Real Time Pedestrian Detection Algorithm by Mean Shift

Qing Tian
College of Information Engineering, North China University of Technology, Beijing, China
Email: tianqingncut@yeah.net

Shuai Qiao
College of Information Engineering, North China University of Technology, Beijing, China

Teng Guo
College of Information Engineering, North China University of Technology, Beijing, China

Yun Wei
Beijing Urban Engineering Design and Research Institute Beijing, China

***Abstract*—Conventional moving objects detection algorithm associated with visible image is often affected by the change of moving objects' shapes, illumination conditions and is also influenced by complex backgrounds, shadow of moving objects, moving objects of self-occlusion or mutual-occlusion phenomenon. This paper presents a method of human detection by mean shift based on depth map. By analyzing and comprehensively applying segmentation method based on height information to extract moving target and remove the background information from depth map, the region of interest (ROI) with moving target should be found, then through mean shift method the goal of real-time objects (pedestrian) detection can be achieved eventually. In this paper, using the depth image pattern recognition is a good way to overcome the difficulties that visible light image pattern recognition often encounters. The depth image pixel value is only related to the distance from the surface of the object to the view window plane. Therefore, depth image has nothing to do with color space and does not suffer from the factors such as illumination, shadow effect. In addition, the mean shift targets detection method with high efficiency and fast speed features can solve the problems of low identification efficiency and poor real-time performance based on traditional pedestrian detection system to a certain extent. Our algorithm using mean shift method based on depth information has been tested on several image sequences and shown to achieve robust and real-time detection.**

***Index Terms*—human detection, depth image, height division, mean shift**

I. Introduction

In recent years with the functions of intelligent monitoring system more and more powerful, the demand for intelligent video monitoring system increases rapidly and it is gradually being applied in almost all walks of life. However, pedestrian detection is the kernel of these monitoring systems. Many methods have been offered for moving objects tracking in image sequences so far. Mean shift algorithm is one of the noted methods [1]. Due to the mean shift target detection tracking algorithm with the advantages of a good real-time performance, low computations, independence of turning and transformation [2], this paper proposed a mean shift method based on depth image to detect Pedestrians. Mean shift algorithm is a kind of statistical probability density gradient algorithm which utilizes kernel function histogram model of rotation and has no sensitive to the edge of blocks, background movement, the target rotation and deformation. Besides, mean shift algorithm has fast and effective characteristics and can also well solve the matching problem between two frames of moving targets. Fukunaga [3] put forward the concept of mean shift for the first time in an essay about probability density gradient function such as estimates in the literature in 1975. But at the beginning of this concept proposed, mean shift was not noticed by people, not mention to the use of actual needs. Yizong Cheng [4] published an important literature about mean shift in 1995. This text did supplements to the basic mean-shift algorithm in two aspects: define a set of kernel functions, which makes the magnitude of contributions of offsets to the mean shift vector follow with the distances from the sampling point to the center of mass; cited a weight coefficient, which leads to different sample points have different importance to the mean shift vector, thus making mean shift algorithm extend the limits of application. Comaniciu [5] and others put mean shift tracking method into practical application and opened the door of mean shift applications in target tracking. Yuan Xiao [6] utilized a single histogram to describe the color characteristics of the object and combined with mean shift algorithm to track moving people. To improve theoretic limitation of mean shift, Zhu Sheng-li [7] proposed an algorithm using mean shift and Kalman filter for fast tracking motion objects. On the basis of this, Baohong Yuan [8] came up

doi:10.4304/jcp.9.6.1341-1346

an improved algorithm and overcame the defect of occlusion in the process of tracking moving object well.

Using the depth image for moving targets detection is a rising technology in recent years. Especially after the Prime Sense Company of Israel launched Kinect (Xbox 360 special external equipment based on 3D measurement technology) for Microsoft in April 2010. More and more scholars pay attention to this new field. Because the depth image pixel value only related to the distance from the surface of the object to the view window plane, it can effectively break through the problems and bottlenecks of optical image recognition. In a certain space range, it can be used to represent the object coordinates in the 3D space [9]. Joshua Fabian [10] developed a "VU-Kinect" block and showed the utility of both the VU-Kinect block and the Kinect itself through a simple 3D object tracking example, which helps fully realize the Kinect' s potential. Junping Zhang [11] described a system for predicting pedestrian counts that significantly extended the utility of statistical learning algorithms.

In this paper, we first use the Kinect motion-sensing camera to obtain depth image. Because there is no shadow in the depth image and it is not affected by illumination change factors of interference, we employ segmentation method based on distance information to remove background and extract the moving object. In terms of the human body detection, we use the mean shift method for detection. Mean shift algorithm is an algorithm based on density gradient rise, which has advantages of low computational complexity, simple implementation and also can ensure the real-time performance of the system [12]. Mean shift pedestrian detection method based on the depth image can not only effectively solve the problem of interference light changes and occlusion, but also receive good detection effect while applying to the high density passenger flow detection in complex environment.

The paper is organized as follows: pedestrian detection method by mean shift algorithm based on depth image is introduced in section 2. Section 3 shows the efficiency and accuracy of the algorithm through the experimental results. Section 4 is the conclusion.

II. Pedestrian Detection Method Based on Depth Image and Mean Shift

Moving target detection technology based on the video has been one of hot topics in the field of computer vision. The target detection is the analysis of video sequences, looking for moving targets in ROI, and making judgments about the targets' location, size, speed and other motions in video sequences so that gaining the moving target real-time state information. It also serves as an initial step of the research on pose estimation, tracking, trajectory analysis, calculation of passengers flow density and speed or activity recognition.

In this paper, we reconstruct target states directly by existing information of target detection in video sequence, finding moving target preliminary location (region of interest) which uses the method of image segmentation based on the distance information, then using mean shift pedestrian detection method to detect moving targets. Our algorithm not only utilizes depth information, but also combines with mean shift pedestrian detection algorithm to obtain faster and more accurate detection results. Fig. 1 shows the whole algorithm flow chart.

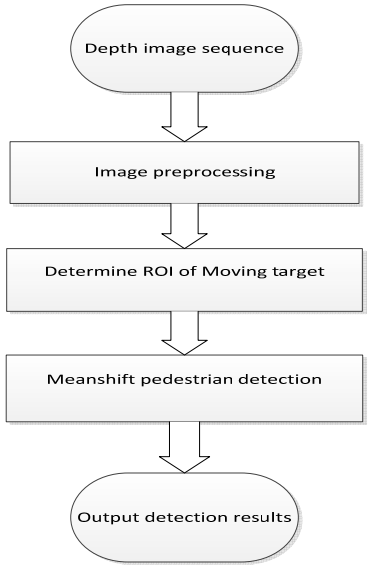

Figure 1. The whole algorithm flow chart

A. Image Preprocessing

Actual image pretreatment process is the process of removing useless information so that improving the algorithm efficiency and speed. It directly affects the computation complexity of next processes. So it is essential to do some pretreatments for raw image sequences. In our algorithm, we firstly smooth the image by median filter. The main purpose is to eliminate noise in image so that improving the image quality and facilitating the subsequent processing and analysis. Median filter [15] is a nonlinear processing method for noise suppression, the basic principle of median filter is using at various points in the field of a point at which the value of the median instead a digital image or the value of the point in the sequence, let the surrounding pixel values close to the real value, thereby eliminating isolated noise points.

For the pixel x, Assuming that there is n neighborhood pixels, put these n pixel gray values in accordance with the order from small to large, then obtain sorted sequence p_1 , p_2 , … , p_n . The gray value of pixel x is set to $p_{\frac{n+1}{2}}$ (n is an odd number)or $\frac{p_{n/2} + p_{n+1/2}}{2}$ (n is an even number), that is:

$$x = med(p_1, p_2, \ldots, p_n) = \begin{cases} p_{\frac{n+1}{2}} & (n\%2 \neq 0) \\ \frac{p_{n/2} + p_{n+1/2}}{2} & (n\%2 = 0) \end{cases} \quad (1)$$

B. Find out the ROI with the Moving Target

This paper removes the background based on segmentation method of height information, and then finds out the ROI with moving targets. For the convenience of data processing, we convert actual distance data to the actual pixel values of depth map for processing, and the camera maximum acquisition range is a known parameter which divides the gray image pixel information into 0 to 255 gray scales, so we can divide effective distance into 255 equal parts, which is corresponding to each pixel level of depth image. After that, the actual depth information is able to be represented by gray values in depth image. We usually get through the practical application scene to determine the camera construction height, and then estimate the distance from pedestrian's head and shoulder to the camera. After that, the next step is to set gray threshold value through conversion relationship of the actual distance and the gray value of depth map. On the assumption that gray value of a height for S:

$$S = H/L_{\max} * 255 \tag{2}$$

The H is the distance from the surface of the object to the view window plane, and $L_{\max}$ is the maximum effective distance of camera. In order to make gray values consistent with the height of pedestrian, we take a reverse operation for pixel values of depth map. After that, if src (x,y) (Original pixels) > threshold; dst(x,y) (Target pixels) = src (x, y), if src (x,y) (Original pixels) < threshold, dst (x,y) (Target pixels) = 0. Through this operation, moving targets can be separated from the background. Supposing filtering gray information under 1000 mm (high threshold), we can convert high threshold to gray threshold through formula (2), and then obtain the image which only contains potential targets. Through this method, interference images in complex background would be removed by setting right height threshold. Then it can determine ROI of depth image (moving target area).

C. Mean shift Pedestrian Detection

Mean shift algorithm's application range is considerably wide, such as in the image smooth aspect, image segmentation aspect and target tracking aspect. Due to the advantages of target histogram features stable, easy to calculate, we choose target gray histogram as the search feature, and through continuous iterative mean shift vector let algorithm converges to the true location of the target, so as to achieve the purpose of detection. Mean shift algorithm based on histogram can not only track the target accurately, but also solve the problem of tracking lost partially caused by occlusion to a certain extent. In addition, it also has good robustness and high efficiency. Assuming that a given Euclidean space R^d , n sample points are x_i (i=1… N), the basic form of the mean shift vector at the x point is defined as:

$$M_h(x) = \frac{1}{k} \sum_{x_i \in S_h} (x_i - x) \tag{3}$$

S_h is an area of a high dimensional ball whose radius is h, and it is an assemblage of y points satisfying the following relations:

$$S_h(x) \equiv \left\{ y : (y-x)^T (y-x) \le h^2 \right\} \tag{4}$$

k represents that at these n point samples x_i ,there are k points in the region S_h . $(x_i - x)$ is the offset vector of the sample points x_i and x . The definition of the mean shift vector $M_h(x)$ in Eq. 3 is to offset vector summation and then average of k sample points (fall into the region S_h) relative to the point x . If sample points x_i are sampled from a probability density function $f(x)$, due to the non-zero probability density gradient direct to the larger probability density direction, the sample points in the area of S_h more fall on the direction of probability density gradient. Therefore, the mean shift vector $M_h(x)$ should point to the direction of probability density gradient.

As is shown in Fig. 2, the range of big circle is, small circles represent sample points falling into the area of S_h ($x_i \in S_h$), and black spot x is the benchmark for the mean shift, the arrows represent offset vectors that the reference point x relative to sample points. Obviously, the average deviation vector will point the area of most samples distributed where it is the gradient direction of probability density function.

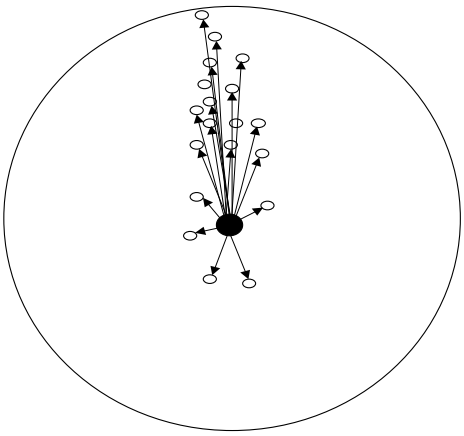

Figure 2. Mean shift schematic plot

During the execution of mean shift pedestrian detection algorithm based on the depth map, our algorithm firstly finds out potential target area (ROI) through background segmentation method based on the height information, then utilizes a rectangular box to calibrate the ROI which is kernel function impact on. In the process of target detection, the first step is selecting the gradient as a feature of target, setting up the target model at the determined target area in the image frames, then the specific method is to calculate the probability of each characteristic value in the feature space of target zone, establish a target model, and then calculate each Eigen value of the feature space in the target area which may exist in subsequent frames, After that, the candidate target model is established. The next step is using mean

shift algorithm to detect targets. In mean shift algorithm, kernel function must satisfy such requirements:

$$\int_{R^d} k(x)dx = 1 \tag{5}$$

$$\int_{R^d} xk(x)dx = 0 \tag{6}$$

$$\lim_{\|x\|\to\infty} \|x\|^d \; k(x) = 0 \tag{7}$$

Epanechnikov kernel [13] is usually selected as kernel function. Eq. (8) shows the expression of this kernel function.

$$K(x) = \begin{cases} \frac{1}{2} c_d^{-1}(d+2)(1-\|x\|^2), & \|x\| < 1 \\ 0 \quad , & others \end{cases} \tag{8}$$

where c_d represents the volume of a unit ball, $\|x\|$ is the norm of x.

After that, the next step of this algorithm is calculating similarity degree current frame between the candidate target model and target model by using the similarity measure function, after a finite time the target in the current will converge to the its true location eventually by iterative calculating the mean shift vector. Specific steps are as follows:

1. The establishment of target model

Make m_0 represents the center of the target area, $\{m_i\}_{i=1,2,\cdots,n}$ represent n pixels fallen into the target zone, n is the number of eigenvalue interval, the target model can be expressed as the following vector form:

$$\hat{k}_n = S\sum_{i=1}^{n} w(\left\|\frac{m_0 - m_i}{h}\right\|^2)\varsigma[t(m_i) - \mu] \tag{9}$$

Among them, S is the corresponding normalization constant, making $\sum_{n}^{a} k_n = 1$, the definition is as follows:

$$S = 1\Big/\sum_{i=1}^{n} w(\left\|\frac{m_0 - m_i}{h}\right\|^2) \tag{10}$$

$w(x)$ is the selected outline function of kernel function. The $w(x)$ assigns a largest value to the center pixel points, but a small weight value far away from the center pixel. Due to the pixels near the center of target model are more reliable than far away from the center point, it has certain inhibitory effect to occlusion and the influence of background. h is the width of window. We use $\left\|\frac{m_0 - m_i}{h}\right\|$ to eliminate the influence of different target calculation in the kernel function of target dimension size change, and we should raise the value of h when increasing the target, on the other hand, lower the value of h. $\varsigma(x)$ is Kronecker function, $\varsigma[t(x_i) - \mu]$'s role is to determine whether pixel color values of x in the target area belong to the n eigenvalue interval, If the pixel belongs to the area the value is 1，Otherwise, the value is 0.

2. The establishment of candidate model

For motion image sequence, the adjacent frames similarity is very large, therefore, in the second frame and other subsequent frames, we usually choose the testing results v_0 of the previous frame as the under test area center, $\{m_i\}_{i=1,2,\cdots,n}$ represent all pixels fallen in the this area, however bandwidth is still h, the target model can be converted to:

$$\bar{r}_n = S_h \sum_{i=1}^{n_i} w(\left\|\frac{v_0 - m_i}{h}\right\|^2)\varsigma[t(m_i) - \mu] \tag{11}$$

Among them, $S_h = 1\Big/\sum_{i=1}^{n} w(\left\|\frac{v_0 - m_i}{h}\right\|^2)$ is the normalized constant, n_i represents the number of target points in candidate regions.

3. The similarity detection

In the ideal situation, the similarity degree of candidate target model and target model should reach 100%. But in the actual application, due to noise and influence of lights, the change of targets' shape, it is not able to guarantee that the two models completely match. Although there are lots of similarity functions, Bhattacharyya coefficient is superior to other similarity function in the mean shift algorithm [14]. Bhattacharyya similarity expression is shown as follows:

$$\hat{\rho}(v) \equiv \rho(\hat{p}((v), \hat{k}_n)) = \sum_{n=1}^{a} \sqrt{\hat{p}(v)\hat{k}_n} \tag{12}$$

Bhattacharyya similarity value's range is [0, 1], and the value represents the similarity degree between two models. The more similar between target model and candidate model, the larger of the coefficient is. Through iteration calculating the Bhattacharyya coefficient of mean shift vector in the current frame, the last candidate region of making maximum can be thought as the final location of target in the frame.

4. The target positioning

In order to calculate the final location of the target in a frame, mean shift algorithm first makes the location of v_0 in the previous frame as the original target center of current frame, then searches for the optimal matching target location since from v_0 . Let v represents the current target center, then calculate the candidate target model character description $\hat{\rho}(v_0)$. We make Taylor series expansion for equation (12) at the position $\hat{\rho}(v_0)$, the Bhattacharyya coefficient can be expressed as:

$$\rho(\hat{p}(v),\hat{k}) = \frac{1}{2}\sum_{n=1}^{a}\sqrt{\hat{p}_n(v_0)\hat{k}_n} + \frac{S_h}{2}\sum_{i=1}^{n_h} z_i w(\left\|\frac{v-m_i}{h}\right\|^2) \quad (13)$$

Among them, $z_i = \sum_{n=1}^{a}\sqrt{\frac{\hat{k}_n}{\hat{p}_n(v_0)}}\varsigma[t(m_i)-\mu]$ is defined as weight coefficient in the process of calculation. As it can be seen from equation (13), the first part is independent of v , and only the second part is concerned with v . The second part can also be represented as a kernel density estimation that the weight is z_i and the expression is shown as the following:

$$f_{n,w} = \frac{S_h}{2}\sum_{i=1}^{n} z_i w(\left\|\frac{v-m_i}{h}\right\|^2) \quad (14)$$

Only Bhattacharyya coefficient is greater than a certain threshold, it can be ensured to take the maximum value, and mean shift vector can be calculated and work out the best corresponding position. The process of looking for maximum value of Bhattacharyya coefficient can be done by an iteration of mean shift vector. Candidate regional centers moving to the real target area v vector can be defined as the following form:

$$d_{h,G}(m) = v_1 - v_0 = \frac{\sum_{i=1}^{n_h} m_i z_i g(\left\|\frac{\hat{v}_0 - m_i}{h}\right\|^2)}{\sum_{i=1}^{n_h} z_i g(\left\|\frac{\hat{v}_0 - m_i}{h}\right\|^2)} - v_0 \quad (15)$$

Among them, $g(m) = -w(m)'$, $d_{h,,G}(m)$ is the mean shift vector which represents the target center moving from the initial point v_0 to the direction of v , thus it can be seen, through an iterative process, mean shift vector is from the initial point v_0 moving towards the direction where the colors of two models are the most similar. This direction is also same as the density gradient direction.

III. The Experimental Results Analysis

In our experiment, we first utilize Kinect camera to collect real-time video which contains depth information of objects, and the image resolution is 640 × 480. In order to maintain the real-time requirements of video image collection and transmission frames rate keeping at 20-30 fps, we use MPEG-4 as image coding and decoding standard. The experiment is performed on PC with Intel(R) Core(TM) i7-2600 CPU 3.4GHz master frequency, 8 thread, 8G memory and windows7 pro-OS. From the experimental results, we know that when the moving targets appear less (0-3) in detection area，each image processing time is between 15 and 40 ms, and when the moving target more (more than 5), each image processing time is around 40-100 ms, which basically satisfies the requirement of real-time.

Our detection system was set up at the top of a teaching building hall. There are many pedestrians in this actual application occasion, so it has a high demand for real-time accuracy of the detection system. The experimental results are shown in Fig. 3.

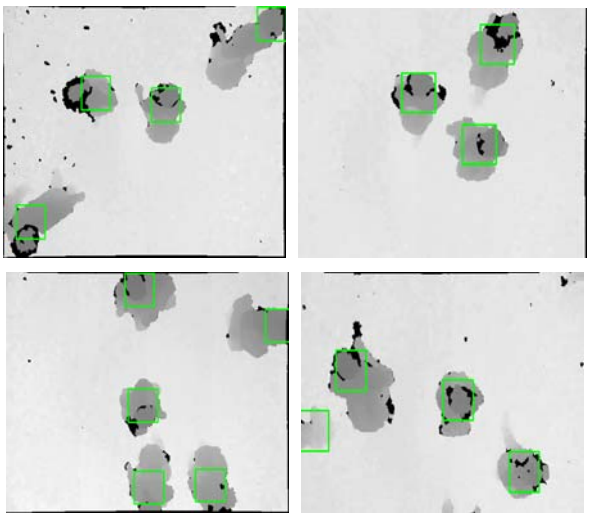

Figure3. The detection results of our algorithm

Then we calculate the detection rate and accuracy through the following formulas:

$$Detection\ rate = \frac{TruePos}{TruePos + FalsePos} \quad (16)$$

$$Accuracy = \frac{TruePos + TrueNeg}{TruePos + TrueNeg + FalseNeg + FalsePos} \quad (17)$$

From the formula (16) (17), the TruePos substitutes for the number of correct detections for pedestrian targets, the FalsePos substitutes for the number of false detections for pedestrian targets, the TrueNeg represents the number of correct pedestrian detections for non pedestrian targets (always 0), the FalseNeg represents the false pedestrian detections for non pedestrian targets. Table I shows the experimental results of our algorithm.

From the experimental results, our algorithm achieves good performance on accuracy and also satisfies requirements of real-time. Especially, comparing with algorithms based on color features, our algorithm can effectively address the case of interference light changes and occlusions.

TABLE I.
THE ACCURACY OF OUR ALGORITHM

Detection result	TruePos	TrueNeg	FalsePos	FalseNeg
	200	0	13	4
Precision	97%			
Accuracy	92.16%			

IV. SUMMARY

This paper mainly introduces a kind of target detection technology by mean shift based on depth image. Because mean shift tracking algorithm just calculates the pixel values of the probability density in target area, and only compares the Bhattacharyya coefficient between target model and candidate target model, the calculation of our algorithm is relatively simple. Besides, the algorithm only counts pixel values in ROI which further reduces the complexity of calculation. Especially, the algorithm just counts pixel values and does not care about changes of target shapes, rotation, small occlusions and so on, so the thesis algorithm has good robustness. Due to pixel gray values in depth image only concerned with the distance between viewing window plane and object surface, depth image has the unique characters which is different from characters of color space: there is no shadow in depth image; it is not affected by illumination change factors interference, so using depth information of image to detect moving targets can overcome some occlusion and overlap problems successfully. At last, through several experiments, the detection algorithm is proved that can be well applied to complex scenes of high density passenger flow detection with great efficiency and accuracy.

ACKNOWLEDGMENT

This work is sponsored by Project of Beijing Municipal Commission of Education (No.KM201210009008) and Natural Science Foundation of China (No. 61103113).

REFERENCES

[1] D. Comaniciu, P. Meer, "Mean Shift Analysis and Applications", *Proc. Seventh International Conference on Computer Vision*, pp. 1197-1203, Sept. 1999.

[2] M. J. Deilamani, R. N. Asli, "Moving Object Tracking Based On Mean Shift Algorithm and Features Fusion", *International Symposium on Artificial Intelligence and Signal Processing*, pp. 48-53, June 2011.

[3] K. Fukunage, L. D. Hostetler. "The estimation of the gradient of a density function with application in pattern recognition", *IEEE Trans. Information Theory*, vol.14, no.3, pp. 32-40, 1975.

[4] Y.Cheng. "Mean shift, mode seeking and clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no.8, pp. 790-799, 1995.

[5] D.Comaniciu, V.Ramesh and P.Meer. "Kernel-based Object Tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.3, pp. 564-577, 2003.

[6] Yuan Xiao,Wang Liping, "Tracking Moving People Based on the Mean Shift Algorithm", *Computer Engineering & Science*，vol. 30, no. 4, pp. 46-49, 2008.

[7] Zhu Shengli, Zhu Shanan and LI Xuchao, "Algorithm for tracking of fast motion objects with Mean shift" , O*pto-Electronic Engineering,* vol.33, no.5, pp. 66-70, 2006.

[8] Yuan Baohong, Zhang Dexiang, Fu Kui and Zhang Lingjun, "Video tracking of human with occlusion based on Mean Shift and Kalman filter"，*Electronic System-Integration Technology Conference (ESTC)*, pp.148-151, Sept 2012.

[9] Lin Peng. "Body Part Recognition Based on Depth Image by Learning". School of Electronic Information and Electrical Engineering, Master thesis, 2012.

[10] J. Fabian, T. Young, J. C. P. Jones, G. M. Clayton, "Integrating the Microsoft Kinect With Simulink: Real-Time Object Tracking Example", *IEEE/ASME Transactions on Mechatronics*, vol.99, pp.1-12, 2012.

[11] Zhang Junping , Tan Ben, Sha Fei, and He Li. "Predicting Pedestrian Counts in Crowded Scenes With Rich and High-Dimensional Features", *IEEE Transactions on Intelligent Transportation Systems,* vol.12, no.4, pp.1037-1046, 2011.

[12] Erik Lilienblum, Bernd Michaelis, "Optical 3D Surface Reconstruction by a Multi-Period Phase Shift Method", *Journal of Computers,* vol. 2, no. 2, pp.73-83, 2007.

[13] Wang Shuai, "A Research of Object Tracking Based on Mean Shift", Shandong University, Master thesis, 2011.

[14] Peng Zhaoyi,, Zhou Yu, Zhu Yanhui, Wen Zhiqiang, "Application of an Improved Mean Shift Algorithm in Real-time Facial Expression Recognition", *Journal of Software,* vol. 6, no.1, pp. 100-107, 2011.

[15] H. L. Eng and K. K. Ma, "Noise adaptive soft-switching median filter", *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 242-251, 2011.

Qing Tian received his PhD degree in electronic science and technology from institute of electronics, Chinese academy of science, Beijing, China, in 2010. He is a distinguished lecturer at North China University of Technology, Beijing, China. His research interests include electrical engineering, intelligent transportation systems, pattern recognition and computer vision.

Shuai Qiao is currently working toward the master's degree in Electronic Science and technology at the College of Information Engineering, North China University of Technology, Beijing, China.

Teng guo is currently working toward the master's degree in Electronic Science and technology at the College of Information Engineering, North China University of Technology, Beijing, China.

Yun Wei received his PhD degree in intelligent transportation systems from Southeast University, Nanjing, China. He is a researcher at Beijing Urban Engineering Design and Research Institute, Beijing. His research interests include intelligent transportation systems and computer vision.