Crowd counting by feature-level fusion of appearance and fluid force

Dingxin Ma
Hangzhou Dianzi University
Hangzhou, China
mdx@hdu.edu.cn

Xuguang Zhang
Hangzhou Dianzi University
Hangzhou, China
zhangxg@hdu.edu.cn

Hui Yu
University of Portsmouth
Portsmouth, UK
hui.yu@port.ac.uk

***Abstract*—Crowd counting is a research hotspot for video surveillance due to its great significance to public safety. The accuracy of crowd counting depends on whether the extracted features can effectively map the number of pedestrians. This paper focuses on this problem by proposing a crowd counting method based on the expression of image appearance and fluid forces. Firstly, Horn-Schunck optical flow method is used to extract the motion crowd. Secondly, based on the motion information of crowd, pedestrians in different directions are distinguished by the k-means clustering algorithm. Then, image appearance features and fluid features are extracted to describe different motion crowd. The image appearance features are gained by calculating the foreground area, foreground perimeter and edge length. The gravity, inertia force, pressure and viscous force are taken as the fluid features. Finally, two kinds of features are combined as the final descriptor and then least squares regression is used to fit features and the number of pedestrians. The experimental results demonstrate that the proposed crowd counting method acquires satisfied performance and outperforms other methods in terms of the mean absolute error and mean square error.**

Keywords—video surveillance, crowd counting, fluid forces, least squares regression

I. INTRODUCTION

Crowd video surveillance provides an important Crowd guarantee for public safety. Crowd video surveillance usually includes crowd detection [1], crowd analysis [2] and crowd counting [3]. Since the number of pedestrians is a key indicator of crowd safety, crowd counting has received great attention from researchers. Generally, crowd counting methods can be divided into two categories: microscopic method and macroscopic method. For the microscopic method, the crowd is regarded as a collection of independent individuals. Then by detecting some body structures and tracking individual trajectories, the number of people is counted. However, this kind of methods is only suitable for a small-scale crowd. Because it is difficult to detect and track the target accurately when the crowd has occlusion. The macroscopic method addresses this problem by treating the crowd as an entirety. After that, some features are extracted and regression model is used to establish the relationship between features and the number of people. For this kind of methods, the key to crowd counting is whether the extracted features accurately reflect the number of pedestrians. Most of the macroscopic methods only describe the number of pedestrians by expression of image appearance, which ignores the motion information of the crowd.

Since the size of crowd can be shown in their motion information, this paper propose a fluid descriptor to characterize the movement pattern of the crowd. The fluid features are described by calculating fluid forces between pedestrians in the scene. The fluid forces model was first proposed to study the motion characteristics of fluid particles in smooth particle hydrodynamics (SPH) [4]. In recent years, researchers have applied it to the study of crowd consistency [5]. As the increase or decrease in the number of pedestrians will lead to changes in fluid forces throughout the scene, we use it as a descriptor for crowd counting. Like many existing methods, we also extract image appearance features, area and perimeter of foreground, length of edge. The fluid features are combined with traditional image appearance features to provide a final descriptor to effectively characterize the crowd from both motion and appearance. The framework of the proposed method is shown in Figure 1.

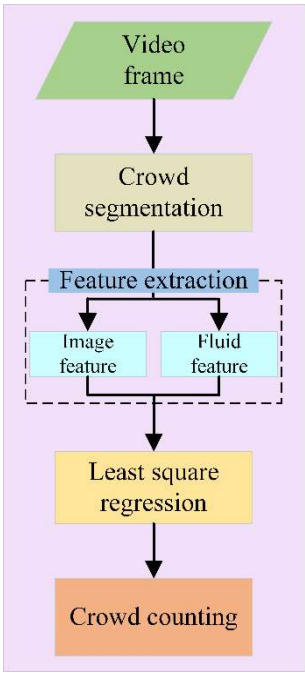

Fig. 1 The result of crowd segmentation

The remainder of this paper is structured as follows: Section 2 provides an overview of related work. Section 3 introduces the method of crowd segmentation. The extraction of image appearance features and fluid features is described in Section 4. Section 5 presents the crowd counting. A detailed experimental results and discussions are given in Section 6. In the end, Section 7 concludes this paper.

II. RELATED WORK

For the microscopic category, the crowd is regarded as a collection of independent individuals. So, the size of crowd is counted by detecting some body structures and tracking individual trajectories. Pätzold et al. used the combination of a shape model and a uniform motion model to detect the head-shoulder region of a human. Then the performance

was improved by tracking the coherent motion detections [6]. Merad et al. extracted the head of each person by using a new head-based detection method from skeleton graph [7]. Ge et al. proposed a Bayesian marked point process (MPP) to model human body shape. And a weighted mixture of Bernoulli distributions was then used to augment the model with intrinsic shape information [8]. Other researchers detect and count individuals in the scene by using face detection [9] and gait recognition [10]. However, this kind of methods is only suitable for small-scale crowd. Because of the occlusion between pedestrians, it is difficult to detect and track the target accurately.

As for the second category, such methods effectively solve the occlusion problem by treating the crowd as an entirety. Some image appearance features, such as foreground features, edge features and texture features, are extracted at first. Then by using regression models, the mapping relationship between features and the number of people can be learned. Marana et al. use the probabilities of grey-level transitions to monitor crowd density features from digitized images of the area, which are applied into a neural network [11]. Marana et al. presented an approach in solving the problem of crowd density estimation by using Minkowski fractal dimension to characterize data texture [12]. Rahmalan et al. proposed a new method called Translation Invariant Orthonormal Chebyshev Moments to extract image features. Then a Self Organizing Map is used to classify the features into a range of density [13]. Liang et al. extracted feature points using the Speed Up Robust Features (SURF) and cluster them to eliminate non-motion feature points. Then these feature points were used to construct crowd eigenvectors, which were trained based on support vector regression machine [14]. In previous research, we used the total number of foreground pixels to map the number of pedestrians based on Least squares fitting [15]. For this kind of methods, the key to crowd counting is whether the extracted features accurately reflect the number of people.

III. CROWD SEGMENTATION

In order to count pedestrians in different directions, motion crowd in different directions should be distinguished at first. Due to background interference, we use Horn-Schunck optical flow method to extract foreground. Then, based on the motion information of crowd flow field, different pedestrians are distinguished by k-means clustering algorithm.

A. Motion crowd extraction

For an image, a set of velocity estimates (u^{n+1},v^{n+1}) of the $(n+1)_{th}$ frame is computed by the following formula:

$$u^{n+1} = \bar{u}^n - \frac{E_x \cdot \bar{u}^n + E_y \cdot \bar{v}^n + E_t}{\alpha^2 + E^2{}_x + E^2{}_y} \cdot E_x$$
$$v^{n+1} = \bar{v}^n - \frac{E_x \cdot \bar{u}^n + E_y \cdot \bar{v}^n + E_t}{\alpha^2 + E^2{}_x + E^2{}_y} \cdot E_y \quad (1)$$

where E_x, E_y and E_t are the estimated derivatives of the image at the position (x, y, t) in the corresponding directions, $\bar{u}^n$ and $\bar{v}^n$ are the average of the previous horizontal and vertical velocity estimates, α^2 is a weighting factor, which plays an important role in reducing the corruption of the quantization error and noise. Then the velocity magnitude

Fig. 2 The result of crowd segmentation

uv and velocity direction θ can be calculated using the following equations:

$$uv = \sqrt{u^2 + v^2}$$
$$\theta = \arctan(\frac{v}{u}) \quad (2)$$

Since the velocity magnitude of background pixel is often small, we can filter the background to obtain the moving pedestrians by setting a threshold η. Pixels with a velocity magnitude less than η are considered as background pixels. Otherwise, pixels are assumed as the foreground.

B. Different motion crowd segmentation

According to the motion information of the crowd, we use k-means clustering algorithm to segment the motion crowd. For k-means clustering algorithm, given a set of data $X = \{X_1, X_2, \ldots, X_n\}$ and the number of data subsets k, the goal of clustering is to cluster data into k subsets that minimizes the within groups sum of squared errors (WGSS) [16], which is formulated as:

$$\textit{Minimise} \quad P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{i,l} d(X_i, Q_l)$$
$$\textit{subject to} \quad \sum_{l=1}^{k} \omega_{i,l} = 1, \quad 1 \le i \le n$$
$$\omega_{i,l} \in \{0, 1\}, \quad 1 \le i \le n, \ 1 \le l \le k \quad (3)$$

where W is an $n \times k$ partition matrix, $Q = \{Q_1, Q_2, \ldots, Q_k\}$ is a set of data in the same subset, $d(\cdot,\cdot)$ is the distance between two data. In this paper, Euclidean distance is used to assess the similarity between pedestrians, the similarity is calculated as:

$$d(X, X_i) = \sqrt{(\theta - \theta_i)^2} \quad (4)$$

Figure 2 presents the result of crowd segmentation. Figure 2(a) shows the input frame. The result of motion crowd extracting is shown in Figure 2(b). We give the result of different motion crowd segmentation in Figure 2(c).

IV. FEATURE EXPRESSION IN DIFFERENT MOTION CROWD

After obtaining the motion crowd in different directions, the image appearance features and fluid features of each motion crowd are extracted for crowd counting. The extraction for image appearance features and fluid features will be discussed in detail in this section.

A. Image appearance features extraction

The number of pedestrians is often reflected in image appearance. Foreground features and edge features are two classic descriptors, which can effectively characterize the appearance of the crowd image. Then three statistics extracted from foreground image and edge image are used to describe the image appearance of motion crowd.

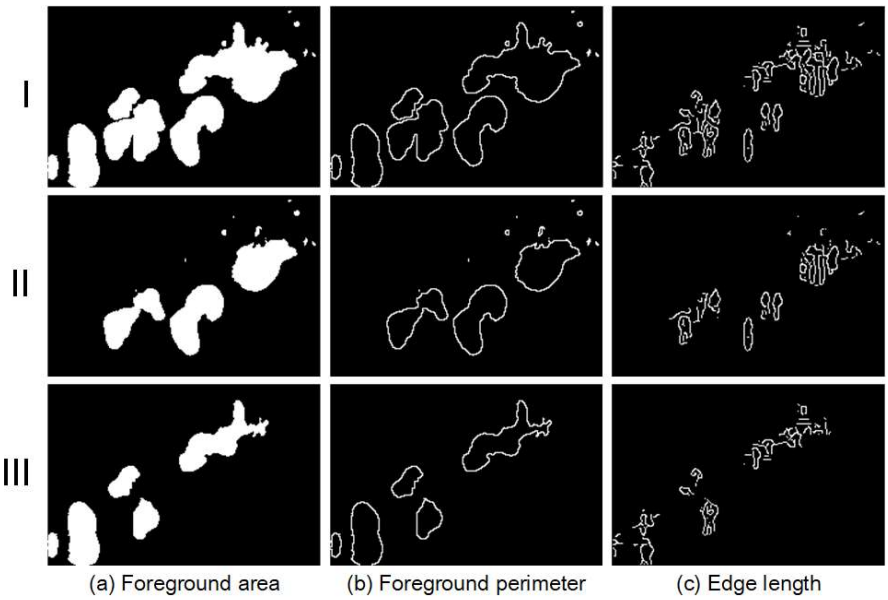

Fig. 3 Examples of foreground area, foreground perimeter and edge length. (I) Entire scene (II) Away from the camera (III) Towards to the camera

In foreground image, foreground area is the number of pixels in foreground, which can be seen in Figure 3(a) .

$$area = \sum_{I(i,j)\in A} I(i,j) \tag{5}$$

where A is a set of pixels in foreground. As shown in Figure 3(b), foreground perimeter is the number of pixels in foreground perimeter.

$$perimeter = \sum_{I(i,j)\in P} I(i,j) \tag{6}$$

where P is a set of pixels in foreground perimeter. In edge image, edge length is the number of pixels in edge within a foreground, which is shown in Figure 3(c).

$$length = \sum_{I(i,j)\in L} I(i,j) \tag{7}$$

where L is a set of pixels in edge within a foreground. It is worth to note that the image appearance features are affected by perspective. Because the numbers of pedestrian pixels are reduced when pedestrian is towards to the camera. Otherwise, the numbers of pedestrian pixels increase. In this paper, this problem is solved by weighting each pixel according to the distance between the camera and the scene.

B. Fluid features extraction

Crowd movement has strong physical properties [17], so we model the crowd flow as a group of interacting particles. The properties of each particle can be described by physical quantities such as pressure, density and speed. Furthermore, the flow field dynamics model among particles is constructed to express the behavior of crowd groups. The proposed fluid feature is based on Smoothed Particle Hydrodynamics (SPH) [4], which has been widely used in many research fields, such as astrophysics, shock explosion and hydrodynamics. In addition, the SPH model has shown good performance in detecting the consistency and anomaly of crowd movement [5, 18].

We analyze the forces in crowd motion by considering the Navier-Stokes equation, which is formulated in follow:

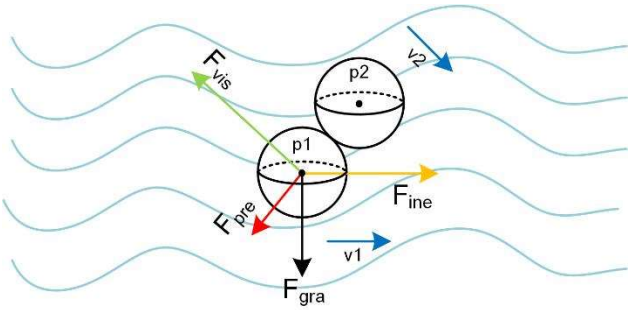

Fig. 4 The forces model of motion crowd

$$\rho\left(\frac{\partial v}{\partial t} + v\cdot\nabla v\right) = F_{gra} + F_{ine} + F_{pre} + F_{vis} \tag{8}$$

According to the N-S equation, we can find that fluid motion is mainly affected by gravity, inertia force, pressure and viscose force, the model is shown in Figure 4. The complete motion pattern of the particles is formed by the combination of these four forces.

Gravity and inertia force are the embodiment of the particle's own motion pattern under the gravitational field and the inertial field, which are respectively formulated as:

$$\begin{aligned} F_{gra} &= \sum_{i=1}^{N} m_i g \cdot K\left(r_c - r_i, \lambda\right) \\ F_{ine} &= \sum_{i=1}^{N} m_i \frac{\left|v_i^{\,t} - v_i^{\,t-1}\right|}{T} \cdot K\left(r_c - r_i, \lambda\right) \end{aligned} \tag{9}$$

where g is gravity acceleration, v_i^t is the speed of current frame, v_i^{t-1} is the speed of previous frame, T is the duration of a frame. $K(r_c\text{-}r_i, \lambda)$ is the smooth kernel function. It can be understood as a weight function of the extent to which other particles affect the study particle over a range of smooth length λ. A Gaussian kernel function is used in this paper.

Pressure reflects the collision between particles. In order to calculate it, the crowd flow is regarded as an ideal fluid and Bernoulli equation is considered, which is formulated as:

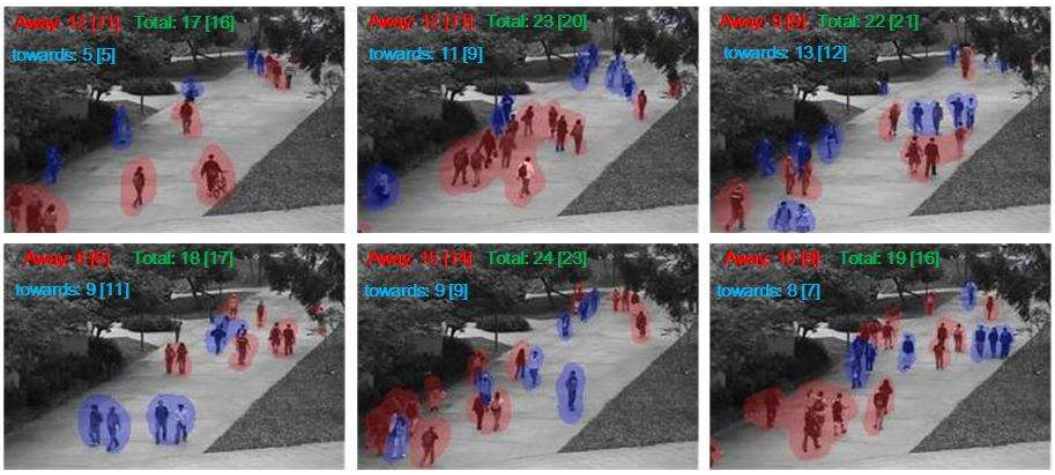

Fig 5. Sample frame of crowd segmentation and pedestrians counting on Peds1: people away from and towards to the camera are marked red and blue. The counting results and ground truth for each motion crowd are shown in the top left.

$$p + \frac{1}{2}\rho v^2 + \rho g h = C \tag{10}$$

where p is pressure, v is velocity, C is a constant. Since the pedestrians are walking on the ground, the gravitational potential energy ρgh is set as 0. So, pressure is formulated as:

$$F_{pre} = \sum_{i=1}^{N}\left(C - \frac{1}{2}m_i v^2\right) S \cdot K\left(r_c - r_i, \lambda\right) \tag{11}$$

The viscous force is produced by the velocity difference, which reflects the mutual friction between particles. It is formulated as:

$$F_{vis} = \sum_{i=1}^{N} \mu S \frac{v_i - v_s}{R} \cdot K(r_c - r_i, \lambda) \tag{12}$$

where μ is fluid viscosity, S is contact area between particles, v_i is velocity of central particle, v_s is velocity of surrounding particle, R is distance between two particles. The kernel function is set as Gaussian kernel function.

V. Crowd Counting

In our work, the image appearance features and fluid features are combined as the final descriptor. In order to establish the relationship between features and the number of pedestrians, the least squares regression is used.

Assume that the data point is (X_{mi}, y_i), where $X_{mi} = (x_{1i}, x_{2i}, \ldots, x_{mi})$ are the features of crowd and y_i is the number of pedestrians. The liner relationship between features and the number of pedestrians is shown as follow:

$$y = a_1 x_{1i} + a_2 x_{2i} + \cdots + a_m x_{mi} + a_{m+1} \tag{13}$$

Base on it, the deviation of each data is $d_i = y_i - y$. And the summed square of residuals is calculated by $S = \sum_{i=1}^{n} d_i$. The target of least squares regression is to minimize the S.

Then we can apply the training data to least squares regression model to fit features and the number of pedestrians. So, the number of pedestrianscan be predicted according to testing data.

VI. Experimental Results

In this paper, UCSD Peds 1 dataset is used to demonstrate the effectiveness of the proposed method. Then, the error of crowd counting is evaluated based on Mean absolute error (MAE) and Mean square error (MSE).

A. Parameter setting

In the proposed method, some parameters should be set. The first one is the threshold to obtain foreground, it is set as $\eta = 0.2$. The second one is the number of clusters. Since there are two different motion directions of the pedestrians in Peds1 dataset, it is set as $k = 2$. Others are the parameters in calculation of fluid features: the smooth length is set as $\lambda = 3$. The mass of each particle is set as $m = 1$. The Bernoulli equation constant is set as $C = 10$. The contact area between particles is set as $S = 1$.

B. Crowd counting for Peds 1

Peds1 contains a large number of pedestrians. There are 20 video sequences for a total of 4000 frames. And the resolution is 238×158. In Peds1, the motion crowd is divided into two categories, the first one is “away” from the camera, and the other is “towards” to the camera. The training data contains 1200 frames (frames 1401-2600) and the remaining 2800 frames is used for testing.

Figure 5 shows some sample frames of crowd segmentation and pedestrians counting on Peds1. As can be seen in figures, pedestrians in different directions can be well separated. And top left of figures presents the counting results of different motion crowd. The number of pedestrians counted by proposed method is very close to the ground truth.

The counting results of the whole test frames are shown in Figure 6. Figure 6(a) shows the results of pedestrians away from the camera. The results of pedestrians towards to the camera can be seen in Figure 6(b). We give the results of total pedestrians in the scene in Figure 6(c). These three curves prove that our method can well describe the real situation of the ground in most test frames.

In order to evaluate the error of crowd counting of our method, we mainly follow two commonly measurements:

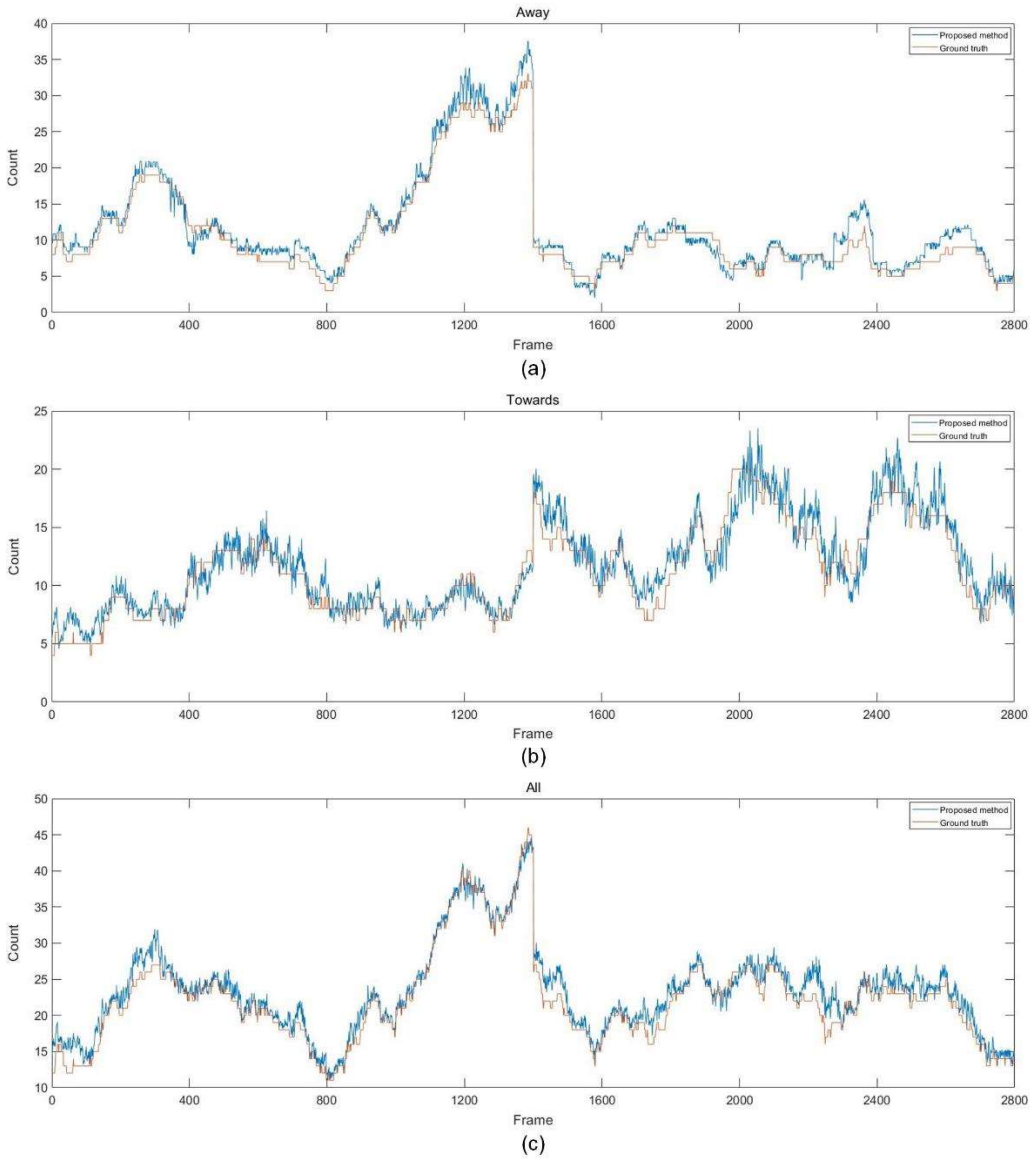

Fig 6. Crowd counting results on Peds1: (a) pedestrians away from the camera (b) pedestrians towards to the camera (c) all the pedestrians in the scene

Table I. Comparison of MAE and MRE on peds1

Method	MAE				MSE			
	away	towards	scene	total	away	towards	scene	total
Linear	1.451	1.324	1.513	4.288	3.335	2.868	3.751	9.953
GPR-l	1.435	1.278	1.489	4.203	3.260	2.692	3.654	9.606
GPR-rr	1.408	1.093	1.551	4.051	2.970	2.029	3.787	8.785
Poisson [19]	1.336	1.360	1.331	4.027	2.917	3.065	3.040	9.022
[20]	1.416	1.418	1.478	4.312	3.264	3.105	3.640	10.010
[21]	1.385	1.339	1.500	4.224	3.118	2.808	3.661	9.587
ours	1.3239	1.1749	1.3460	3.8448	2.6260	2.2403	3.1383	8.0046

mean absolute error (MAE) and mean square error (MSE).

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\left|E(i)-T(i)\right| \qquad MSE = \frac{1}{M}\sum_{i=1}^{M}(E(i)-T(i))^2 \tag{14}$$

Based on MAE and MSE, we compare the proposed method with other contributions [17-19]. The counting error for each crowd motion (away, towards and scene) is shown in Table I. Furthermore, we also calculate the total error to assess overall performance of each method. It can be seen in Table I, BPR-rr achieves the best performance among all the methods. The total MAE and MSE are 3.654 and 7.412. The proposed method is also competitive, our error rate is less than most methods. And the MAE and MSE for each motion crowd and total error are very close to BPR-rr.

VII. Conclusion

In this paper, we propose a crowd counting method based on the expression of image appearance and fluid forces. Firstly, Horn-Schunck optical flow method is used to obtain motion crowd. Secondly, pedestrians in different directions are segmented by k-means clustering algorithm. Then, image appearance features and fluid features are extracted and fused to describe different crowd motions. The image appearance features are gained by calculating the foreground area, foreground perimeter and edge length. The gravity, inertia force, pressure and viscous force are taken as the fluid features. Finally, two kinds of features are combined as the final descriptor and least squares regression is used to fit features and the number of pedestrians. The experimental results demonstrate that the proposed crowd counting method is competitive in comparison with other methods. In future works, we will test our method to more datasets and enhance its adaptability in different scenes.

Acknowledgments

This research was supported by National Natural Science Foundation of China (no. 61771418).

References

[1] X. Zhang, X. Shu, Z. He, "Crowd Panic State Detection using Entropy of The Distribution of Enthalpy," Physica A: Statistical Mechanics and its Applications, vol. 525, pp. 935-945, 2019.

[2] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, X. Yang, "Data-Driven Crowd Understanding: A Baseline for a Large-Scale Crowd Dataset," IEEE Transactions on Multimedia, vol. 18, no. 6, pp. 1048-1061, June 2016.

[3] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M. Cheng, G. Zheng, "Crowd Counting with Deep Negative Correlation Learning," IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 5382-5390, 2018.

[4] R. A. Gingold, J. J. Monaghan, "Smoothed particle hydrodynamics: theory and application to non-spherical stars," Monthly Notices of the Royal Astronomical Society, vol. 181, no. 3, pp. 375-389, 1977.

[5] H. Ullah, M. Uzair, M. Ullah, A. Khan, A. Ahmad, W. Khan, "Density independent hydrodynamics model for crowd coherency detection," Neurocomputing, vol. 242, pp. 28-39, 2017.

[6] M. Pätzold, R. H. Evangelio, T. Sikora, "Counting People in Crowded Environments by Fusion of Shape and Motion Information," IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, pp. 157-164, 2010.

[7] D. Merad, K. Aziz, N. Thome, "Fast People Counting Using Head Detection from Skeleton Graph," IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, pp. 233-240, 2010.

[8] W. Ge, R. T. Collins, "Marked point processes for crowd counting," IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, pp. 2913-2920, 2009.

[9] V. Paul, M. J. Jones, "Robust Real-Time Face Detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004.

[10] Q. Yang, D. Xue, "Gait recognition based on sparse representation and segmented frame difference energy image," Information and Control, vol. 42, no. 1, pp. 27-32, 2013.

[11] A. N. Marana, L. Da Fontoura Costa, R. A. Lotufo, S. A. Velastin, "Estimating crowd density with Minkowski fractal dimension," IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, USA, vol. 6, pp. 3521-3524, 1999.

[12] A. N. Marana, S. A. Velastin, L. F. Costa, R. A. Lotufo, "Estimation of crowd density using image processing," IET Colloquium on Image Processing for Security Applications, London, UK, pp. 11/1-11/8, 1997.

[13] H. Rahmalan, M. S. Nixon, J. N. Carter, "On Crowd Density Estimation for Surveillance," IET Conference on Crime and Security, London, pp. 540-545, 2006.

[14] R. Liang, Y. Zhu, H. Wang, "Counting crowd flow based on feature points," Neurocomputing, vol. 133, pp. 377-384, 2014.

[15] X. Zhang, H. He, S. Cao, H. Liu, "Flow field texture representation-based motion segmentation for crowd counting," Machine Vision and Applications, vol. 26, no. 7, pp. 871-883, 2015.

[16] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values." Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.

[17] X. Zhang, Q. Yu, H. Yu, "Physics Inspired Methods for Crowd Video Surveillance and Analysis: A Survey," IEEE Access, vol. 6, pp. 66816-66830, 2018.

[18] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, B. Stevens, "Scene perception guided crowd anomaly detection," Neurocomputing, vol. 414, pp. 291-302, 2020.

[19] A. B. Chan, N. Vasconcelos, "Counting People With Low-Level Features and Bayesian Regression," IEEE Transactions on Image Processing, vol. 21, no. 4, pp. 2160-2177, 2012.

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

[21] D. Kong, D. Gray, Hai Tao, "A Viewpoint Invariant Approach for Crowd Counting," International Conference on Pattern Recognition, Hong Kong, pp. 1187-1190, 2006.