Density Estimation of Unidirectional Crowds*

Liang Zhang, Tao Deng, Yaling Song, Yi Fan
Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, Tianjin, 300300, China
l-zhang@cauc.edu.cn, t_deng@sina.cn

Abstract - An improved pixel-based method is proposed in this paper aiming at density estimation for unidirectional crowds. It is a common situation that the crowd gathered facing towards the same orientation where hazardous events may be prone to occur due to high density, such as crowds gathered in front of a rostrum, crowds waiting at an entrance, flowing crowds in a passageway, etc. In these situations, the crowd density has a linear relationship with both the proportion of facial region to the foreground and the proportion of inner edges to the foreground edges. Based on above observations, we get a model and solve its parameters by multivariate statistical regression. The experiment shows the proposed method improves the accuracy of the commonly used pixel-based method especially for high-density crowds.

Index Terms - estimation of crowd density, unidirectional crowd, statistical regression, pixel-based method

1. Introduction

Crowd density estimations based on video analysis is of great significance to public safety, crowd management, and virtual environment design, etc. Based on computer vision technology, the methods of crowd density estimation can be roughly divided into three categories [1]: i) pixel-based analysis, ii) texture-based analysis, and iii) object-level analysis. Some representative approaches of the three categories are presented below.

Pixel-based methods rely on local features, such as individual pixel analysis obtained through background subtraction models or edge detection, to estimate the density of crowd. Davies et al. [2] was one of the pioneer teems in the use of pixel-based method. They assumed the linear models to map foreground pixels (or edges) to the number of the people in the scene, then used statistical regression method to obtain the model of the crowd density estimation. Ma et al. [3] proposed an approach based on the number of foreground pixels to estimate the crowd density, which includes the geometric correction due to the perspective of the camera. But because of serious occlusions in the high dense crowd, the crowd density no longer keeps the linear relationship with foreground pixels (or edges). Different density groups showing different texture features are the principle of the method based on texture for crowd density estimation. For example, Marana et al. [4] assumed that images of low-density crowd tend to present a coarse texture, while the images of dense crowd tend to present fine textures. Through the use of texture feature descriptor for image analysis, the density of the crowd was estimated. Rahmanlan et al. [5] made a deep comparison of three techniques used in the texture analysis to address the problem of the crowd density estimation. The three techniques were the gray-level dependence matrix, the Minkowski fractal dimension, and Chebyshev orthogonal moments. The analysis indicated that the methods based on the gray-level dependence matrix or the Minkowski fractal dimension presented the better results. Although some efforts using texture for estimation of crowd density were made by many researchers, they all faced the same problem that the low accuracy in the low-density crowd. The object-level analysis method tries to identify an individual in the crowd. It often combines the local and global features to search the individual target and tends to make more accuracy results when compared to pixel-based approach or texture-based analysis. For example, Lin et al. [6] used the Haar wavelet transform to search for object with head-like contour in the image space. Zhao et al. [7] make use of three-dimensional human models to represent people in the scene. However, it is difficult to separate the individual from the crowd in dense crowd.

In recent years, some novel methods [8-10] emerge in the domain of the crowd density estimation. For example, Li et al. [8] calculated the ratio of foreground blobs to the whole image with a mechanism of threshold segmentation, then a regression algorithm based on the local features was used to analyze images below the threshold and a classification algorithm based on global features images was used to analyze images above the threshold, and Jones et al. [9] designed a classifier to detect pedestrians by using spatiotemporal information. Their approach seems to successfully differentiate pedestrians from vehicles, but is not suitable for dense scenarios.

In general, the pixel-based approach shows the lowest computational complexity and it is the most appropriate for the application in real-time crowd density estimation. However, obvious defects of the low accuracy will be encountered when it comes to relatively dense scenarios. In fact, the situation that the crowds facing towards the same orientation appears frequently in some public occasions where high-density crowds and hazardous events may be prone to occur, such as crowds gathered in front of the rostrum, waiting crowds at the counter or entrance, flowing crowds in passenger channel and so on. In these scenes, the proportion of the facial region in the foreground and that of the inner edges in the foreground contours will increase with the increasing of crowd density. Hence, more efficient and accurate methods based on the unidirectional characteristics for crowd density estimation is required. According to the defects that the algorithms for estimating the density of crowds in [2] showing lower detection precision in a high population density , an novel

* This work was supported by the National Natural Science Foundation of China under Grant 61179045.

approach is proposed which can improve the accuracy of results gained by using algorithms in [2] in some degree.

The paper is organized as follows. Section 2 introduces the principle and process of the proposed algorithm. Section 3 expounds the acquisition of foreground objects and the extraction process of face region and edge pixels. Section 4 describes the process of building crowd density estimation model using sample data training. Section 5 compares the accuracy of different algorithm for crowd density estimation in one-way channel, and analyzes the advantages of this algorithm. Section 6 is the conclusion.

2. Algorithm Overview

In [2], a pixel-based crowd density estimation model was proposed, which assumed that there is a linear increasing relation between crowd density and the number of foreground pixels or thinned edge pixels. And relying on the linear regression statistical method, the estimated model function is generated as follow:

$$Z = m \cdot X + b \tag{1}$$

where Z is the number of pixels (e.g. number of non-background pixels or number of thinned edge pixels), X is the number of people in the scene, m and b are the linear regression coefficients. Fig.1 shows the changing relationship between the number of people in a scene and the proportion of foreground pixels to the total image pixels, which was performed by the author in [8] using the UCSD database [10]. It is obvious that the method based on (1) has good detection performance under the condition of sparse crowd. However the model is no longer applicable when the crowd density exceeds a certain threshold.

Based on the above algorithm and taking into account the characteristics of the population occlusion phenomena, this paper introduces the innovative concept called "the inner edge". The edges of foreground are divided into two categories, one is called “the inner edge”, and the other is called “the non-inner edge”. As shown in Fig.2, there are two people in (a) where parts of the red people’s body are blocked by the blue one. (b) is the edge iamge gained from (a). The black edges are defined as “the non-inner edge”, and the red edges which represent the boundaries of bodies are called “the inner edge”. Obviously, the proportion of the inner edges to edges of foreground will increase with the crowd density increasing. The algorithm proposed in this paper is mainly based on two ideas:

(1) Because the probability of occlusion in human bodies is higher than that of the heads, the proportion of skin colour region dominated by human faces to the whole foreground will increase linearly with the increase of the crowd density.

(2) As the crowd density increases, the proportion of the inner edge pixels to foreground edge pixels shows a linear increase trend.

In view of the above two points, we re-define the model for estimating the density of crowd as:

$$P = m \cdot Y + b \tag{2}$$

$$m = k \cdot F + c \cdot E + d \tag{3}$$

where P is the number of people, Y is the number of non-background pixels, F is the proportion of facial region to foreground, E is the ratio of the inner edges to the foreground edges, m, b, c, and d are coefficients obtained from the experimental data by linear regression. It can be seen that this method improves the accuracy by modifying the results gained from using algorithms in [2] in some degree. Fig.3 is a flowchart of the algorithm in this paper. Moreover, it can address the problem called near-far effect, since people near to the camera will occupy more area than people of the same size far from the camera.

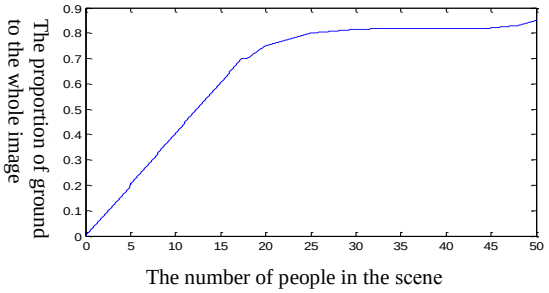
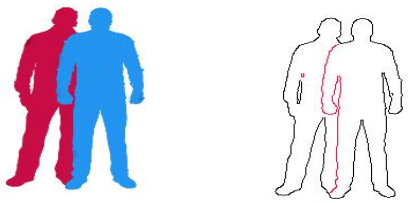


Fig.1 Relationship between the density of crowd and the proportion of foreground pixels to total pixels



(a) (b)

Fig.2 Classification of edges. The red edges in (b) are the “inner edges”.

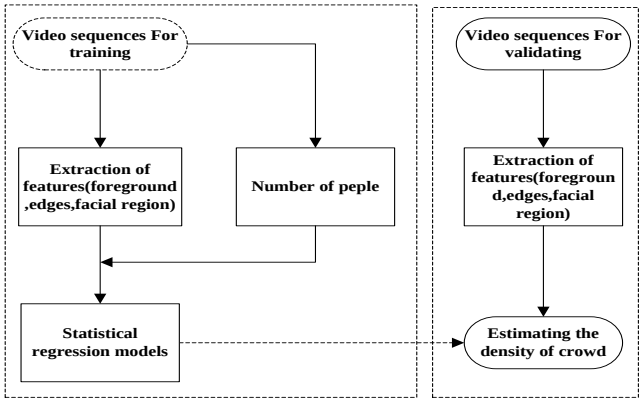


Fig.3 Flow chart of the algorithm for estimating density of crowd

3. Extraction of Features

A. Foreground Pixels Extraction

The extraction of foreground means segmenting the change areas of image sequences from the background. At present, the background model method was commonly used to detect moving objects. For example, Stauffer [11] employed Gaussian mixture model for adaptive background modeling, which modeled each pixel as a mixture of Gaussians to determine whether or not a pixel is part of the background. The method can separate the background and foreground effectively and can be used for real-time tracking. In theory, any complex scene can be generated by a certain number of Gaussian distributions. Unfortunately, with the increase of the number (K) of Gaussian distributions, computational complexity also increases. Therefore, it becomes impossible to simulate all scenes. Typically K is often set to 3 to 5 and the adaptive Gaussian mixture model approach is used to extract foreground image as well.

Although the mixture Gaussian background model can overcome interferences of illumination and shadows to some extent, the situation that misjudging the foreground as the background is unavoidable because the target movement speed is too slow or the color of foreground is similar to the background. As a result, the obtained foreground mask always covers some kinds of interferences, such as cracks, cavities and star points, which have bad effects on foreground edge extraction. It is requisite to preprocess the foreground mask before detecting edges.

In order to filter out the interferences of star points, the median filter is used to deal with the foreground mask firstly. Then the method of dilation operation based on mathematical morphology method is used to remove the crack phenomenon, and then performs the corrosion operation to keep the original silhouette features of foreground. Next, according to the characteristics of connected domain, we perform the polygon fitting operation on outer silhouette, and then label and fill the internal cavity regions of foreground. Figure 4 shows the comparison of foreground mask before and after treatment, the left is the foreground mask image obtained by using Gaussian mixture model directly, and the right is obtained after processing. Obviously, the interferences such as cracks and cavities in original foreground masks have been removed. Then foreground image can be obtained by mapping the foreground masks to original images.

B. Face Region Extraction

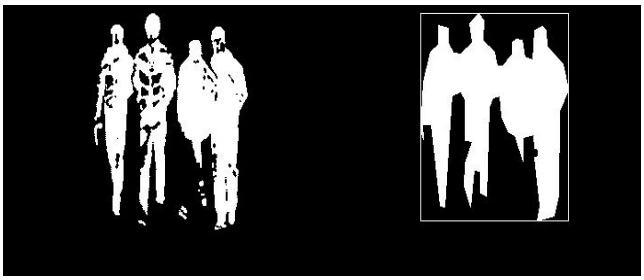
The method does not need to get accurate information about face region fortunately, but only to get the area of the facial region. So the facial region can be extracted by the feature of skin colour.

The distribution of skin color in the color space is relatively concentrated, but it tends to be affected by light intensity. Usually, the method of converting RGB space to a certain color space, such as YCrCb or HSV, is used to reduce the influence. The method described in [12] is used in our scheme to extract the face regions. Fig.5 shows the result of the facial region extraction. It only takes the CrCb plane of the YCrCb space into consideration, and sets the color space in rectangular area within Cr ∈ [133,173], Cb ∈ [77,127] as skin color area.

C. Edges Extraction

The proportion of the inner edges to the edges of the foreground can effectively reflect the degree of crowd density. The extracted edges are the physical boundaries between the foreground and background as well as human targets.

The main difference between inner and non-inner edges lies in their different positions among the edges of foreground. Both sides of non-inner edges must have one side belonging to the background region, but both sides of the inner edges are foreground region. According to this feature, a decision that whether all the eight pixels surrounding the pixel of edges are in the foreground or not is made to distinguish between inner and non-inner edges as Fig.6.



(a) (b)

Fig.4 comparison of foreground mask before and after treatment

Fig.5 Extraction of facial region. the “left” is the original image ,the “right” is the facial region extracted from foreground.

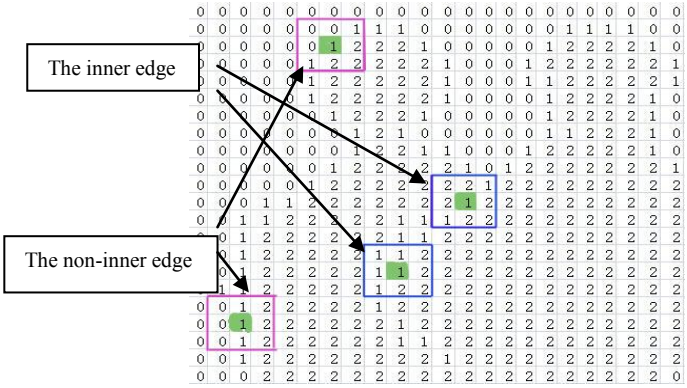


Fig.6 Extraction of edges. ”0” represents the background; ”1” represents the edges;”2”represents the foreground

4. Training the Model of Crowd Density Estimation

The density estimation model for unidirectional crowd constructed in this paper is based on statistical regression method. In order to construct an estimation model with better robustness, the chosen samples for training should contain various crowd densities as far as possible. 1000 frames from nearly 20,000 monitor images were selected as training samples. To test the quality of the fitting degree of multivariate statistical regression models, many factors mainly including the goodness of fit test, significance tests of the regression equation and parameter estimation should be covered. Among them, the goodness of fit test is determined by the coefficient of determination R^2, and the closer R^2 is to 1, the higher fitting degree the multiple linear regression equation will show. Significance test of Regression equation is examined relying on statistic F and threshold f .Generally the bigger test value of F revels the better performance , and F> f. To prevent existence of excess independent variables in the equation, p associated with significance probability should satisfy p < alpha where alpha represents the significance level and its default is 0.05.

When the capacity of training sample is set to 1000 and alpha is set to 0.05, the values of the above parameters are got as that $R^2 = 0.9724$, F = 394.3535, f = 0, and p=0.0396. R^2 is very close to 1, which indicates that the dependent variable explained very well by the explanatory variables. In other words, the multivariate linear equation performs well; F is large enough and F> f, which illustrates that significant regression equation was established; P <0.05, and it illustrates that each independent variable selected of the regression equation is significant. In summary, the estimation model is effective.

5. Experimental Results

The proposed algorithm is used to cope with the problem of crowd density estimation for unidirectional crowd. But no public database is available to train the estimation model. The video sequences used in this paper were shot on a flyover of school or in front of teaching building by ourselves.

The crowd density can be divided into four levels by the number of people in the monitoring scene: low (1 to 5), medium (6 to 10), mid- high (11 to 15), high (16 or more). Further, there is little difference between two adjacent frames for the reasons that moving speed of crowd is very slow and a video camera can shoot 25 frames per second, so it is unnecessary to estimate each frame image during the real-time video sequences. Therefore, one frame every 15 frames is extracted to estimating the density of crowd. Table I lists the estimation accuracy of three methods for unidirectional crowd density.

6. Conclusions

A method of unidirectional crowd density estimation is presented, in which the feature of people facing the same direction in unidirectional crowd are considered. It takes advantage of the characteristics that the density of crowd has the linear relationship with both the proportion of facial region to the foreground and the proportion of inner edges to foreground edges, which improves the accuracy of the results only based on foreground pixels. Experiments demonstrate that the proposed method not only shows better estimation accuracy for low-density crowds, but also effectively improves the high-density crowd estimation accuracy.

TABLE I Comparison of Accuracy in Estimation

Method	Accuracy Rate			
	low	medium	mid-high	high
Proposed Method	93.12%	90.32%	86.46%	84.36%
Foreground pixels method [2]	91.95%	87.45%	70.76%	53.63%
Gray-level dependence matrix method [5]	76.34%	79.53%	83.91%	87.86%

References

[1] Jacques Junior J C S, Musse S R, Jung C R. Crowd analysis using computer vision techniques. Signal Processing Magazine, IEEE, 2010, 27(5): 66-77.

[2] Davies A C, Yin J H, Velastin S A. Crowd monitoring using image processing. Electronics & Communication Engineering Journal, 1995, 7(1): 37-47.

[3] Ma R, Li L, Huang W, et al. On pixel count based crowd density estimation for visual surveillance//Cybernetics and Intelligent Systems, 2004 IEEE Conference on. IEEE, 2004, 1: 170-173.

[4] Marana A N, Costa L F, Lotufo R A, et al. On the efficacy of texture analysis for crowd monitoring//Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI'98. International Symposium on. IEEE, 1998: 354-361.

[5] Rahmalan H, Nixon M S, Carter J N. On crowd density estimation for surveillance//Crime and Security, 2006. The Institution of Engineering and Technology Conference on. IET, 2006: 540-545.

[6] Lin S F, Chen J Y, Chao H X. Estimation of number of people in crowded scenes using perspective transformation. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2001, 31(6): 645-654.

[7] Zhao T, Nevatia R. Bayesian human segmentation in crowded situations//Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. IEEE, 2003, 2: II-459-66 vol. 2.

[8] Li Y, Wang G J, Lin X G. Crowd density estimation algorithm combining local and global features. Journal of Tsinghua University. 2013: 542-545, 549.

[9] Jones M J, Snow D. Pedestrian detection using boosted features over many frames//Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008: 1-4.

[10] Chan B. Liang S. UCSD dataset [Z/OL]. [2013-09-13]. http://www.svcl.ucsd.edu/projects/peoplecnt.

[11] Stauffer C, Grimson W E L. Adaptive background mixture models for real-time tracking//Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. IEEE, 1999, 2.

[12] Chai D, Ngan K N. Face segmentation using skin-color map in videophone applications. Circuits and Systems for Video Technology, IEEE Transactions on, 1999, 9(4): 551-564.