Crowd Counting Method Based on Cross-Source Multi-View and Multi-Granularity Video Fusion

Liu Bai

Soochow University

Cheng Wu (✉ cwu@suda.edu.cn)

Soochow University

Yufeng Lin

Soochow University

Jin Zhang

Soochow University

Jie Sheng

Soochow University

Feng Xie

Soochow University

Yiming Wang

Soochow University

Research

Keywords: Crowd gathering safety, Computer vision, Multi-granularity data fusion, Distributed computing

Posted Date: March 4th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-16036/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Crowd Counting Method Based on Cross-Source Multi-View and Multi-Granularity Video Fusion

Liu Bai¹, Cheng Wu^{1*}, Yufeng Lin², Jin Zhang¹, Jie Sheng¹, Feng Xie¹ and Yiming Wang¹

Abstract

With the rapid growth of the world's population and the rapid development of urbanization, the issue of crowd gathering safety has aroused widespread concern in society. Extensive video surveillance systems provide rich data support for dense crowd management. Video-based crowd counting and density estimation methods are the core technologies to ensure the safety of crowd gathering. Different from single-view video analysis, cross-source multi-view multi-granularity video contains more cross-information. The complementary sharing of information is of great help to solve the problems such as occlusion in the current crowd counting. Therefore, this article proposes a crowd counting method based on cross-source multi-view and multi-granularity video distributed information fusion. By establishing a distributed structure from different cameras that matches low-altitude and high-altitude views, it uses fine-grained from low-altitude images. The high-resolution local information corrects and supplements the global information from high-altitude images, so as to calculate a more accurate and global number and density of people. This method is actually applied to the landmark building of Suzhou Life City Square. The changes in the number of people and the movement situation during the evacuation are analyzed and evaluated, and good results are obtained.

Keywords: Crowd gathering safety; Computer vision; Multi-granularity data fusion; Distributed computing

1 Introduction

In recent years, with the rapid growth of the world population and the rapid development of urbanization, frequent sports events, entertainment rallies, cultural exchanges and other activities have led to the actual number of people being much larger than the capacity of the venue. This causes a series of sudden safety issues and becomes a social focus [1]. In this case, the safety of places can only be improved by accurate crowd counting and density estimation. Traditional crowd counting often relies on the help of artificial experience, whose estimation results lack accuracy and real-time performance. At present, due to the rapid development of the Internet of Things technology [2], especially the widespread use of ubiquitous video surveillance systems, interconnected video capture devices have been able to provide a full range of dynamic coverage capabilities of perspectives, providing more data support abilities for crowd density estimates [3], which makes it possible to accurately count the crowds in dense places.

Crowd counting and density estimation based on continuous video frames are widely used in various

fields [4], such as city planning, traffic monitoring, public safety and so on. The design of large-scale urban complexes and crowd gathering infrastructure often involves the concept of "humanization", and the pattern and density of crowds have become one of the criteria for site design [5]. Some public places such as bus station also need real-time crowd counting [6] to online monitoring and early warning [7]. Some scholars are also dedicated to the development of automated systems or solutions for crowd counting [8], making public safety towards intelligent development. Compared with these works about crowd counting, the simulation design of crowd evacuation also focuses on the density distribution and accurate counting of the crowd. Ji et al. [9] aimed at the characteristics of high density crowd evacuation, proposed a new triangular mesh cellular automaton model and more accurately simulated the evacuation process of high-density people. Ben et al. [10] developed an agent-based cellular automaton (CA) environment model method, which simulated four different evacuation scenarios and effectively guided crowd evacuation.

Recent researches about crowd counting and density estimation are various. But we can see that there are still some core challenges:

*Correspondence: cwu@suda.edu.cn

¹School of Rail Transportation, Soochow University, Suzhou, China
Full list of author information is available at the end of the article

Figure 1 Challenges in crowd density estimation. Images are from NWPU-Crowd datasets [11].

- **Crowd occlusion:** Normally, occlusion has always been the main difficulty. Because the crowd is highly concentrated, the camera cannot accurately identify overlapping human bodies. It would reduce the estimation accuracy, as shown in Figure 1 (a).
- **Video perspective and depth of field (DOF):** Perspective can be described that the farther away from the camera, the larger the body size in the lens, and vice versa. It makes individual models not be measured with the same scale. DOF can be described as a distance in the field of view of the camera. The human body in the distance range is clear, and the human body outside the range is blurred, which is not conducive to the extraction of human characteristics, as Figure 1 (b) and (c).
- **Affected by the content and form of the event, terrain and other factors,** the crowd tends to gather in areas that are conducive to participating in activities, such as at outdoor music festivals, the crowd usually likes to gather on steps near and far from the stage. Because these places have better viewing horizons, see Figure 1 (e). This phenomenon is in stark contrast to the uniform arrangement shown in Figure 1 (d).
- **Non-uniform illumination of the environment:** The change of ambient light in a short time may increase more interference noise, make the target features complicated and fuzzy, and greatly reduce the counting accuracy, as Figure 1 (f).

These challenges make crowd counting extremely difficult and hard to apply effectively in real-world scenarios.

We found that there is cross-information between multiple cameras with the same granularity across sources, that is, multiple cameras in the same global environment share some local information, which complement each other. This is the crowd in a single camera. Problems such as occlusion provide new solutions. In addition, the current research methods and implementation methods calculate the number of people from video data perceived from camera equipment installed at low altitude (near the ground), which cannot meet the requirements for calculating the global number or density of public places in large cities. This is because the global camera equipment video data installed at high altitude (away from the ground) cannot provide the rich and specific crowd characteristics contained in the low-altitude video data. As shown in the first line in Figure 2, the crowd characteristics in the low-altitude video are very obvious. It is good for extraction calculations. The global picture shown in the second line in Figure 2 shows that the crowd is highly concentrated and the features are extremely insignificant. However, the global picture can provide the spatial and temporal flow trend of the crowd, which cannot be achieved by low-altitude video. For estimating the density of a wide-area population, if the multi-granular global camera and local camera information can be complementary, this allows the global population distribution to be decomposed into a multi-scale

Figure 2 Comparison of crowd situation in local viewing angle and global viewing angle. Images are from NWPU-Crowd datasets [11] and Shanghaitech dataset [21].

fusion of all local information, which is likely to make up for the uneven population distribution. The problem is that the estimation accuracy is inaccurate due to the uneven illumination with the environment.

In response to the above problems, this paper proposes a distributed cross-source multi-granularity collaborative real-time dynamic monitoring information fusion crowd counting method. By establishing a distributed structure of low-altitude perspective and high-altitude perspective, the integrity of low-altitude information is used to supplement the high-altitude information. To count the final global numbers. Among them, the estimation of the number of local people at low altitudes is based on a multi-column convolutional neural network model (MCNN), which extracts head features of different sizes, and finally normalizes to obtain the actual number of people. After several experiments, it proves that its accuracy is within the range of error tolerance. Crowd situational awareness at high altitudes uses a hybrid Gaussian model (GMM) to extract foreground images as key moving points, considers moving point spacing and assigns density values. Finally, the fusion analysis is performed in the actual video of multiple periods and multiple angles, and compared with the actual situation, the effect is significant.

The structure of the article is as follows: Section 2 introduces the related work of crowd density detection and multi-angle fusion from the field of machine vision. Section 3 describes a multi-column convolutional neural network model and a hybrid Gaussian model. Based on this, a multi-angle distributed information fusion model is established. Section 4 applies our model in

actual scenes, gives the experimental results, and conducts a comparative analysis; We discuss the results in Section 5.

2 Related work

2.1 Research on crowd density detection

In the field of machine vision, scholars have reviewed the methods of crowd counting and crowd density detection. Saleh et al. [12] summarized two methods of crowd density estimation and crowd counting, including object-based target detection and feature extraction (pixel, texture, corner) analysis. Yogameena et al. [13] analyzed various methods for studying crowd anomalies in crowded scenes, and made a further supplement on the basis of Saleh to propose crowd counting and density methods including direct and indirect methods.

The direct method based on target detection is to label the individuals in the image and finally add up. Lin et al. [14] used wavelet transform to extract the feature regions of beaded contours, and then used support vector machines to distinguish the features into head contours or non-head contours. Finally, they used perspective transformation technology to accurately estimate the number of people. Zhao et al. [15] proposed a model-based method to explain the observation results of multiple partially occluded human hypotheses in a Bayesian frame, using a data-driven Markov chain Monte Carlo method to integrate knowledge of human shapes, camera models, and image cues into the framework to estimate the number of people. Sidla et al. [16] introduced a vision-based pedestrian detection and tracking system that extracts shape information from video frames to detect individuals and count

people in very crowded situations. The main disadvantage of this method is that although individuals can be counted correctly by segmenting the individuals correctly based on the markers, it is not possible to accurately estimate the number of people in a densely populated high density crowd.

The indirect method based on feature extraction is to extract human features as training parameters, and finally normalize to get the total number of people. Liang et al. [17] proposed a statistical method of crowd flow based on feature points to obtain the characteristics of crowd flow in the scene, including crowd orientation and number counting. Support vector regression machine was used to train data and provide a higher accuracy of crowd flow statistics. Marana et al. [18] introduced a new crowd density estimation technique based on the Minkowski fractal dimension, which is used to characterize a large number of data textures in physical and biological sciences, and also reflects the crowdedness of people in crowd images. Kok et al. [19] learned to distinguish crowd characteristics from particles, which conform to the contour between the crowd and the background (ie, non-population) regions for density estimation. Fradi et al. [20] proposed a new crowd density measurement method in which pixel-level local information replaces the global crowd level or the number of people per frame, and used local features as observations of the probability density function to generate an automatic crowd density map. Zhang et al. [21] proposed a multi column convolutional neural network (MCNN) structure to map the image to its crowd density map. Zhang et al. [22] proposed a multi-resolution attention convolution neural network (MRA-CNN) to estimate the population density. In the training process, the density level classification task was added, combined with the characteristics of multi-scale and multi context, to respond to the uneven distribution of the crowd. Due to the lack of detailed extraction of human features, such methods cannot distinguish human features from environmental noise well in complex changing scenes, causing inevitable estimation errors.

In recent years, deep learning has become the mainstream of feature extraction methods [4], including convolutional neural networks (CNN), Contextual Pyramid CNNs (CP-CNN) [23], CSRNet [24] et al. Fu et al. [25] designed two ConvNet cascaded classifiers to improve the accuracy and speed of the classifier, and to optimize the crowd density by optimizing the convolutional neural network. Hu et al. [26] used the ConvNet structure to extract crowd characteristics from static images such as video images, and then used two types of monitoring signals: crowd count and crowd density to learn crowd characteristics and estimate specific numbers. Jiang et al. [27] increased the diversity

of deep neural networks with additional multi-scale labels, and achieved high-performance crowd density map estimation. Zhang et al. [28] introduced a label distribution learning strategy based on the regression model when training video data, which made up for the shortcomings of insufficient training data in the regression model and improved the estimation accuracy. Miao et al. [29] proposed a spatio-temporal convolutional neural network, which considers the inherent temporal correlation between adjacent frames of video, and uses the advantages of spatiotemporal information for crowd counting. Song et al. [30] proposed a method for estimating the crowd density of a video scene invariant crowd using a geographic information system (GIS) to monitor large-scale crowd sizes, which strengthened the adaptiveness of the camera's estimation model. In fact, with the widespread popularity of surveillance cameras, dynamic surveillance videos from different viewing angles at high and low altitudes can not only meet the requirements of model marking on human details, but also meet the needs of feature extraction on group textures. Therefore, the research center is shifting to the fusion of the two technical methods.

2.2 Research on multi-view video fusion

At present, the video analysis of single device, single view and single dimension no longer meets the work requirements, and the actual situation is more inclined to multi view video joint analysis. Xu et al. [31] et al. classified fusion into primary, intermediate and advanced multi viewing angle fusion according to the hierarchical classification of fusion information.

The first type is primary fusion. The idea is to start with the video image of a single camera. When the system predicts that the video at the current viewing angle does not meet the analysis needs, it switches to a camera with another viewing angle. Cai et al. [32] proposed a framework for simultaneous gray-scale image tracking of moving human body in indoor environment, which was used for multiple fixed cameras to shoot, automatically switched the camera equipment with different viewing angles, selected the camera that provided the best viewing angle and needed the least switching to continue tracking, and finally gave the experimental results to evaluate the performance of the tracking system. Mehrube et al. [33] used two synchronous cameras to obtain independent video streams, detected moving objects from two different viewing angles, and improved the robustness of tracking process. Because there are no related features between different cameras, this type of fusion information is very limited and is in the primary fusion stage.

Figure 3 Spatial relationship between local viewing angle and global viewing angle. The blue area is a high-altitude global field of view and fully covers the crowd space, and the red area is a low-altitude local field of view that is still within the global field of vision.

The second type is intermediate-level fusion. The idea is to extract the features of all individual camera devices and fuse them into a global feature sum. Kang et al. [34] proposed a method for multiple cameras to continuously track moving objects, which can process video streams from still and pan tilt zoom cameras at the same time, allow more accurate motion measurement of objects watched by two cameras at the same time, and automatically handle occlusion, detection error and camera switching. Xu et al. [35] proposed a system structure and method for tracking people, which solved the problem of single view background adaptive and multi view data fusion accuracy respectively, and was verified in calculating the real-time position of football players in the game. Cao et al. [36] proposed a video-based crowd density analysis and prediction system for wide-area surveillance applications. The cumulative mosaic image difference method is used to extract crowd areas with irregular motion, and a multi-camera network is used to predict crowd density. Although this fusion method attempts to solve the occlusion problem, it is still not ideal because these features are extracted from multiple camera views and there is great difference.

The third type is advanced fusion. The idea is that a single camera no longer extracts features, but provides foreground bitmap information to the fusion center. Xu et al. [37] proposed a depth information-guided crowd counting (DigCrowd) method to deal with crowded scenes for crowded scenes with unbalanced crowd distribution. Firstly, the scene was di-

vided into a distant view area and a near view area using depth information of the image. The distant view area was then mapped to its crowd density map, and the detection method was used to count the crowd in the near scenic area. Mousse et al. [38] proposed a crowd counting system based on overlapping cameras, which integrated all single view foreground information to locate everyone in the scene, and verified the real-time performance of the system on the public data set, improving the detection accuracy of crowd estimation. This type of method utilizes foreground bitmap information from multiple cameras to reduce occlusion interference. However, there will be unavoidable losses when mapping the foreground image, including: the network bandwidth performance may not meet the real-time requirements; the processing speed is too slow during mapping. Real-time monitoring systems may be less suitable.

For the above problems, we adopt a distributed approach. The multi-column convolutional neural network model is trained on the local video provided by the low altitude camera equipment to extract the head features of different sizes, and then the number of people is normalized, which greatly reduces the problem of inaccurate estimation accuracy caused by video angle; In the global video provided by the high-altitude camera, we use the Gaussian mixed model to extract the foreground image of the moving crowd, calculate the crowd density based on the key moving points in the foreground image, and establish a mapping relationship with the field of vision of the low-altitude

Figure 4 The structure of multi-column convolutional neural network model (MCNN).

camera in the global crowd situation to calculate the global headcount. Because the number of local crowds and the global crowd situation are processed in a distributed manner, the load on the network is reduced, and the operating speed is also increased. In addition, only the camera position needs to be fixed to meet the requirements of real-time monitoring.

3 Methodology

In this section, we will introduce the distributed local view counting method and global view density sensing method, and from the perspective of multi-view video fusion, the high and low viewing angle information is complementary, and finally the total number of people in the global viewing angle is obtained. The spatial relationship is shown in Figure 3.

3.1 Crowd counting from a local viewing angle

Due to the phenomenon of different sizes in the visual field of the crowd, the training target using a uniform scale neural network has a large deviation. Therefore, the head features of each individual are divided into three types: large, medium, and small. Convolution analysis is performed on each type of head feature, which determines that we use a multi-column convolutional neural network to calculate the number of people, as shown in Figure 4.

The input of the model is the original video image. Three columns of convolution kernel networks of different sizes extract the head features of different sizes in parallel. Among them, the first column is taken as

an example. The original image passes through three filters of sizes 5×5 , 7×7 and 9×9 to extract the head features under the convolution kernel, then passes the 2×2 maximum pooling layer, reducing the resolution of the upper layer image to the original $1/4$, to reduce the number of parameters. The activation function is a linear rectification function. Finally, a three-column feature is linearly weighted through a 1×1 filter and normalized. The density values are added to obtain the total number of people. The standardization process mainly depends on the gaussian kernel function:

$$F(x) = \sum_{i=1}^M \delta_i * G(x, \sigma_i) \quad (1)$$

where δ_i represents the impulse function of the i -th personal head, M represents the total number of heads, σ_i represents the adaptive average head distance within a certain range, and $\sigma_i = \alpha * \arg(d_{i,j})$ where α is the weight value of the adaptive range. The loss function is the adjusted euclidean distance loss function:

$$L(\beta) = \frac{1}{2N} \sum_{i=1}^N \|F(X_i, \beta) - F_i\|_2^2. \quad (2)$$

where β is the target parameter, N is the total number of pictures, X_i represents the i -th image of the input, F_i is the truth map of the density map corresponding to X_i , $F(X_i, \beta)$ represents the normalized density map.

Figure 5 Multi-granularity distributed data fusion structure diagram.

3.2 Crowd density from a global viewing angle

Due to the global video comes from the camera equipment installed at a high altitude, the individual's body size in the global viewing angle is very small, and even the individuals with large motion range appear point or columnar in the video. The problem of crowd density also pays more attention to group behavior, and no longer focuses on the characteristics of people themselves. This means that we can approximate an individual as a number of pixels, the individual's motion can be approximated as the change of position of the pixels, and the density of the crowd can be approximated as the distance between pixels.

The characteristic pixel points of the crowd can be obtained by extracting the foreground image of the global video. The extraction process uses a mixed Gaussian model. Assuming that the mixed Gaussian model consists of K Gaussian models, its probability density function is as follows:

$$p(w) = \sum_{k=1}^K p(k)p(w|k) = \sum_{k=1}^K \pi_k N(w|\pi_k, \sum k) \quad (3)$$

where $p(w|k) = N(w|\pi_k, \sum k)$ is the probability density function of the k -th Gaussian model. That is, after selecting the k -th model, the probability that the model produces w . $p(k) = \pi_k$ is the weight of the k -th Gaussian model, that is, the prior probability of selecting the k -th model, and $\sum_{k=1}^K \pi_k = 1$. Then the

model result is morphologically denoised, and finally the foreground image is obtained, which is the set of feature pixels.

The next step is to perform density classification processing on the acquired feature pixels. Assume that there are Ld pixels in total, and (m_j, n_j) is the j -th pixel coordinate. We can calculate the distance between the remaining points and the j -th pixel point, $d = \sqrt{(m_j - m_l)^2 + (n_j - n_l)^2}$, where $j \neq l, l \in [1, Ld]$. If only pixels with a distance less than the pre-set value r are retained, the density at that point is $D = dd/\pi * r^2$, where dd is the number of pixels with a distance less than r .

3.3 Multi-granularity distributed data fusion

We have performed feature analysis on local and global videos separately. However, if we want to get the final global population, we also need to perform information fusion on the distributed results of these two parts. The specific framework is shown in Figure 5. The video data comes from the same cloud server, and the videos from different perspectives are in a synchronized state to ensure the convergence of the video.

The distributed information fusion process is as follows: Assume that there are three low-altitude camera devices in the crowd gathering place, denoted as A_1, A_2, A_3 , and one high-altitude camera device B_1 . At time t , all camera equipment videos are selected simultaneously. The low-altitude video is calculated

Figure 6 Gray-scaled training and validation sets.

Figure 7 The result picture after the original picture is input into the model.

by the MCNN, and the actual number is recorded as n_1 , n_2 and n_3 . For high-altitude video, the density of the crowd pixels in the high-altitude video is obtained through the GMM and density grading algorithms. Through the field calibration method, we can mark the position and shape of the low-altitude field of view in the high-altitude video, and then obtain the density distribution of the three local fields of view under the global field of view. Assuming that the three local density values are d_1 , d_2 , and d_3 , there is an ND proportionality coefficient that can be calculated using

$$\alpha_i = \frac{n_i}{d_i}, \quad (4)$$

where $i = 1, 2, 3$. Here it represents the mapping relationship between the number of people and the density of people in this perspective. Finally, by assigning the weights of the three ND proportional coefficients, we

can calculate the global population

$$G = \sum \delta * \alpha * D = \sum_{j=1}^L d \sum_{i=1}^3 \epsilon_i \alpha_i * D_j \quad (5)$$

where δ represents the weight coefficient, α is the ND scale factor, D is the pixel density value, i is the low-altitude view number, j is the j -th pixel, and Ld is the number of pixels, ϵ_1 is the weight coefficient of the first camera device, which is determined by the distance between the pixel point and the center points of the three perspectives,

$$\epsilon_1 = \frac{d_{j1}}{(d_{j1} + d_{j2} + d_{j3})}, \quad (6)$$

where d_{j1} represents the j -th The distance from the pixel to the center of the field of view A_1 , d_{j2} , d_{j3} is the same.

Figure 8 Model training and comparative results. The left figure is the of convergence trends MAE and MSE. The right figure is the numerical comparison of actual numbers and model estimated numbers.

Figure 9 Original video frame image and density map after density classification.

4 Results and Discussion

We have applied and verified our proposed global crowd calculation method in real scenes. The City Life Square, located next to Jinji Lake in Suzhou, Jiangsu Province, China, has a total area of 4,300 square meters. During the musical fountain show, the crowd gathered at a large scale, which once attracted the attention of relevant security departments. Our data comes from surveillance videos covering multiple angles of the square, including about 40 low viewing angle

videos and 2 high viewing angle videos. The video was collected on October 1, 2018, from 18 : 00 to 21 : 00.

We first marked more than 500 video frames containing clear human bodies for videos at low viewing angles. In order to ensure the richness of the training set, we cut each picture into 8 images of equal size and grayed it out, including 3500 as the training data set and 500 as the verification data set, as shown in Figure 6. We selected the trained model and take the video frames of low-altitude views as input, and get the result picture shown in Figure 7, which shows the

Table 1 Related parameters of multi-granular distributed fusion method.

Time Frame	n_1	n_2	n_3	d_1	d_2	d_3	α_1	α_2	α_3
20	26	11.7	43.94	69.40	30.93	115.08	0.374	0.378	0.381
21	23.58	13.97	36.44	57.88	50.25	118.07	0.407	0.278	0.308
22	20.56	9.74	42.55	37.73	33.95	118.37	0.545	0.287	0.359
23	28.74	7.32	28.47	23.60	31.10	88.01	1.218	0.235	0.323
24	31.26	7.98	28.66	55.86	24.38	43.47	0.560	0.327	0.660
25	25.81	8.93	34.34	48.21	24.24	52.5	~	~	~

denseness of the head. The convergence trend of MAE and MSE during training is shown in the left figure of Figure 8. We applied it to more than 500 low-altitude actual video frame pictures, and compared the actual number of true value people with the estimated number of people in the model, as shown in the right figure of Figure 8. We found that when the low altitude local field of vision is large, the number of people is large, then, the model has a large error. However, when the field of vision is small, it maintains a high accuracy. This is because when the local viewing angle reaches a certain height, the human characteristics are no longer obvious, which is why we have always mentioned that the overall crowd estimation problem needs the fusion of local viewing angle and global viewing angle. Therefore, in practical applications, it should be ensured that people in the field of vision at a local viewing angle maintain clear human characteristics, that is, the installation height of the camera equipment at a local viewing angle should be within a reasonable range.

We select a high-altitude camera device, numbered B_1 , as the research object of global viewing angle, and select 3 near-ground cameras as local viewing angle, which are numbered A_1 , A_2 , A_3 , and the unified video

segment time is 20 : 30 ~ 21 : 00. The Gaussian mixed model is used to extract the foreground image of the global image, and the density distribution of the moving pixel points of the crowd is obtained through density classification, as shown in Figure 9. The video frame images at 20 : 30 and 20 : 58 are selected, which are the time when the crowd starts to evacuate and the time when the evacuation is about to end. After eliminating the environmental interference, the density grading processing is carried out. It can be seen that the density change is very obvious, which is also in line with the actual situation. Calculate the number of people in three low-altitude video frames and their positions in the global viewing angle, as shown in Figure 10. The red area in the picture in the last column of the figure is the corresponding location marked.

We select a time period of 20 : 53 : 20 ~ 20 : 53 : 24 video segment, and use the multi-granular distributed fusion method to calculate the mapping relationship between the actual number of people at low altitude and the density at high altitude. The relevant parameters are shown in Table 1. n_1 , n_2 , and n_3 are the actual numbers of people at the three angles, and d_1 , d_2 , and d_3 are the total density values of the three

Figure 11 Multi-granularity distributed fusion results. (a) Comparison of actual number of people and estimated number of fusions of A_1 viewing angle; (b) Comparison of actual number of people and estimated number of fusions of A_2 viewing angle; (c) Comparison of actual number of people and estimated number of fusions of A_3 viewing angle; (d) Global crowd change trends.

angles at the global viewing. We use the average of the 5s ND scale factor to predict the actual number of people in the 6-th second. That is, at 25 seconds, the estimated number of people in the A_1 viewing angle is: $(0.374+0.407+0.545+1.218+0.560)/5 \times 48.21 = 29.93$, and the actual number at 25 seconds is 25.81. The calculation method of the other viewing angles is the same. We test multiple times and performed statistical analysis, as shown in Figure 11. The changes in the number of people from the three angles from 20 : 33 to 20 : 53 are consistent, the estimated number and the actual number also maintain a high degree of matching. This verifies that our multi-granular distributed fusion method has high accuracy and practicability.

Finally, we estimate the number of people in the global viewing angle based on the three-part weights determined by the ND scale factor and pixel distance after the fusion, and count the total number of people every 2 minutes from 20 : 33 to 20 : 53, as shown in Figure 11 (d). Since this is the time after the fountain show ended, the crowd continue to evacuate, showing a downward trend, basically in line with the actual situation of the crowd in the video.

5 Conclusions

This article discusses the social focus of crowds safety with potential hazards, introduces the application and

techniques of crowd density detection in various fields, and summarizes crowd density detection methods from the perspective of machine vision. It also addresses the problem of crowd detection from an aerial perspective, and summarizes current multi-view fusion methods. Based on these, this article proposes a cross-source multi-view multi-granularity distributed fusion crowd counting method, which calculates the actual number of low-altitude and high-altitude crowds in a distributed manner, fuses low-altitude and high-altitude crowd information to complement and share their knowledge, and completes the accurate estimation of the global population. Finally, our method was successfully applied at the City Life Square in Suzhou City. Through the acquisition of video images with high-precision camera equipment, we analyze the number of people and the flow of the square when the crowd is evacuated, which provides rich data support for security control on crowd gathering places.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the scientific research project funding of Suzhou Rail Transit Group Co., Ltd. under Grand SZZG06YJ6000017.

Author's contributions

These authors contributed equally to this work.

Acknowledgements

We would like to express our sincere gratitude to the editor and anonymous reviewers for their valuable comments, which have greatly improved this article.

Author details

¹School of Rail Transportation, Soochow University, Suzhou, China.

²Command Center of Public Security Branch of Suzhou Industrial Park, Suzhou, China.

References

1. Soomaroo, L., Murray, V.: Disasters at mass gatherings: lessons from history. *PLoS currents* **4** (2012)
2. Solmaz, G., Wu, F.-J., Cirillo, F., Kovacs, E., Santana, J.R., Sánchez, L., Sotres, P., Munoz, L.: Toward understanding crowd mobility in smart cities through the internet of things. *IEEE Communications Magazine* **57**(4), 40–46 (2019)
3. Shao, J., Kang, K., Change Loy, C., Wang, X.: Deeply learned attributes for crowded scene understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4657–4666 (2015)
4. Sindagi, V.A., Patel, V.M.: A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters* **107**, 3–16 (2018)
5. Zhou, B., Tang, X., Wang, X.: Learning collective crowd behaviors with dynamic pedestrian-agents. *International Journal of Computer Vision* **111**(1), 50–68 (2015)
6. Chow, T.W., Yam, J.-F., Cho, S.-Y.: Fast training algorithm for feedforward neural networks: application to crowd estimation at underground stations. *Artificial Intelligence in Engineering* **13**(3), 301–307 (1999)
7. Cho, S.-Y., Chow, T.W.: A fast neural learning vision system for crowd estimation at underground stations platform. *Neural processing letters* **10**(2), 111–120 (1999)
8. Aziz, M.W., Naem, F., Alizai, M.H., Khan, K.B.: Automated solutions for crowd size estimation. *Social Science Computer Review* **36**(5), 610–631 (2018)
9. Ji, J., Lu, L., Jin, Z., Wei, S., Ni, L.: A cellular automata model for high-density crowd evacuation using triangle grids. *Physica A: Statistical Mechanics and its Applications* **509**, 1034–1045 (2018)
10. Ben, X., Huang, X., Zhuang, Z., Yan, R., Xu, S.: Agent-based approach for crowded pedestrian evacuation simulation. *IET Intelligent Transport Systems* **7**(1), 55–67 (2013)
11. Wang, Q., Gao, J., Lin, W., Li, X.: Nwpu-crowd: A large-scale benchmark for crowd counting. *arXiv preprint arXiv:2001.03360* (2020)
12. 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* **41**, 103–114 (2015)
13. Yogameena, B., Nagananthini, C.: Computer vision based crowd disaster avoidance system: A survey. *International journal of disaster risk reduction* **22**, 95–129 (2017)
14. 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, and Cybernetics-Part A: Systems and Humans* **31**(6), 645–654 (2001)
15. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. *IEEE transactions on pattern analysis and machine intelligence* **30**(7), 1198–1211 (2008)
16. Sidla, O., Lypetsky, Y., Brandle, N., Seer, S.: Pedestrian detection and tracking for counting applications in crowded situations. In: *2006 IEEE International Conference on Video and Signal Based Surveillance*, pp. 70–70 (2006). IEEE
17. Liang, R., Zhu, Y., Wang, H.: Counting crowd flow based on feature points. *Neurocomputing* **133**, 377–384 (2014)
18. Marana, A.N., Costa, L.D.F., Lotufo, R., Velastin, S.A.: Estimating crowd density with minkowski fractal dimension. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 6, pp. 3521–3524 (1999). IEEE
19. Kok, V.J., Chan, C.S.: Granular-based dense crowd density estimation. *Multimedia Tools and Applications* **77**(15), 20227–20246 (2018)
20. Fradi, H., Dugelay, J.-L.: Towards crowd density-aware video surveillance applications. *Information Fusion* **24**, 3–15 (2015)
21. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597 (2016)
22. Zhang, Y., Zhou, C., Chang, F., Kot, A.C.: Multi-resolution attention convolutional neural network for crowd counting. *Neurocomputing* **329**, 144–152 (2019)
23. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1861–1870 (2017)
24. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100 (2018)
25. Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C.: Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence* **43**, 81–88 (2015)
26. Hu, Y., Chang, H., Nian, F., Wang, Y., Li, T.: Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation* **38**, 530–539 (2016)
27. Jiang, H., Jin, W.: Effective use of convolutional neural networks and diverse deep supervision for better crowd counting. *Applied Intelligence* **49**(7), 2415–2433 (2019)
28. Zhang, Z., Wang, M., Geng, X.: Crowd counting in public video surveillance by label distribution learning. *Neurocomputing* **166**, 151–163 (2015)
29. Miao, Y., Han, J., Gao, Y., Zhang, B.: St-cnn: Spatial-temporal convolutional neural network for crowd counting in videos. *Pattern Recognition Letters* **125**, 113–118 (2019)
30. Hongquan, S., Xuejun, L., Guonian, L., Xingguo, Z., Feng, W.: Video scene invariant crowd density estimation using geographic information systems. *China Communications* **11**(11), 80–89 (2014)
31. Xu, M., Ren, J., Chen, D., Smith, J., Wang, G.: Real-time detection via homography mapping of foreground polygons from multiple cameras. In: *2011 18th IEEE International Conference on Image Processing*, pp. 3593–3596 (2011). IEEE
32. Cai, Q., Aggarwal, J.K.: Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In: *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pp. 356–362 (1998). IEEE
33. Mehrupeoglu, M., Rojas, D., McLauchlan, L.: Object tracking using multiple camera video streams. In: *Real-Time Image and Video Processing 2010*, vol. 7724, p. 77240 (2010). International Society for Optics and Photonics
34. Kang, J., Cohen, I., Medioni, G.: Continuous tracking within and across camera streams. In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1, p. (2003). IEEE
35. Xu, M., Orwell, J., Lowey, L., Thirde, D.: Architecture and algorithms for tracking football players with multiple cameras. *IEE Proceedings-Vision, Image and Signal Processing* **152**(2), 232–241 (2005)
36. Lijun, C., Kaiqi, H.: Video-based crowd density estimation and prediction system for wide-area surveillance. *China Communications* **10**(5), 79–88 (2013)
37. Xu, M., Ge, Z., Jiang, X., Cui, G., Lv, P., Zhou, B., Xu, C.: Depth information guided crowd counting for complex crowd scenes. *Pattern Recognition Letters* **125**, 563–569 (2019)
38. Mousse, M.A., Motamed, C., Ezin, E.C.: People counting via multiple views using a fast information fusion approach. *Multimedia Tools and Applications* **76**(5), 6801–6819 (2017)

Figures

Figure 1

Challenges in crowd density estimation. Images are from NWPU-Crowd datasets [11].

Figure 2

Comparison of crowd situation in local viewing angle and global viewing angle. Images are from NWPU-Crowd datasets [11] and Shanghaitech dataset [21].

Figure 3

Spatial relationship between local viewing angle and global viewing angle. The blue area is a high-altitude global field of view and fully covers the crowd space, and the red area is a low-altitude local field of view that is still within the global field of vision.

Figure 4

The structure of multi-column convolutional neural network model (MCNN).

Figure 5

Multi-granularity distributed data fusion structure diagram.

Figure 6

Gray-scaled training and validation sets.

Figure 7

The result picture after the original picture is input into the model.

Figure 8

Model training and comparative results. The left figure is the of convergence trends MAE and MSE. The right figure is the numerical comparison of actual numbers and model estimated numbers.

Figure 9

Original video frame image and density map after density classification.

Figure 10

Correspondence of low viewing angle in global viewing angle.

Figure 11

Multi-granularity distributed fusion results. (a) Comparison of actual number of people and estimated number of fusions of A1 viewing angle; (b) Comparison of actual number of people and estimated number of fusions of A2 viewing angle; (c) Comparison of actual number of people and estimated number of fusions of A3 viewing angle; (d) Global crowd change trends.