (3) (1987) 413–427.
[9] X.R. Chen, A. Yuille, Detecting and reading text in natural scenes, in: Proceedings of the CVPR, 2004.
[10] M.M. Chun, Y. Jiang, Contextual cueing: implicit learning and memory of visual context guides spatial attention, Cogn. Psychol. 36 (1998) 28–71.
[11] P. De Graef, D. Christiaens, G. d'Ydewalle, Perceptual effects of scene context on object identification, Psychol. Res. 52 (1990) 317–329.
[12] M.J. Egenhofer, R. Franzosa, Point-set topological spatial relations, Int. J. Geograph. Inform. Syst. 5 (2) (1991) 161–174.
[13] S. Fischer, G. Cristobal, R. Redondo, Sparse edge coding using overcomplete Gabor wavelets, Proc. ICIP 1 (2005) I-85-8.
[14] C. Garcia, M. Delakis, Convolutional face finder: a neural architecture for fast and robust face detection, IEEE-PAMI 26 (11) (2004) 1408–1423.
[15] M.A. Giese, D.A. Leopold, Physiologically inspired neural model for the encoding of face spaces, Neurocomputing 65–66 (2005) 93–101.
[16] V.A. Gudivada, V.V. Raghavan, Design and evaluation of algorithms for image retrieval by spatial similarity, ACM Trans. Inform. Syst. 13 (2) (1995).
[17] A. Hyvarinen, P.O. Hoyer, A two-layer sparse coding model learn simple and complex cell receptive fields and topography from natural images, Vision Res. 41 (18) (2002) 2413–2423.
[18] H. Kruppa, M. Santana, B. Schiele, Fast and robust face finding via local context, in: Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'03), Nice, France, October 2003.
[19] D.D. Lee, H.S. Seung, Learning the parts of objects with nonnegative matrix factorization, Nature 401 (1999) 788–791.
[20] J. Li, N. Allinson, D. Tao, X. Li, Multi-training support vector machine for image retrieval, IEEE Trans. Image Process. 15 (11) (2006) 3597–3601.
[21] E. Mavritsaki, D. Heinke, G. Humphreys, G. Deco, Suppressive effects in visual search: a neurocomputational analysis of preview search, Neurocomputing 70 (2007) 1925–1931.
[22] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1?, Vision Res. 37 (1997) 3313–3325.
[23] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, Proc. CVPR 3 (1997) 130–136.

[24] L. Paletta, C. Greindl, Context based object detection from video, in: LNCS 2626-Proceedings of the International Conference on Computer Vision Systems, 2003, pp. 502–512.
[25] S.E. Palmer, The effects of contextual scenes on the identification of objects, Memory Cogn. 3 (1975) 519–526.
[26] C.P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: Proceedings of the International Conference on Computer Vision, January 1998, pp. 555–562.
[27] M.C. Potter, Meaning in visual search, Science 187 (1975) 965–966.
[28] I.A. Rybak, V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, N.A. Shevtsova, A model of attention-guided visual perception and recognition, Vision Res. 8 (1998) 2387–2400.
[29] H. Schneiderman, T. Kanade, A statistical method for 3D object detection applied to faces and cars, Proc. CVPR 1 (2000) 746–751.
[30] T.M. Strat, M.A. Fischler, Context-based vision: recognizing objects using information from both 2D and 3D imagery, IEEE-PAMI 13 (10) (1991) 1050–1065.
[31] K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Trans. Pattern Anal. Machine Intell. 20 (1) (1998) 39–51.
[32] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machine-based relevance feedback in image retrieval, IEEE Trans. Pattern Anal. Machine Intell. 28 (7) (2006) 1088–1099.
[33] D. Tao, X. Li, S. Maybank, Negative samples analysis in relevance feedback, IEEE Trans. Knowledge Data Eng. 19 (4) (2007) 568–580.
[34] D. Tao, X. Li, X. Wu, S. Maybank, General tensor discriminant analysis and Gabor features for gait recognition, IEEE Trans. Pattern Anal. Machine Intell. 29 (10) (2007) 1700–1715.
[35] The Face Database of the University of Bern, 〈ftp://iamftp.unibe.ch/pub/Images/FaceImages/〉, 2008.
[36] A. Torralba, Modeling global scene factors in attention, J. Opt. Soc. Am. A (special issue on Bayesian and Statistical Approaches to Vision) 20 (7) (2003) 1407–1418.
[37] A. Torralba, P. Sinha, Statistical context priming for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2001.
[38] A. Torralba, K.P. Murphy, W.T. Freeman, Contextual models for object detection using boosted random fields, Adv. Neural Inform. Process. Syst., 2004.
[39] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vision 57 (2) (2004) 137–154.
[40] M.P. Yong, S. Yamane, Sparse population coding of faces in the inferotemporal cortex, Science 256 (1) (1992) 1327–1330.

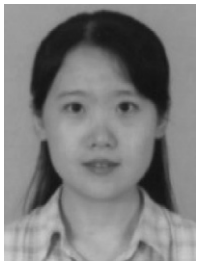
Jun Miao is with the Institute of Computing Technology, Chinese Academy of Sciences, China. He received his Ph.D. in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2005. His research interests include artificial intelligence, neural networks, image understanding and biological vision. His current research focuses on visual neural networks.

Lijuan Duan is currently with the College of Computer Science and Technology, Beijing University of Technology, China. She received her Ph.D. in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2003. Her research interests include artificial intelligence, image processing and machine vision.

Laiyun Qing is with the School of Information Science and Engineering, Graduate University of the Chinese Academy of Sciences, China. She received her Ph.D. in computer science from Chinese Academy of Sciences in 2005. Her research interests include pattern recognition, image processing and statistical learning. Her current research focuses on neural information processing.

Wen Gao received the M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1985 and in 1988, respectively, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo, in 1992, and a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 1993. From 1994 to 1995, he was a Visiting Professor at the AI Lab, Massachusetts Institute of Technology, Cambridge. Currently, he is a Professor with the School of Electronic Engineering and Computer Science, Peking University, and a Professor in computer science at Harbin Institute of Technology. He is also the guest Professor at the Insititute of Computing Technology, Chinese Academy of Sciences and the Honor Professor in computer science at the City University of Hong Kong. He is the External Fellow of International Computer Science Institute, University of California, Berkeley. He has published seven books and over 200 scientific papers. His research interests are in the areas of signal processing, image and video communication, computer vision, and artificial intelligence. Dr. Gao is an Associate Editor of the *IEEE Transactions on Circuits and Systems for Video Technology*, Editor-in-Chief of the *Journal of Computer* (in Chinese), and Editor of the *Journal of Visual Communication and Image Representation*.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively. He was a Professor at the Harbin Institute of Technology from 1999 to 2005. He was a Visiting Scholar at Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2004. He joined the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in August 2004. His research interests include image processing, pattern recognition, computer vision, and multimodal interface. Dr. Chen has served as a program committee member for more than 20 international and national conferences. He has received several awards, including the China's State Scientific and Technological Progress Award in 2000, 2003, and 2005, for his research work.



Yuan Yuan is currently a Lecturer at the Aston University, UK. She received her BEng degree from the University of Science and Technology of China and Ph.D. degree from the University of Bath, UK. She published more than 20 papers in journals and conferences on visual information processing, compression, retrieval, etc. She is an Associate Editor of the *International Journal of Image and Graphics* (World Scientific). She was on program committees of several IEEE/ACM conferences. She is a reviewer for several IEEE transactions, other international journals and conferences. She is a member of the IEEE.