Location and Fusion Algorithm of High-Rise Building Rescue Drill Scene Based on Binocular Vision

Jia Ma and Zhiguo Shi(✉)

University of Science and Technology Beijing, Beijing 100083, China
szg@ustb.edu.cn

Abstract. In the emergency rescue exercise of high-rise buildings, mastering the accurate position of the participants is an important means for coaches to arrange tactics, evaluate the efficiency of rescue aid, evaluate the effect and ensure the safety of the participants. Video location is a more accurate positioning method, using personnel detection, personnel tracking can lock the position of personnel in the monitoring, but once occlusive, personnel can not be detected, will cause the loss of personnel identity information, another problem is that the current technology is difficult to stably identify the identity of personnel through signs. Therefore, this paper studies the fusion algorithm based on the characteristics that the most widely used WiFi fingerprint location can provide rough position information and personnel identity information. The detection with identity information is obtained by matching the personnel information provided by the WiFi fingerprint location system with the detected personnel in the video. At the same time, the location result of WiFi fingerprint can provide reference position when occlusive for a long time. Aiming at the characteristics of fixed number of participants and fixed identity information in emergency rescue exercise, this paper proposes a personnel tracking algorithm based on appearance and motion characteristics. This algorithm reduces the incidence of identity exchange problem when the personnel are very close, and records the representation information of the participants for a long time, which can make the personnel can be rerecognized after a long period of disappearance, and avoid the problem of matching error caused by multiple matching of WiFi fingerprint information and video location information.

Keywords: Location · Binocular vision · Person detection · Tracking algorithm

1 Introduction

In the emergency rescue, the rescue of high-rise buildings is the most typical. In recent years, with the increase of high-rise buildings, high-rise building accidents bring more and more challenges to rescue, especially high-rise building fires, high-rise building post-earthquake rescue. In the environment of such emergency rescue drill, the high-precision position information of the personnel is required to be used as the reference

H. Ning (Ed.): CyberDI 2019/CyberLife 2019, CCIS 1137, pp. 450–469, 2019.
https://doi.org/10.1007/978-981-15-1922-2_32

basis for the tactical arrangement, the evaluation of the drill effect, and the risk assessment.

At present, multi-use wireless communication mode, including RFID, WiFi Fingerprint, Bluetooth, Ultra-Wideband, etc. [2–4], can be used to meet the indoor positioning needs of most public places. Ultra-Wideband and Bluetooth are the most accurate methods, while RFID and WiFi Fingerprints [5] are the second. However, Ultra-Wideband positioning needs expensive devices, Bluetooth needs to replace Bluetooth label batteries regularly, and the maintenance cost is very high. WiFi, as the most commonly used wireless signal in daily life, has wide coverage and cheap equipment layout, but if used as an emergency rescue exercise, it is not accurate enough.

With the development of artificial intelligence technology, target recognition for image has been widely used. Image recognition technology can be used to calibrate the position of people in the graph directly, so as to carry out accurate positioning. However, image recognition is difficult to accurately identify people in a wide range of areas, so the calibrated personnel position can not correspond to the actual personnel identity. In this paper, a fixed scene target location method based on multi-view video image is proposed. Later, the research status, the transformation from image coordinates to real coordinates, how to obtain the identity of the detected target, and how to achieve stable video location will be carried out.

2 Related Work

2.1 Indoor Positioning Theory Based on WiFi

WiFi is one of the most common wireless signals in daily life. It generally follows IEEE 802.11b/g/n protocol and works at a frequency of 2.4 GHz. Each WiFi signal is sent by a wireless ap (access point), often a wireless router. Each wireless ap has its own unique global code, that is, mac address, and generally these ap do not move frequently. WiFi fingerprint location is a convenient and accurate indoor location method.

Using WiFi fingerprinting method to locate is divided into two stages, the first stage is the collection of WiFi fingerprints in a certain area. The so-called WiFi fingerprint refers to a set of key value pairs composed of a group of RSSI from AP and the corresponding real coordinates collected by the terminal at a certain point. This associates the signal strength with the location. The collected fingerprint information is then stored in the database, and each set of fingerprints is unique. In general, the points we collect are not random, and the coordinate system is built with a certain point as the origin, so that the coordinates of each point can be obtained. With the increase of the density of acquisition points, the accuracy of positioning will also increase accordingly, but there will also be certain upper limit. When RSSI near several points are very similar or even overlapping, the increase of acquisition points will not increase the accuracy.

In the process of collecting RSSI, the signal strength of each ap is time-varying, but the whole fluctuates around a range, and we need to collect and average it many times at one point as the estimated value of RSSI at a certain point.

The second phase is the positioning phase. Locate the terminal to a certain location in the area. The RSSI, of ap collected by the terminal is compared with the WiFi fingerprint in the database at the same time to find the closest several fingerprints. Here are two ways, one is the minimum distance, including the Euclidean distance, the Manhattan distance, the Mahalanobis distance, or the maximum similarity, including the cosine similarity, the Spearman similarity, and so on. The most commonly used cosine similarity is used in this paper.

2.2 Pedestrian Detection Method

At present, pedestrian detection methods are mainly divided into three categories, one is based on motion detection, the second is mainly based on machine learning, and the third is deep learning. Among them, based on motion detection such as Gaussian mixture model, frame difference method, vibe algorithm [6, 7] and so on, the idea of these background modeling algorithms is to get a background model through the previous frame learning, and then compare the current frame with the background model to get the moving target. These algorithms are simple to implement and fast to implement, but these algorithms only use pixel-level information and do not make use of the high-level semantics of the image, so the following problems exist: they can only detect moving targets; they are greatly affected by light; if multiple target adhesions can not be dealt with; they are vulnerable to bad weather and so on. The machine learning algorithm and the depth learning algorithm improve the above problems from the advanced semantics of the image.

Navneet Dalal proposed a pedestrian detection algorithm based on hog SVM on 2005 CVPR [8]. HOG (directional gradient histogram) feature is a feature operator used for object detection in computer vision and image processing. HOG uses the gradient histogram of the local region to form the feature by calculating and counting the gradient histogram of the local region, which makes use of the orientation and intensity information of the edge. The method of HOG is to calculate the gradient of fixed size picture, then divide the picture into grid points, then calculate the gradient orientation and intensity of each grid point, then form the gradient direction distribution histogram of all pixels in the grid, and finally summarize the whole histogram feature. This feature describes the shape and appearance information of pedestrians, and is insensitive to light changes and small spatial translation.

In view of the fact that HOG features only focus on edge and shape information, and it is difficult to deal with occlusion and sensitive forehead problems, some researchers have proposed the integral channel feature (ICF) [9]. The integral channel features include 10 channels: gradient histogram in 6 directions, 3 luv color channels, and a gradient amplitude. By combining ICF with AdaBoost, the author carries out cascade classification training. Instead of zooming the picture to a fixed size, he designed several common scale classifiers. For pedestrians of other sizes, the prediction results of typical classifiers were used to approximate the difference, and the accuracy was higher than that of direct image scaling.

In order to solve the problem of occlusion, a method (DMP) [10] for the detection of parts is proposed, and the human body is divided into the parts such as the head, the trunk, the limbs and other components. These parts are detected respectively, and the detected results are combined. DMP includes two parts: root model and component model. The root model (Root-Filter) is mainly aimed at the potential region of the object to obtain the position of the possible object, but whether there is really the desired object needs to be further confirmed after the calculation combined with the component model (Part-Filter). In addition, DMP algorithm also uses Latent-SVM classifiers with strong discrimination ability, which makes it achieve good results in human body detection.

Methods such as Faster-RCNN, SSD, Yolo, FPN [11, 12], etc., in the field of depth learning can be used to detect pedestrians, whose accuracy is significantly higher than the SVM and Adaboost classifier. But because the scene illumination in the training set is monotonous, the target in the figure is relatively sparse, and the problem of occlusion and lighting in the pedestrian detection can not be well processed. The Liliang zhang's team improved [13] the Faster-RCNN, only reserved the RPN network for candidate area extraction, and changed the classification network into a random forest, which improved the problem that the CNN network is too sparse for small target extraction features.In addition DMP method is also used, so they get a good effect.

The Institute of Artificial Intelligence of the Origin of the United Arab Emirates (UAE) proposed a detection idea without anchor frame [14], which directly convolution the picture without sliding window to predict the center point and scale of the target. They achieved a very good results.

2.3 Multi-target Tracking Based on Personnel Detection

Personnel tracking algorithm is an effective means to improve the efficiency of personnel detection and reduce the false detection rate in video. In this paper, multi-target tracking is mainly studied, occlusion is still one of the difficulties to be solved in this field. At present, the method of deep learning has gradually surpassed the probability method and machine learning method in this field, and has become the mainstream of research.

Bergmann et al. proposed to convert the target detector into a tracker, and use rerecognition and motion prediction to complete the tracking task. In this method, the boundary box regression of the object detector is used to predict the new position of the object in the next frame, and the new position of the object in the next frame is extended by simple re-recognition and camera motion compensation [15]. Due to the limited effect of pedestrian recognition in scenes with a large number of people, the tracking effect of this model in more complex scenes is poor.

On the basis of sort algorithm, Nicolai et al. proposed that Deep Sort, applies the idea of Cascade Matching to the matching of multi-target tracking, which effectively reduces the probability of target identity switching when occlusion occurs [16]. Although this method also uses the advanced semantics of the image and the motion information of the target, the image feature extraction network is too simple, resulting in the extracted features sometimes can not be used to determine whether the detection

target is the object you want to track. This paper will improve this method based on the existing scenarios.

SenseTime Technology proposes a multi-target tracking framework which can capture long-term and short-term clues. The switching perception separator in data association is used to improve the robustness of identity switching matching in multi-target tracking. At the same time, a simple but effective method is introduced to retrieve potential classifiers [17].

In addition, Milan et al. proposed a novel multi-class multi-target tracking (MCMOT) framework, which combines detection response and variable point detection (CPD) algorithm to carry out infinite multi-target tracking. The effect of this framework is better than that of the most advanced video tracking technology [18]. Lee uses CNN-based target detector and KLT (Lucas-Kanede Tracker)-based motion detector to calculate the likelihood probability of the foreground as the detection response of different categories of targets [19].

2.4 Projection of Image Coordinates to World Coordinates

In order to locate the target in the image, it is necessary to convert the image coordinates of pedestrians in the image to the real coordinates. In this paper, the stereo matching algorithm of binocular camera is used to solve this problem.

In general, the binocular camera is a wide-angle lens, the imaging is distorted, and the imaging surface of the two cameras may not be coplanar, which causes interference to the subsequent stereo matching. The camera needs to be calibrated before the stereo matching algorithm is carried out. The calibration is divided into two parts, namely the calibration of the single camera and the calibration of the double camera [20].

In this paper, Zhang Zhengyou calibration algorithm is used to calibrate a single camera: multiple groups of chessboard lattice maps are taken, corner detection and sub-pixel information extraction are carried out by OpenCV library function. Using this information, the internal parameter matrix M and distortion matrix J of the camera, as well as the rotation matrix R and the shift matrix T of each image can be obtained. Through multiple iterations, more accurate M and J can be obtained, and they can be input into the corresponding camera correction function as parameters, and the calibration of the single camera can be completed.

Binocular calibration is based on the calibration of monocular camera. In addition to obtaining the internal parameter matrix and distortion coefficient matrix, the additional parameters that need to be calibrated are eigenmatrix E, basic matrix F, rotation matrix R and shift matrix T. The R and T of the binocular camera can be calculated by the following formula:

$$\begin{cases} R = R_r R_l^T \\ T = T_r - R_r R_l^T T_l \end{cases} \tag{1}$$

R_r and R_l are the rotation matrices of the right camera and the left camera, respectively, and T_r and T_l are the translation matrices of the right camera and the left camera, respectively.

The intrinsic matrix E and the basic matrix F of the corresponding binocular camera can be obtained by bringing the inner parameter matrix M and the distortion matrix J of the single camera and the whole rotation matrix R and the translation matrix R into the library function, and then the E and F are brought into the library function to realize the calibration of the binocular camera.

There is the following relationship between image coordinates and world coordinates:

$$Z_w = \frac{bf_x}{d} \tag{2}$$

$$X_w = \frac{b(u - u_0)}{d} \tag{3}$$

$$Y_w = \frac{b(v - v_0)}{d} \tag{4}$$

Where $(Z_w,\ X_w,\ Y_w)$ represents the world coordinates of a certain point, (u, v) represents its image coordinates. b represents the distance between the two cameras' center of light, and d represents parallax. Among them, b can be obtained by measurement, and the parallax d can be obtained by *SGBM* algorithm, so that the image coordinates of people in video can be transformed into the real world coordinates.

3 Rescue Scene Location Algorithm Based on WiFi Location and Video Image Location

3.1 The Overall Framework of Video Image Location Algorithm

The video image, as the carrier of the target, carries the identity information and the position information of the target, in particular the position information which has a considerable accuracy. However, when there are many people in the image, the method of target recognition cannot accurately distinguish the different people's identity, and the simple position information is not worth. So that We use the combination of WiFi fingerprint and video image location for fusion localization. The WiFi fingerprint positioning devices are easy to be arranged, and can provide the characteristics of the double information of the person's identity and position, but the accuracy of the location information provided is not high. If we combine the WiFi fingerprint positioning with the video image location, give full play to the image positioning accuracy and the characteristic that the WiFi fingerprint positioning has the identity information, the indoor positioning accuracy of the two technologies will be further improved [21, 22].

The following steps are required for the location of people in a video image:

(1) Detection: obtain size and position of all personnel in a frame, which is represented by a box;
(2) Obtaining the pixel coordinates which around feet of each identified pedestrian;
(3) Convert the coordinates of pixels in the image to the coordinates in the real world.

After these three steps, you can get the actual location of everyone in the video (Fig. 1).

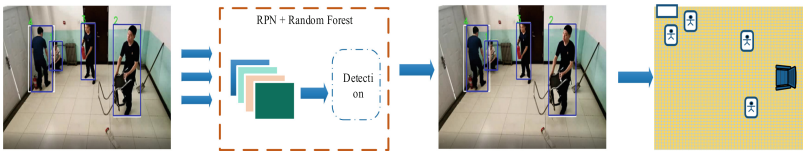


Fig. 1. Flow chart of video image location

While identifying and obtaining pedestrian location, a timestamp is recorded, and the positioning results of WiFi fingerprints for indoor personnel are obtained from the database. The video location results and WiFi location results are matched one by one according to the similarity of the them, so that the pedestrians in the video can obtain the identity information in the WiFi fingerprint location results.

In the practical application, due to the existence of the reasons such as occlusion and light change, the detection link of the pedestrian often has missed detection and false detection, resulting in the failure of the association between the video positioning information and the WiFi fingerprint positioning information. To this end, the above problems will be improved by using the target tracking algorithm (Fig. 2).

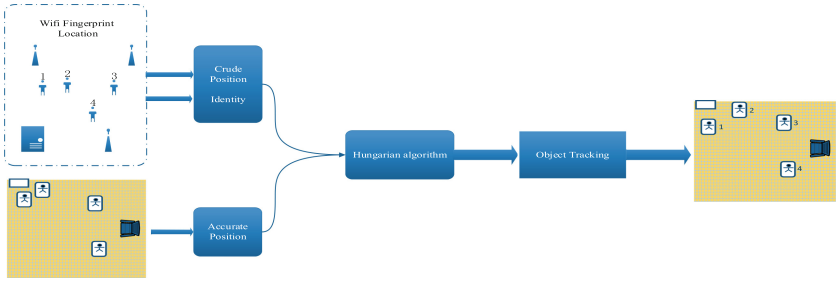


Fig. 2. Block diagram of WiFi fingerprint and video stream fusion localization algorithm

3.2 Fusion of Multi-target Position Information

The position coordinates of the target in reality can be obtained by using WiFi fingerprint, and the coordinates of the target in reality can be obtained by video image. However, the target obtained by image is only location information, there is no identity information, the information obtained by WiFi has both location information and identity information, but the location information is not accurate. If the two can be correlated with each other, a fusion positioning system with both identity information and accurate location information can be generated.

For the multi-objective association problem of matching multi-person WiFi information and image information in a fixed scene, it can be regarded as an assignment problem, that is, assuming that there are m tasks, m personnel, each person gets a task, and solve the problem of minimizing. In this problem, WiFi produces n personnel information, and the image recognizes the corresponding n personnel information, but in practical application, sometimes the image will produce missed detection or false detection, resulting in the resulting personnel information greater than or less than n, which is beyond the scope of the traditional assignment problem.

The Hungarian method is a very good method for the traditional assignment problem. Its basic idea is to change the original value coefficient matrix into a new value matrix with many 0 elements through certain operations, while maintaining the optimal solution of the original problem. In this paper, we use the extended Hungarian algorithm to increase the number of virtual elements to supplement the number of people when the information generated by the video image is not enough to solve the problem that the information generated by the video image may be different from that of the WiFi information.

In the video image detection, it is assumed that n detection results are obtained, and it is recorded as follows:

$$IP_i = x_i, y_i; i = 1, 2 \ldots n \tag{5}$$

At the same time, WiFi also located m test results, which are as follows:

$$\mathrm{RP}_j = x_j, \mathrm{y}_j, \mathrm{z}_j; j = 1, 2 \ldots m \tag{6}$$

The matrix P is constructed, and its dimension is $n \times m$, matrix represents the Euclidean distance between the WiFi detection results and the video image detection results.

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & pnm \end{pmatrix} \tag{7}$$

Among them, the calculation formula of each element is obtained by the following formula:

$$p_{ij} = IP_i - RP_j \tag{8}$$

In the actual calculation, use the matrix of L_{d*d}, $d = max(n, m)$. For elements with dimensions greater than n rows and m columns in L are set to 0, and the other elements are the same as in matrix P. After the coefficient matrix is constructed, it can be solved according to the Hungarian algorithm. The results of the solution can be divided into the following cases:

For each pair that matches successfully $(IP_i,\ RP_j)$, the result of IP_i is used as the final location result, and the identity information carried by RP_j is used as the final fusion identity information.

For the matching result, RP_j has no matching object, that is, n < m, then the positioning result of RP_j is used as the positioning result of the target.

If *IP* does not have a matching object in the matching result, it is considered to be an image recognition error, because WiFi location contains all the personnel information, and more personnel information appears, it is an error. Ignore *IP* that does not match.

After this calculation, we can get m location results with identity:

$$P_{\mathrm{j}} = \begin{cases} (x_i, y_i, z_j) & \text{If } IP_i \text{ matches } RP_j \text{ successfully} \\ (x_j, y_j, z_j) & \text{If } IP_i \text{ matches } RP_j \text{ unsuccessfully} \end{cases} j = 1, 2..m \tag{9}$$

Through the above steps, the position information of all the people in one frame is obtained, and the continuous target movement information can be obtained by frame detection. However, due to the inaccurate location of WiFi and the instability of video detection, the matching error occurs, which affects the stability of matching.

3.3 Information Fusion Location Algorithm Based on Target Tracking

In the process of video detection, due to the existence of occlusion, people out of the camera field of view and other problems, resulting in unstable video pedestrian recognition, which leads to unstable matching between video and WiFi positioning, this paper proposes a combination of target tracking and WiFi positioning algorithm to improve this problem.

The tracking algorithm is divided into the following phases and Fig. 3 shows the flow chart:

(1) Creation and trajectory prediction for target.
(2) Matching between detection results and tracking targets.
(3) using cascade matching to solve the problem of target identity exchange when occlusion occurs.
(4) IOU matching is used again for detection objects and tracking targets that do not match successfully after cascade matching.

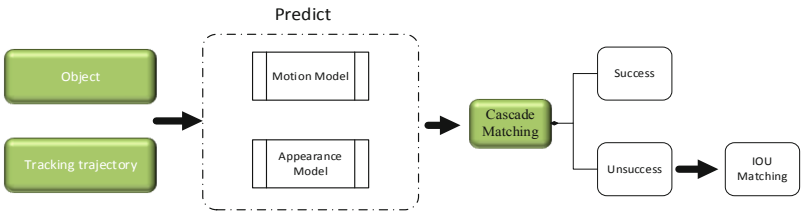


Fig. 3. The algorithm flow chart of target tracking.

First of all, introduce the measurement method of association. We detect a frame in the video, and the detected target is selected by using the anchor box (bounding box), and the format of the description anchor box is $(\mathrm{u}, \mathrm{v}, \mathrm{r}, \mathrm{h}, \dot{x}, \dot{y}, \dot{r}, \dot{h})$. $(\mathrm{u}, \mathrm{v}, \mathrm{r}, \mathrm{h})$ indicates the position and size of the anchor frame, $(\dot{x}, \dot{y}, \dot{r}, \dot{h})$ is the coordinate of the center of the anchor frame, $\dot{r}$ is the ratio of the length to width of the anchor frame, and $\dot{h}$ is the left length of the anchor frame. $(\dot{x}, \dot{y}, \dot{r}, \dot{h})$ represents the velocity information of each variable in $(\mathrm{u}, \mathrm{v}, \mathrm{r}, \mathrm{h})$ described in the image. We first use Kalman filter algorithm to predict the position of the detected anchor frame, where the pedestrian motion model is assumed to be uniform motion and the observation model is a linear model. The following experiments prove the rationality of the hypothesis. Here, the prediction results are recorded as $d_j = (\mathrm{u}, \mathrm{v}, \mathrm{r}, \mathrm{h})$. Once the target is lost for a long time, this paper sets to more than 20 frames (the setting of this parameter is related to the number of frames taken by the camera and the movement speed of the actual scene), and the WiFi fingerprint system shows that the missing tracker is still in this room, then the positioning result of WiFi fingerprint is used as the basis of the prediction.

The newly detected target is recorded as $d_i = (\mathrm{u}, \mathrm{v}, \mathrm{r}, \mathrm{h})$, and we want the newly detected target to match the predicted result based on the last detection in order to give the newly detected target identity information. The traditional methods to measure the two objectives are Euclidean distance, Pap distance, cosine similarity and Mahalanobis distance. Because of the perspective distortion in the image, the size of the object will change with the far and near angle. Therefore, the metric size of the anchor frame at different distances from the camera is different, and the influence of different metric scales can be effectively eliminated by using the Mahalanobis distance.

$$d_{i,j}^{1} = \left(d_j - d_i\right)^T S^{-1} \left(d_j - d_i\right) i = 1, 2 \ldots n,\ j = 1, 2 \ldots m \tag{10}$$

Where d_j represents the No. j tracking target determined after the last detection, d_i represents the No. i detection target of the current frame, and $d_{i,j}^{1}$ represents the Mahalanobis distance between the anchor frame of the previous frame detected target and the current frame detection target.

In addition to the motion model-based metrics, the appearance information contained in the image is also measured. The new Dense block structure is used for the extraction of the appearance model. The structure has the characteristics of less parameters and strong expression. Here the last classification layer is replaced by a convolution layer of $1 * 1$ instead of the full connection, so that the feature vector of the target is obtained. The network structure is as shown in the following table (Table 1):

The network is pre-trained on the large pedestrian detection data set *Person Re − Identification* and contains a total of 1 million parameters. The network runs at 30 ms on NVIDIA GTX1080, meeting the speed requirements of real-time processing.

The appearance extraction network extracts the detected person and outputs a feature vector of a 49-dimension. We compare this feature vector with the feature

Table 1. CNN network structure used in appearance extraction

Name	Patch size/stride	Output size
Conv 1	$7 * 7/2$	112×112
Pooling	$3 * 3$ max pool/2	56×56
Dense block (1)	$\begin{pmatrix} 1 * 1\text{conv} \\ 3 * 3\text{conv} \end{pmatrix} \times 6$	56×56
Translation layer (1)	$1 * 1\text{conv}$ $2 * 2$ averagepool/2	56×56 28×28
Dense block (2)	$\begin{pmatrix} 1 * 1\text{conv} \\ 3 * 3\text{conv} \end{pmatrix} \times 12$	28×28
Translation layer (2)	$1 * 1\text{conv}$ $2 * 2$ averagepool/2	28×28 14×14
Dense block (3)	$\begin{pmatrix} 1 * 1\text{conv} \\ 3 * 3\text{conv} \end{pmatrix} \times 24$	$64 \times 32 \times 16$
Translation layer (3)	$1 * 1\text{conv}$ $2 * 2$ averagepool/2	14×14 7×7
Dense block (4)	$\begin{pmatrix} 1 * 1\text{conv} \\ 3 * 3\text{conv} \end{pmatrix} \times 16$	7×7
Conv 2	$1 * 1$ conv	49×1

vector corresponding to the tracking target in the previous frame, and then judge whether they are the same target. The measure method here adopts the cosine similarity, and the calculation formula is as follows:

$$d_{i,j}^2 = \min\left(1 - r_j^T r_k^{(i)} | r_k^{(i)} \in R_i\right) \tag{11}$$

Where r_j is the feature vector of the newly detected target, and $r_k^{(i)}$ is the set of k frame eigenvectors of the No. i tracking object in the past. R_i is a collection of all trace objects in the past.

$$R_i = \left\{ r_k^{(i)} \right\}_{K=1}^{L_K} \tag{12}$$

The matching result $d_{i,j}^1$ based on motion information is merged with the matching result $d_{i,j}^2$ based on appearance information, and a new matching result is obtained. The fusion expression is as follows:

$$c_{(i,j)} = \mu d_{i,j}^1 + (1 - \mu) d_{i,j}^2 \tag{13}$$

The following rules are used to detect whether an object becomes a tracking object:

(1) the appearance of n trackers is detected and recorded by multiple frames before the positioning begins.

(2) the target is not detected in 100 consecutive frames, and according to the WiFi information, it is determined that the target leaves the current monitoring area to stop matching the tracking target until the WiFi positioning information returns to the current scene.

In the process of target tracking, occlusion will inevitably occur, assuming that when a tracking target a is obscured by target b, it will not be detected by the detector. Because the predicted value of a and b is very close, it is very likely that the detection value of b will be matched with a, resulting in the matching exchange phenomenon (ID switch). For this reason, the cascade matching method is used to match here.

The pseudo code for cascade matching is as follows:

Cascade Matching

Enter: the serial number of the tracking target $T = \{1,2 \ldots . N\}$, The serial number of the detection target. $D = 1,2 \ldots M$

1. Calculation cost matrix $C = [c_{i,j}]$
2. Set up a set M to represent the matched tracking target and the detected target, and initialize it.
3. Create a collection u to indicate that there is no matching successful collection in the detection target and initialize it
4. For n in （1，$\ldots Age_{max}$）：
5. {
6. Select $T_n(T_n \in T, n \in Age)$according to the countdown of the order of disappearing tracking targets
7. $x_{i,j} \leftarrow \min\_cost\_matching\{C, T_n, U\}$
8. $M \leftarrow M \cup \{(i,j)|b_{i,j} \cdot x_{i,j} > 0\}$
9. $end\ For$
10. return M, U

Algorithm 1. The cascade matching algorithm

After cascade matching, we obtain tracking and detection targets M and U that have been matched successfully and not matched successfully. The U and $(T - T_n)$ are processed by sort mechanism standard. Calculate the IOU between them and using Hungarian algorithm to match. The following results were obtained:

$$P_{IOU} = \begin{cases} (T_i, D_j) \\ T_i \\ D_j \end{cases} \quad i,j = 1,2..N \tag{14}$$

Among them, P_{IOU} is the result of IOU matching, (T_i, D_j) indicates the successfully matched tracking target and detected target, T_i indicates that there are unmatched tracking targets, and D_j indicates that there are only detected targets in the matching results, most of which are due to false detection.

Because the number of tracking targets can be determined according to the WiFi fingerprint system, it belongs to constant, but the detected people may be occlusive and

disappear, so some tracking may not be able to match the corresponding detection targets after this step. For how to deal with this kind of target will be explained later.

Finally, the Kalman prediction matrix is updated to delete the tracking target that is not in this scene (according to the scene information provided by WiFi fingerprint system). It is expressed as follows:

$$T_{new} = M_T + P_T^{IOU} - WiFi_{state=0} \tag{15}$$

T_{new} represents updated tracking, M_T indicates the tracking of successful matching after cascade matching, P_T^{IOU} indicates the tracking of successfully matching after cascade matching, and $WiFi_{state=0}$ indicates that the personnel information of this scene does not exist in the WiFi fingerprint system.

In addition, we need to update ID (label) of the tracking target and then the tracking of one frame is completed. The beginning of the next frame is still the prediction of the existing target by Kalman filter.

The target tracking can lock the position of the target in the image stably for a period of time, and the position information of the personnel can be accurately determined by the transformation of image coordinates to world coordinates. Two information will be fused below.

Firstly, still using the target position information fusion algorithm proposed in 3.2 to match the detected target of video with the target detected by WiFi fingerprint. The matching results are as follows:

$$P_{matching} = \begin{cases} (x_i, y_i, z_i)\ If\ IP_i\ matches\ RP_j\ successfully \\ (x_i, y_i, z_i)\ If\ IP_i\ matches\ RP_j\ unsuccessfully \end{cases} \quad j = 1, 2..N$$

Where (x_j, y_j) represents the result of the video localization, (x_j, y_j) represents the result of the WiFi fingerprint positioning, and z_j is the person information carried by the WiFi fingerprint system. At this time, the target in the video has the identity information of the personnel in the WiFi fingerprint system. Next, go to the section that follows the location:

$$P_i^{result} = \begin{cases} P_{track} S_{t-1} = 1, S_t = 1; \\ P_{kalman} S_{t-3} = 1, S_t = 0; \\ P_{WiFi} S_{t-3} = 0, S_t = 0; \\ P_{track} S_{t-1} = 0, S_t = 1; \end{cases} \qquad i = 1, 2. \ldots .N \tag{16}$$

Among them, P_i^{result} is the result of tracking and positioning, P_{kalman} is the predicted position obtained by Kalman filter according to the previous frame positioning results, and P_{WiFi} is the result of WiFi fingerprint location. In the process of tracking and positioning, occlusion may occur, so that the target can not be detected. Here, S_{t-1} is used to represent the tracking state of target I at the previous time, S_{t-3} represents the tracking state of the first three frames, and S_t represents the tracking state of the current time.

That is to say, in the process of target tracking, if the tracking results are converted to the world coordinates, the accurate position of the target in the room can be easily

obtained. Once the tracking fails in a certain frame, the target information will not be obtained in the video, we will use the Kalman filter to predict the position based information of the previous frame, and the predicted position will be used as the detection result of the lost frame. Because the number of frames taken by the general video surveillance does not exceed 30 frames per second, and the position change of the normal operating personnel usually does not change obviously within 0.1 s, the predicted results are used as the positioning results within 3 frames. If the target is lost for a long time, the prediction results can no longer meet the positioning accuracy, and then switch to the WiFi fingerprint system, using the results of WiFi fingerprint location as the detection value of the current position of the target. Once the target reappears, a new round of identity matching will be carried out on the lost target, the actual identity of the target will be given to the target, and the process of positioning will continue to be followed.

In addition, the problem of identity information exchange is inevitable in the course of target tracking, so that a monitoring threshold is needed between the result of the video positioning of each frame and the positioning result of the WiFi fingerprint. Once the difference between the positioning distance between the two is greater than the threshold value, we will re-matching the target with the problem, and adopt the position of the WiFi fingerprint positioning as predict results before the matching is completed. The setting is as follows:

$$P_i^{result} = \begin{cases} P_i^{result} \left(P_i^{result} - p_i^{WiFi} \right) > gate \\ p_i^{WiFi} \left(P_i^{result} - p_i^{WiFi} \right) > gate \end{cases} i = 1, 2 \ldots N \tag{17}$$

The p_i^{WiFi} means the location information of the WiFi fingerprint location of the No. i person, left $\left(P_i^{result} - p_i^{WiFi} \right)$ representing the difference between the image location and the WiFi fingerprint, and the *gate* represents a threshold value that determines whether a identity needs to be rematched.

4 Experimental Verification

4.1 Experimental Method and Environment Configuration

The experimental system is divided into two parts, one is the location system of WiFi fingerprint: four routers are set up in the four corners of the room, the mobile phone is used as the location label to detect the intensity of WiFi signal, collecting information of signal strength with interval of 1 m, the collected data is stored in the database, and using the Hungarian algorithm to establish the WiFi fingerprint.

The other part is the image acquisition and processing system, which uses a fixed camera to collect the images of personnel in the area, and sends the images to the computer for pedestrian detection, tracking processing and matching processing. The camera adopts wide-angle lens, the computer is configured as i7-6600k, and the video card is GTX1080.

4.2 Experimental Analysis and Related Algorithms

According to the above algorithm flow, we first test the part of personnel detection and tracking, and the test data is a video shot in a fixed space. The main test content is that when the rescue exercise of high-rise building is carried out, the rescue staff produce the influence of various posture and occlusion on the performance of the algorithm. The main concern is the problem of identity exchange (id switch), and the other is the problem of the loss caused by the occlusion.

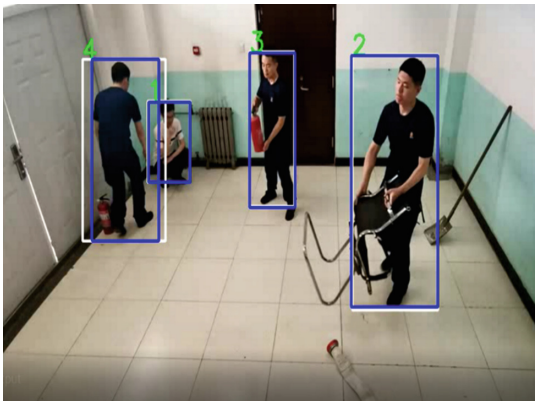


Fig. 4. Results of frame 30 processing

Figure 4 above shows the tracking results of frame 30, with four people in the image, using the label on the anchor box to distinguish the identity information of the four people, the target of No. 1 is the trapped person, and the rest is the rescue personnel. When the algorithm runs, it is good for the recognition of people of all kinds of posture.

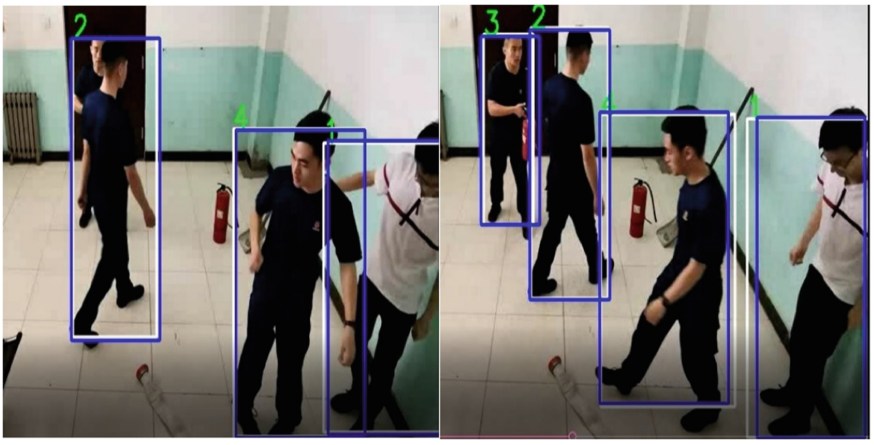
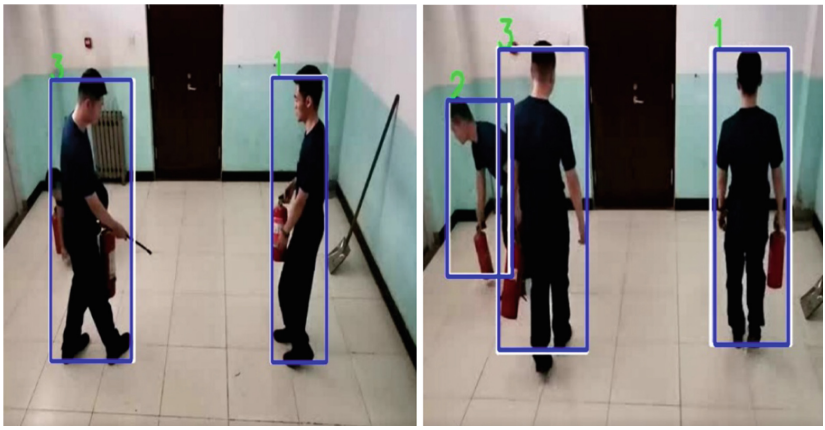


Fig. 5. Results of 90-frame and 95-frame processing when transient occlusion occurs

Figure 5(a) is the detection result of frame 90. It can be found that due to occlusion, the detection box can not select the whole pedestrian accurately. (b) is the result of frame 95 processing. After a brief 0.3 s reappearance of the tracking object, the missing information is quickly retraced.



(a) (b)

Fig. 6. A case where a long period of severe occlusion takes place

In order to verify the influence of long time occlusion on tracking, we also test the special scene. Ask three rescuers deliberately block each other in the scene. Figure 6(a) shows that person 2 is blocked by person 3, it can be seen that person 2 is almost completely disappeared and cannot be detected. About 3 s after frame 112, person 2 got up and was normally traced. About 3 s after frame 112, personnel 2 got up and was normally traced. It shows that the algorithm has certain ability to deal with the long-term loss of tracking objects in this scene.

In order to show the change of the position of the personnel clearly, the paper chooses the two-person position data to draw the graph. Figure 7 is a graph of the location of the WiFi fingerprint of the two people in the room. In the smaller indoor space, the accuracy of WiFi fingerprint location is not high, we can see that on the one hand, the path in the image is tortuous, on the other hand, the accumulation of sampling points will occur in the place where the inflection point is slow.

Figure 8 is the fusion location result. It can be seen that the curve is smooth, in accordance with the law of motion, and there is no jump phenomenon, which shows that the algorithm has a great improvement on the indoor location results of WiFi fingerprints.

4.3 Algorithm Availability and Performance Analysis

The fusion localization algorithm proposed in this paper is online mode, which is divided into four stages: personnel detection in video, fusion of personnel information

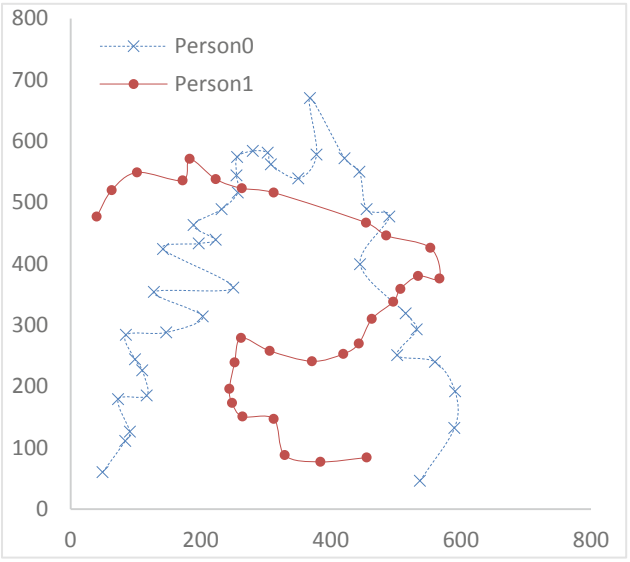


Fig. 7. Results of WiFi fingerprint location

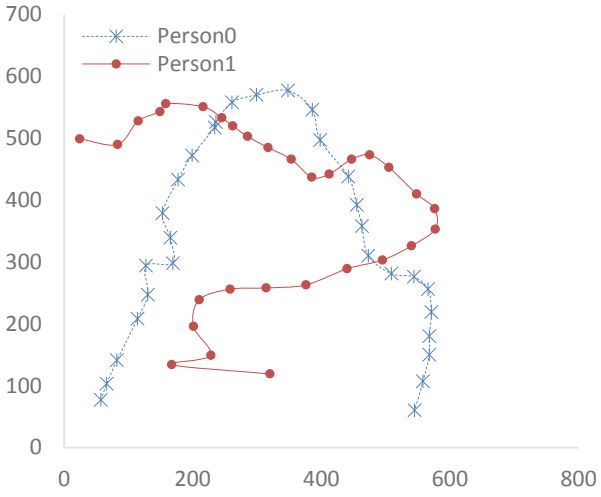


Fig. 8. Results of fusion positioning

and WiFi fingerprint information, video detection and tracking, and WiFi correction of lost targets. Under the experimental conditions, the comprehensive detection speed can reach 20 frames per second. Because the personnel move slowly indoors, in practice, the camera frame rate is adjusted to 15 frames per second, so that the detection speed is basically synchronized with the monitoring video speed. However, too high video frame rate will lead to the lag of detection results and can not achieve online positioning.

The positioning continuity has been analyzed in the above experiment. In addition to the random motion, the paper also carries out the specified route movement to study the accuracy of the positioning, and the following data is obtained through the experiment (Table 2):

Table 2. Error of regular route positioning

Target state	Unobstructed persons	Obstructed person
Average error (m)	0.16 + 10.0	0.25 + 15.0

It can be found that the positioning accuracy of occlusive personnel is much larger than that of unocclusive people, mainly because the positioning accuracy of WiFi fingerprint is lower, once the target is lost for a long time, WiFi fingerprint tracking will be switched, covering for a long time during testing, indirectly simulating the situation when the number of occlusive people is large, in addition, tens of pixels offset may occur when occlusive occurs, and the farther away from the camera. The more serious the deviation, the more serious the deviation.

In this paper, the performance of tracking algorithm in this scene is tested for a long time. The experimental data are 20 m^2 indoor, 5 segments of video moving around by 4 people, each video length of 1 min, video playback speed of 20 frames per second, a total of 1200 frames per video, a total of 6000 frames and 240000 anchor frames. The comparison with the algorithm in general scenario is as follows (Fig. 9):

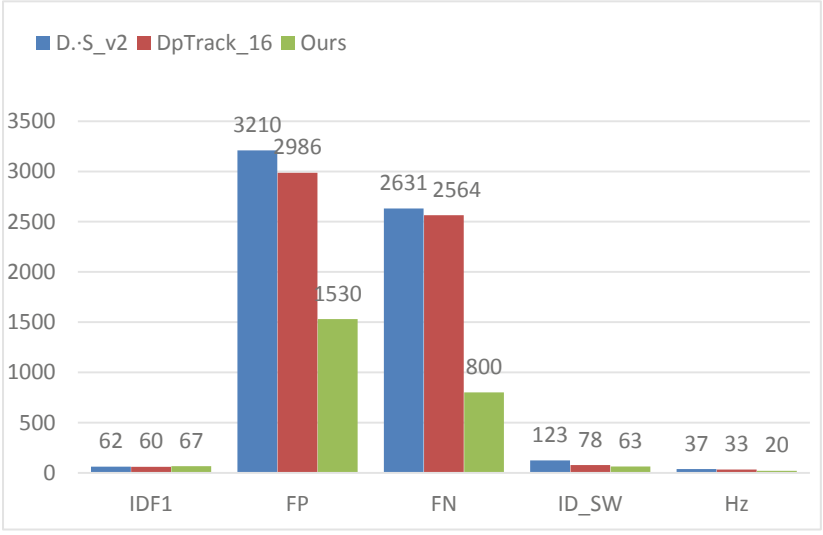


Fig. 9. Performance comparison diagram of personnel tracking algorithm

It can be seen from the above table that the fusion algorithm has good tracking performance in this scene, especially on MT and ML, mainly because of the fixed number of scenes. Once the detected target and the existing tracking do not meet the threshold of motion matching or representation matching, these unmatched targets will be removed from the threshold limit, and according to the distance value between the detected target and the tracking target to run secondary matching. The idea of cascade matching also greatly reduces the probability of target identity exchange. Because the number of people in the overall monitoring is not large, so the omission of tracking is not much, FN gets a good value, but the smooth wall occasionally appears the shadow of people, resulting in new tracking, which brings instability to the tracking part (Table 3).

Table 3. Index comparison results of personnel tracking algorithm

Name	MOTA	IDF1	MT	ML	FP	FN	ID_SW	Hz
D. • S_v2	73.6	62	54%	30%	3210	2631	123	37
DpTrack_16	70.3	60	61%	19%	2986	2564	78	33
Ours	90.5	67	82%	10%	1530	800	63	20

5 Conclusion

Based on the common WiFi equipment and monitoring system, this paper combines the personnel information carried by WiFi fingerprint location with the high precision of image positioning, and the proposed fusion location algorithm can greatly improve the effect of WiFi fingerprint location under the condition of low cost, and meet the needs of high precision positioning of participants in the scene of emergency rescue exercise. Although multi-camera can be used to avoid occlusion to the greatest extent, it is difficult to avoid the problem of poor light in practical applications, especially in darker rooms, the problem of video missed inspectors will be particularly prominent, which also limits the application of this algorithm in more scenes. In the next step of this paper, the research direction of pure image location without WiFi fingerprint will be studied, and the tracking and location of people will be completed by using the advanced semantic features of the characters in the image.

Acknowledgement. This study was supported by State's Key Project of Research and Development Plan (No. 2018YFC0810601, No. 2016YFC0901303). The work was conducted at University of Science and Technology Beijing.

References

1. Wenjuan, L.: The 13th five-year plan for the construction of emergency response system will establish a unified framework of emergency management standard system. Stand. Eng. Constr. **2**(5), 99–110 (2017)
2. Polito, S., Biondo, D.: Performance evaluation of active RFID location systems based on RF power measures. In: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (2007)

3. Cruz, O., Ramos, E., Ramírez, M.: 3D indoor location and navigation system based on Bluetooth. In: International Conference on Electrical Communications & Computers (2011)
4. Yu, K., Montillet, J.P., Rabbachin, A.: UWB location and tracking for wireless embedded networks. Signal Process. **86**(9), 2153–2171 (2006)
5. Zheng, Y., Wu, C., Liu, Y.: Locating in fingerprint space: wireless indoor localization with little human intervention. In: International Conference on Mobile Computing & Networking (2012)
6. Barnich, O., Van Droogenbroeck, M.: ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans. Image Process. **20**(6), 1709–1724 (2011)
7. Hofmann, M., Tiefenbacher, P., Rigoll, G.: Background segmentation with feedback: the pixel-based adaptive segmenter. In: Computer Vision & Pattern Recognition Workshops (2012)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision & Pattern Recognition (2005)
9. Vidhyalakshmi, M.K., Poovammal, E.: A survey on face detection and person re-identification. **1**, 283–292 (2016)
10. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision & Pattern Recognition (2005)
11. Li, J., Liang, X., Shen, S.M.: Scale-aware fast R-CNN for pedestrian detection. IEEE Trans. Multimed. **PP**(99), 1 (2015)
12. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 443–457. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_28
13. Mao, J., Xiao, T., Jiang, Y.: What can help pedestrian detection? In: Computer Vision & Pattern Recognition (2017)
14. Liu, W., Liao, S., Ren, W.: High-level semantic feature detection: a new perspective for pedestrian detection. arXiv:1904.02948 [cs.CV]
15. Feng, W., Hu, Z., Wu, W.: Multi-object tracking with multiple cues and switcher-aware classification. arXiv:1901.06129 [cs.CV]
16. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. arXiv:1903.0562 5 [cs.CV]
17. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. arXiv:1703.07402 [cs.CV]
18. Lee, B., Erdenee, E., Jin, S., Nam, M.Y., Jung, Y.G., Rhee, P.K.: Multi-class multi-object tracking using changing point detection. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 68–83. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_6
19. Sawhney, H.S., Kumar, R.: True multi-image alignment and its application to mosaicing and lens distortion correction. In: Conference on Computer Vision & Pattern Recognition (1997)
20. Yoneyama, S., Kikuta, H.: Lens distortion correction for digital image correlation by measuring rigid body displacement. Opt. Eng. **45**(2), 409–411 (2006)
21. Miyaki, T., Yamasaki, T., Aizawa, K.: Visual tracking of pedestrians jointly using Wi-Fi location system on distributed camera network. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 1762–1765. IEEE (2007)
22. Rafiee, M.: Improving indoor security surveillance by fusing data from BIM, UWB and Video. Concordia University (2014)