A Method for People Counting Using Low-level Features Based on SVR with PSO Optimization

Jiaojiao Yuan[1], Hong Bao[1], Haitao Lou[1], and Cheng Xu[1, 2]
[1]Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing and 100101, China
[2]Institute of Network Technology, Beijing University of Posts and Telecommunication, Beijing and 100876, China
Email: baohong@buu.edu.cn

Abstract—People counting is an important part of video surveillance. In recent years, significant progress has been made in the field using the method of feature regression. In this context, feature extraction and using a machine learning algorithm to establish the relationship of extracted feature and the number of people are two basic steps. To extract the feature of crowd, methods in the literature either using the statistics values of the foreground pixels, or using the number of corners. In this paper, in order to obtain a better description of crowd, both of the two kinds of features are obtained respectively by using the FAST algorithm and VIBE algorithm, and a processing of normalization is done to solve the problem of perspective distortion. Then, the correspondence between these features and the number of people is studied by SVR. In addition, in order to avoid the improper selection of parameters of SVR, the PSO algorithm is used to select the relevant parameters in SVR. The method has been tested on the PETS2009 datasets and the self-shooting datasets, and the experimental results show the effectiveness of the method. And, the method has been extensively compared with the algorithm by Albiol *et al*, which provided the highest performance at the PETS 2009 contest on people counting. The results confirm that the proposed method improves the accuracy and robustness.

Index Terms—FAST, VIBE, PSO, SVR, People counting

I. Introduction

Crowd density analysis is a very important research direction in the field of pattern recognition, and related technology is now widely used in video surveillance field. People counting is a very effective method to analyze crowd density. There are two different ways to count people. In the direct approach(also called detection-based), each person in the scene is individually individually detected, using some form of segmentation [1] and object detection; This problem has been addressed by adopting part-based detectors [2], or by detecting only heads [3] or the Ω -shape formed by heads and shoulders. This method is not affected by the perspective distortion, and it can obtain a higher accuracy in low density crowd. However, the detection of people itself is a challenging task, especially in occlusion and congestion scenes where accurate detection of the target becomes difficult and the results are also not credible. Recent methods typically bypass the task of detecting people and instead focus on the indirect way.

The indirect way is based on feature extraction and regression [4]-[7], using the regression algorithm to learn the correspondence between the features and numbers of people, so as to estimate the number of people in the video scenes [8]-[10]. To analysis low and middle density crowd, some statistics values of foreground pixels (for example, the number of foreground pixels, the number of edge pixels of foreground pixels; etc.) or the information of the crowd(the number or distribution of corner points) are often used. In another perspective, some other works focus on the study of regression algorithms (e.g. Gaussian Process regression, neural network in [11]). Indirect method is simpler than direct method, and has better robustness.

In this paper, a indirect method is proposed to count the people. The main contributions are the following two aspects: (1) Unlike other literature, which only extract one kind of features of the crowd, or one kind of the features is treated as the constraint of another, the method constructs feature vectors using both the statistical value of the pixels and corners, and a step of weighting the features is done to handle the problem of perspective distortion. (2)Instead of manually setting the parameters in the SVR, PSO is used to optimize the parameter selection in SVR to ensure the accuracy of the model. Due to the above steps, the method has achieved good experimental results.

The remainder of the paper is organized as follows: In the next Section II, a taxonomy of relevant works to people counting is presented. Then, the approach for people counting is introduced in Section III. A detailed experimental results and analysis follows in Section IV. Finally, a briefly conclude and an outlook to future work are given in Section V.

II. Releated Work

Existing counting methods can be classified into three categories: counting by detection [12]-[14], counting by clustering [15] and counting by regression[16]-[18]. At present, counting by regression is the most popular method.

Manuscript received July 25, 2017; revised November 20, 2017.
This work was supported by the National Natural Science Foundation of China (Grant No.91420202 and No.61271370); The Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions under project IDHT: 20140508, and the Science and Technology Plan Projects of Beijing Municipal Institutions (D161100003516003).
Corresponding author: xc-f4@163.com
doi:10.12720/jcm.12.11.617-622

Feature extraction is an important step for regression algorithm. For low density crowd, the number of pixels or corners is usually used to form the feature vector. Davies was one of the earliest researchers who used the statistical information of foreground pixels to estimate the number of people[19]. In his paper, background subtraction is used to obtain the information of foreground pixels of the moving crowd, and a view is presented that the number of people and the number of foreground pixels present a linear relationship in low density. His method was applied for CCTV in London. Albiol *et al.* found a simple linear relationship between the number of people and the number of Harris corners belong to the crowd, and the method provides the highest performance at the PETS 2009 contest on people counting. Fradi detected the corner points using FAST, and add a step to track corner points using Robust Local Optical Flow (RLOF), the crowd density map is generated to analysis crowd. It is worth noting that most of the methods including mentioned above, using only one kind of features to describe the crowd. The method is also only using the corner as a constraint on the number of pixels, rather than using all of them to form the feature vector. It means that some research work can be carried out in this regard.

Ideally, there is a simple linear relationship between features and the number of people; However, this relationship is usually affected by perspective distortion, namely, the value of the feature varies with the distance of the target from the camera, which affects the accuracy of the people counting in a certain degree. Therefore, differently from the above aforementioned works, some other methods take into account the effects of perspective distortions. Fradi weighted the foreground pixels to correct the perspective distortion, the results obtained a significant improvement than Albiol's algorithm which does not consider the perspective distortion. Conte using the Inverse Perspective Mapping (IPM) to solve the problem of perspective distortion, greatly improving the accuracy of the results. Weighting the foreground pixels is simple and real-time, and the perspective distortion has been solved in a certain degree; IPM need to calibrate the parameters of camera, which is complex to calculate, but it owns a better accuracy.

The classifier is also an important part of improving the prediction results. Chow proposed a method which using RBF neural network to fit the relationship between the number of people and the statistical values of foreground pixels, the result owns a higher accuracy than Davies's method. The setting of parameter of the classifier is very important. In general, PSO and GA can be used to optimize the parameters of the classifier [20], [21].

III. Proposed Method

The approach proposed in this paper is based on regression. An illustration of the proposed method is shown in Fig. 1. The remainder of this section describes each of these system components.

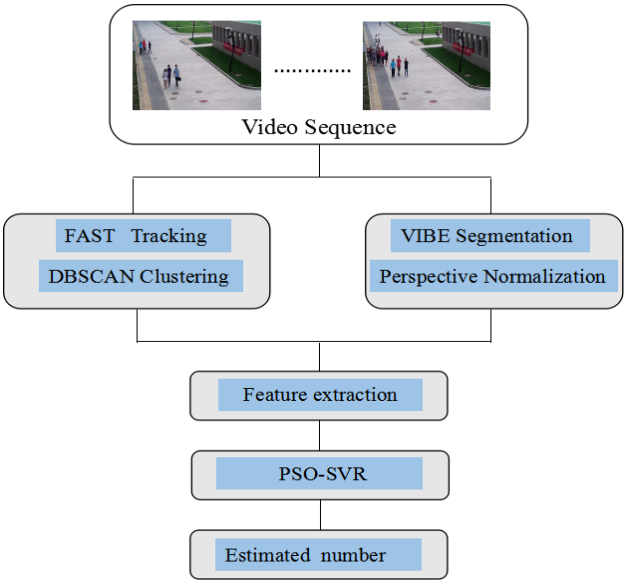


Fig. 1. The flow chart of the proposed method

As shown in Fig. 1, the method respectively using two threads to extract pixels and corner points. The two threads work in the following way: According to the result of clustering feature points, crowd is divided into different regions, and then, the number of pixels in the same region that generated by VIBE is calculated. Then, a feature vector of the region is formed by using the above values. Finally, the feature vectors are input into the regression algorithm to predict the number of people.

A. Moving FAST Points Detection

Detecting corners is a basic step to count people. Harris, SIFT and FAST are usually used to detect corners. In order to ensure the real-time and reliability of the system, the selected corner should have high computational performance and reliability. For reliability, FAST is used in [22], [23] to detect crowds from aerial images, and the results show an excellent performance. For computational performance, an experiment is done in this section to show the time taken by different algorithms. The result is shown in Table I.

TABLE I: Comparison of the Time Taken by Different Algorithms to Detect Corners

Algorithm	Time (ms)
SIFT	284.51
SUFT	266.37
FAST	30.25

The conditions for the experiment are as follows: the size of the image is 768*576; Intel Core CPU 2.9GHz; RAM: 8G.

The result of the Table I shows that FAST can detect corner points in real time. So, under the condition that both of the real-time and robustness of the algorithm are considered, FAST is used to detect corner points.

The extracted FAST features are distributed on the background and crowd, but we only need the feature

points from the crowd. Hence, we need to add in our system a separation step between foreground and background. It is done by optical flow [24] to detect local features in order to distinguish the moving and static ones. The static and moving interest points are distinguished on the basis of the following rule:

$$p(x,y)=\begin{cases} moving \quad point \quad if\left|v(x,y)\right|>\alpha \\ static \quad point \quad if\left|v(x,y)\right|\le\alpha \end{cases} \tag{1}$$

where p(x, y) is the interest point at the x, y coordinates, $\left|v(x,y)\right|$ is the magnitude of the motion vector calculated in x, y with respect to the previous frame; α is a bias value (in our experiment, α =0.5).

B. FAST point Clustering

Correct clustering of the feature points can help to construct the feature vector effectively, and then influences the accuracy of the regression model. Therefore, the clustering process is a key step in this paper. Xu [25] using OPTICS algorithm to cluster the feature points to generate the corresponding reachable graphs. Conte clustering of feature points using graph theory whose parameters is insensitive to the particular application [7].

In this paper, DBSCAN algorithm is used to cluster feature points. DBSCAN algorithm is based on spatial density, clustering the regions with higher spatial density into clusters. As shown in Fig. 2, point P and point M are reachable, point M and point Q are reachable, so point P and point Q are also reachable. The circle they owns can be treat as a cluster. DBSCAN runs fast and it can find the noise in the data, which is the traditional clustering algorithm can not do. The result of DBSCAN algorithm on some images is shown in Fig. 3, and it shows the effectiveness of DBSCAN.

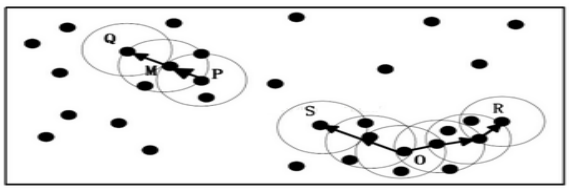


Fig. 2. DBSCAN algorithm schematic

(A)

(B)

Fig. 3. Feature point clustering results using DBSCAN algorithm: (A) Beijing Union University campus, middle density crowd; (B) PETS2009 S1.L1.13-57(view 1), middle density crowd

C. Foreground Segmentation and Perspective Normalization

In this paper, the VIBE [26]-[28] is used to segment the crowd. VIBE only needs one frame to complete the process of model initialization; In addition, it is sensitive to noise and needs fewer computational time. At the same time, it is very stable for the changes in light and camera shake and other effects. The result of the VIBE is shown in Fig. 4.

Fig. 4. VIBE Foreground segmentation

After the foreground objects are obtained, it is necessary to compute the statistical information of the foreground pixels. And, the method mentioned in [29] is used to solve the problem of perspective distortion, which weighting each foreground pixel according to a perspective map with assigning larger weight for farther points in the scene. Perspective correction based on linear interpolation is shown in Fig. 5.

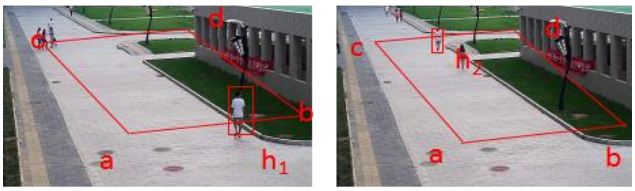


Fig. 5. Perspective Correction Based on Linear Interpolation

In Fig. 5, the weight on the line ab is 1, and the weight on the line cd is w_n, then the weights of the pixels between line ab and line cd are calculated as follows:

$$w_i=\frac{h_1\times b_2\times\left|CD\right|+h_2\times b_1\times\left|AB\right|}{(h_1+h_2)b_2\left|CD\right|} \tag{2}$$

where b2,b1 is the distance of line li(the row of the object located)between line cd and ab. The number of pixels on each line is calculated as follows:

$$N_i=\sum_{y=1}^{Y}W_p(y)*N_T(y) \tag{3}$$

where N_i is the total number of foreground pixels in the y-th row.

D. Feature Extraction and Regression

In this step, the feature vector is extracted and put into a regressor. According to the each cluster generated by clustering, the number of corners detected by FAST and the weighted pixel statistics generated by VIBE algorithm are extracted as feature vector and the number of people in each frame is given. The SVR optimized by PSO [30] is used to learn the relationship between the number of people and feature vector. The output of regression is

estimated number of people in the group represented by the cluster. The relationship between the feature vector and the number of people can be formulated as:

$$n_{people} = \sum_{i=1}^{m} f(n_{points}, n_{pixels}) \quad (4)$$

where n_{people} is the estimated number of people, m is the total number of cluster in a frame, n_{points} is the number of interest points within a cluster; n_{pixels} is the number of pixels within a cluster. The process of training is shown in Fig. 6.

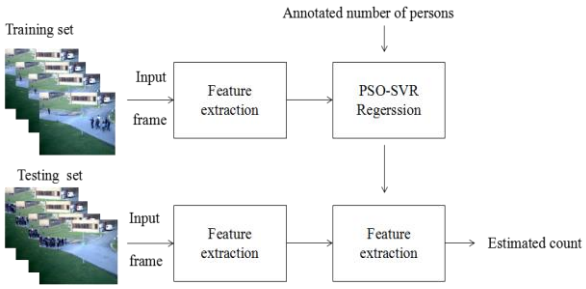


Fig. 6. The process of training

IV. Experimental Results

A. Datasets

The datasets include PETS2009 dataset and the Beijing Union University campus datasets. The size of the frame is 768*576. First of all, Polus divided the crowd density into five levels. In this paper, the definition of the crowd density levels is shown in Table II. The condition of dataset is shown in Table III.

To assess our method, we compare the estimated number of persons to the ground truth using the Mean Absolute Error (MAE) and the Mean Relative Error (MRE) metrics which are defined as:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|E(i) - G(i)| \quad (5)$$

$$MRE = \frac{1}{M}\sum_{i=1}^{M}\frac{|E(i) - G(i)|}{G(i)} \quad (6)$$

where M is the total number of frames in a video sequence. E(i) is the estimated number of people in the i-th frame, G(i) is the ground-truth number of persons in the i−th frame. MAE and MRE are also used in [6], [7].

TABLE II: Crowd Density Level Table

level	Range of Density(people/m2)	Range of People
very low	0~0.2	0~10
low	0.2~0.5	10~25
middle	0.5~1	25~50

TABLE III: Experimental Dataset

Dataset Video Sequence	Length(frame)	Conditions	Number of people
PETS2009 S1.L1.13-57(view 1)	221	Medium density crowd	5~34
PETS2009 S1.L1.13-59(view 1)	241	Medium density crowd	3~26
PETS2009 S1.L3.14-17(view 1)	91	Medium density crowd	6~41
PETS2009 S1.L1.13-57(view 2)	221	Medium density crowd	8~46
PETS2009 S3.MF.12-43(view 2)	108	Very low density crowd	1~7
BUU Video1	1250	Medium density crowd	3~21
BUU Video2	2800	Medium density crowd	15~52
BUU Video3	1475	Very low density crowd	1~8

B. Experiments Results and Analysis

For each dataset, 20% are training set, the rest are testing set. Comparison of different methods is shown in Table IV.

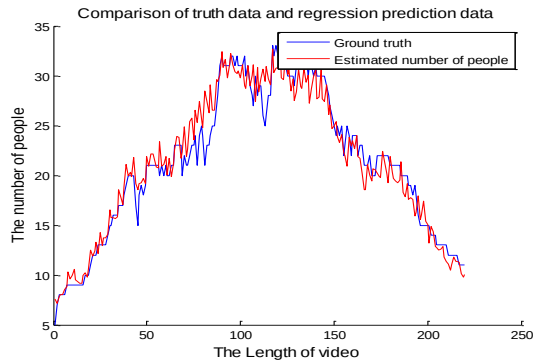
It can be seen from the above analysis that the method of this paper is superior to the method of [4] on the specified dataset, and has achieved the similar performance like the method in [7]. The main reasons are the following two points: (1) The pixels and corners are fused to obtain a richer and more precise description of the crowd, and perspective distortion is studied. (2) The PSO algorithm is used to optimize the SVR, which improves the accuracy of the model. The result on BUU is shown in Table V. The result on some dataset is shown in Fig. 7.

TABLE IV: Comparison of Different Methods on PETS2009

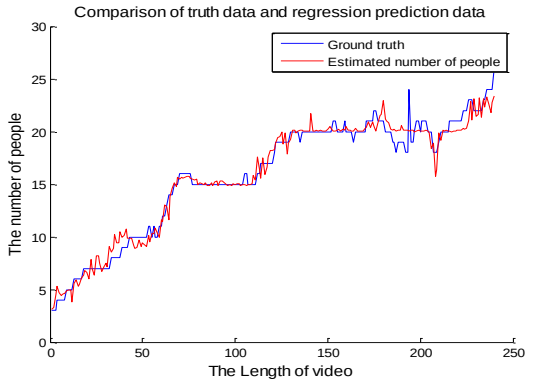
Video Sequence	Albiol et al.[4]		Conte et al[7]		Our approach	
	MAE	MRE	MAE	MRE	MAE	MRE
S1.L1.13-57(view1)	2.80	12.6%	1.92	8.7%	1.87	9.58%
S1.L1.13-59(view1)	3.86	24.9%	2.24	17.3%	2.34	15.4%
S1.L1.14-17(view1)	2.64	14.0%	1.75	9.2%	2.08	10.6%
S1.L1.13-57(view2)	29.45	106.0%	11.76	30.0%	11.54	30.5%
S3.MF.12-43(view2)	12.34	311.9%	0.63	18.8%	2.19	34.8%

TABLE V: MAE/MRE ON BUU DATASET

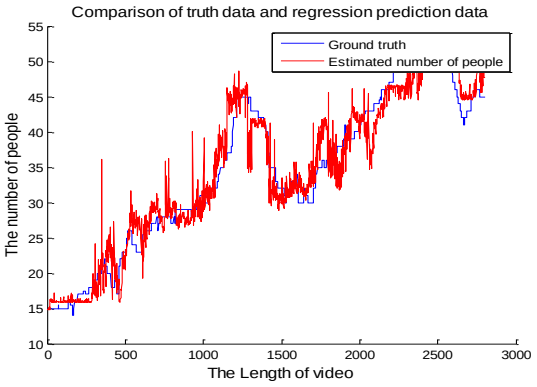
Video Sequence	MAE	MRE
Video1	0.60	7.4%
Video2	2.23	0.8%
Video3	0.43	7.4%



(A) PETS2009 S1.L1.13-57(view 1)



(B) PETS2009 S1.L1.35-59(view 1)



(C) BUU video2

Fig. 7. Compartion curves of estimated number and the number of ground truth on some video sequences in experiment dataset.

V. CONCLUSION

In this paper, we propose a method for population counting in video surveillance. This method has better effect on self-shooting datasets and the PETS 2009 database. The result of the experimentation is better than the algorithm by Albiol *et al*., This has been confirmed that the proposed method owns a higher accuracy. As a future work, a more extensive experimentation will be performed, trying to use deep learning to fit the relationship between features and numbers.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No.91420202 and No.61271370). And, the Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five–year Plan (IDHT20170511), and the Science and Technology Plan Projects of Beijing Municipal Institutions (D161100003516003).

REFERENCES

[1] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conference on Computer Vision & Pattern Recognition*, 2007, pp.1-6.

[2] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proc. Tenth IEEE International Conference on Computer Vision IEEE Computer Society*, 2005, pp. 90-97.

[3] S. Lin, J. Chen, and H. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions Systems Man, and Cybernetics-Part A Systems and Humans*, vol. 31, no. 6, pp. 645–654, 2001.

[4] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi, "Video analysis using corner motion statistics," in *Proc. IEEE Int. Workshop on Performance Evaluation of Tracking & Surveillance*, 2009, pp. 31–38.

[5] H. Fradi and J. L. Dugelay, "Crowd density map estimation based on feature tracks," in *Proc. International Workshop on Multimedia Signal Processing*, 2013, pp. 40-45.

[6] H. Fradi and J. L. Dugelay, "Low level crowd analysis using frame-wise normalized feature for people counting," in *Proc. IEEE International Workshop on Information Forensics and Security*, 2012, pp.246-251.

[7] D. Conte, *et al*., "A Method for Counting People in Crowded Scenes," in *Proc. IEEE International Conference on Advanced Video & Signal Based Surveillance IEEE Computer Society*, 2010, pp. 225-232.

[8] D. Huang and T. W. S. Chow, "A people-counting system using a hybrid RBF neural network," *Neural Processing Letters*, vol. 18, no. 2, pp. 97-113, 2003.

[9] K. Chen, *et al.*, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, 2013, pp. 2467-2474.

[10] L. Fiaschi, *et al*., "Learning to count with regression forest and structured labels," pp. 2685-2688, 2012.

[11] S. Yi, *et al*., "L0 regularized stationary time estimation for crowd group analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2219-2226.

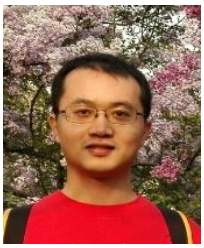
[12] Z. Wang, *et al*., "Crowd density estimation based on local binary pattern co-occurrence matrix," in *Proc. IEEE International Conference on Multimedia and Expo Workshops*, 2012, pp. 372-377.

[13] M. Li, *et al*., "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *Proc. International Conference on Pattern Recognition*, 2008, pp. 1-4.
[14] M. Rodriguez, *et al*., "Density-aware person detection and tracking in crowds," in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 2423-2430.
[15] B. Wang, *et al*., "Crowd density estimation based on texture feature extraction," *Journal of Multimedia*, vol. 8, no. 4, pp. 331-337, 2013.
[16] Y. Zhang and D. Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. IEEE Conference on Computer Vision & Pattern Recognition*, pp. 2622-2629, 2010.
[17] V. Lempitsky, S. Victor, and A. Zisserman. "Learning to count objects in images," in *Proc. Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2010, pp. 1591-1591.
[18] C. L. Chen, S. Gong, *et al*., "From semi-supervised to transfer counting of crowds," in *Proc. IEEE International Conference on Computer Vision*, 2013, pp. 2256-2263.
[19] A. C. Davies, J. H. Yin, *et al*., "Crowd monitoring using image processing," *Electronics & Communication Engineering Journal*, vol. 7, no. 1, pp. 37-47, 1995.
[20] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. 9th European Conference on Computer Vision*, 2006, pp. 430-443.
[21] L. Xie and P. Wang, "Method for estimation of crowd density using neural network with PSO optimization based on gray level co-occurrence matrix," *International Journal of Machine Learning & Computing*, vol. 3, no. 6, pp. 520-523, 2013.
[22] O. Meynberg and G. Kuschk, "Airborne crowd density estimation," in *Proc. Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-3/W, 2013, pp. 49-54.
[23] M. Butenuth, *et al*., "Integrating pedestrian simulation, tracking and event detection for crowd analysis," in *Proc. IEEE Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, 2011, pp. 150-157.
[24] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade Meets Horn/Schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211-231, 2005.
[25] X. Cheng, *et al*., "Crowd density estimation based on improved harris & optics algorithm," *Journal of Computers*, vol. 9, no. 5, 2014.
[26] K. Kang and X. Wang "Fully convolutional neural networks for crowd segmentation," *Computer Science*, vol. 49, no. 1, pp. 25-30, 2014.
[27] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 20, no. 6, pp. 1709-24, 2011.
[28] O. Barnich and M. V. Droogenbroeck, "ViBE: A powerful random technique to estimate the background in video sequences," in *Proc. IEEE International Conference on Acoustics IEEE*, 2009, pp. 945-948.
[29] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog*, 2008, pp. 1–7.
[30] X. Zhang and Y. Guo, "Optimization of SVM parameters based on PSO algorithm," in *Proc. International Conference on Natural Computation*, Tianjian, China, August 14-16, 2009, vol. 6, pp. 536-539.

Jiaojiao Yuan, is currently a master at the Information Institute in Beijing Union University, China. Her research interests include Pattern recognition and machine learning.

Hong Bao, received his Ph.D. degree from school of computer and information technology, Beijing Jiao tong University Beijing, China. He is a professor of Beijing Union University. His current research interests include intelligent control and intelligent vehicle.

Haitao Lou, is a lecture of Beijing Union University. His current research interests include digital image processing, artificial intelligence and pattern recognition and intelligent vehicle.



Cheng Xu, is currently a Ph.D. at the State Key Laboratory of networking and switching technology in Beijing University of Posts and Telecommunications (BUPT), China. His research interests include wireless security and internet of vehicle