# Real-Time Surveillance of People on an Embedded DSP-Platform

Qifeng Qiao*, Yu Peng, and Dali Zhang

*Abstract* — **This paper presents a set of techniques used in a real-time visual surveillance system. The system is implemented on a low-cost embedded DSP platform that is designed to work with stationary video sources. It consists of detection, a tracking and a classification module. The detector uses a statistical method to establish the background model and extract the foreground pixels. These pixels are grouped into blobs which are classified into single person, people in a group and other objects by the dynamic periodicity analysis. The tracking module uses mean shift algorithm to locate the target position. The system aims to control the human density in the surveilled scene and detect what happens abnormally. The major advantage of this system is the real-time capability and it only requires a video stream without other additional sensors. We evaluate the system in the real application, for example monitoring the subway entrance and the building hall, and the results prove the system's superior performance.**

*Index Terms* — **Surveillance, object tracking, background modeling and embedded system**

## I. INTRODUCTION

The real-time surveillance is a hot application field for computer vision. Security and safety are major concerns facing all organizations today. In response, millions of cameras have been installed worldwide to control access, secure circumference and monitor for dangerous activities [1]. However, theses cameras are watched by special personnel. The demand for the accurate and intelligent surveillance systems becomes stronger. For the computer station, the video stream should be extracted from the cameras and sent in a compressed form to a central node, hence requiring both the communication and computation infrastructure. Consequently, the embedded systems are more convenient and cheaper than general computers and suitable for installation in harsh environments.

There have been many projects on detecting and tracking people. The systems have different processing technologies and functionality. Pfinder [2] solves the problem of person tracking in complex scenes that has a single unconcluded person and fixed camera. W4 [3] is a real time visual surveillance system for detecting and tracking multiple people and monitoring their activities in an outdoor environment. Both Pfinder and W4 use a statistical background model to locate people. However, Pfinder uses a single Gaussian distribution of color at each pixel, while W4 uses a bimodal distribution of intensity at each pixel. CMUproject [4] uses a distributed network of active video sensors to provide continuous coverage of people and vehicles in a cluttered environment. It can detect and track multiple people and vehicles in the complex scenes and monitor their activities over a long time. The computed locations feed into a higher level tracking module that tasks multiple sensors with variable pan, tilt and zoom to cooperatively and continuously track an object through the scene.

Most surveillance systems contain motion detection and tracking which depends on the result of the motion detection. The detection maintains the temporal association of tracked objects. Some techniques are proposed for comprehensive survey of background subtraction [5], [6]. To detect a target, adaptive multimodal background models are frequently used, since the historical scenes are maintained to promote the accuracy. This method has some limitation in the real-time embedded platform for the computation and storage restriction. In this paper, a modified method is used to achieve faster execution and less storage while keeping the comparable accuracy.

The aim of an object tracker is to generate the trajectory of an object over a time by locating its position in every frame of the video. After the detector provides the possible object regions, the tracker builds the correspondence between objects across frames. Tracking methods can be generalized into three groups: point tracking, kernel tracking and silhouette tracking. Mean shift algorithm [7] is one kind of kernel based tracking. We use this method, since it reduces the computation by eliminating the exhaustive search and is robust to partial occlusion, clutters, non-rigid motions and abrupt scene changes.

The first contribution of our system is to provide an accurate and lower computation complex method for the adaptive modeling of background and detection. Moreover, the cheap cameras can meet the requirement of the processing system, which broadens its application field. The mean shift tracking for multiple objects is also novel in current surveillance research.

The paper is organized as follows. First of all, the overview of the embedded system will be introduced in section 2. We present the background modeling and the target detection method in the third section. The background model is the reformed multimodal model but with less computation cost. The mean shift for tracking multiple objects is discussed in section 4. The experiments in section 5 prove the sound performance of our system in real-time working.

## II. SYSTEM ARCHITECTURE

The embedded system uses Texas Instruments DM642 fixed point DSP as the process chip. With performance of up to 5760 million instructions per second at a clock rate of 720MHZ, the device offers cost-effective solutions to high-performance

Manuscript received Oct. 9, 2007.
This work was supported in part by the Automation Department in Tsinghua University.
* Qifeng Qiao is with the Automation Department, Tsinghua Univsity, Beijing, P.R.China(corresponding author to provide phone: 86-010-62773435; e-mail:qqf05@ mails.tsinghua.edu.cn).
Yu Peng is with the Automation Department, Tsinghua University, Beijing, P.R.China (e-mail: pengy05@ mails.tsinghua.edu.cn).
Dali Zhang is the professor in Automation Department, Tsinghua University, Beijing, P.R.China (e-mail:zdl5@ mail.tsinghua.edu.cn).

programming challenges. It has the SDRAM memory block of 256MB. The embedded system allows for connecting any analog or digital video sources. Figure 1 describes the structure of the network based surveillance system. The embedded system has efficient connection with the user interface. Figure 2 illustrates a framework for the embedded system.

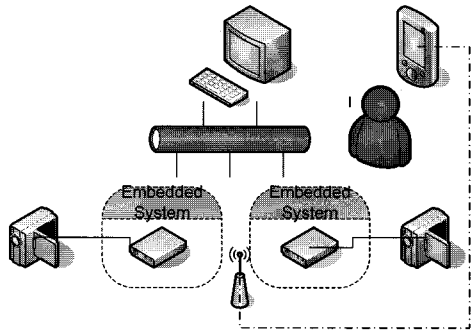


Fig. 1. The interface structure for the whole network based surveillance system.

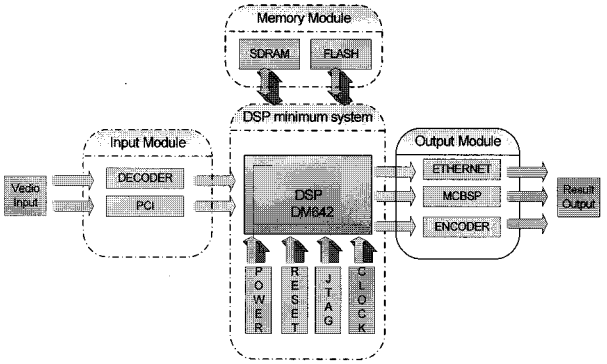


Fig. 2. The framework for the embedded system

The framework consists of three modules. First, the targets are detected in the input frames. After processing, the tracker locates the target positions by iteratively computation, in order to optimize the detection process and establish the relation between subsequent frames. Then the objects are classified to improve accurate detection for pedestrians. Finally, the results are exported to the terminal.

## III. BACKGROUND MODEL

All tracking methods require a detection mechanism to get the targets. There are some traditional ways for object detection. Temporal differencing is very adaptive to dynamic environments, but hardly extracts all relevant feature pixels. Background subtraction provides the most complete feature data, but is very sensitive to dynamic scene changes caused by lighting changes and extraneous events.

The methods of Mixed Gaussians [8] and Wallflower [9] are robust to scene changes by storing multimodal representations of backgrounds. The dynamic scene information is contained in the model. These approaches demand significant computation resources and storages.

For application in real-time embedded system, the update of Mixed Gaussians is transformed into a linear scheme, escaping from nonlinear update of weights and pixel values.

The background modeling is called multimodal mean [10], in which each background pixel is represented by an average of history values. A box of cells describes whether a pixel belongs to the background.

### 3.1 Learning initial background model

Pixel values are in RGB vector. The background model for a pixel is depicted by a set of representations, called cells. Each cell $C_{i,n,x}$ contains sums of $S_{i,n,x}$ for each color component and a count number $Num_{i,n}$ that indicates how many times a pixel is observed in n frames. $I_{x,n}$ represents the $x$ color component of a pixel in $n\ th$ frame. At the beginning, the background model can be established regardless moving foreground objects in the scene. The median filter over time is applied to several seconds of video to distinguish moving pixels from stationary pixels. The initial model parameters are realized in this stage. In our model, we set four cells for each pixel and use an RGB color representation. For the cell $C_{i,n,x}$, the mean color component value is computed as

$$\mu_{i,n,x} = S_{i,n,x} / Num_{i,n} \tag{1}$$

### 3.2 Model updating and foreground detection

The background should be adaptive to the environment changes, such as the illumination changes and the new object appearance. The background model is updated by assimilating the new information from each new frame. Foreground objects are segmented from the background in each frame by a four stage process: threshold and region segmentation, noise cleaning, morphological filtering and a connected component analysis.

A pixel is a background pixel if it matches a cell in every color component for the following conditions.

$$\left|I_{n,x} - \mu_{i,n-1,x}\right| \le T_x \ and\ Num_{i,n-1} > T_f\ ,\ x = R,G,B \tag{2}$$

where $T_x$ is the predefined threshold for corresponding color, $T_f$ is the predefined threshold that prevents the pixel to be regarded as the background too easily. If the times of matching $Num_{i,n}$ are smaller than $T_f$, the pixel is still considered to be a foreground pixel. The model updating uses periodical decimation of the cells to enable long-term adaptation to scene changes. The updating process is generalized as follows.

- Step1: Search for cell matching by function(2)
- Step2: If the pixel $I_{x,n}$ has a matched cell $C_{i,n,x}$, the corresponding component $S_{i,n,x}$ and $Num_{i,n}$ are updated by this pixel.

$$S_{i,n,x} = (S_{i,n-1,x} + I_{i,x})/2^b$$
$$Num_{i,n} = (Num_{i,n-1} + 1)/2^b$$
$$if\ t \bmod b = 0,\ b = 1;\ else\ b = 0$$

- Step3: If the pixel $I_{x,n}$ has no matching, it is labeled as the foreground pixel. A new background cell will be integrated into the cell box, which contains the new information. If the box has full places, the lowest ranked cell is replaced. The ranking depends on comparing $Num_{i,n}$ and a recent indicator $R_{i,n}$.

$R_{i,n}$ is the sum of a vector $(r_{i,n}, s_{i,tn})$, in which $r_{i,n}$ starts at zero, is incremented when $C_{i,n,x}$ is matched, and is reset every w frames. $s_{i,n}$ saves the previous maximum value of $r_{i,n}$. $R_{i,n}$ suggests how often a cell $C_{i,n,x}$ is observed with a recent window. The least recent cell is first considered to be replaced. Otherwise the lowest $C_{i,n,x}$ cell is to be replaced.

The post processing uses region-based noise cleaning to eliminate noise. The system then generates a set of shape and appearance features for each detected foreground object. The features are used to distinguish the people from the other objects. The foreground detection is generalized in Figure 3.

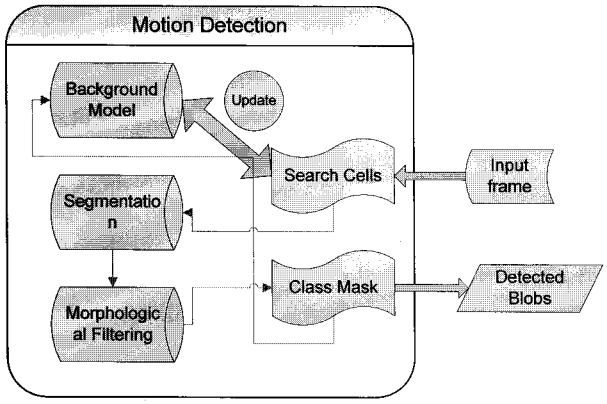


Fig. 3. The motion detection module

## IV. Object Tracking

The multiple objects tracking depends on the modeling of whole image as a set of layers and the iterative execution of mean shift estimation for each object. The representation includes a single background layer and one layer for each object. The multiple objects tracking is disassembled into times of single tracking. After creating the confidence map, the mean shift tracking is executed. It iteratively searches the closest mode of a sample distribution by locating the zeros of the gradient function of density estimation.

The core of mean shift algorithm [7] is to translate the target location x by mean shift vector $\Delta x$ :

$$m_{h,G}(x) = \Delta x = \frac{\sum_{i=1}^{n_h} x_i \omega_i g\left(\left\|\frac{y_0 - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n_h} \omega_i g\left(\left\|\frac{y_0 - x_i}{h}\right\|^2\right)} - x \qquad (3)$$

$g(x)$ is the derivative of $k(x)$ for all x points, except for a finite set of points. $y_0$ is the target centre.

The colour histograms can be used to characterize the target appearance. The reference target is represented by its probability density function in the histogram feature space. The target model $\tilde{q}_u$ and candidate $\tilde{p}_u$ can be represented as following functions.

$$\tilde{q}_u = C\sum_{i=1}^{n} k\left(\left\|\frac{\vec{x}_0 - x_i}{h}\right\|^2\right)\delta\left[b(x_i) - u\right] \qquad (4)$$

$$\tilde{p}_u(y) = C_h\sum_{i=1}^{n_h} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right)\delta\left[b(x_i) - u\right] \qquad (5)$$

The normalization constant $C$ and $C_h$ make sure that the sum of probability distribution is one. The function b(x) associates to the pixel at x and maps the pixel to its bin space. $\delta$ is the Kronecker delta function.

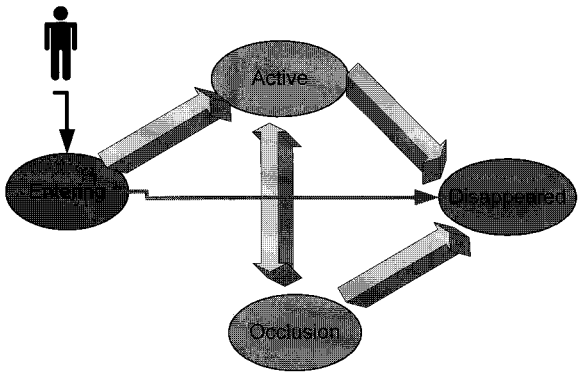


Fig. 4. Tracking state definition and transformation

Because the peripheral pixels are always contaminated by fast moving, clutters and disturbance of the background pixels, the confidences of the pixels should be different. The kernel with a convex and monotonic decreasing profile assigns smaller weights to peripheral pixels, which emphasizes more reliable pixels around the centre and reduces the disturbance of contaminated peripheral pixels.

Each tracker has a state association, which records how the system treats the tracker. The state has different definitions and it is vividly displayed in Figure 4.

- Entering: the target is first detected and the tracking should be continuously executed. If some frames having no detection results in deleting the corresponding tracker.
- Active: the target is well tracked for successive frames. It means the normal tracking state.
- Occlusion: the target is well tracked in previous frames but lost in the last frame. The estimated position of the target is in the valid area of the surveillance scene. Otherwise, we treat it as the disappeared state. Tracks in this state will switch back to the active state once relo-

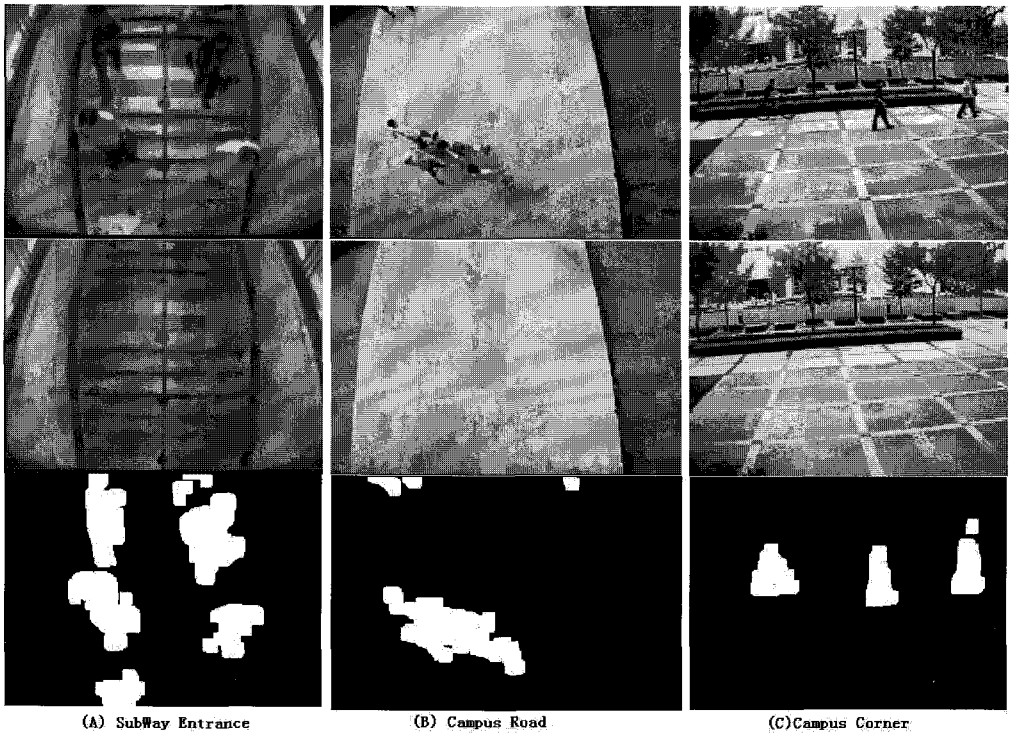


Fig. 5. Experiments on three kinds of video resources. (a) The video source is the camera in a subway entrance; (b) the video source is a camera in the campus monitoring network; (c) the video source is a home digital video-camera. All the cameras are fixed and low cost. The camera views are different: vertical, acclivitous and horizontal.

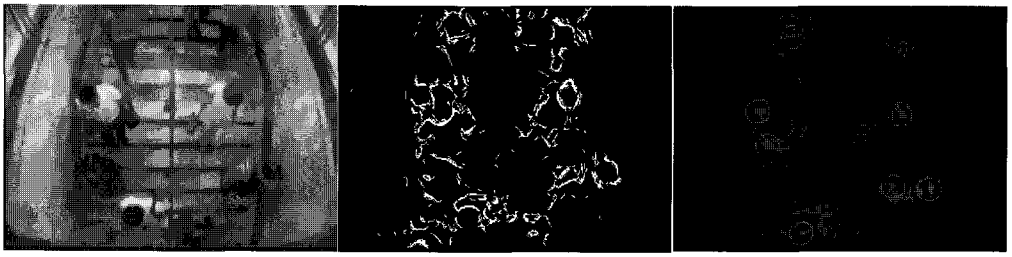

Fig. 6. Hough transformation result for human heads. The left image is the source frame, the middle is the extracted silhouette and the right is the Hough transformed result with the red circle labeling the head position.

Fig. 7. The tracking experiment on the surveillance of a hall with a stationary camera. Two persons walk together and the third person appears in the south-east corner. They meet at a point near the centre of the image, and then they are dispersed in different directions. Partial occlusions happen when they meet. The images show that the system is robust to partial occlusions. Frame indexes, starting from the top left corner, are 6, 24, 39, 74, 80, 87, 95, and 115.

cating the target, or they will be deleted if they lose the target for the timeout threshold.

Disappeared: the estimated target position is out of the valid position and the detection no longer exits. The corresponding trackers are deleted permanently.

The system combines two observations to classify people and the other objects. Human body shape is symmetric and people exhibit periodic motion while they are moving. Hough transformation is used to detect human head, which is around the circular silhouette from the top view. It can help excluding the non-circular shape objects. In our system, we discard the traditional Hough transformation and apply the reformed method. The new Hough transformation saves computation ability and keeps good detection. With the uncertain radius, we still limit the parameter space in two dimensions. The circular center is located in the line defined by the combination of foreground pixels and their orientation information.

With the mapped pixels classified into the symmetric and non-symmetric groups, the shape periodicity of each non-symmetric region is computed individually. A non-symmetric region which has no significant periodicity is considered as an object carried by a person. The non-symmetric region which has significant periodicity is considered as the body part and merged with nearest head region.

## V. EXPERIMENTS

Figure 5 is the experiment of our system on the normal fixed camera resources. It shows the performance of background modeling and the foreground detection. Figure 5(a) is the result for the video source of a subway entrance. There are many people continually walking, or running through the camera view. The passing people scarcely stay in the entrance and include the condition of both sporadic persons and the crowds. The performances in a campus road and a corner are shown separately in Figure 5(b) and Figure 5(c). The top row contains the original frames, the middle row is the background estimation and the bottom row shows the post-processed foreground.

Figure 5 proves the effective performance of our background modeling and target detection. The method provides a clean and accurate target extraction for the further processing. The detection errors are exhibited.

Table 1 generalizes the counting result of people entering the subway gate. In the video of subway entrance, the system aims to count the number of people entering the subway. Real number counted by the human labor is 732 times, while the mean counting number and the mean error are computed by averaging ten times of results of running on the same video. The times of experiments demonstrate that the counting ability of our system has low error rate, which is controlled less than 5%.

The Hough transformation used to detect the human head provides a cue to identify the human existence. A recording of Hough result is shown in Figure 6. The highlighted position is the estimated head centre, which is also considered as a human label.

Table 1 . The number of people counted through the scene

| Source | Mean Count | True No. | Mean Error |
|---|---|---|---|
| Subway | 752 | 732 | 3.4% |

The test sequence is the video of subway entrance. We can find that the Hough detection locates the human head as searching the circular shape object, while some false noise occurs at the same time. Depending on the tracking estimation and silhouette analysis, we can further smooth the noise.

Figure 7 is a set of images from the tracking results in a building hall. The video source comes from the CAVIAR project, which is the context aware vision using image-based active recognition and funded by the EC's information society technology's program. The results prove our system's superiority in adaptive background modeling and robustness to partial occlusions. The person, appearing from the south-east corner and labeled by the box with number 1, experiences the light changing when he passes by the sunny area in the 39$^{th}$ frame. Partial occlusions occur while the people are meeting, for example in the 74$^{th}$, 80$^{th}$ and 87$^{th}$ frames. The person targets are tracked correctly in spite of such difficulties. As the target withdraws from the camera view, the corresponding tracker is destroyed.

## VI. CONCLUSIONS

We have presented a real-time visual surveillance system operated on the embedded platform. Apart from the traditional image processing methods, the new novel techniques are designed in our system. The multimodal mean background modeling improves the adaptive competency for changing environment and has fast computation. Mean shift based tracking strengthens the system to be more robust to the changes and occlusions. Some special modifications reduce the algorithms' computation complexity and realize the efficient performance on the embedded system. The real applications have proved the well working of our system, while the further improvement will be studied in the recognition related module.

## REFERENCES

[1] J.Aguilera, D.Thirde, M.Kampel, M.Borg, G.Fernandez and J.Ferryman, *Visual Surveillance for Airport Monitoring Application,* 11$^{th}$ Computer Vision Winter Workshop 2006

[2] C.Wren, A.Azarbayejani, T.Drarell, and A.Pentland, *Pfinder: Real-Time Tracking of the Human Body*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, No. 7, July,1997

[3] Ismail Hartitaoglu, David Harwood, Larry S.Davis, *$W^4$:Real-Time Surveillance of People and Their Activities*, Trans. Pattern Analysis and Machine Intelligence, vol. 22, No. 8, August,2000

[4] Robert T. Collins, Alan J.Lipton, etc. *A System for Video Surveillance and Monitoring*, A report in Carnegie Mellon University, 2000

[5] Radke, R.J., Andra, S., Al- Kofahi, O., Roysam, B., *Image change detection algorithm: A systemaic survey*, IEEE Trans. on Image Processing, 14(3) pp.294-307, March, 2005

[6] Piccardi, M., *Background subtraction techniques: a review*, IEEE International Conference on Systems, Man and Cybernetics, vol 4., pp 3099-3104, October, 2004

[7] D. Comaniciu, V. Ramesh, and P. Meer, *Kernel-based object tracking*. IEEE Trans. Pattern Analysis and Machine Intelligence, 25(5), 564–577, 2003

[8] C. Stauffer and W.E.L Grimson, *Adaptive background mixture models for real-time tracking*, Computer Vision and Pattern Recognition, pp246-252, June, 1999

[9] Toyama, K., Krumm, J., Brummitt, B., and Meyers, B., *Wallflower:Principles and Practices of Background Maintenance*, ICCV, pp 255-261, 1999

[10] S.Apewokin, B.Valentine, L.Wills, S.Wills, A.Gentile, *Multimodal Mean Adaptive Backgrounding for Embedded Real-Time Video Surveillance,* IEEE, International Conference of Computer Vision and Pattern Recognition, 2007

**Qifeng Qiao** received the B.S. degree in electrical engineering from Beijing University of Aeronautics and Astronautics in 2004. After graduation, he was admitted for M.S. candidate in School of Information Science in Tsinghua University. He is now interested in video image processing, motion tracking, and activity recognition. His project experience includes optical coherence tomography, visual surveillance, image processing and DSP system.

**Yu Peng** received the B.S. degree in electrical engineering from China Jiliang University in 1999. After graduation, he worked in China Second Artillery Corps. In 2005, he was admitted for M.S. candidate in Automation Department in Tsinghua University. His research interest contains multi-instance learning and image retrieval.

**Dali Zhang** received the Ph.D. degree in electrical engineering from Stuttgart University in Germany in 1992. He is now a professor in Automation department in Tsinghua University. He is also a member in standing committee of China Society of Image and Graphics. His research interest contains image processing, image analysis, pattern recognition, embedded system and computer vision. He has published more than sixty papers in journals and international conferences. His projects include the national 973 research funding.