Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 8703576, 11 pages
https://doi.org/10.1155/2018/8703576

Research Article

Predicting Pedestrian Counts for Crossing Scenario Based on Fused Infrared-Visual Videos

Shize Huang, Wei Chen, Rongjie Yu, Xiaolu Yang, and Decun Dong

Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, 201804, China

Correspondence should be addressed to Rongjie Yu; yurongjie@tongji.edu.cn

Received 10 August 2018; Revised 26 October 2018; Accepted 27 November 2018; Published 4 December 2018

Academic Editor: Krzysztof Okarma

Estimating the number of pedestrians based upon surveillance videos and images has been a critical task while implementing intelligent signal controls at intersections. However, this has been a difficult task considering the pedestrian waiting area is an outdoor scenario with complex and time-varying surrounding environment. In this study, a method for estimating pedestrian counts based on multisource video data has been proposed. First, the partial least squares regression (PLSR) model is developed to estimate the number of pedestrians from single-source video (either visible light video or infrared video). Meanwhile, the temporal feature of the scenario (daytime or nighttime) is identified based on visible light video. According to the recognized time periods, pedestrian count detection results from the visible light and infrared video data can be obtained with preset corresponding confidence levels. The empirical experiments showed that this fusion method based on environment perception holds the benefits of 24-hour monitoring for outdoor scenarios at the pedestrian waiting area and substantially improved accuracy of pedestrian counting.

1. Introduction

Estimating the number of pedestrians is critical within the intelligent transportation system. The pedestrian counts have been a vital input for intersection signal control [1], the guidance of passenger flow, and early warning of large-scale crowd gathering [2, 3]. However, the approach of estimating pedestrian counts under outdoor scenarios, such as the pedestrian waiting area, is still an unsolved challenge.

Generally, there are two main approaches to estimate the number of pedestrians. One kind is based on reliable tracking of individual pedestrians, which achieves the purpose of counting pedestrians through identifying each individual pedestrian based on image data [4–7]. However, this method is suitable for the case where the pedestrian density is low. If the pedestrian density is high and there is severe pedestrian overlapping, the performance of the method will be deteriorated. The other approach extracts feature from image data and applies regression analysis techniques to estimate the pedestrian counts rather than trying to identify each pedestrian in the image. This method is concluded to be more flexible since there is no need to track each pedestrian in the image.

The surveillance video data have been frequently adopted to estimate the number of pedestrians, which can be further divided into visible light video and infrared video. Infrared video is mostly used to determine whether there are people at scene or whether the target is a human being [8–11]. But, they were barely used for estimating the pedestrian counts. On the contrary, tremendous efforts have been investigated on the estimations of pedestrian counts using visible light video. For instance, Davies et al. [12] used geometric features such as areas and perimeters to estimate the number of pedestrians in the image. He [13] proposed a two-region learning algorithm, applying improved aggregate channel feature detection and Gaussian process regression to estimate the number of pedestrians. Chan [14] segmented the image, extracted the features of each segmentation region, and then used Gaussian process regression to learn the correspondence between the features and the number of pedestrians in each segment. Zhang [15] applied dimensionality reduction techniques to process high latitude features of images and performed regression analysis. Li [16] proposed a feature description operator combining wavelet transform and gray level cooccurrence matrix and used SVM to obtain the pedestrian density model.

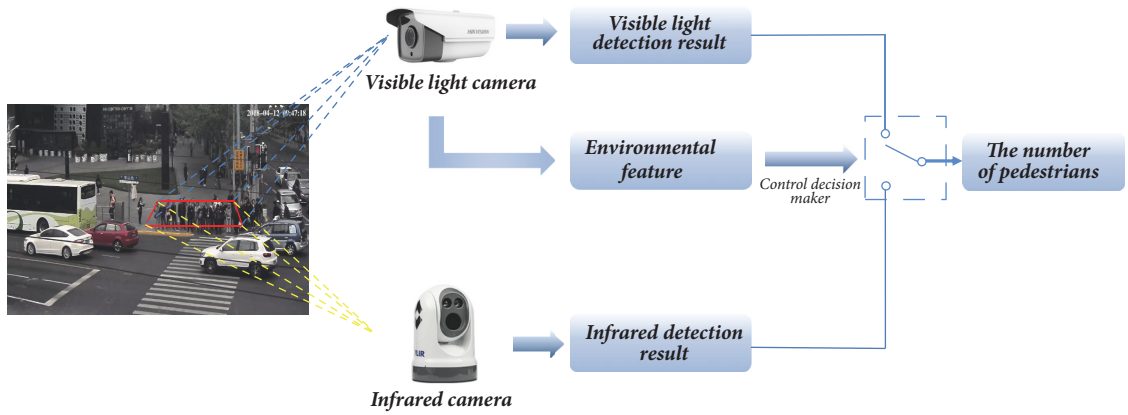


Figure 1: Pedestrian-counting framework.

Yan [17] used the simile classifier to optimize the subimage and then used the regression analysis model to establish the relationship between subimage blocks and the number of pedestrians. However, the abovementioned studies are based on visible light video which is sensitive to lighting conditions and cannot be implemented for monitoring the pedestrian waiting area for the whole day.

In this study, we propose a pedestrian number estimation method which is dependent on fusion of visible light video and infrared video based on environment perception, in order to realize 24-hour pedestrian counts detection for the pedestrian waiting area. First, partial least squares regression (PLSR) was employed to obtain the number of pedestrians from the image based upon visible light video and infrared video, respectively. Then, based on the environmental feature obtained from the visible light video, an information fusion model is established to obtain the number of pedestrians in the image. The specific schematic diagram is shown in Figure 1.

The remaining of this paper is organized as follows: in Section 2, we describe the image processing and how to extract features from images. Then, Section 3 describes the establishment of the pedestrian count estimation model and how to fuse the result of visible light detection with the result of infrared detection. And the report and analyses of the experimental results are given in Section 4 while Section 5 summarizes the work and discusses future directions.

2. Image Processing

In this section, visible light image processing and infrared image processing procedures are introduced correspondingly.

2.1. Visible Light Image Processing. The most important task in image processing is to extract the foreground of the motion from the image. For visible light images, background difference method was adopted to obtain the motion foreground in the image.

Since the background image would gradually change along the time in the actual scene, the background image needs to be updated in real time. Kalman filter was used to update the background here. To be specific, the background image at the time t is determined by the background image at time $t-1$ and the real-time image at time t, which includes both prediction and update. The forecast formula is as follows:

$$\begin{aligned} \widehat{B}(x,y,t) &= B(x,y,t-1) \\ \widehat{P}(x,y,t) &= P(x,y,t-1) + Q \end{aligned} \tag{1}$$

where $B(x,y,t-1)$ is the background optimal value at time $t-1$, $\widehat{B}(x,y,t)$ is the background prediction value at time t, $P(x,y,t-1)$ is the covariance at time $t-1$, $\widehat{P}(x,y,t)$ is the prediction of covariance at time t, and Q is the systematic process error.

The background update formula for time t is as follows:

$$\begin{aligned} k(x,y,t) &= \frac{\widehat{P}(x,y,t)}{\left(\widehat{P}(x,y,t)+R\right)} \\ B(x,y,t) &= \widehat{B}(x,y,t) \\ &\quad + k(x,y,t)\left[I_{vi}(x,y,t) - \widehat{B}(x,y,t)\right] \\ P(x,y,t) &= \left(1-k(x,y,t)\right)\widehat{P}(x,y,t) \end{aligned} \tag{2}$$

where R is the system measurement error, $k(x,y,t)$ is the system gain, $I_{vi}(x,y,t)$ is the gray image acquired by the visible light camera at time t, and $B(x,y,t)$ is the time t background. The optimal value $P(x,y,t)$ is the covariance at time t.

The visible light image $I_{vi}(x,y,t)$ is differentiated from the corresponding background image $B(x,y,t)$. The background difference result $F_{vi}(x,y,t)$ is

$$F_{vi}(x,y,t) = \left|I_{vi}(x,y,t) - B(x,y,t)\right| \tag{3}$$

Then, a binary region-of-interest (ROI) mask proposed by Chan [18] was applied to $F_{vi}(x,y,t)$, which not only reduces the amount of subsequent calculations, but also prevents some interference in noninterest areas. After applying the

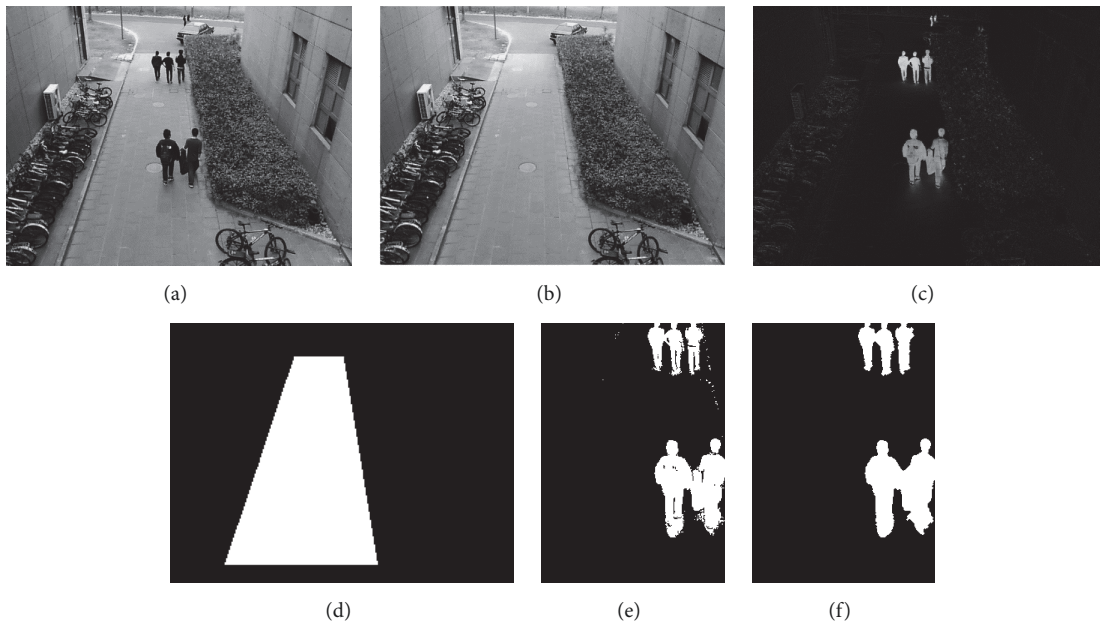


Figure 2: Visible light image processing (a) Original image. (b) Background image. (c) Background difference result. (d) ROI mask. (e) Foreground image. (f) Final result.

ROI mask, the binary foreground of visible light images $F'_{vi}(x, y, t)$ is calculated by

$$F'_{vi}(x, y, t) = \begin{cases} 1, & F_{vi}(x, y, t) > Th_{vi} \\ 0, & F_{vi}(x, y, t) \le Th_{vi} \end{cases} \quad (4)$$

where Th_{vi} is the threshold used for binary processing. In our experiments, we set $Th_{vi} = 45$.

For the image $F'_{vi}(x, y, t)$, the closed operation (dilation followed by erosion operation) is to fill the small holes in the connected domain, connect adjacent objects, and smooth the boundary [19]. Then, it analyzes the connected domain and eliminates the connected domain with smaller area to remove noise [20]. The final result of visible light image processing is the image $M_{vi}(x, y, t)$. Then the set of blobs in $M_{vi}(x, y, t)$ is

$$B_{vi} = \{bvi_1, \cdots, bvi_k, \cdots, bvi_n\} \quad (5)$$

where bvi_k is the k-th blob in the image $M_{vi}(x, y, t)$ and n is the total number of blobs in the image $M_{vi}(x, y, t)$.

For example, Figures 2(a) and 2(b) are the original and background images, respectively. Figure 2(c) shows the background difference result $F_{vi}(x, y, t)$. Figure 2(d) is the ROI mask. Figure 2(f) is the final result $M_{vi}(x, y, t)$.

2.2. Infrared Image Processing. The infrared video data are imaged by thermal radiation, which is not sensitive to ambient light. Since the pedestrian generally appears as a highlighted area in the infrared image, we extract the foreground of the image by the gray value of the image. First, the projection images of the infrared images on the R, G, and B color channels are analyzed to find the projection image which has the greatest difference between pedestrians and the surrounding environment. Figure 3 illustrates that, in the projection image on the G color channel, the characteristics of the pedestrian are the most prominent and easier to distinguish. This projection image is defined as the grayscale image $I_{in}(x, y, t)$.

With the application of the ROI mask, the binary foreground of infrared images $F'_{in}(x, y, t)$ is calculated by

$$F'_{vi}(x, y, t) = \begin{cases} 1, & I_{in}(x, y, t) > Th_{in} \\ 0, & I_{in}(x, y, t) \le Th_{in} \end{cases} \quad (6)$$

where Th_{in} is the threshold used for binary processing. In our experiments, we set $Th_{in} = 120$.

For the image $F'_{vi}(x, y, t)$, the closed operation and connected domain analysis are also performed to remove the noise and ensure the integrity of the pedestrian. The final result of the infrared image is $M_{in}(x, y, t)$. Then the set of blobs B_{in} in $M_{in}(x, y, t)$ is

$$B_{in} = \{bin_1, \cdots, bin_k, \cdots, bin_n\} \quad (7)$$

where bin_k is the k-th blob in the image $M_{in}(x, y, t)$ and n is the total number of blobs in the image $M_{in}(x, y, t)$.

For example, Figure 4(a) is the image $I_{in}(x, y, t)$. Figure 4(b) is the ROI mask and Figure 4(d) is the final result $M_{in}(x, y, t)$.

2.3. Feature Extraction. Here the visible light image feature extraction procedure was taken as an example, while the feature extraction of infrared images is similar. The contained features of blobs and the inferred number of pedestrians were further extracted. Take the blob bvi_k as an example to calculate its geometric features and positional features using the following steps:

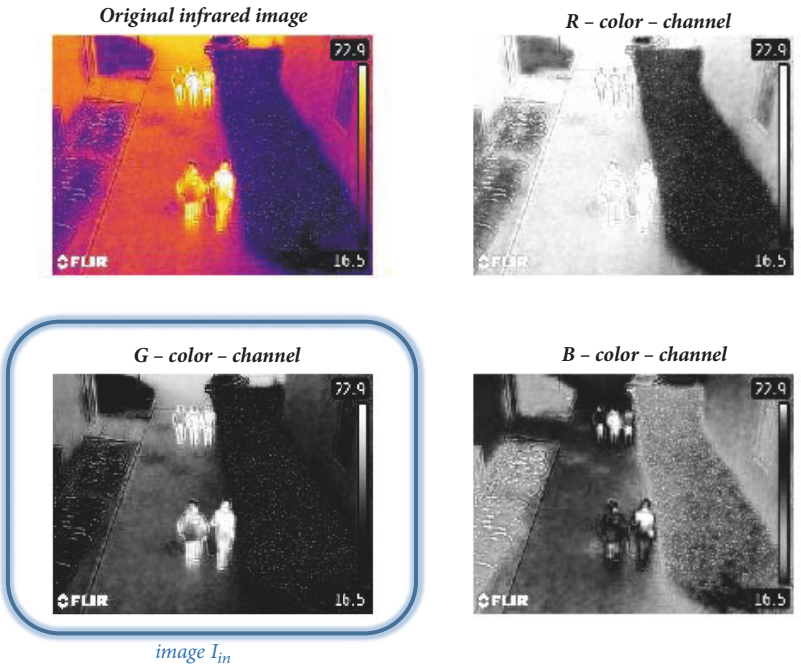


Figure 3: The RGB analysis of infrared images.

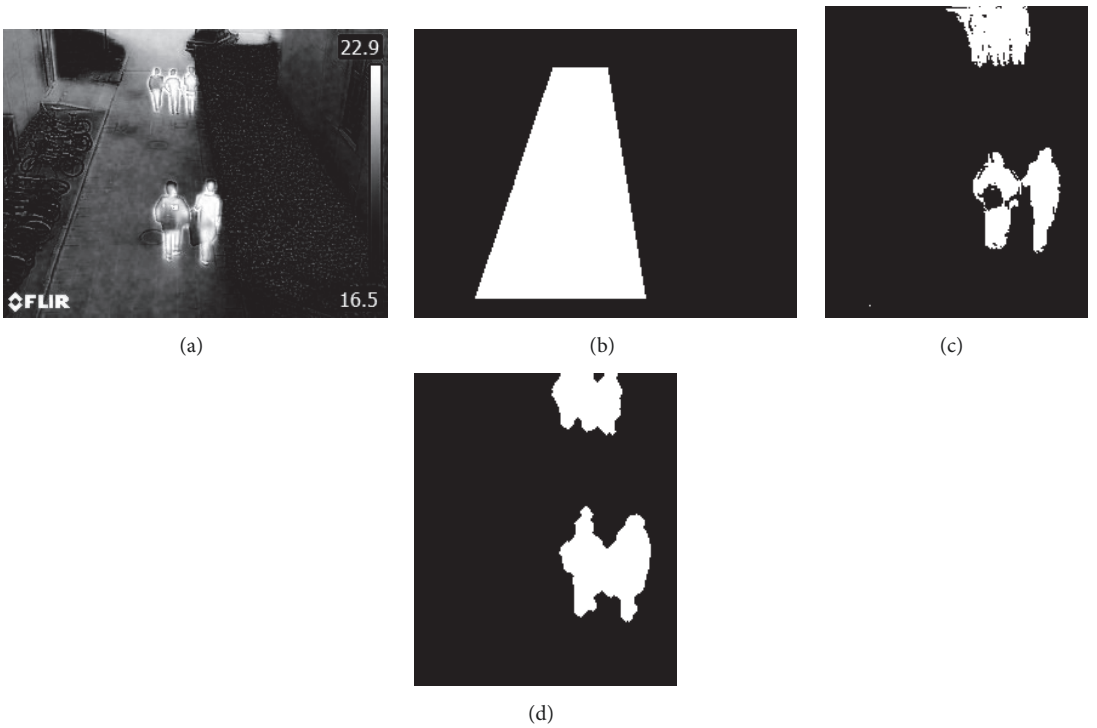


Figure 4: Infrared image processing (a) Original image. (b) ROI mask. (c) Foreground image. (d) Final result.

(1) Area A_k, which is the weighted sum of all pixels in the blob,

$$A_k = \sum_{x=1}^{X} \sum_{y=1}^{Y} bvi_k(x, y) \tag{8}$$

(2) Number of edge points EP_k, which is the weighted sum of pixels on the boundaries of the blob,

$$EP_k = \sum_{x=1}^{X} \sum_{y=1}^{Y} \varepsilon \left[bvi_k(x, y) \right] \tag{9}$$

where $\varepsilon[bvi_k(x, y)]$ denotes the edge image that is generated by the Sobel edge detector on the image bvi_k.

(3) Length of the spot L_k, which is the maximum number of pixels in the horizontal direction of the blob,

$$L_k = \max\left(\sum_{y=1}^{Y} bvi_k\left(1 : end, y\right)\right) \tag{10}$$

(4) Height of the spot H_k, which is the maximum number of pixels in the vertical direction of the blob,

$$H_k = \max\left(\sum_{x=1}^{X} bvi_k\left(x, 1 : end\right)\right) \tag{11}$$

(5) Horizontal position PX_k, which is the horizontal position of the center pixel of the blob in image M_{vi} (for infrared images it is M_{in}),

$$PX_k = \frac{\max\left(\sigma\left[bvi_k\right]\right) + \min\left(\sigma\left[bvi_k\right]\right)}{2}) \tag{12}$$

where $\sigma[bvi_k]$ denotes a horizontal position set of the pixels of the spot bvi_k in the image M_{vi}.

(6) Vertical position PY_k, which is the vertical position of the center pixel of the blob in image M_{vi} (for infrared images it is M_{in}),

$$PY_k = \frac{\max\left(\beta\left[bvi_k\right]\right) + \min\left(\beta\left[bvi_k\right]\right)}{2}) \tag{13}$$

where $\beta[bvi_k]$ denotes a vertical position set of the pixels of the spot bvi_k in the image M_{vi}.

Features A, EP, L, and H have strong correlations with pedestrian crowd density. In general, at the same position of the image, the larger the values of A, EP, L, and H, the more the number of pedestrians included in the blob. And the further away a pedestrian is from the camera lens, the smaller he is in the image. Therefore, we use position features PX and PY to record the positional relationship between pedestrians and the camera lens to ensure the accuracy of pedestrian counting.

Since the final decision result is based on visible light detection results and infrared detection results, an indicator was further introduced to selectively believe based on the distinct detection methods in different situations. In this study, the ambient brightness BR from the visible light image is the indicator.

$$BR_t = \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y} I_{vi}\left(x, y, t\right)}{xy} \tag{14}$$

where BR_t denotes the ambient brightness at time t.

3. Model Establishment

This section focuses on how to infer the number of pedestrians from the extracted features. There are mainly two tasks being carried out: (1) a pedestrian count estimation model was developed based on the features of single-source video to establish; (2) then information fusion model was established based on the detection results of multisource video.

3.1. Pedestrian Count Estimation Model. In order to estimate the number of pedestrians in the blob and prevent the problem that overaggregated data might fail to reveal the true correlation between variables, we apply partial least squares regression (PLSR) [21, 22]. PLSR is a method for multivariate statistical analysis. It draws on the idea of extracting information from explanatory variables in principal component regression, and can effectively solve the multiple correlation problem between variables.

The independent variable $\{x_1, \cdots, x_p\}$ contains p elements and the dependent variable $\{y_1, \cdots, y_q\}$ contains q elements. In order to study the statistical relationships between the dependent variable and the independent variables, assume there are n sample observations, which constitute the independent variable set $X = [x_1, \cdots, x_p]_{n\times p}$ and the dependent variable set $Y = [y_1, \cdots, y_q]_{n\times q}$. The normalization results of X and Y are $E_0 = (E_{01}, \cdots, E_{0p})_{n\times p}$ and $F_0 = (F_{01}, \cdots, F_{0q})_{n\times q}$.

First, the main components are extracted in E_0 and F_0. t_1 and u_1 are the first component of X and Y. Then t_1 and u_1 need to meet the following conditions:

$$\begin{aligned} &t_1 = E_0 w_1 \quad \|w_1\| \\ &u_1 = F_0 c_1 \quad \|c_1\| = 1 \\ &\max_{w_1, c_1} \quad \operatorname{cov}\left(t_1, u_1\right) = \operatorname{cov}\left(E_0 w_1, F_0 c_1\right) \\ &\text{s.t} \quad \begin{cases} w_1' w_1 = 1 \\ c_1' c_1 = 1 \end{cases}, = 1 \end{aligned} \tag{15}$$

where $\operatorname{cov}(t_1, u_1)$ denotes the covariance between t_1 and u_1.

After the first components t_1 and u_1 are extracted, the regression of X versus t_1 and the regression of Y versus u_1 are performed, respectively. If the regression equation has reached a satisfactory accuracy, the algorithm terminates; otherwise, the second round of component extraction will be performed using the residual information of X and Y. So reciprocate until a satisfactory accuracy is achieved. If we finally extract a total of m components $t_1, \cdots, t_m$, PLSR will be implemented by implementing y_k regression of $t_1, \cdots, t_m$ and then expressed as y_k regression equations for the original variables $x_1, \cdots, x_p$ $(k = 1, 2, \cdots, q)$.

Take the visible light image $M_{vi}(x, y, t)$ as an example. Based on PLSR, we establish the pedestrian estimation model where the feature set $\{A_k, EP_k, L_k, H_k, PX_k, PY_k\}$ of the blob bvi_k is an input and the number of pedestrians PN_k included in the spot bvi_k is an output.

$$PN_k = f\left(A_k, EP_k, L_k, H_k, PX_k, PY_k\right) \tag{16}$$

Based on the above model, the number of pedestrians included in each blob in the image $M_{vi}(x, y, t)$ is calculated. The total number of pedestrians PN_{vi} in the image $M_{vi}(x, y, t)$ is

$$PN_{vi} = \sum_{k=1}^{n} \lfloor PN_k + 0.5 \rfloor \tag{17}$$

where $\lfloor PN_k + 0.5 \rfloor$ denotes rounding of PN_k.

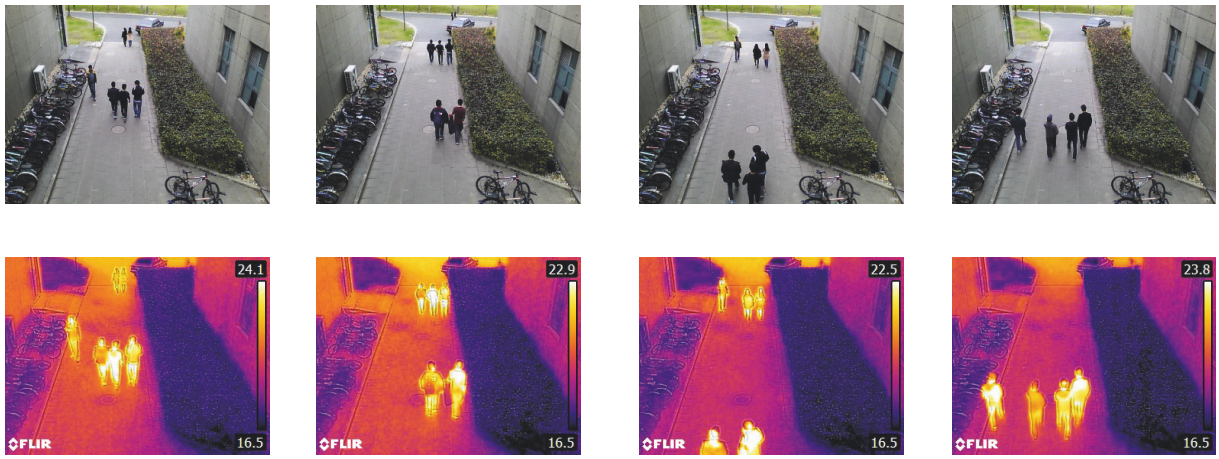


Figure 5: Examples of experimental data.

PN_{vi} is the detection result of visible light. Using the same method, we can get the infrared detection result PN_{in}.

3.2. Information Fusion Model. The environment of outdoor scenarios like the pedestrian waiting area varies substantially along the daytime due to the lighting conditions, temperature, etc. In order to ensure the accuracy of the pedestrian count estimations, a method of combining the visible light detection result with the infrared detection result was proposed with its advantages of applying feasibility in different scenarios. First, the current scenario (day or night) is identified based on the ambient brightness BR_t obtained above. Then, according to the recognition result of the scenario, a corresponding confidence level is set for the detection result of the visible light and the detection result of the infrared. In the case of good daylight and good light, we believe the detection result of visible light; otherwise we believe the detection result of infrared. Therefore, the information fusion result PN_t at time t is

$$PN_t = \alpha_{vi} PN_{vi} + \alpha_{in} PN_{in}$$

$$\text{s.t} \quad \alpha_{vi} = \begin{cases} 0, & BR_t < Th_{br} \\ 1, & else, \end{cases} \tag{18}$$

$$\alpha_{in} = \begin{cases} 1, & BR_t < Th_{br} \\ 0, & else \end{cases}$$

where α_{vi} is the confidence level of visible light detection result and α_{in} is the confidence level of infrared detection result. Th_{br} is the environment segmentation threshold and we set it $Th_{br} = 75$.

4. Empirical Analysis

The empirical analysis was conducted at the campus of Tongji University. A total of 106 groups of daytime images and 18 groups of night images (as shown in Figure 5) were collected. The visible image is 640∗480 pixels, and the infrared image is 320∗240 pixels. This section uses 8-fold cross validation to divide the image set into a training set and a test set and then to check the accuracy of the proposed method.

4.1. Daytime Scenario. For the subset of daytime images, the visible light detection results are shown in Table 1 and the infrared detection results are shown in Table 2. Figure 6 is a schematic diagram of information fusion in a daytime scenario. It can be seen from Figure 6 that the visible light image is clearer and the noise in the processing result of the visible light image is smaller. This is because the resolution of the visible light image is higher than that of the infrared image. Therefore, the result of information fusion gives credibility to the detection result of visible light, which is consistent with the actual situation.

4.2. Nighttime Scenario. For a group of night images, since there are no street lights near the experimental site, this would cause the visible light detection complete failure. Therefore, the visible light detection result is 0. The infrared detection results are shown in Table 3. Figure 7 is a schematic diagram of information fusion in a night scenario. Since the ambient brightness at this time is very low, the result of the information fusion is selected to believe the infrared detection result, which is consistent with the actual situation.

4.3. Influence of Thresholds Th_{vi} and Th_{in}. The thresholds Th_{vi} and Th_{in} are key parameters in this study, which were used to distinguish pedestrians from the background in the image. If Th_{vi} and Th_{in} are too large, a large number of pixels representing the pedestrians in the image will be misjudged as the background, which will result in incomplete motion foreground. As a consequence, the final pedestrian count result will be small. If Th_{vi} and Th_{in} are too small, a large number of pixels representing the background in the image will be misjudged as pedestrians, so that the foreground of the motion will contain a lot of noise. And the final pedestrian count result will be large.

Here different thresholds Th_{vi} were performed as an example, and the threshold Th_{in} is similar. For the same

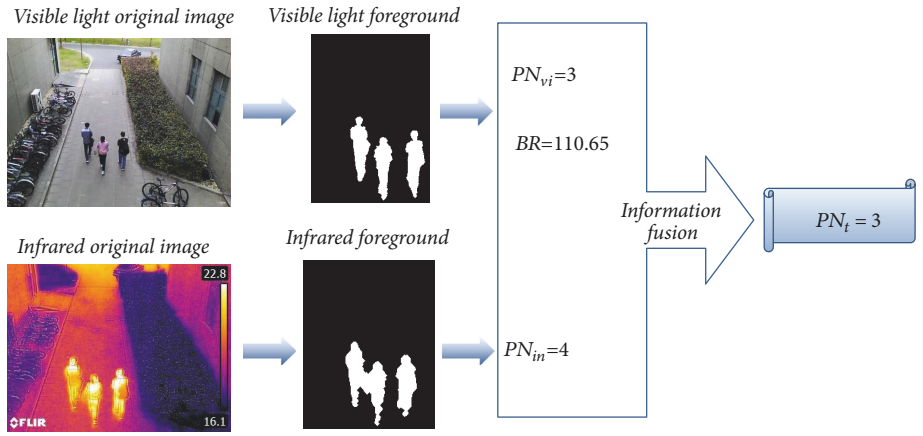


Figure 6: Schematic diagram of information fusion in a daytime scenario.

Table 1: Visible light detection results in a daytime scenario.

Blob number k	Input						Output	
	A_k	EP_k	L_k	H_k	PX_k	PY_k	PN_k	Rounded results
1	2478	271	38	120	213.5	92.5	0.97255	1
2	2295	261	35	117	250	135	0.76629	1
3	2581	279	37	126	241.5	194	0.80096	1
Visible light detection result PN_{vi}								3

Table 2: Infrared detection results in a daytime scenario.

Blob number k	Input						Output	
	A_k	EP_k	L_k	H_k	PX_k	PY_k	PN_k	Rounded results
1	2504	343	55	103	152	84	2.5982	3
2	1345	173	28	76	154.5	136.5	0.78543	1
Infrared detection result PN_{in}								4

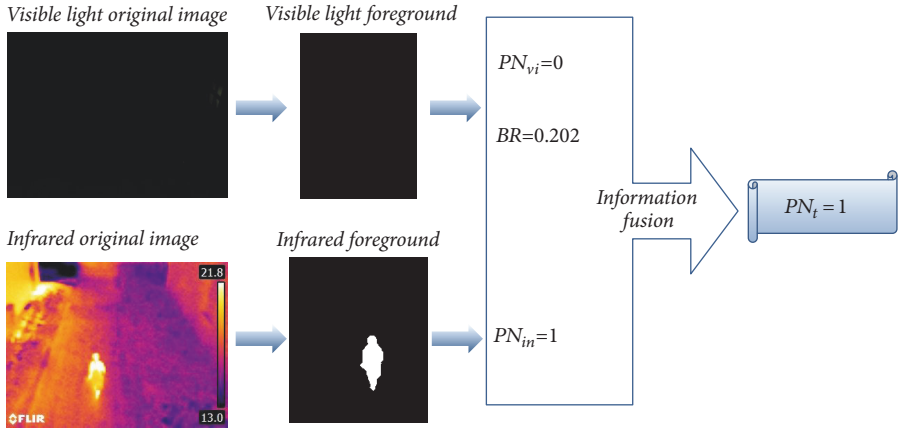
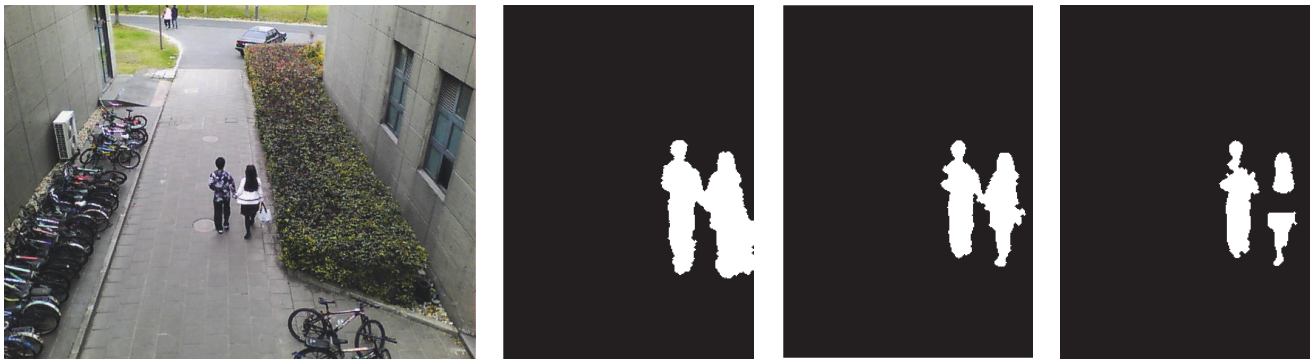


Figure 7: Schematic diagram of information fusion in a night scenario.



(a) *Original image* (b) Th_{in} = 30 (c) Th_{in} = 45 (d) Th_{in} = 70

Figure 8: The results of motion foreground extraction with different threshold Th_{vi}.

Table 3: Infrared detection results in a nighttime scenario.

Blob number *k*	**Input**						**Output**	
	A_k	EP_k	L_k	H_k	PX_k	PY_k	PN_k	Rounded results
1	1159	154	29	69	133	104	0.91752	1
Infrared detection result PN_{in}								1

Table 4: The results of pedestrian counting with different threshold Th_{vi}.

Th_{vi}	**Blob number *k***	**Input**						**Output**	**Detection result**
		A_k	EP_k	L_k	H_k	PX_k	PY_k	PN_k Rounded results	
30	1	5840	534	82	122	181.5	180.5	2.73 3	3
45	2	4045	517	76	112	176.5	177.5	2.37 2	2
70	1	1864	263	36	104	172.5	158.5	1.04 1	1
	2	613	130	26	48	207.5	196.5	0.22 0	
	3	521	91	20	36	149.5	197.5	0.36 0	

visible image, we set the threshold Th_{vi} to 30, 45, and 70, respectively. The results of the extraction of the motion foreground are shown in Figure 8 and the results of the pedestrian detection are shown in Table 4. According to Figure 8 and Table 4, we can find that when the threshold Th_{vi} is too small (Th_{vi}=30), the motion foreground contains more noise, and the final pedestrian count result is too large. When threshold Th_{vi} is too large (Th_{vi}=70), the motion foreground is incomplete and the final pedestrian count results are small. Therefore, the thresholds Th_{vi} and Th_{in} need to be set according to the characteristics of the data and the actual situation.

4.4. Contribution of the Features. The method in this paper is based on six features (Section 2.3 Feature Extraction). In order to evaluate the contribution of these features to the final result, the average elastic coefficient is introduced. The bigger the average elastic coefficient of the feature, the greater contribution to the final result. And the average elastic coefficient $\overline{E}$ is

$$\overline{E} = \frac{\Delta y/\overline{y}}{\Delta x/\overline{x}} \tag{19}$$

where $\overline{x}$ is the average of the independent variables and $\overline{y}$ is the average of the dependent variables.

In the visible light model and the infrared model, the average elastic coefficient of each feature is calculated separately. The calculation results are shown in Figure 9. We have found that the feature PX is the most influential feature of the final result in both the visible light model and the infrared model, because PX is the most important parameter to represent the distance from the pedestrian to the lens in the testing scenario of this paper. On the other hand, the features A, EP, and L are reasonable predictors of crowd density, which reflects the number of pedestrians from different angles. One possible explanation for the low contribution of features H and PY is that the camera's field of view is parallel to the road, not vertical or oblique in the testing scenario of this paper.

4.5. The Efficiency of Background Update. Visible light video detection is based on background differences to obtain motion foreground. Since the environment around the pedestrian waiting area varies greatly in a day, real-time background update is a must. Here, the efficiency of the background update method based on Kalman filter is tested. Three rounds of tests were performed on 102 images. The

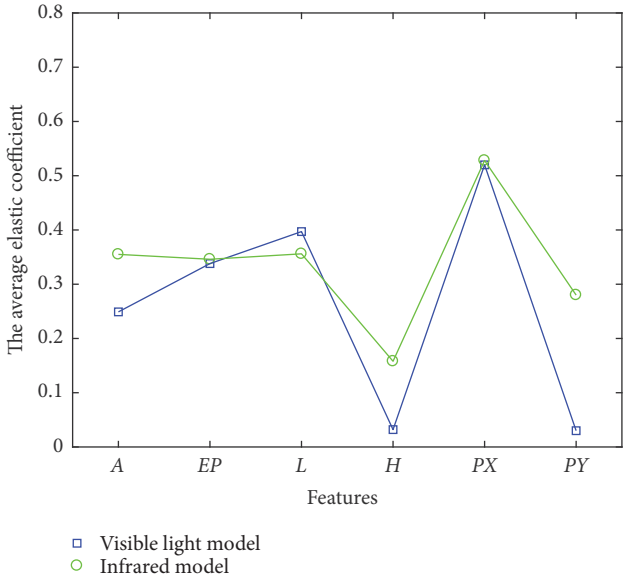


Figure 9: The calculation results of average elastic coefficient.

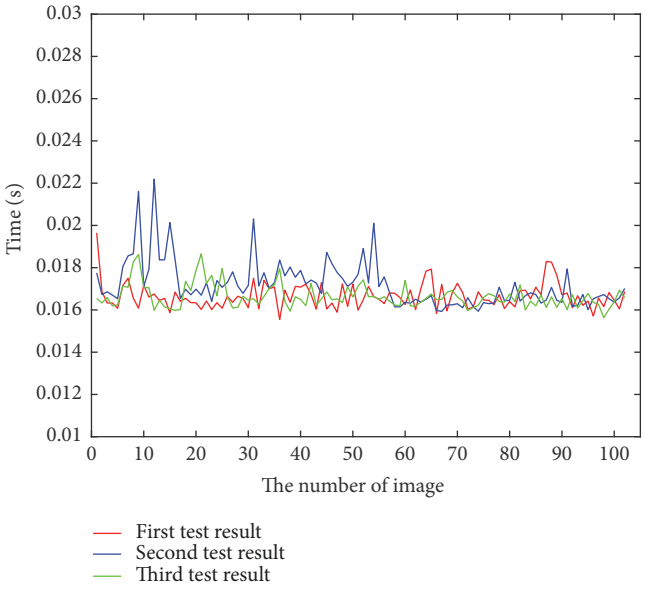


Figure 10: The test results of background update efficiency.

results are shown in Figure 10 and Table 5. (Note: this test was performed on a laptop and the test software is MATLAB 2017b.)

According to Table 5 and Figure 10, it can be found that the average time of background update is about 0.0168s. The result is ideal and can meet the needs of practical applications.

4.6. Accuracy Verification. The accuracy of the proposed method is verified based on 8-fold cross validation. 124 images are randomly divided into 8 groups. Each group is in turn used as a test set for the model, and the remaining 7 groups serve as a training set for the model. The experimental results are shown in Figure 11 and Table 6. As can be seen from Figure 11, the accuracy of the individual visible light detection

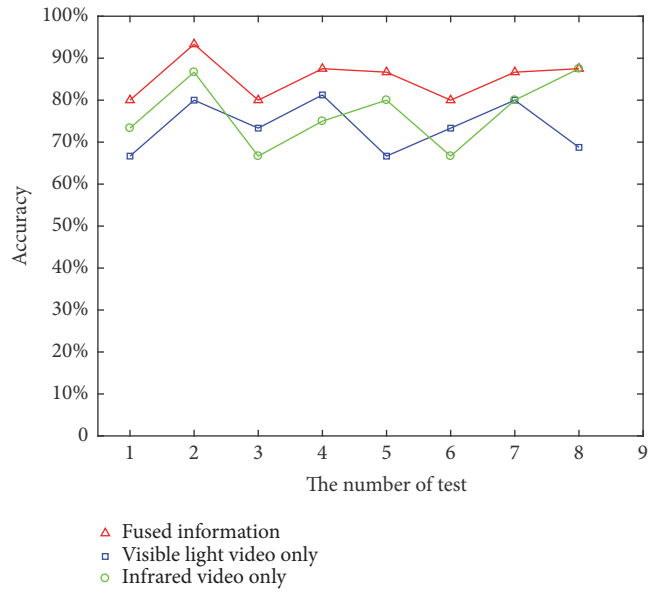


Figure 11: The results of 8-fold cross validation.

Table 5: The efficiency of background update.

Number	Total time/s	Average time/s
1	1.6985	0.0167
2	1.7535	0.0172
3	1.6967	0.0166

is sometimes higher than that of the individual infrared detection, and sometimes lower. However, the accuracy of information fusion detection is always the highest. Combined with Table 6, the average accuracy of information fusion detection is higher than that of the individual visible light and infrared detections, while considering both daytime scenarios and nighttime scenarios.

Moreover, since there is no public dataset containing both infrared and visible light images, we test other methods on the dataset of this paper to show the advantage of the proposed method. Table 7 listed the prediction accuracy comparisons. It can be seen that the fusion method could provide better performance with lower MSE and higher accuracy as compared to the existing methods. Therefore, for 24-hour pedestrian counting in outdoor scenarios, the fusion method between visible light video and infrared video from the perspective of environment perception is more effective than the single video (visual videos or infrared videos).

5. Conclusion

In this study, a fused method between visible light video and infrared video based on environment perception for estimating the number of pedestrians has been proposed. And the method is intended to combine visual light information with infrared information to enable pedestrian counting techniques for complex outdoor scenarios. The proposed

Table 6: The results of 8-fold cross validation.

The number of test	Visible light video only	Infrared video only	Fused information
1	66.67%	73.33%	80.00%
2	80.00%	86.67%	93.33%
3	73.33%	66.67%	80.00%
4	81.25%	75.00%	87.50%
5	66.67%	80.00%	86.67%
6	73.33%	66.67%	80.00%
7	80.00%	80.00%	86.67%
8	68.75%	87.50%	87.50%
Average accuracy	73.75%	76.98%	85.21%

Table 7: The comparison of MSE and accuracy.

Paper	MSE	Accuracy
[12]	0. 726	62.50%
[15]	0.412	76.88%
Proposed method in this study	0.246	85.21%

approach is depending on two aspects: the estimation of the number of pedestrians based on single-source video and the information fusion based on multisource detection results. First, PLSR was applied to combine the dimensionality reduction analysis with the regression analysis to establish the pedestrian number estimation model based on single-source video. The method holds the advantages of reducing the redundancy of the data in the feature set and effectively solving the multiple correlations between variables. Meanwhile, the ambient brightness was employed to identify the scene of images and integrate the visible light detection result and the infrared detection result. The empirical analyses showed that, for 24-hour pedestrian counting in outdoor scenarios, the proposed method has better performance than the method using single information source, which expands the application scenario of pedestrian counting and provides reference for relevant research.

As for future analyses, one thing that needs to be expanded is the sample size of the empirical analyses and test the feasibility of utilizing deep learning networks to identify different scenarios (day, night, rain, fog, etc.). Besides, being under heavy fog or rain conditions will substantially increase the noise of video, and how to reduce the interference of these noises on pedestrian count would be a challenging issue in the future to be investigated. In addition, continued improvements of the information fusion model and the feasibility of employing new sensing equipment (such as laser scanners) to estimate the number of pedestrians will be tested.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by National Key R&D Program of China (2016YFB1200402), National Natural Science Foundation of China (61703308; 71771174), and the Fundamental Research Funds for the Central Universities.

References

[1] K. H. Guan, *An adaptive optimization method of pedestrian crossing time based on video detection*, Jilin University, 2015.

[2] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[3] Y.-C. Chiu and P. B. Mirchandani, "Online behavior-robust feedback information routing strategy for mass evacuation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 2, pp. 264–274, 2008.

[4] P. Viola and M. Jones, "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade," *Advances in Neural Information Processing Systems*, vol. 14, pp. 1311–1318, 2001.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, June 2005.

[6] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07*, USA, June 2007.

[7] O. Barinova, V. Lempitsky, and P. Kholi, "On detection of multiple object instances using hough transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1773–1784, 2012.

[8] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1679–1697, 2004.

[9] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, "Pedestrian detection in infrared images," in *Proceedings of the*

2003 IEEE Intelligent Vehicles Symposium, IV 2003, pp. 662–667, USA, June 2003.

[10] R. O'Malley, E. Jones, and M. Glavin, "Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation," *Infrared Physics & Technology*, vol. 53, no. 6, pp. 439–449, 2010.

[11] S. Biswas K and P. Milanfar, "Linear Support Tensor Machine: Pedestrian Detection in Thermal Infrared Images," *IEEE Transactions on Image Processing*, p. 99, 2016.

[12] A. C. Davies, J. H. Yin, and S. A. Velastin, "Crowd monitoring using image processing," *Electronics & Communication Engineering Journal*, vol. 7, no. 1, pp. 37–47, 1995.

[13] G. He, Q. Chen, D. Jiang, X. Lu, and Y. Yuan, "A double-region learning algorithm for counting the number of pedestrians in subway surveillance videos," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 302–314, 2017.

[14] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1–8, 2008.

[15] J. Zhang, B. Tan, F. Sha, and L. He, "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.

[16] Y. Li, G. J. Wang, and H. G. Lin, "Crowd density estimation algorithm combining local and global features," *Journal of Tsinghua University (Science and Technology)*, vol. 4, pp. 542–545, 2013.

[17] Q. Xinhui, W. Xiufei, and X. Zhou, "Counting people in various crowed density scenes using support vector regression," *Journal of Image and Graphics*, vol. 18, no. 4, pp. 392–398, 2013.

[18] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2008.

[19] Z. W. Kim, "Real time object tracking based on dynamic feature grouping with background subtraction," in *Proceeding of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, AL, USA, June 2008.

[20] K. Suzuki, I. Horiba, and N. Sugie, "Linear-time connected-component labeling based on sequential local operations," *Computer Vision and Image Understanding*, vol. 89, no. 1, pp. 1–23, 2003.

[21] R. Yu, M. Quddus, X. Wang et al., "Impact of data aggregation approaches on the relationships between operating speed and traffic safety," *Accident Analysis & Prevention*, vol. 120, pp. 304–310, 2018.

[22] S. D. Jong and A. Phatak, "Partial least squares regression," in *Proceedings of the International Workshop on Recent Advances in Total Least Squares Techniques and Errors-In-Variables Modeling*, pp. 25–36, Society for Industrial and Applied Mathematics, 1997.