Crowd counting method on sparse scene

Huaiming Li, Fei Wang, Fangfang Song, Lianqing Wang
Faculty of Management and Economics,
Dalian University of Technology
Dalian, China
Email:956816152@qq.com

***Abstract*—In recent years, With the development of science and technology to promote the popularity of video surveillance, computer techniques have a great value on obtaining the crowd counting information of surveillance video automatically, but perspective effects, mutual occlusion between people and other factors make crowd counting difficult. This paper presents a crowd counting method on sparse scene. Firstly, analyzing the characteristics of the surveillance video to access available prior knowledge; Secondly, combining with prior knowledge to extract the characteristics of target prospects block; Finally, support vector regression machine is employed to estimate the number. Experiments show that the method improves the situation of pedestrians occlusion crowd counting estimation accuracy.**

Keywords—Crowd Density; Prior Knowledge; Perspective correction; Exercise Intensity

I. INTRODUCTION

China is the most populous country in the world, the rapid development of urban construction has led to the lack of overall planning in the process of construction, resulting in inadequate urban population control, in particular, the population and industry of the large cities is too gathered, which brought lots of unsafe factors, a variety of safety accidents mortality rate climbed. How can we live in a safer place, how to build a strong security network to ensure the safety of the city. At the beginning of 21st century, the country put forward the concept of "safe city", in the promotion of safe city, as the core part of the video surveillance in public or private environment coverage rate showed explosive growth, tens of thousands of surveillance video only rely on the eyes of the monitoring staff has been difficult to make a timely feedback on the abnormal phenomenon. In order to liberate the eyes of the monitoring personnel and improve the work efficiency, more and more researchers committed to research intelligent video[1], they use the computers to help people obtain and analyze the number of person and the crowd behavior characteristics in the video surveillance, which has improved the efficiency of monitoring personnel.

In this paper, we study the algorithm of population number estimation at the present stage, now the mainstream crowd counting method can be roughly divided into two categories[2]:One kind is the direct method of counting the population[3, 4], which is based on the target detection method, by a good pedestrian detection algorithm to detect whether the moving object in the video is a pedestrian, and then we can directly count the number of pedestrians. In order to accurately count the number of pedestrians, these population estimation methods need to be very accurately detected the pedestrians in the video, but when the number of the people are relatively large or there are a lot of occlusion, simply cannot be accurate for the crowd of pedestrians moving target single detection and segmentation. Another kind is the indirect method of counting the population[5-9], which is based on the absence of a single target detection, through the methods of statistical learning(such as neural network, SVM) to establish the decision rules by the foreground features of the extracted moving target, and then we can estimate the number of pedestrians. This method is more efficient than the direct estimation method, so this paper chooses the indirect method to estimate the number of people.

The selection of image features and perspective effect are important factors to influence the estimation of crowd numbers. There are some commonly used image feature, such as pixel feature[6](the total area of the prospect, the edge pixel points of the prospect, the perimeter area ratio, etc.), texture features[9](gray level co-occurrence matrix, histogram, etc.), feature points[10](Harris, SIFT, etc.). The crowd number estimation method based on the statistical features of the pixel is simple to use and computing speed fast. This method can meet the needs of real-time operation, and the effect is better in the case of fewer people and no crowd, but this method is influenced perspective effect and shielding factor; The crowd number estimation method based on texture features is very suitable for the detection of high population density, and it is not affected by the occlusion and perspective or other factors, but its operation is very complex; The method based on feature points is similar to the method based on the statistical features of the pixel, it is also to establish a functional relationship between the feature points and the number of crowd, and then to solve this function, but it is very complex to select and extract the feature points. Pixel feature can be more intuitive to reflect the changes of the crowd numbers in sparse scene, the method is simple and effective. So this paper mainly uses the pixel feature.

Due to the perspective effect, it is difficult to establish the relationship model between the features and the number of crowd, this paper proposes a crowd density estimation method based on prior knowledge. In real life, when the crowd density become greater, the relative space is limited, the movement of people will be smaller. Taking into account the prior knowledge, the author defines a feature which describes the features of movement of the crowd, named as the exercise intensity. This method see prospects block as a unit, using the background subtraction and frame difference method to extract the features of exercise intensity of target prospects block, and combined with the statistical features of the pixel, and then

analyzed these features by the method of SVR (support vector regression). In addition, in view of the problem of perspective effect in video surveillance, this paper carries out a perspective correction of the extracted effective crowd features. Specific steps are as follows:

(1) According to the scene of the surveillance video, analyze the characteristics of the surveillance video under the scene, to extract valuable prior knowledge.

(2) Input video, transformed into an image sequence, and preprocessed image.

(3) Using the background subtraction method to extract the effective prospects object, and labeled the prospects block.

(4) Extracted the features of each prospects block: the number of prospects pixels, the number of edge pixels, the features of exercise intensity, the width of the rectangle of the prospects block, the gravity point coordinates.

(5) Correct the image features by the method of statistical regression.

(6) Construct the feature vector, and analyze the statistical features of the pixel of crowd. Then estimate the number of crowd of the prospects block, and add the number of estimates from every prospects block in each frame to get the number of scenes in each frame.

Fig.1 is the flow chart of the method:

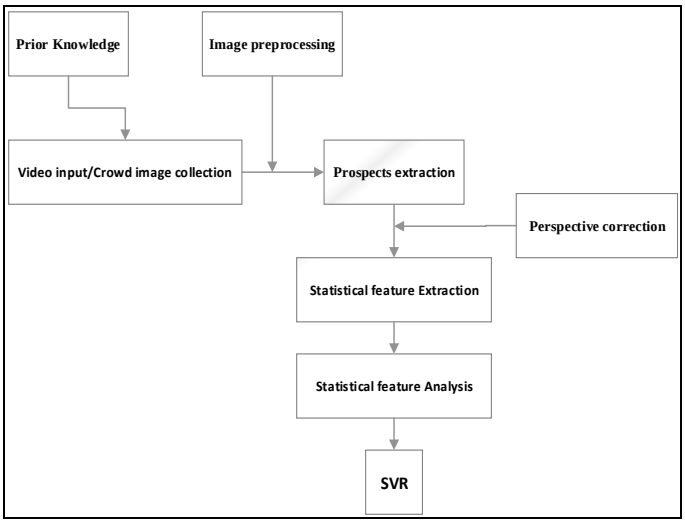


Fig.1. Algorithm Flow Chart

II. PERSPECTIVE CORRECTION BASED ON STATISTICAL REGRESSION

The camera can convert world from three-dimensional into two-dimensional by the lens and inevitably bring the perspective effect. The perspective effect refers to the Nearer distance from the camera, the greater the man, on the contrary, the farther distance from the camera, the smaller the man, Crowd prospect blocks of the same size at different distances from the camera to estimate population number may appear different results, So it is an important factor to affect the accuracy of crowd counting estimation[11].Before foreground block feature is extracted, it is very important to choose suitable perspective correction algorithm.

At present, There are two types of perspective correction method: One kind is the camera parameter calibration, which Lin [12] put forward in 2001, this algorithm with high accuracy. However, calculation is complicated and camera has changes need to calibration again. So later some scholars put forward another method, linear interpolation method[13] , Draw a reference line by using the method of geometric mathematical , according to the fixed point theorem, get the perspective correction parameter of the designated area corresponding , this parameter can reflect the quantitative differences from the same man and the different distances. This method is more simple operation and lower complexity, in the case of the serious perspective effect, this method can ease effect to some extent, but the accuracy is not high.

Inspired by the linear interpolation method, this paper found a linear relationship between the perspective correction parameters and the position people in the image. On this basis, this paper puts forward a perspective correction algorithm based on statistical regression. Compared with the above two methods, the method is simple, high precision, and the scene fit well.

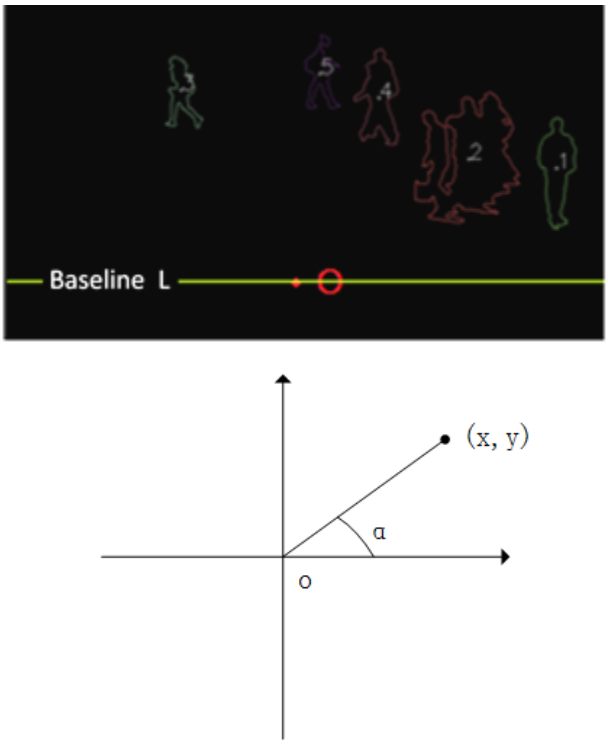


Fig.2. Perspective correction Description

Assume in the current frame a foreground block has a perspective correction parameter w, the actual area of the current frame for this foreground block can be multiplied by the correction parameters w to mitigate the effects of the perspective effect, as shown in (1). And correction parameter w is associated with marked foreground block of the center of gravity point coordinates, as shown in (2).

$$S_A = w * S_C \quad (1)$$

$$w = f(x, y, \alpha, d) \quad (2)$$

As shown in Fig.2, Assume that S_A is the prospects block area which is already correct, S_C is the prospects block area

TABLE I Statistics between the candidate variable and the corrective parameters

Index \ Predictive Variables / Numerical Value	Abscissa	Ordinate	Angle	Distance	Abscissa+ Ordinate	Angle + Distance
Residual standard error	0.7916	0.02659	0.7791	0.6829	0.02618	0.6545
Multiple R-squared	0.009962	0.99155	0.04098	0.2631	0.99157	0.3239
Adjusted R-squared	0.008788	0.98317	0.03984	0.2623	0.98356	0.3223
P-value	0.00368	<2.2e-16	2.894e-09	<2.2e-16	<2.2e-16	<2.2e-16

current frame measured , w is correct parameters, The center of gravity point coordinates $g\ (x, y)$, α is g and horizontal Angle of the origin O, d is the Euclidean distance between p and o. Input a video, extract the prospects, and obtained the corresponding $g\ (x, y)$ and the total number of pixels in every prospects blocks. Defined baseline L, when moving object through the baseline, record the current foreground block's gravity point coordinates (x_A, y_A) and the total number of pixels S_A.When motion target move to another location, also record the prospects block's gravity point coordinates (x_C, y_C) as well as the total number of pixels S_C. By the formula (1.0) will get the actual correction parameters $w = S_A/S_C$.

This paper chooses a few forecast, such as abscissa, ordinate, Angle, distance, abscissa and ordinate, Angle and distance, use these candidate variables to fit with the correct parameters w, and get statistics to select the appropriate forecast independent variable. Relevant numerical value as shown in TABLE I

By analysis of data in TABLE I, the fitness between ordinate and the correction parameter is the best, so choose the ordinate as independent variables in least-squares regression. After repeated experiments and verification, the linear relationship between ordinate and correction parameters is $: w = f(y) = my + n$, Which m and n as a parameter. Correct parameter and ordinate fitting a scatter diagram as shown below:

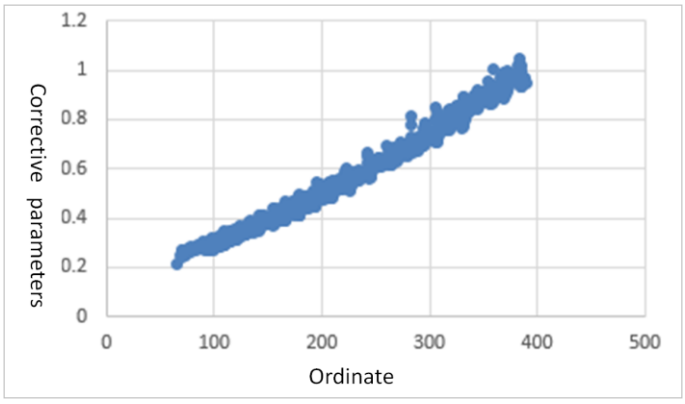


Fig.3. Scatter plot

Randomly intercept a piece from a video, respectively based on the formula (1) obtained actual correction parameters and according to the linear expression obtained estimation. Then draw the line chart to compare, as shown in Fig.4: you can see, the fitting of the correction parameters are consistent with the actual parameters approximately, the error is small, very close to real data, so this correction method is used for crowd counting estimation method in this paper.

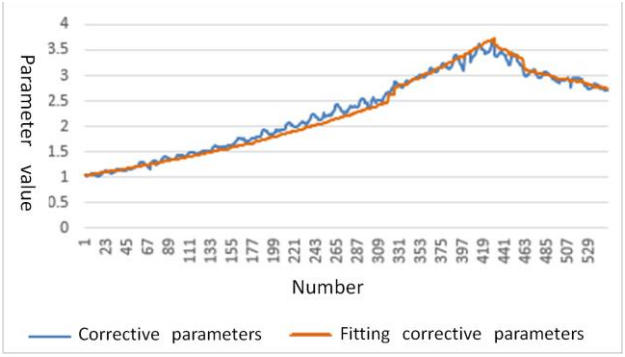


Fig.4. The experimental results of correction algorithm

III. RESEARCH ON POPULATION ESTIMATION METHOD BASED ON PRIOR KNOWLEDGE

A. *Extraction of prior knowledge in video surveillance*

Prior to the concept originated in the field of philosophy, "prior knowledge" in this paper means through analyzing the characteristics of scene in video surveillance to get valid information about the moving targets. Generally speaking, the effective prior knowledge can improve the performance of the algorithm .To extract effective prior knowledge and prior knowledge in the universal applicability of different monitoring environment has always been the focus of scholars attention. At present the main research of scholars in time context and space, using 3D information to obtain some effective prior knowledge. Such as: in some of the monitoring area (the road, the corridor), the crowd of area are fixed; Some monitoring area (the sky, the water), it is impossible to appear the crowd, and so on.

Although different monitoring video may have different prior knowledge, but generally we think in most surveillance video the closer distance between people, restricted by the relative space, movement range will be smaller, walking speed will slow down. In this paper, add this prior knowledge to estimate crowd. Although using some prior knowledge, but the authors believe that more effective prior knowledge is not found, by extracting a priori knowledge, the overall performance and speed of the algorithm can further improve.

B. *Extraction Definition and extraction of exercise intensity feature in prior knowledge*

According to prior knowledge, that is, as the crowd density is more and more big, the distance between people more and more small, and movement speed also will be more and more

slow. combined with this knowledge, this paper proposes a new features based on pixels - move intensity E . The feature mainly describes the movement intensity of moving target.

And it is obtained by using interframe difference method and background subtraction method, specific steps are as follows:

(1) Given a period of video, convert video into image sequence and preprocess image. Then get binary image sequence $B = \{b_1, b_2, \ldots, b_n\}$ based on codebook background modeling. And use the interframe difference method to get another set of binary image sequence $F = \{f_1, f_2, \ldots, f_n\}$.

(2) The i frame background subtraction method to get the picture b_i and inter frame difference method to get the picture f_i , b_i add f_i to get an image sequence. Then analyze the connectivity of the foreground block in these image sequence , record the position of the minimum circumscribed rectangle and size as shown in Fig.5(c), and then subtract the (i–1) frame b_{i-1}, get a set of image S_i, according to the Fig.5(c) record the position of rectangular box in the S_i draw rectangle of the same size as shown in Fig.5(d), each rectangle represents a foreground block.

(c) combine operation

(d) S_i

Fig.5. preparation of Feature extraction

(3) Foreground block as a unit, calculate width d of every foreground block (the width of the minimum circumscribed rectangle).

(4) To assume that in S_i foreground block has n rows of pixels, each row has C_k pixel points, $k=1,2,\ldots,n$, n is integer; to sort C_k, get $C_1 \le C_2 \le \cdots \le C_k$, after repeated observations and experiments have proved that remove 30% of the highest and the lowest 30% , and 40% more representative, the middle can represent the entire foreground block movement situation, therefore, select the middle 40% on average, as shown in equation (3) :

$$R = \frac{\sum_{k=30\%n}^{70\%n} C_k}{40\%n} \tag{3}$$

(5) Finally get move intensity E:

$$E = R / d \tag{4}$$

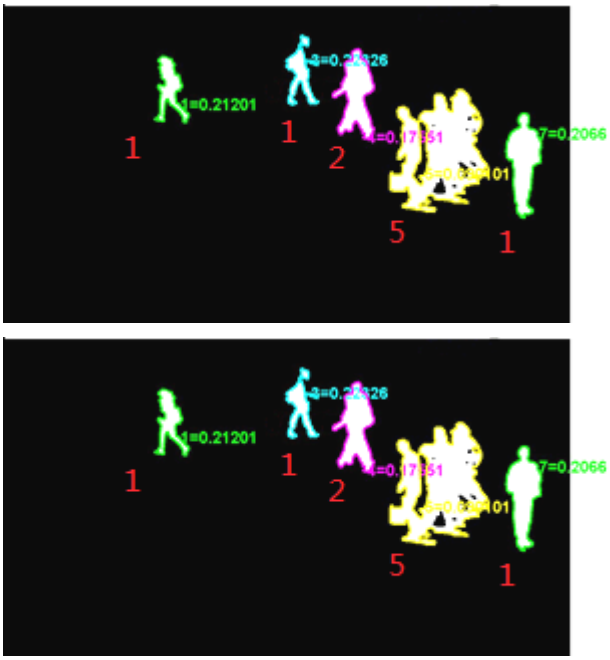


Fig.6. Result of E

TABLE II Data of E

Number	Exercise Intensity	Number	Exercise Intensity
1	0.23051	4	0.12461
1	0.22326	4	0.13224
2	0.17551	5	0.11534
3	0.16244	5	0.090101
3	0.15036	7	0.086346

In Fig.6, red Numbers mean the number of people in foreground block by manual annotation, TABLE II lists some typical move intensity values. Can be seen from Fig.6, and TABLE II when foreground block has one person ,E values between 0.2 to 0.3, when people increased to four, value dropped to 0.12451, when increased to 7 person, value reduced again. So the final experimental results verified expectations, when the more people foreground block has, the smaller move intensity is.

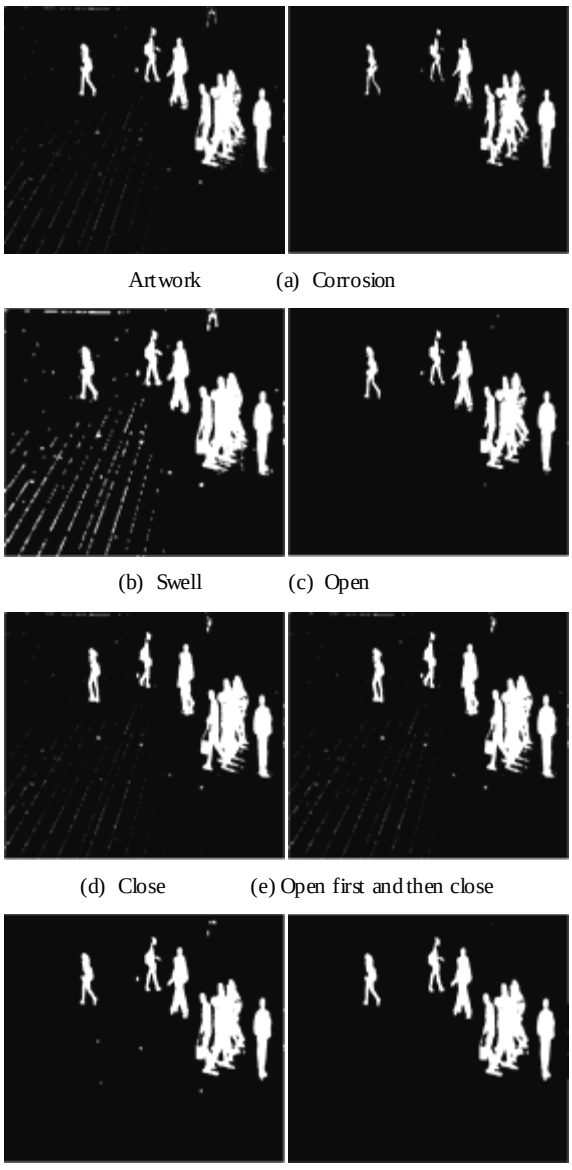
IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experiment is realized with the Microsoft Visual Studio 2010 and opencv2.4.9, the experiment data sets come from the actual shooting with Sony and the scene is a small front square.

Specific experimental procedure is as follows:

(1) Select a video from a period data set, firstly, convert the video into image sequences and then make them to be grayscale, and do de-noising by mean filtration for the

Gaussian noise. Then use a rectangular core of 3 * 3 to do the following image morphological operations, the effect is shown in Fig.7, in the figure (a) (c)and(e) ,there is better performance in terms of removing some isolated noise , But it will lose part of the image information, for example, make person's head and body separated to form two parts and body incomplete; By contrast, from the point of the clear of noise and the integrity of prospect, the best mean is the digital image morphological processing in the way of first close and then open. On the basis of the way of being first close and then open, Fig.7 (g) shows the final results plus the threshold to filter smaller points of the prospects.



Artwork (a) Corrosion

(b) Swell (c) Open

(d) Close (e) Open first and then close

(f) First close and then open (g) Threshold filtering

Fig.7. Digital image morphology processing

After pretreatment, foreground is extracted by using background subtraction based on codebook background modeling method.

(2) Do extraction of the statistical characteristics of the pixel based on foreground block, through analysis of the connectivity; regard each connected region as a foreground block, as shown in the Fig.8.

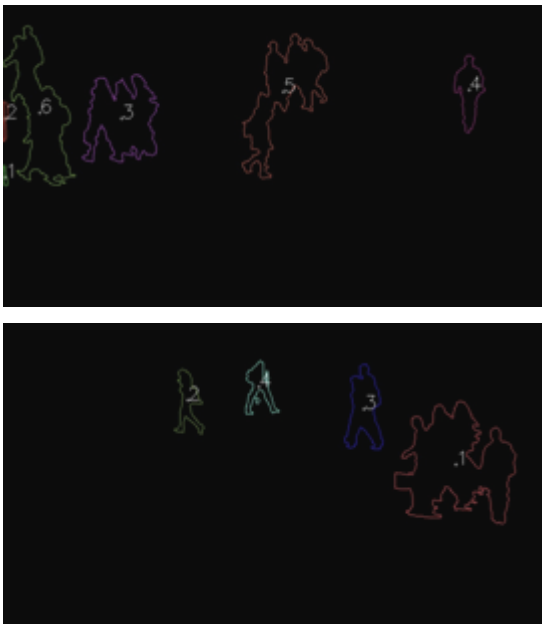


Fig.8. Label for foreground block

(3) In Fig.8, extract the statistical characteristics of pixel for each foreground block, as the same as last chapter, foreground block feature is corrected by the perspective correction algorithm based on statistical regression which is proposed in this paper. Finally , get a set of feature vector F = {the total number of prospect pixels, the total number of edge pixels, the ratio of total number of foreground pixels and the edge pixel, the length of the minimum circumscribed rectangle of foreground block, the width of the minimum circumscribed rectangle of foreground block, the feature of exercise intensity}.

(4) The statistics feature to be divided into two parts, a part to support vector regression machine to train regression, then use the trained classifier to predicted the number of another part of people to get the estimated number of prospects each block, and then all people number in foreground block are added in a frame, that is the estimated number of the frame at the scene.

As following Fig.9 shown, the blue line represents the true number of people in the scene, orange represents the estimated number of persons without added feature of the exercise intensity at the scene, and the gray line represents the number in the scene by using the method proposed in this chapter. By comparing the three fold lines, we can draw a conclusion intuitively, between 47-53 and 77-89, when we did not add the exercise intensity characteristics, since the occlusion of pedestrians, the estimated number of people is not accurate, and there is a big error; the gray line shows the positive effect because adding exercise intensity characteristics reduce the influence of occlusion, improve the accuracy of the estimation method for population density. The results support the view

presented in this paper show that the method has practical significance.

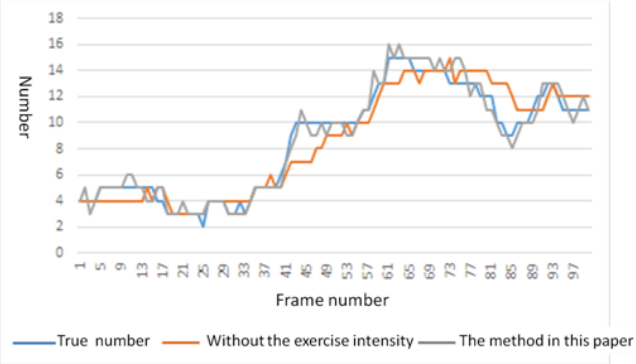


Fig.9. The experimental results of crowd density

V. CONCLUSION

This paper proposes a perspective correction method based on statistical regression and defines the exercise intensity based on prior knowledge, and applies them to count the crowd. Experiments show that the algorithm in sparse scene can achieve good results when the crowd appeared occlusion problems, which indicates that the proposed method of perspective correction for crowd count is effective with exercise intensity combined. In the follow-up study, we will do further research on addressing in the severely blocked phenomenon of a high-density population and broaden the scope of application of the algorithm to enhance the robustness of the algorithm.

REFERENCES

[1] J. Kexue Dai etc, Video Mining A Survey. Journal of Image and Graphics, vol. 4, pp.451-457,2006.

[2] J. Saleh S A M, Suandi S A, Ibrahim H. Recent survey on crowd density estimation and counting for visual surveillance. Engineering Applications of Artificial Intelligence, vol. 41, pp.103-114,2015.

[3] M. Khatoon R, Saqlain S M, Bibi S. A robust and enhanced approach for human detection in crowd. International Multitopic Conference. pp. 215-221,2012.

[4] J. Xing J, Ai H, Liu L, et al. Robust crowd counting using detection flow, vol. 1,pp. 2061-2064, 2011.

[5] J. Hou Y L, Pang G K H. People Counting and Human Detection in a Challenging Situation. IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, vol. 41, pp.24-33,2011.

[6] J. Davies A C, Yin J H, Velastin S A. Crowd monitoring using image processing. Electronics & Communication Engineering Journal, vol. 7, pp. 37-47,1995.

[7] J. WANG Qiang, SUN Hong. Crowd Density Estimation Based on Pixel and Texture. Electronic, vol. 28, pp. 129-132+136,2015.

[8] J. LI Yin, WANG Guijin, LIN Xinggang. Crowd density estimation algorithm combining local and global features. J Tsinghua Univ,vol. 53, pp. 542-545+549,2013.

[9] J. LIN Qin, ZHANG Li. Crowed Abnormal Detection Based on GLCM and Optical Flow. Computer and Modernization. 2014,3:114-118.

[10] J. Albiol A, Mar á J, Silla A, et al. Video analysis using corner motion statistics. Proc.of the IEEE Int.workshop on Performance Evaluation of Tracking & Surveillance –38 Tools Appl, 2009.

[11] M. Su Hang. The large scale crowd analysis based on high definition video. Shanghai Jiao Tong University. 2009.

[12] J. Lin S F, Chen J Y, Chao H X. Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems Man & Cybernetics Part A Systems & Humans, vol. 31, pp. 645-654,2001.

[13] C. Chan A B, Morrow M, Vasconcelos N. Analysis of crowded scenes using holistic properties[C]//Performance Evaluation of Tracking and Surveillance workshop at CVPR., pp. 101-108,2009.