Object Tracking Based on Visual Attention Model and Particle Filter

Long-Fei Zhang[1], Yuan-Da Cao[2], Ming-Jie Zhang[3], Yi-Zhuo Wang[4]

School of Computer Science and Engineering,
Beijing Institute of Technology,
Beijing, China. 100081
{longfeizhang[1], ydcao[2], frankwyz[4]}@bit.edu.cn, zhangmingjie@yeah.net[3]

Abstract

An object tracking method based on visual attention model and particle filter is presented. An improved visual attention model is employed to measure the similarity between tracked objects and candidate objects. Gaussian weighted color, intensity, orientation and motion saliency map are calculated with strategy to compose the attention value, which can be used to measure the similarity of the objects. This similarity measurement is more accurate than others used in object tracking algorithms. Experimental results show that both single object and multiple-objects could be tracked efficiently.

Keyword: visual attention model, particle filter, object tracking.

I. Introduction

In recent years, more and more researchers use particle filtering to track moving objects in image sequences [1-5] . As even the object motion model and the observation model are nonlinear and the noise is non-Gaussian, the object can be tracked well. Particle filtering is a technique for implementing a recursive Bayesian filter by sequential Monte Carlo simulations [6]. The key idea is to represent the required posterior density function by a set of random samples with associated weights and to compute estimates based on these samples and weights. The weights of particles are related to the observation of the object in the current frame.

Almost all the features such as color[5] and contour[4] are exploited to present the tracked objects. But low level features can not provide enough information for object tracking.

In other words, object tracking is a kind of visual tracking. Visual tracking technology offers an intimate and immediate way of interpreting users' behaviors to help a computer search the moving objects with the same feature. The gaze behavior of participants is compared with data obtained through a model of Visual Attention (VA) [7] to detect differences in behavior arising from varying video content.

This paper exploits visual attention model to measure the similarity among these objects, using particle filtering to track the object in image sequence. Section II presents visual attention model and the similarity measurement of objects. Section III briefly introduces the particle filtering. Section IV describes the object tracking algorithm. Section V presents the experimental results in detail; The last section is the conclusion.

II. Visual attention model

Attention is a neurobiological conception. It could imply the concentration of mental power upon an object by observing or listening [9]. Computational attention makes us represent the multimedia information so close to the sense of human being. Itti et.al [7,8] reviewed the recent works on computational models of visual attention, and present an useful spatial salience based visual attention model. Sun[9] raises another object based visual attention model to simulate the gaze of human eyes. Ma[10] uses attention model to summarize the video clips. In this paper, an improved spatial salience based visual attention model is presented to establish the object features for object tracking.

A. *Visual Attention Model for Object*

Definition: The visual attention model for Object is defined as a set of attention objects:

$$AO_i = \{ ROI_i , AV_i, MPS \}, \ 0<i<N \tag{1}$$

$$ROI_i = rectangle\left(\sum_{0<k<2^{2(N+1)}} B_k^i \right) \tag{2}$$

$$AV_i = \{ AVC_i, AVI_i , AVO_i, AVM_i \} \tag{3}$$

Where:

ROI_i	Region-Of-Interest of AO_i
MPS_i	Minimal Perceptible Size of AO_i
AV_i	Attention Value of AO_i
B_k^i	Blocks by separate the AO_i with the size of MPS_i
AVC_i	Attention Value of AO_i caused by color feature
AVI_i	Attention Value of AO_i caused by color intensity feature
AVO_i	Attention Value of AO_i caused by orientation feature
AVM_i	Attention Value of AO_i caused by motion feature

We chose a pyramid structure to represent the original image for a scalable computing. In the pyramid structure, the original image is decomposed into sets of blocks. These blocks can be lowpass and bandpass components via Gaussian and Laplacian pyramids[11], respectively. The Gaussian pyramid consists of lowpass filtered (LPF) versions of the input image, with each stage of the pyramid computed by lowpass filtering and subsampling of the previous stage image. But in this paper, we decompose the frame(image) into N scale, $2^{2(N+1)}$ blocks B_k with the size of MPS_i . It is more accurate to fit the attention distribution by its isotropy characteristic.

B. *Color and Intensity Salience*

HSV color space is more easily perceived by the human eye. Thus, we exploit the mean normalized *HSV* color to present the primary color feature of the object. We separate the frame F to blocks B_k. We could calculate the color feature of each blocks to form the pyramid image, which introduced in section *A*. So that we could get the color distribution of the frame and could measure the primary attention of the frame.

The methods to extract the color and intensity features is as follows:
Step 1. Calculate mean color of each blocks as :

$$\mathrm{C}(B_k)=\left(H_{B_k},S_{B_k},V_{B_k}\right)=\left(\mathrm{mean}\left(\mathrm{h}(B_k)\right),\mathrm{mean}\left(\mathrm{s}(B_k)\right),\mathrm{mean}\left(\mathrm{v}(B_k)\right)\right) \tag{4}$$

Step 2. Compute the chess-board distance between B_k and the ROI_i (regarded as B_j) by (5):

$$\nabla_d(B_k,B_j)=\left\|B_k-B_j\right\|=\mathrm{MAX}\left(\left|x(B_k)-x(B_j)\right|,\left|y(B_k)-y(B_j)\right|\right) \tag{5}$$

where:(x(B),y(B)) is the coordinate of the center of block B;

Step 3. Calculate color contrast between block B_k and the ROI_i (regarded as B_j) with color distance equation(6):

$$\nabla_c(B_k,B_j)=1-\frac{1}{\sqrt{5}}\sqrt{\left(S_{B_j}V_{B_j}\cos H_{B_j}-S_{B_k}V_{B_k}\cos H_{B_k}\right)^2+\left(S_{B_j}V_{B_j}\sin H_{B_j}-S_{B_k}V_{B_k}\sin H_{B_k}\right)^2+\left(V_{B_j}-V_{B_k}\right)^2} \tag{6}$$

Step 4. The color feature attention value AVC_i is:

$$AVC_i=\frac{\sum_{B_k^i\in ROI_i}\mathrm{W}_{\mathrm{gauss}}(\nabla_d(ROI_i,B_k^i))\cdot\mathrm{C}\left(B_k^i\right)}{\sum_{B_k^i\in ROI_i}\mathrm{W}_{\mathrm{gauss}}(\nabla_d(ROI_i,B_k^i))} \tag{7}$$

Step 5. Calculate the color variance between blocks in previous frame F_{t-1} and this frame F_t at the same location by equation (6) and record it as

$$\nabla_m(B_k)=\nabla_c(B_k^t,B_k^{t-1}) \tag{8}$$

$\mathrm{AVM}(B_i)$ is calculate by (9):

$$AVM_i=\frac{\sum_{B_k^i\in ROI_i}\mathrm{W}_{\mathrm{gauss}}(\nabla_d(ROI_i,B_k^i))\cdot W_m\cdot\nabla_m\left(B_k^i\right)}{\sum_{B_k^i\in ROI_i}\mathrm{W}_{\mathrm{gauss}}(\nabla_d(ROI_i,B_k^i))} \tag{9}$$

Step 6. Then, we get the color distribution and the color value of blocks in the objects. The intensity of attention value of the object can be calculated by (10):

$$AVI_i=\frac{\sum_{B_k^i\in ROI_i}W_{gauss}(\nabla_d(ROI_i,B_k^i))\cdot\nabla_c(B_k,B_j)}{\sum_{B_k^i\in ROI_i}W_{gauss}(\nabla_d(ROI_i,B_k^i))} \tag{10}$$

Where W_m is the weight of motion, $\mathrm{W}_{\mathrm{gauss}}(.)$ is the weight of its position defined with normalized Gaussian conversation kernel centered in the object. The weight selection is shown in figure 1.

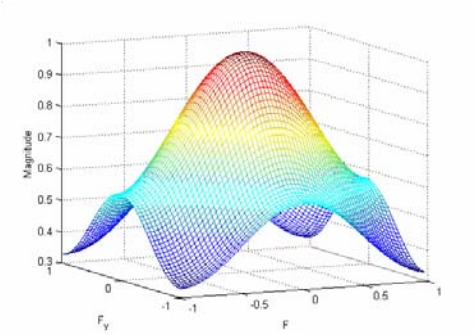


Fig. 1. Attention intensity distribution. The convolution template valued by normalized Gaussian ($W_{gauss}(.)$) fits *N* scales pyramid images (k=5，$2^{2(N+1)}$ =64*64 blocks).

C. Orientation Salience

The object's local orientation information is obtained from intensity using oriented Gabor pyramids $OP(\zeta,\theta)$,where ζ represents the scale and θ is the orientation. (Gabor filters, which are the product of a cosine grating and a 2D Gaussian envelope, approximate the receptive field sensitivity profile (impulse response) of orientation-selective neurons in primary visual cortex [7,9].)

We define $\overline{\theta}_{P_1,P_2}$ as the orientation difference between pixels p_1 and p_2. Let $V_{p_1}(\theta)$ and $V_{p_2}(\phi)$ be the orientation vectors of p_1 and p_2 in the current orientation pyramid respectively.

Note that V, θ, and ϕ themselves all consist of multiple components. For example, if we have four preferred orientations, then:

$$V_{p_1}(\theta)=[V_{p_1}(0),V_{p_1}(\pi/2),V_{p_1}(\pi),V_{p_1}(3\pi/4)] \tag{11}$$

We define the orientation salience $\nabla_o(p_1,p_2)$ as:

$$\nabla_o(p_1,p_2)=W_{gauss}\left(\nabla_d(p_1,p_2)\right)\cdot\sin\left(\overline{\theta}_{p_1,p_2}\right) \tag{12}$$

Sinusoid function is a nonlinear and monotonically increasing function from 0 to 1 over the range $[0, \pi/2]$ and symmetric in $[0, \pi]$, thus we choose sinusoid as the orientation salience factor.

Then $\overline{\theta}_{p_1,p_2}$ can be given by equation (13):

$$\overline{\theta}_{p_1,p_2}=\frac{\int_0^{\pi}\phi\left[\int_0^{\pi}V_{p_1}(\theta)V_{p_2}((\theta+\phi)\bmod\pi)d\theta\right]d\phi}{\iint_{\pi}V_{p_1}(\theta)V_{p_2}((\theta+\phi)\bmod\pi)d\theta d\phi} \tag{13}$$

The discrete form is:

$$\overline{\theta}_{p_1,p_2}=\frac{\sum_{j=0}^{\zeta-1}j\varphi\sum_{i=0}^{\zeta-1}V_{p_1}(i\varphi)V_{p_1}((i\varphi+j\varphi)\bmod\pi)}{\sum_{j=0}^{\zeta-1}\sum_{i=0}^{\zeta-1}V_{p_1}(i\varphi)V_{p_2}((i\varphi+j\varphi)\bmod\pi)} \tag{14}$$

Where *i* and *j* is the pyramid level, ζ is defined as the number of orientation pyramids or preferred orientations. ϕ is $j\varphi$ and θ is $i\varphi$, where $\varphi = \pi/\zeta$;

Thus, we could get the orientation attention AVO_i :

$$AVO_i = \frac{\sum_{\zeta} \sum_{k=0}^{n(\zeta-1)} \nabla_o \left(center(ROI_i), p_k \right) W_{gauss} \left(\nabla_d \left(center(ROI_i), p_k \right) \right)}{\sum_{\zeta} \sum_{k=0}^{n(\zeta-1)} W_{gauss} \left(\nabla_d \left(center(ROI_i), p_k \right) \right)} \tag{15}$$

where *n* is the neighborhood pixel of a pixel, and n is 8; $8(\zeta-1)$ is the number of neighborhood pixels of object.

D. Visual Attention Similarity

By the equation (3), we could get the AV_i of AO_i:

$$AV_i = W_C \cdot AVC_i + W_I \cdot AVI_i + W_m \cdot AVM_i + W_O \cdot AVO_i \tag{16}$$

where W_c ,W_i ,W_m and W_o are the weighting coefficients of color attention, intensity attention, motion attention and orientation attention, respectively. In this paper, we set them to 1/4.

Because the tracked objects will not change much, such as *ROI* and color, we could use attention model to measure the similarity between the target and candidate objects.

We use ROI_i^{t-1} of the target AO_i in F^{t-1} (the frame in *t-1*) instead of $ROI_i{}^t$ in (7) and (9~10) to calculate the AVC_j and AVM_j of candidate objects AO_j, and we could get AV_j by (16).

Finally, we can calculate the similarity between the two objects by (17):

$$Sim(AO_i, AO_j) = \left| AV_i - AV_j \right| \tag{17}$$

III. Particle Filtering

Particle filtering is a technique for implementing a recursive Bayesian filter by sequential Monte Carlo simulations [6]. The key idea is to represent the required posterior density function by a set of random samples with associated weights and to compute estimates based on these samples and weights.

In recent years, particle filtering has been used to track objects in a clutter, in which the posterior density $p(X_t/Z_t)$ and the observation density $p(Z_t/X_t)$ are often non-Gaussian, and the system model is nonlinear. Where X_t denotes the state vector of the tracked object at time *t*, and Z_t denotes all the observation vector of the tracked object $\{z_1, \ldots, z_t\}$ up to time *t*.

Particle filtering uses a weighted sample (particle) set S_t to approximate the probability distribution of the object state $p(X_t/Z_t)$, where *N* is the number of particles in the particle set and *t* is the observation time. Each sample consists of an element *s* which represents the hypothetical state of the object and a corresponding discrete sampling probability *w*.

where $S_t = \left\{ \left(s_t^{(n)}, w_t^{(n)} \right) \mid n = 1, \ldots, N \right\}$, and $\sum_{n=1}^{N} w_t^{(n)} = 1$.

Every particle in the sample set evolves according to a system model, producing the sample set

$$S_{t+1} = \left\{ \left(s_{t+1}^{(n)}, w_{t+1}^{(n)} \right) | \, n = 1, \ldots, N \right\} \tag{18}$$

$$w_{t+1}^{(n)} = K \frac{p\left(z_{t+1} | \, s_{t+1}^{(n)} \right) p\left(s_{t+1}^{(n)} | \, s_t^{(n)} \right)}{q\left(s_{t+1}^{(n)} | \, s_t^{(n)}, z_{t+1} \right)} \tag{19}$$

Where: K is the normalization factor to ensure: $\sum_{n=1}^{N} w_{t+1}^{(n)} = 1$, and $q(\cdot)$ is the importance density operator.

Particle filtering models uncertainty, and considers the multiple state hypotheses simultaneously, so it provides a robust tracking framework.

IV. Object tracking method

Suppose the region (we use *O* stands for the region) containing the tracked object in image sequence is known. The Object tracking method based on particle filtering is:

Input: the tracked object *O;*
Output：the sequence of the estimated state of *O;*

- $t = 0$, Initialization:
  Generate the initial particle set:

$$S_0 = \left\{ \left(s_0^{(n)}, w_0^{(n)} \right) | \, n = 1, \ldots, N \right\} \tag{20}$$

where $w_0^{(n)} = 1/N$; $s_0^{(n)}$ is the sample drawn from Gaussian distribution $g\left(x ; X_0, \Sigma \right)$, whose mean value is X_0, is the initial position of the tracked object, $\sum$ is the covariance matrix.

- *for* $t = 1,2,\ldots,F$ (where F is the frame number)
  Get the current state of the tracked object X_t, using the algorithm (Visual attention model).
  - Calculate the velocity (v_{t-1}) of the tracked object:

$$v_{t-1} = X_t - X_{t-1} \tag{21}$$

  - Update particles in S_{t-1} and its weight

$$s_t^{(n)} = s_{t-1}^{(n)} + v_{t-1} + \omega_{t-1} \tag{22}$$

  Where ω_{t-1} is the process noise .

$$w_t^{(n)} = \exp\left\{ -\frac{\sum_{j=1}^{N_X} \left(X_t^j - s_t^{(n),j} \right)}{2\sigma^2} \right\} \tag{23}$$

  N_X is dimension number of the state vector.

- Resampling:

(a) calculate the normalized cumulative probabilities

$$c_1 = 0$$

$$c_i = c_{i-1} + w_t^{(n)},\ i = 2,...,N.$$

(b) generate a uniformly distributed random number $u_1 \in U[0,1/N]$.

set $i = 1$.

(c) *For* $j = 1,...,N$

$u_j = u_1 + (j-1)/N$;

while $u_j > c_i$

$i = i + 1$;

end while;

$s_{t+1}^{(j)} = \tilde{s}_{t+1}^{(i)}$;

end for;

- Calculate the sum of weights:

$$Sum = \sum_{n=1}^{N} w_t^{(n)} \tag{24}$$

- Normalize the weights:

$$w_t^{(n)} = w_t^{(n)} / Sum \tag{25}$$

- Estimate the state of the object:

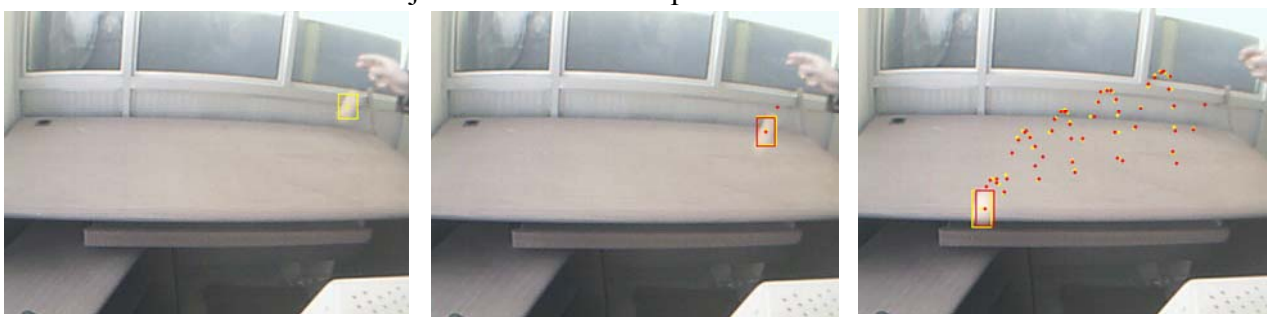
$$X_t = \sum_{n=1}^{N_s} w_t^{(n)} s_t^{(n)} \tag{26}$$

- *end for*

Using the object tracking algorithm listed above, the selected object in the image sequence can be efficiently tracked.

V. Experiment

The image sequences used in our experiments are captured by a PC camera at 15 frames/sec with a resolution of 320×240 pixels. They consist of 168 frames of a table tennis ball moving in a typical laboratory environment. Every particle in the particle set is a 4D vector, which represents the region which contains the candidate object. The number of particles is 2000.



(a)frame #73 (b)frame #74 (c)frame #126

Fig. 2. Single object tracking

Figure 2 (a) shows the initial position of the tracked object, the yellow rectangle is the region which contains the selected object, the yellow point is the center of the rectangle. Figure 3 shows the spatial saliency map of visual attention of frame #73, which is showed in figure 2(a). Figure 2 (b) and (c) show the trajectory of the tracked object. The yellow rectangle is the observed position of the region which contains the tracked object, the points is the center of the rectangle. The yellow points are the observed center of the rectangles, and the red points are the estimated center of the rectangles.

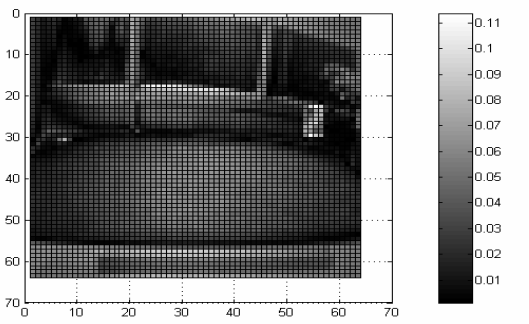
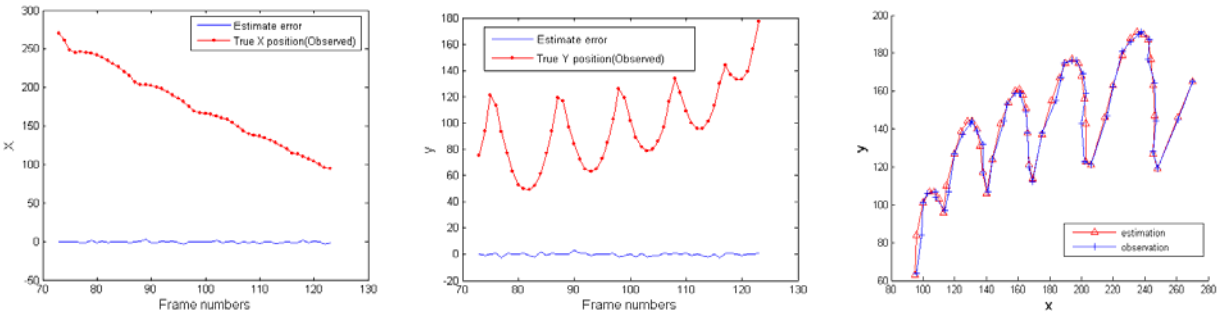


Fig. 3. Saliency map of Spatial visual attention in the frame #73. The white part on the middle of the left is the target object . The color bar scales the visual attention value.



(a) (b) (c)

Fig. 4. The estimated and observed tracking curve on (x-t),(y-t),and (x-y) axis

Figure.4 shows the estimation curve and the observation curve of the object from frame 73 to 126. (a) and (b) shows the error curve of the estimated table tennis centers in blue, along with the true (observed) table tennis centers shown as lines with dots in red. (c) shows the estimated and observed position of object center's position.

The average error of estimated position on the x axis is 0.235 and on the y axis is 0.137. The results showed a good performance for object tracking.

(a)frame #299 (b) frame #305 (c)frame #317

Fig. 5. Multi-object tracking

Figure.5 shows the second experiment: table tennis balls rolled from one side of the desk to the other. (a) shows the selected objects to be tracked in different rectangle. The points in different color in (b) and (c) shows the trajectory of different objects.

The two experiment show that the proposed object tracking method in this paper can track the selected objects efficiently.

VI. Conclusion

An object tracking method based on a computational visual attention model and particle filter is presented. An improved visual attention model is proposed to present the tracked object sensitively and measure the similarity of objects accurately. Instead of RGB color space, HSV, which is more coincident to human's perceiving, is used in visual attention model. Gaussian weighted color space salience map with color intensity and orientation are used to estimate the parameters of particle filter more accurately than color histogram which is widely used in object tracking.

We employed a particle filter method for using the visual attention-based object tracking. Both single and multiple objects can be tracked well. Experimental results show that the proposed method can yield a good result.

References

[1] Zaveri, M.A.; Desai, U.B.; Merchant, S.N.; Automated model selection based tracking of multiple targets using particle filtering. TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, Vol.2 (2003) 831 - 835

[2] Boers, Y.; Driessen, J.N. Multitarget particle filter track before detect application[C] Radar, Sonar and Navigation, IEE Proceedings, Vol.151, Iss.6 (2004) 351-357

[3] T.Hong; S.K.Wang; Z.Q.Wang; Frontal motion tracking based on image features analysis and particle filter.[C] Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, Vol.7 (2004) 3995-3998

[4] S.h.k.Zhou, Chellappa, R.; Moghaddam, B Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Transactions on Image Processing, Vol.13 , Iss.11 (2004) 1491–1506

[5] Yoshinori Satoh, Takayuki Okatani and Koichiro Deguchi. A Color-based Tracking by Kalman Particle Filter. Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04). 1051-4651/04 (2004) 502-505

[6] A. Doucet, N. D. Freitas, and N. Gordon, Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag (2001).

[7] Itti L, C. Koch, and E.Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Ingtelligence, Vol.20, No.11 (1998) 1254- 1259

[8] Itti L, Koch C. Computational Modeling of Visual Attention. Nature Reviews Neuroscience 2(3) .(2001) 194-203

[9] Y.R. Sun, Robert Fisher. Object-based visual attention for computer vision. Artificial Intelligence Vol.146, Iss.1 (2003) 77-123

[10] Y.F. Ma, X.S. Hua, L. Lu, H.J. Zhang. User Attention Model based Video Summarization. IEEE Transactions on Multimedia (2004)

[11] Greenspan, H.; Belongie, S.; Goodman, R. etc. Overcomplete steerable pyramid filters and rotation invariance. Proceedings CVPR '94.(1994) 222-228

Longfei Zhang received the BS. degree from the department of Computer Science and Technology, Henan University, China, in 2000. He received the MS. and Ph.D. degrees from the Department of Computer Science and Engineering at the Beijing Institute of Technology(BIT) in 2005. He has been with the School of Computer Science and Engineering, BIT. His research interests include Computer Vision, Machine Learning, and Pattern Recognition.

Yuanda Cao received the Diploma in electrical engineering form the Beijing Institute of Technology in 1969. He has been with school of Computer Science and Engineering, Beijing Institute of Technology. He is a fellow of China Computer Federation, director of intelligent information network special commission, Chinese Association for Artificial Intelligence. He is professor and Ph.D. advisor. His major research interests include Article Intelligent, Pattern recognition and Network Security.

Mingjie Zhang received the BS. degree from the department of Computer Science and Technology, Zhengzhou University, China, in 2000. He received the MS. and Ph.D. degrees from the Department of Computer Science and Engineering at the Beijing Institute of Technology in 2005. His research interests include Computer Vision, Image Processing, and Pattern Recognition.

Yizhuo Wang received the BS. degree from the department of Computer Science and Technology, Henan University, China, in 2000. He received the MS. and Ph.D. degrees from the Department of Computer Science and Engineering at the Beijing Institute of Technology in 2005. He has been with the School of Computer Science and Engineering, Beijing Institute of Technology. His research interests include Image Processing, Data Compressing.