Attention Driven Face Recognition, Learning from Human Vision System

Fang Fang[1], Laiyun Qing[2], Chongxiu Wang[2], Jun Miao[3], Xilin Chen[3] and Wen Gao[4]

**[1] School of Computer Science and Technology, Harbin Institute of Technology
Harbin, 150001, China**

**[2] Graduate University of Chinese Academy of Science
Beijing, 100049, China**

**[3] Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, 100190, China**

**[4] Institute of Digital Media, Peking University
Beijing, 100871, China**

Abstract

This paper proposes a novel face recognition algorithm inspired by Human Visual System (HVS). Firstly, we learn where people look by recording observers' eye movements when they are viewing face images. We find that the observers are consistent in the regions fixated and such fixated regions are selected as the salient regions. Secondly, we represent the face images by four scales of Local Binary Gabor Patterns (LGBPs) for the salient regions whereas one scale LGBPs for the others, inspired by the fact that fovea of HVS has a higher spatial acuity than the periphery. Thirdly, we integrate the global information of face images in face recognition. The experimental results demonstrate that the proposed method learning from human beings is comparable with those learned with machine learning algorithms, which shows that the characteristics of the HVS provide valuable insights into face recognition.

Keywords: *Face Recognition, Selective Attention, Human Visual System.*

1. Introduction

Face recognition has become a popular area of research in computer vision and one of the most successful applications of image analysis and understanding [1, 2] over the past several decades. Its' wide range of potential uses include security applications, intelligence-computer interaction and so on. The nature and scientific challenges of face recognition decides that not only computer science researchers are interested in it, but also neuroscientists and psychologists.

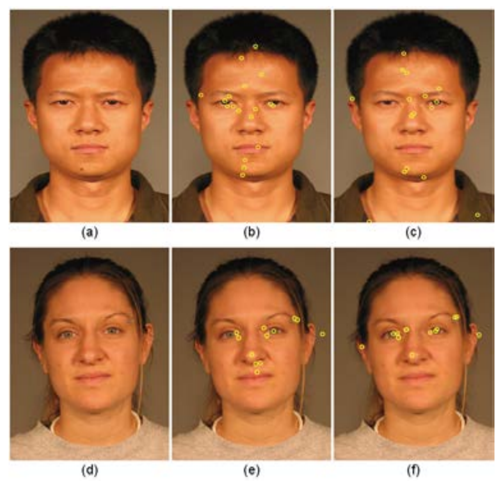

Fig. 1 Some examples of the recorded fixations, of which (b) and (e) are fixations of one observer, and (c) and (f) are fixations of another observer. The yellow circles represent the fixations.

Much progress has been made in the past two decades and numerous face recognition algorithms [3] have been developed. The most popular ones are based on machine learning, such as Eigenfaces [4] and Fisherfaces [5], SVM [6] and AdaBoost [7]. However, most of the statistical methods suffer from the generalizations problem.

It is the general opinion that advances in computer vision research will provide useful insights to neuroscientists and psychologists into how human brain works, and vice versa. Several algorithms [8, 9, 10] are based on Gabor filters, which are known to model the responses of the simple cells in vision cortex.

In this paper, we argue that we can take a further step following such direction as mimicking human vision systems. Though many face recognition systems sample the image in a grid-like fashion, human beings are able to select the most interesting regions to focus on (fixations) and jump (saccade) between them. some examples of the fixations when a observer is viewing a face image is shown in 1. This selection process is an important aspect of attention, and it has a profound impact on our perception [11]. Moreover, the spatial resolution of the human visual system decreases dramatically away from the point of fixation, which is attributed mainly to the fact that the ganglion cells are packed densely at the center of the retina (i.e. the foveola), and the sampling rate drops rapidly [12]. Therefore, the eye gathers most information during the fixations while little information is gathered during the saccades.

Inspired by the fact that HVS has a much higher spatial acuity than the periphery, a face recognition representation based on spatial variant sampling is proposed to simulate such foveated imaging phenomenon. We sample the image in a retinal way. The fixated regions (fovea) are filled with data from more Gabor filters, while less Gabor filters are reserved for the outside of the fovea. The output of the Gabor filters is further encoded by the Local Gabor Binary Patterns (LGBPs) as in [10].

The visual psychophysicists' research also has shown that human observers are able to obtain the layout of a scene within a short glance before any fixations [13]. Within such a glance, the grasped information provides useful information about spatial configuration and scene category. Therefore the global information of face images coded as the low frequency components gotten by Fourier transform is integrated into the face recognition system.

The remaining parts of this paper are organized as follows. Section 2 learns the salient regions of face images from the eye movement data. In section 3, face representation based on attention modeling which integrated both the information of salient regions and glance is presented. Extensive experiments are conducted in Section 4, followed by conclusions and discussions in the last section.

2. Selective Attention on Face Images

Though many popular face representations are based on grid sampling, human visual system works in a different way. In order to build a detailed representation of a scene, the human visual system actively scans the visual environment using discrete fixations linked by ballistic saccadic eye movements [14]. The eye gathers most information during the fixations while little information is gathered during the saccade. Such eye movement has been thought as the explicit form of visual attention.

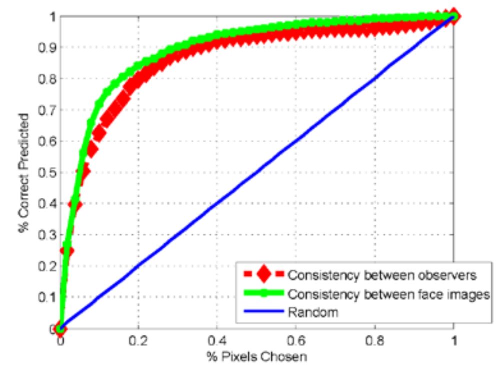

Fig. 2 Consistency of different observers viewing the same face image, and same observers viewing different face images. The consistency is defined as inter-subject consistency. We use fixations of all-except-one observers to generate a "saliency map" to predict fixations of the excluded observer. Y axis is the correct predicting rate and X axis is the selected region proportion. The curve of the consistency is much higher than random.

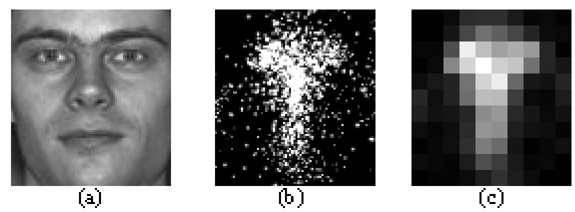

Fig. 3 (a) A cropped face image. (b) The statistical distribution of fixation positions. (c) The probability of each local region.

We use EyeLink II system [20] to record the eye fixations when observers are viewing face images. In the experiments, four male college students who are naive to the purpose of the experiments are selected as subjects. We present each face image to the subjects at the center of a 19 inch high refresh rate CRT monitor which has the resolution of 1600×1200. Each image is presented to the subjects for five seconds. Every subject views 100 frontal view face images from FRGC [2] dataset.

Some examples of the recorded fixations are shown in Fig. 1. It can be seen from Fig. 1 that, though the fixations vary among the observers and images, the regions which are most likely to be fixated are highly consistent.

We examine the inter-observer consistency among subjects' fixations, the operation is similar to "leave-one" method: for each stimulus, we use fixations of all-except-one observers to generate a "saliency map" to predict fixations of the excluded observer. The results averaged over subjects and stimulus are shown in Fig. 2. Similarly, the consistency of one observer for different face images is

shown in Fig. 2. It can be seen from Fig. 2 that the curve of consistency is much higher than random.
We get the distribution of the fixations from the samples of the fixations of the four observers on the 100 frontal view face images, as shown in Fig. 3(b). Because our face representation is based on regions and there are total 10 × 11 regions, we add the probability in each region and get Fig. 3(c), where greater possibility being fixated is indicated by whiter value. Fig. 3 demonstrates that HVS puts more importance on regions like two eyes, nose and mouth, which is well consistent with our intuition.

3. Face Recognition System Based on attention

According to the researches of the psychophysicists [13], human understand a scene by firstly glancing the scene, followed by fixating some interesting regions in it. We develop a face recognition system based on such selective attention mechanism, which integrate both of the information gathered by fixations and glance.

3.1 Face Recognition Based on Fixations

Despite a large field of view, HVS processes only a tiny central region (the fovea) with very great detail while the resolution decreases rapidly with increasing eccentricity [12]. We employ the multi-scale and multi-orientation Gabor filters to simulate such foveated phenomenon. The fixated regions are convoluted with more scales' Gabor filters to mimic the fovea. While in the non-fixated regions, less scales' Gabor filters are used to imitate the periphery, which forces the periphery contains much less information than the fovea.
We only list the parameters of the Gabor filters as follows:

$$\psi_{\mu,\upsilon} = \frac{\left\|k_{\mu,\upsilon}\right\|}{\sigma^2} e^{\left(-\left\|k_{\mu,\upsilon}\right\|^2 \left\|z\right\|^2 / 2\sigma^2\right)} \left[e^{ik_{\mu,\upsilon}z} - e^{-\sigma^2/2} \right] \quad (1)$$

Where μ and υ define the orientation and scale of the Gabor filters, $z = (x, y)$, $\left\|\cdot\right\|$ denotes the norm operator, and the wave vector $k_{\mu,\upsilon} = k_\upsilon e^{i\phi_\mu}$,where $k_\upsilon = k_{\max} / \lambda^v$ and $\phi_u = \pi\mu/8$. λ is the spacing factor between filters in the frequency domain.
In order to further enhance the information in Gabor filters, the magnitude values of the Gabor filters are further encoded with Local Binary Pattern (LBP) operator [15], namely LGBP [10]. The details of LBP are omitted for the limited space. Please refer to [10] for more details.
The input face images are normalized to 80 × 88 pixels and each image is divided into 10 × 11 non-overlapping regions. We choose 20 percent of the regions that have larger probability being fixated as the salient regions, as shown in Fig. 4(a), indicated by white value. All the others are non-salient regions.

Table 1: The accuracies of different choices of scales on the FERET datasets

Methods	***fb***	***fc***	***DupI***	***DupII***
LGBP[10]	**0.96**	**0.96**	**0.69**	**0.61**
WLGBP[10]	**0.98**	**0.97**	**0.74**	**0.71**
ν1=0	0.88	0.87	0.39	0.21
V1=1	0.94	0.96	0.56	0.41
V1=2	0.95	0.96	0.66	0.56
V1=3	0.96	0.96	0.70	0.64
V1=4	0.94	0.94	0.67	0.61
V1=3,V2=1,2,4	**0.97**	**0.97**	**0.74**	**0.72**

In the original LGBP [10], five scales $\nu \in \{0, 1, \cdots, 4\}$ and eight orientations $\mu \in \{0, 1, \cdots, 7\}$ Gabor filters are used, which means 5 × 8 × 110 = 4, 400 Gabor filters are used. In our experiment, in order to reserve more information in fovea and less information in periphery, four scales $\nu \in \{1, 2, 3, 4\}$ and eight orientations Gabor filters are employed for the 22 salient regions, while for non-salient regions, only one scale $\nu \in \{3\}$ Gabor filters are used. The reasons for choosing such scales can be found from Table 1, which will be discussed in Section 4. Therefore there are totally 32 × 22+8 × (110-22) = 1; 408 Gabor filters are used, only 32 percent of the original LGBP [10].
Histogram intersection as in [10] is used as similarity measure when we compare two face images.

3.2 Face Recognition Based on Glance

As discussed above, human are able to obtain the layout of a scene within a short glance before any fixations. The information gathered during glance are the global information of the scene. We represent such global information by the lower frequencies gotten by 2-D Discrete Fourier Transform (DFT), though some other global features such as GIST can also be used for this purpose. These features are further processed by Fisher's Linear Discriminant Analysis (FLDA) [16] to reduce the dimension.
Finally, the cosine of the angle between two feature vectors based on glance is computed as the similarity measurement.

3.3 Face Recognition by Integration

The information gathered by fixations and glance, both of which are based on attention mechanism, are integrated into the final face recognition system.
Similarity based on glance between two input face images can be represented as $\eta(m_1, m_2)$, where m_1 and m_2 are feature vectors based on glance of two input face images respectively. And the similarity based on fixations between them can be represented as $\psi(H_1, H_2)$, where H_1 and H_2 are feature vectors based on fixations of two input face images respectively. Let ω_G denotes the weight of the similarity based on glance, and then the combined similarity of the two face images can be represented as follows:

$$C = \omega_G \times \eta\left(m_1, m_2\right) + \left(1 - \omega_G\right) \times \psi\left(H_1, H_2\right) \quad (2)$$

4. Experimental Results

Experiments are conducted on FERET database [17] to evaluate the proposed method. FERET database includes 1; 196 objects, where the gallery set has 1; 196 face images, and the four probe sets are fb, fc, DupI and DupII, which has 1; 195, 194, 722 and 234 face images respectively. fb and fc were taken the same time as gallery with different expression and lighting respectively. The DupI were obtained anywhere between one minute and 1; 031 days after their respective gallery matches. The DupII were those taken only at least 18 months after their gallery entries.

4.1 Experimental Results of Fixations

Several choices of the scales are used to represent the non-salient regions are tested. Table 1 shows that the experimental results on FERET database, where ν_1 represents the scales used in all regions and ν_2 represents the scales used only in salient regions. It can be seen from Table 1 that, the 3rd scale has the best performance, and then followed by the 4th, 2nd and 1st scales, which is consisted with the studies that middle spatial frequencies provide information that is more useful for face recognition than are other frequency ranges [19].
Therefore we use the 3rd scale at all the regions of the face images and use the 4th, 2nd, and 1st scales in fixated regions additionally. The performance is better than original LGBP [10], which uses five scales in all the regions of the face images. The performance is also comparable with Weighted Local Gabor Binary Pattern (WLGBP) [10], which not only uses five scales in all the regions of the face images, but also different weights of different regions of each Gabor filter learned by Linear Discriminant Analysis.

We perform another two sets of experiments to further illustrate that the salient regions selected by human vision systems are reasonable. One set includes two randomly selected salient regions (see in Fig. 4(b) (c)).The accuracies of the randomly selected regions on FERET database are shown in Table 2. The performance with the regions fixated by human beings is better than those with randomly selected regions, which is consistent with our

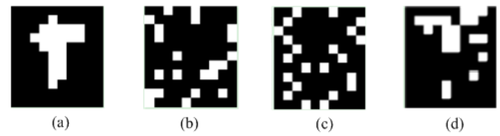

Fig. 4 Different choices of salient regions. (a) Salient regions selected by human vision systems. (b) Randomly selected salient regions set 1. (c) Randomly selected salient regions set 2. (d) Salient regions selected by AdaBoost.

Table 2: Different choices of salient regions random on the FERET datasets

Methods	*fb*	*fc*	*DupI*	*DupII*
Random Regions 1	0.95	0.96	0.68	0.61
Random Regions 2	0.95	0.96	0.69	0.62
AdaBoost	0.97	0.99	0.74	0.70
Adaboost Selected Regions	0.98	0.99	0.74	0.70
Fixations	**0.97**	**0.97**	**0.74**	**0.72**

Table 3: The accuracies of integration on the FERET datasets

Methods	*fb*	*fc*	*DupI*	*DupII*
WLGBP[10]	0.98	0.97	0.74	0.71
Integration	**0.99**	**0.99**	**0.76**	**0.71**

intuition. The other set experiments are performed by AdaBoost[18]. The results of AdaBoost and those of the most largerest weights regions selected by AdaBoost (Fig. 4(d)) are also shown in Table 2, both use the same LGBPs features. It can be seen that the performances of human selected regions are comparable to those of AdaBoost and AdaBoost selected salient regions. Please note that the face images used to collect eye fixations are not included in FERET datasets. Actually we hope that human beings are not so sensitive to datasets compared with machine learning algorithms (please note that AdaBoost are a little worse in the case of DupII, which maybe not been included in training set.)

4.2 Experimental Results of Integration

As mentioned in section 3, classifiers based on fixations and glance are combined to form the integrated classifier.

The weight of similarity based on glance $\omega_G = 0.13$. The choice of ω_G is according to experience, and the experimental results are not sensitive to ω_G range from 0.10 to 0.15. The experimental results on FERET dataset are shown in Table 3.
It can be seen from the results that, by combining the feature based on glance, the performance is better than WLGBP [10]. Experimental results demonstrate that the feature based on fixations and glance are indeed complementary for distinguishing faces, which is consistent with HVS.

5. Conlusions

This paper proposes a novel face recognition algorithm inspired by the selective attention of HVS. The regions are most likely to be fixated by human eyes are selected as salient regions. Motivated by the fact that fovea has a much higher spatial acuity than the periphery, a face representation based on spatial variant sampling is proposed to simulate such foveated imaging phenomenon of human eyes, where more information is reserved for the salient regions. Moreover, the information based on glance which adopts the low spatial frequency components of the image is integrated into the face recognition system to elicit a percept that occurs before any fixations.
The experimental results on FERET database demonstrate that the proposed method achieves comparable performance with those of machine learning algorithm such as AdaBoost, which shows that the characteristics of the HVS provide valuable insights into face recognition.
Though many valuable works have been done in selective attention and many computing models are proposed, there are few works that introduce such attention mechanism into face recognition and other object recognition. This paper is just an attempt to do such things. Please note that some components of the proposed framework can be modified. For example, LGBP [10] can be replaced with other alternative features to mimic fixations.

Acknowledgments

This research is partially sponsored by National Basic Research Program of China (No.2009CB320902), Beijing Natural Science Foundation (No. 4102013), Natural Science Foundation of China (No.60970087, 61070116, 61070149 and 61175115) and President Fund of Graduate University of Chinese Academy of Sciences (No.085102HN00).

References

[1] M. H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, 2002.
[2] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone, “Face recognition vendor test 2002: Evaluation report,” Tech. Rep. NISTIR 6965, National Institute of Standards and Technology, 2003. [Online]. Available: http://www.frvt.org.
[3] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” ACM Comput. Surv., vol. 35, no. 4, pp. 399-458, 2003.
[4] M. Turk, A. Pentland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience,vol. 3, no. 1, 1991.
[5] P. Belhumer, P. Hespanha, and D. Kriegman, “Eigenfaecs vs. fisherfaces: Recognition using class specific linear projection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, 1997.
[6] K. R. M¨uller, S. Mika, G. R¨atsch, K. Tsuda, and B. Sch¨olkopf, “An introduction to Kernel-based learning algorithms,” IEEE Transactions on Neural Networks, vol. 12, no. 2, pp. 181-201, 2001.
[7] P. Yang, S. G. Shan, W. Gao, S. Z. Li, D. Zhang, “Face recognition using Ada-Boosted Gabor features,” Automatic Face and Gesture Recognition, pp 356-361, 2004.
[8] L. Wiskott, J. M. Fellous, N. Kruger, C. Malsburg, “Face Recognition by Elastic Bunch Graph Matching,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 775-779, 1997.
[9] C. Liu and H. Wechsler, “Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition”, IEEE Transaction on Image Processing, vol. 11, no. 5, pp. 467-476, 2002.
[10] W. C. Zhang, S. G. Shan, W. Gao, X. L. Chen, H. M. Zhang, “Local Gabor Binary Pattern Histogram Sequence (LGBPHS): a Novel Non-Statistical Model for Face Representation and Recognition,” International Conference on Computer Vision, pp 786-791, 2005.
[11] G. Rizzolatti, L. Riggio, I. Dascola, C. Umilt_a, “Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention,” Neuropsychologia, vol. 25, pp. 31-40, 1987.
[12] Andrew T. Duchowski, “Eye Tracking Methodology: Theory and Practice,” Springer, 2nd edition, August 3, 2007, ISBN:1846286085.
[13] M. C. Potter, “Meaning in visual search,” Science, vol. 187, no. 4180, pp. 965-966, 1975.
[14] A. Yarbus, “Eye Movements and Vision”, New York: Plenum, 1967.
[15] T. Ojala, M. Pietik¨ainen, and T. M¨aenp¨aa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987, 2002.
[16] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, “Fisher Discriminant Analysis with Kernels,” IEEE Conference on Neural Networks for Signal Processing, 1999.
[17] P. J. Phillips, H. M. Syed, A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, pp. 1090-1034, 2000.

[18] Y. Freund, R. E. Schapire. “A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting,” In Computational Learning Theory: Eurocolt ’95, pp. 23–37, 1995.
[19] I. Boutet, C. Collin, J. Faubert, “Configural face encoding and spatial frequency information,” Perception and Psychophysics, vol. 65, pp. 1078-1093, 2003.
[20] http://.www.sr-research.com/.

Fang Fang received the B.S. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2004. She is currently working towards the Ph.D. degree at the School of Computer Science and Technology, HIT. Her research interests mainly include face recognition and image Saliency detection.

Laiyun Qing (S’03–M’09) received the B.Sc. and M.Sc. degrees in computer science from Northeastern University, Shenyang, China, in 1996 and 1999, respectively. She received the Ph.D. degree in computer science from Chinese Academy of Sciences, Beijing, in 2005.
She is currently an Associated Professor at the School of Information Science and Engineering, Graduate University of the Chinese Academy of Sciences, Beijing. Her research interests include pattern recognition, image processing, and statistical learning. Her current research focuses on visual perception and cognition.

Chongxiu Wang received the B.S. degree in Computer Science and Technology from Beijing Institute University (BIT), Beijing, China, in 2009. She is now working towards the M.S. degree at the Graduate University of the Chinese Academy of Sciences, Beijing, China. Her research mainly includes face recognition and object recognition.

Jun Miao (S’00–M’04) received the B.Sc. and M.Sc. degrees in computer science from Beijing University of Technology, Beijing, China, in 1993 and 1999, respectively. He received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2005.
He is currently an Associated Professor at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include artificial intelligence, neural networks, neural information processing, image understanding, and biological vision. He has published more than 30 research articles in refereed journals and proceedings on face detection, visual neural networks, visual neural information coding, neural oscillation, image segmentation, visual perception, and cognition. His two main contributions are the technique of Human Face Gravity-Center Template for face detection and the model of Visual Perceiving and Eyeball-Motion Controlling Neural Network for visual search, respectively.
Dr. Miao is a member of the China Computer Federation and a member of the Chinese Society for Neuroscience. He is the recipient of Microsoft Fellowship Award in 2000, a recipient of 2003 Shanghai Science and Technology Progress Awards (the First Award), and a recipient of 2005 National Science and Technology Progress Awards of China (the Second Award).

Xilin Chen (M’00–SM’09) 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 with the Harbin Institute of Technology from 1999 to 2005. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2004. Since August 2004, he has been with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, where he is also with the Key Laboratory of Intelligent Information Processing and the ICT-ISVISION Joint Research and Development Laboratory for Face Recognition. His research interests include image processing, pattern recognition, computer vision, and multimodal interfaces.
Dr. Chen has served as a program committee member for more than 20 international and national conferences. He has received several awards, including the State Scientific and Technological Progress Award in 2000, 2003, and 2005 in China for his research work.

Wen Gao (M’92–SM’05–F’09) received the B.Sc. and M.Sc. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985 and 1988, respectively. He received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.
He is currently a Professor of Computer Science at Peking University, Beijing, China. Before joining Peking University, he was a Professor of Computer Science at Harbin Institute of Technology from 1991 to 1995, and a professor at the Institute of Computing Technology of Chinese Academy of Sciences. He has published extensively, including four books and over 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.
Dr. Gao serves on the editorial board for several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FORVIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, the EURASIP Journal of Image Communications, and the Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.