Cognitive Computation (2018) 10:94–104
https://doi.org/10.1007/s12559-017-9529-6

Cognitive Fusion of Thermal and Visible Imagery for Effective Detection and Tracking of Pedestrians in Videos

Yijun Yan[1] · Jinchang Ren[1] · Huimin Zhao[2,3] · Genyun Sun[4] · Zheng Wang[5] · Jiangbin Zheng[6] · Stephen Marshall[1] · John Soraghan[1]

Received: 19 June 2017 / Accepted: 21 November 2017 / Published online: 4 December 2017

Abstract

In this paper, we present an efficient framework to cognitively detect and track salient objects from videos. In general, colored visible image in red-green-blue (RGB) has better distinguishability in human visual perception, yet it suffers from the effect of illumination noise and shadows. On the contrary, the thermal image is less sensitive to these noise effects though its distinguishability varies according to environmental settings. To this end, cognitive fusion of these two modalities provides an effective solution to tackle this problem. First, a background model is extracted followed by a two-stage background subtraction for foreground detection in visible and thermal images. To deal with cases of occlusion or overlap, knowledge-based forward tracking and backward tracking are employed to identify separate objects even the foreground detection fails. To evaluate the proposed method, a publicly available color-thermal benchmark dataset Object Tracking and Classification in and Beyond the Visible Spectrum is employed here. For our foreground detection evaluation, objective and subjective analysis against several state-of-the-art methods have been done on our manually segmented ground truth. For our object tracking evaluation, comprehensive qualitative experiments have also been done on all video sequences. Promising results have shown that the proposed fusion-based approach can successfully detect and track multiple human objects in most scenes regardless of any light change or occlusion problem.

Keywords Multiple object detection · Pedestrian detection/tracking · Cognitive fusion · Visible image · Thermal image

Introduction

In the past decades, detection and tracking of video objects has always been a major task in the computer vision field [1–3]. As one subset of video object tracking, pedestrian detection and tracking has drawn massive research attention and been applied to many applications such as visual surveillance [4–8], driver-assistance systems [9–11], human activity recognition [12–14], and others [15, 16]. For pedestrian detection and tracking, visible camera and thermal imagery are two popularly used sources of image modalities, though not necessarily in a combined solution [17–19]. However, either visible image or thermal image has their advantages and disadvantages. Visible image can show detailed color information; however, it really suffer from lighting variations, cluttered backgrounds, artificial appearances, i.e., shadows, and etc. Since the object is detected by its temperature and radiated heat, thermal image can eliminate the influence of color and illumination changes on the objects' appearance [20] in any weather conditions and at both day and night time. However, in some cases, e.g., occlusions, the thermal camera may fail to

✉ Jinchang Ren
jinchang.ren@strath.ac.uk

✉ Huimin Zhao
zhaohuimin@gpnu.edu.cn

✉ Genyun Sun
genyunsun@163.com

✉ Zheng Wang
wzheng@tju.edu.cn

[1] Department of Electronic and Electrical Engineering, University of Strathclyde, Royal College Building, 204 George Street, Glasgow, UK

[2] School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

[3] The Guangzhou Key Laboratory of Digital Content Processing and Security Technologies, Guangzhou, China

[4] School of Geosciences, China University of Petroleum, Qingdao, China

[5] School of Computer Software, Tianjin University, Tianjin, China

[6] School of Software and Microelectronic, Northwestern Polytechnical University, Xi'an, China

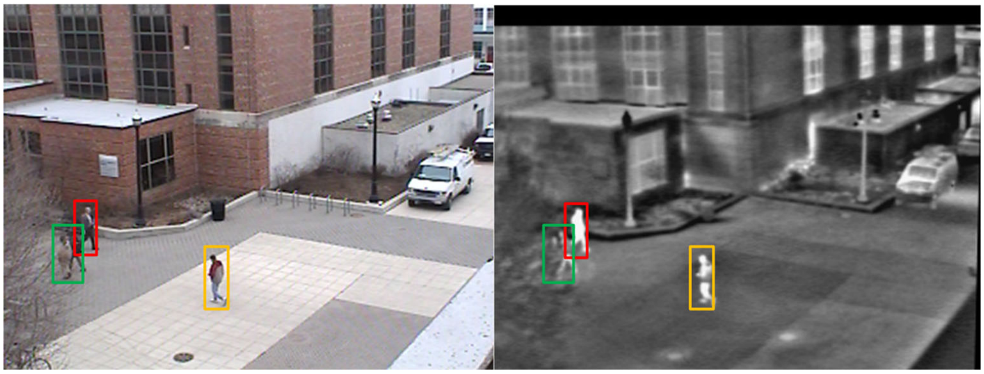

Fig. 1 Visible image of a scene (left) and thermal image of the same scene (right)

detect the object properly. In Fig. 1, there are three pedestrian templates; for the one with a yellow rectangle, both visible and thermal image can detect it very well since it has high contrast to the background in the visible domain and human temperature in the thermal domain. For the template in the red rectangle, it has a compact shape in the thermal image. However, in the visible image, we can just identify it coarsely due to the similar appearance in color of the background and the person's cloth. The one in green rectangle can be seen in the visible image but hardly observed in the corresponding thermal image. This is because thermography is only able to directly detect surface temperatures, and it cannot work well when the object is (partially) occluded. Moreover, it will detect any objects (e.g., windows and cars in Fig. 1) with surface temperature.

For the purpose of object detection, background subtraction plays an important role in it. Due to its significance, a large number of background subtraction algorithms have been proposed in recent years. Andrew et al. [21] proposed a single-camera statistical segmentation algorithm where a combination of statistical background image estimation and Bayesian-based segmentation is used to achieve foreground detection. Domenico and Luca [22] proposed a fast background subtraction method based on a clustering algorithm with a condition-based mechanism. Zhao et al. [23] proposed a background modeling method for motion detection in dynamic scenes based on type-2 fuzzy Gaussian mixture model [24] and Markov random field (MRF) [25]. In [26], authors introduced a background subtraction framework based on texture feature. Furthermore, color cues are clustered by the codebook scheme in order to refine the texture-based detection. Pierre-Luc points out in [27] that most background subtraction methods do not pay attention to the spatial or spatio-temporal relationship of each analyzed pixel, and also suffer in complexity, computation cost, and versatility. Therefore, he proposed a spatiotemporal-based background subtraction algorithm which has been proved low-cost and highly efficient. In addition, he also proposed another one using spatiotemporal feature descriptors in [28] in order to build an adaptive and flexible model rather than tuning parameters in different scenarios for optimal performance. In [29], a background subtraction model based on independent component analysis and principal component analysis is proposed to detect multiple moving objects under complex outdoor scenes such as bad weather or dynamic background. In [30], based on the assumption that moving objects are usually small and sparse, a collaborative low-rank and sparse separation model is proposed to robustly detect moving objects with different sizes. However, background regions which have the similar color/intensity as the foreground may be detected as foreground by mistake. In Wang et al. [31], a coarse-to-fine pedestrian detection method is proposed for visual surveillance, which can solve the problem in detecting small pedestrians. By using pan-tilt-zoom control, it also helps to achieve real-time tracking, though the performance depends on specified sensor settings.

However, due to lack of cognitive knowledge, some of their methods have good objective performance; their subjective performance is not satisfied (detailed in "Experimental Results"). Besides, existing approaches mainly rely on color image for pedestrians' detection and tracking, using different features such as color and texture for modeling. In our paper, thermal images are also used, which have neither color nor texture information but just intensity instead. Unlike color images, thermal images are robust to any weather or illumination conditions though they are sensitive to surface temperature. As a result, it is necessary to find a new path to process both visible and thermal image based on their characteristics. Inspired by several multi-modality image fusion approaches [32–34], where color and infrared images are integrated for saliency-based image fusion [32, 34] and image registration [33], the fusion of the two image modalities (RGB and thermal) offers new insights for the supplementary information they can provide. This has proven to be a success in determining the refined foreground map by the fusion of both visible and thermal binary maps. By combining cognitive models from different levels and aspects, we have proposed a generic model for effective detection and tracking of pedestrians from color and thermal videos.

In our proposed approach, different levels of cognitive models are integrated together for effective detection and tracking of pedestrians from color and thermal videos. These include color- and intensity-based cognitive models of human visual perception for robust background estimation and foreground detection, cognitive models of object priors for shape-constrained morphological filtering in determining the refined foreground maps, and cognitive model of motion for motion consistency-constrained mean shift in extracting single persons from a group. By systematically integrating these cognitive models together, an effective model is developed and proven to be the best when benchmarking with several state-of-the-art techniques. It is believed the proposed approach can be also applied in other areas of object detection and tracking, e.g., medical imaging for improved performance.

The main contributions in this paper can be highlighted in the following three aspects:

- As color and intensity information plays important roles in the cognitive models of our human visual perception, an adaptive Gaussian mixture model is proposed to measure the distribution of such information in multi-modality images (color and thermal) before deriving the estimated background for foreground detection.
- Based on the prior knowledge of the human objects to be detected, shape constraints are fused in combination with morphological filtering for determining the refined foreground maps.
- Inspired by cognitive model of motion, motion consistency is applied in a constrained mean-shift scheme for the extraction of single persons from a group.

The rest of the paper is organized as follows: The "Overview of the Proposed System" illustrates the framework of the proposed method. The "Foreground Detection" describes the foreground detection approach. The "Object Tracking" elaborates the object tracking method. Experimental results are presented and discussed in the "Experimental Results." Finally, some concluding remarks and future work are summarized in the "Conclusion."

Overview of the Proposed System

In this paper, we proposed a two-stage background subtraction procedure based on human cognition knowledge on both visible and thermal images for fusion-based pedestrian detection, and four modules are included in Fig. 2. In the first stage, we predict the background model by computing the median value of randomly selected frames in the videos (module 1), and apply an adaptive threshold to detect binary foreground map along with knowledge-based morphological refinement (module 2). In the second stage, we use the results from module 1 as prior frames and employ learning-based adaptive Gaussian mixture model to estimate the background model and generate the binary foreground map (module 3). Then the initial and Gaussian-based foreground maps of both visible and thermal images will be refined by shape-constrained morphological filtering and further fused together to get the final foreground map (module 4). In the performance evaluation (module 5), the proposed background subtraction method is compared against a number of state-of-the-art methods on a widely used publicly available video sequences. Some widely used evaluation criteria such as precision, recall, and F measure are used for quantitative assessment. In addition, we also proposed constrained mean-shift tracking method to have a capability of scale change and identify the individual pedestrian template from a pedestrian group more efficiently (detailed in "Object Tracking"). Furthermore, the performance of object tracking is also evaluated by qualitative assessment. Detailed results are reported in the "Experimental Results."

Foreground Detection

In this section, a two-stage foreground detection method is applied for both visible and thermal images. Eventually, the desired foreground map is fused by the foreground detection results of two types of images with cognition-based morphological process.

Random Median Background Subtraction

To capture the initial region of pedestrians in visible and thermal image, we first estimate the background model by computing a median map (Fig. 2 module 1) of N frames randomly selected from the video sequence. And initial background subtraction process for each visible or thermal frame is defined as:

$$\mathrm{BS_{ini}}(\mathrm{x},\mathrm{y}) = |\mathrm{I}(\mathrm{x},\mathrm{y}) - \mathrm{I_{med}}(\mathrm{x},\mathrm{y})| \tag{1}$$

After that, we binarize the $\mathrm{BS_{ini}}$ with an adaptive threshold, i.e., OTSU [35] to get a binary image $\mathrm{I_{bi}}$ with coarse human body region (Fig. 2 module 2). However, $\mathrm{I_{bi}}$ contains many ambiguous contents and some objects that should be detected as a whole are fractured. Therefore, a cognitive-based morphology refinement is applied here to filter the insignificant region and integrate the potential objects. Since the object that we want to detect is pedestrian, and we can assume the shape of the pedestrians is an ellipse or a rectangle based on our cognition, so that its major axis length is usually larger than minor axis length. Therefore, in our morphology refinement, we define a rectangle-shaped structuring element to connect separated regions together to be a whole object. The width and height of the rectangle is defined as $2\mathrm{n}+1$ and $2\mathrm{n}+3$ $\left(n\in\mathrm{Z}_0^+\right)$, respectively. Here, we set n as 1. Furthermore, as the size of the pedestrian in the video will not be small, we filter those noise regions by an empirical threshold T. From Fig. 3, we can see in the refinement result $\mathrm{I_r}$, the noise regions with the small area have been removed and every object has been integrated.

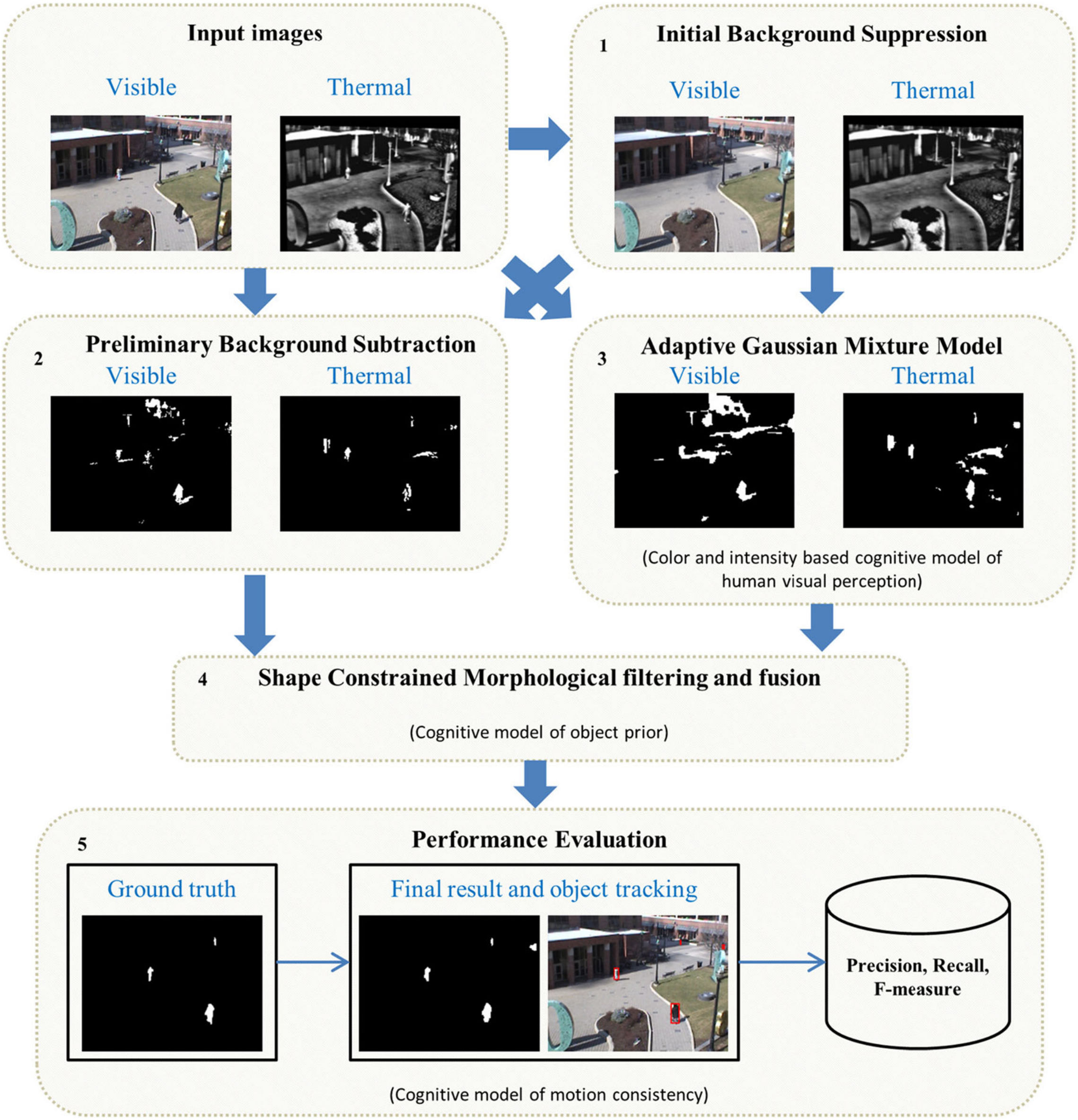
Input images
Visible
Thermal
1
Initial Background Suppression
Visible
Thermal
2
Preliminary Background Subtraction
Visible
Thermal
3
Adaptive Gaussian Mixture Model
Visible
Thermal
(Color and intensity based cognitive model of human visual perception)
4
Shape Constrained Morphological filtering and fusion
(Cognitive model of object prior)
5
Performance Evaluation
Ground truth
Final result and object tracking
Precision, Recall, F-measure
(Cognitive model of motion consistency)

Fig. 2 Proposed framework within five modules

Adaptive Mixture Background Subtraction

Although the random median background subtraction module can detect some potential objects, it still contains many false alarms due to lack of the analysis of the scene changes, lighting changes, moving object, and etc. Therefore, a learning-based background mixture model is employed here to estimate the foreground map

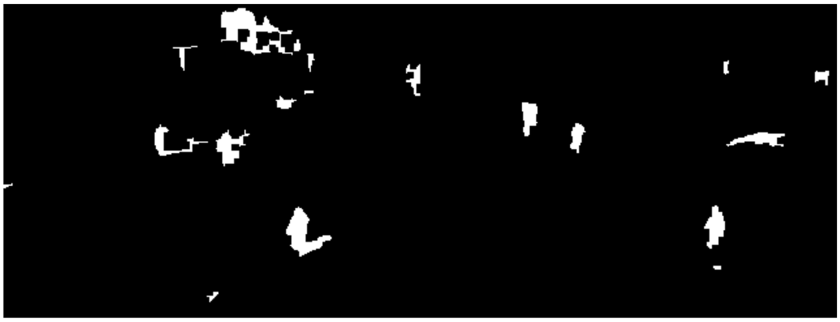

Fig. 3 The refined initial background subtraction results of visible (left) and thermal (right) images

under a real scene. For a particular surface under particular lighting, a single Gaussian per pixel is sufficient to represent the pixel value. However, in practice, there are multiple surfaces due to the lighting condition change. Thus, in order to fit the real world situation and our human cognition, multiple adaptive Gaussians are necessary. In this paper, we model each pixel by a mixture of K Gaussian distributions. The probability of observed pixel X_t at time t can be written as:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta\left(X_t; \mu_{i,t}, \Sigma_{i,t}\right) \tag{2}$$

where $\omega_{i,\ t}$ is the weight parameter of the i^{th} Gaussian in the mixture at time t, $\mu_{i,\ t}$ and $\Sigma_{i,t} = \sigma_i^2 I$ are the mean and covariance value of the i^{th} Gaussian in the mixture at time t. $\eta(*)$ is the normal distribution of the i^{th} Gaussian component.

$$\eta\left(X_t; \mu_{i,t}; \Sigma_{i,t}\right) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_{i,t}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(X_t-\mu_{i,t}\right)^T \Sigma_{i,t}^{-1}\left(X_t-\mu_{i,t}\right)} \tag{3}$$

The first B distributions are chosen as the background model

$$B = argmin_b\left(\sum_{i=1}^{b} \omega_i > T\right) \tag{4}$$

T is the minimum portion of the data that should be counted as background. For any new observed pixel value, X_t will be considered as foreground if it is more than 2.5 standard deviations away from existing B distributions. And the first Gaussian component that matches the new observed pixel value will be updated by the following progress:

$$\omega_{i,t} = (1-\alpha)\omega_{i,t-1} + \alpha\hat{p}\left(\omega_{i,t}|X_t\right) \tag{5}$$

$$\mu_{i,t} = (1-\alpha)\mu_{i,t-1} + \rho X_t\big) \tag{6}$$

$$\Sigma_{i,t} = (1-\alpha)\Sigma_{i,t-1} + \rho\left(X_t-\mu_{i,t}\right)\left(X_t-\mu_{i,t}\right)^T \tag{7}$$

$$\rho = \alpha\eta\left(X_t; \mu_{i,t}, \Sigma_{i,t}\right) \tag{8}$$

$$\hat{p}\left(\omega_{i,t}|X_t\right) = \begin{cases} 1 & ; if\ \omega_{i,t} matches\ first\ Gaussian\ component \\ 0 & ; otherwise \end{cases} \tag{9}$$

where α is the learning rate.

In addition, we use ten random median background subtraction results to predict the initial value of the parameters $\omega_{i,\ t}$, $\mu_{i,\ t}$, and $\Sigma_{i,\ t}$ for better performance. After the adaptive background mixture model is done, we can get the foreground map $I_a{}^{vis}$ and $I_a{}^{thm}$ of visible and thermal images (Fig. 2 module 3).

Fusion Strategy

In order to generate the final foreground map and make the fusion result close to human perception, we put a shape-constrained morphological refinement to the results from the previous stage and integrate them together. For $I_r{}^{vis}$, $I_a{}^{vis}$, $I_r{}^{thm}$, and $I_a{}^{thm}$, we define a function $D(\cdot)$ that can dilate all the potential objects with a shape-based structuring element. And we set $n = 0$ because we just want to smooth the edge for each object and connect the small gap between some object pieces. By doing so, the shape of the object will have continuity, which matches human perceptions. Then the final foreground map (Fig. 2 module 5) can be built by fusion strategy as follows:

$$I_{vis} = \left(I_a{}^{vis} \cap D\left(I_r{}^{vis}\right)\right) \cup \left(I_r{}^{vis} \cap D\left(I_a{}^{vis}\right)\right) \tag{10}$$

$$I_{thm} = \left(I_a{}^{thm} \cap D\left(I_r{}^{thm}\right)\right) \cup \left(I_r{}^{thm} \cap D\left(I_a{}^{thm}\right)\right) \tag{11}$$

$$I_{final} = \left(I_{vis} \cap D\left(I_{thm}\right)\right) \cup \left(I_{thm} \cap D\left(I_{vis}\right)\right) \tag{12}$$

Object Tracking

For any continuous frames, if the later frame has fewer objects than the former frame, there will be only two situations. The first situation is one or more objects in the former frame have been out of the later frame, and the other situation is some individual objects in the later frame are detected as a whole in the foreground detection stage due to the inevitable overlap and occlusion problem. Figure 4 (right and middle) shows the detection detail of two adjacent frames where there should be three pedestrian patterns detected in both frames, but for the frame in the middle, the object detection method considers the left two patterns as one object because they are close to each other. Therefore, in this section, an improved mean-shift method is proposed to track the individual objects in the second situation.

Conventional mean-shift method [36] mainly has two drawbacks. The first one is that it tracks the object mostly based on the color and texture feature, but does not take too much account of the spatial relationship of the object. Therefore, if the object has the similar color with surrounding background, the tracker will probably locate the object at the background region in the following frame. The second one is the similarity computation for two probability density functions (PDFs). In [36], it defines the distance between two PDFs as

$$\mathrm{d(y)} = \sqrt{1-\rho\left[\hat{\mathrm{p}}(\mathrm{y}), \hat{\mathrm{q}}\right]} \tag{13}$$

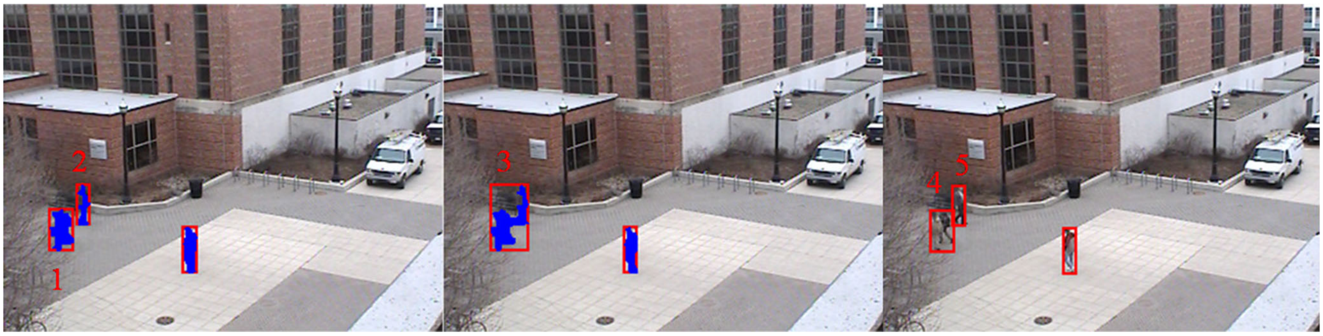

Fig. 4 Initial detection result of frame 1 (left) and frame 2 (middle), and updated detection result of frame 2 (right) after mean-shift tracking progress

$$\rho\left[\hat{p}(y), \hat{q}\right] = \sum_{u=1}^{m} \sqrt{\hat{p}_u(y)\hat{q}_u} \tag{14}$$

where $\rho[\cdot]$ is the Bhattacharyya coefficient, $\hat{q} = \{\hat{q}_u\}_{u=1\ldots m}$ (with $\sum_{u=1}^{m} \hat{q}_u = 1$) is the discrete density from the m-bin histogram of the object model, $\hat{p}(y) = \{\hat{p}_u(y)\}_{u=1\ldots m}$ (with $\sum_{u=1}^{m} \hat{p}_u(y) = 1$) is estimated as a given location y from the m-bin histogram of the object candidate. However, $\hat{q}$ does not change with time which is not fit with human cognition because the surrounding of the object cannot be always the same in the real scene. On the other hand, unchangeable $\hat{q}$ will also increase the convergence cost because it will take more time to match object candidate and object model within difference background.

To overcome two problems mentioned above, we propose constrained mean-shift method where two improvements are introduced in the following. Firstly, the object model is updated in each frame in order to get the real-time $\hat{q}$. Thus, the size of the $\hat{q}$ will change with the scale changing of the object. Meanwhile, the pedestrians usually move slowly which means their surrounding background in adjacent frames will not be changed too much. In this case, $\hat{p}(y)$ can be quickly matched with $\hat{q}$ in each frame. Secondly, we limit the shift range with the spatial information of the objects in adjacent frames. We define F_{n-1} and F_nd are frame n-1 and frame n, R^i_{n-1} is the region i in F_{n-1}, and R^j_n is the region j in F_n, $X^{i,1}_{n-1}$, $X^{i,2}_{n-1}, Y^{i,1}_{n-1}, \textit{and}\ Y^{i,2}_{n-1}$ which are the location elements of R^i_{n-1}, and $X^{j,1}_n, X^{j,2}_n, Y^{j,1}_n, \textit{and}\ Y^{j,2}_n$ are the location elements of R^j_n. After the location of R^i_{n-1} candidate in F_n (expressed as $X^{i,1}_n, X^{i,2}_n, Y^{i,1}_n, \textit{and}\ Y^{i,2}_n$) is determined by conventional mean-shift algorithm in every iteration, we further refine this location by a displacement term represented as λ_x, λ_y.

Let $\lambda^{i,1}_x = X^{j,1}_n - X^{i,1}_n, \lambda^{i,2}_x = X^{j,2}_n - X^{i,2}_n, \lambda^{i,1}_y = Y^{j,1}_n - Y^{i,1}_n$, *and* $\lambda^{i,2}_y = Y^{j,2}_n - Y^{i,2}_n$ be the displacement terms, the new position of the object can be determined by using these displacement terms as follows:

$$\begin{cases} X^i_n = X^i_n + \lambda^{i,1}_x, \textit{if}\,\lambda^{i,1}_x > 0 \\ X^i_n = X^i_n + \lambda^{i,2}_x, \textit{if}\,\lambda^{i,2}_x < 0 \end{cases} \tag{15}$$

$$\begin{cases} Y^i_n = Y^i_n + \lambda^{i,1}_y, \textit{if}\,\lambda^{i,1}_y > 0 \\ Y^i_n = Y^i_n + \lambda^{i,2}_y, \textit{if}\,\lambda^{i,2}_y < 0 \end{cases} \tag{16}$$

As can be seen from Fig. 4, region 1 and region 2 in frame 1 are two individual object models, the corresponding object candidate should be limited in region 3 in frame 2. In this case, the object group in frame can be tracked separately in regions 4 and 5 (shown in the right image in Fig. 4).

Experimental Results

Dataset Description and Evaluation Criteria

To evaluate the performance of our foreground detection and object tracking methods, a publicly available database 03 OSU Color-Thermal Database from OTCBVS are employed here. Thermal sequences are captured by Raytheon PalmIR 250D thermal sensor and color sequence are captured by Sony TRV87 Handycam color sensor. All the frames in both sequences have a spatial resolution of 320 × 240 pixels. The number of frames in each video sequence is Sequence-1:2107, Sequence-2:1201, Sequence-3:3399, Sequence-4:3011, Sequence-5:4061, and Sequence-6:3303, respectively. Figure 5 shows some visible and thermal frames and the results of our foreground detection method. For our foreground detection method, we do both qualitative (Fig. 6) and quantitative (Table 4) analysis against six state-of-the-art methods, i.e., GMG [21], IMBS [22], LOBSTER [27], MultiCue [26], SuBSENSE [28], and T2FMRF [23] on some manually segmented silhouettes. For our object tracking method, we do comprehensive qualitative experiments on all video sequences (Fig. 7).

For quantitative performance assessment of the proposed foreground detection algorithm, several commonly used metrics are adopted in our experiments, which include the precision, recall, and F measure. The precision value P and recall value R are determined by $P = \frac{T_p}{T_p+F_p}$, $R = \frac{T_p}{T_p+F_p}$, where T_p, F_p, and F_n, respectively, refer to the number of correctly detected foreground pixels of the pedestrians, incorrectly

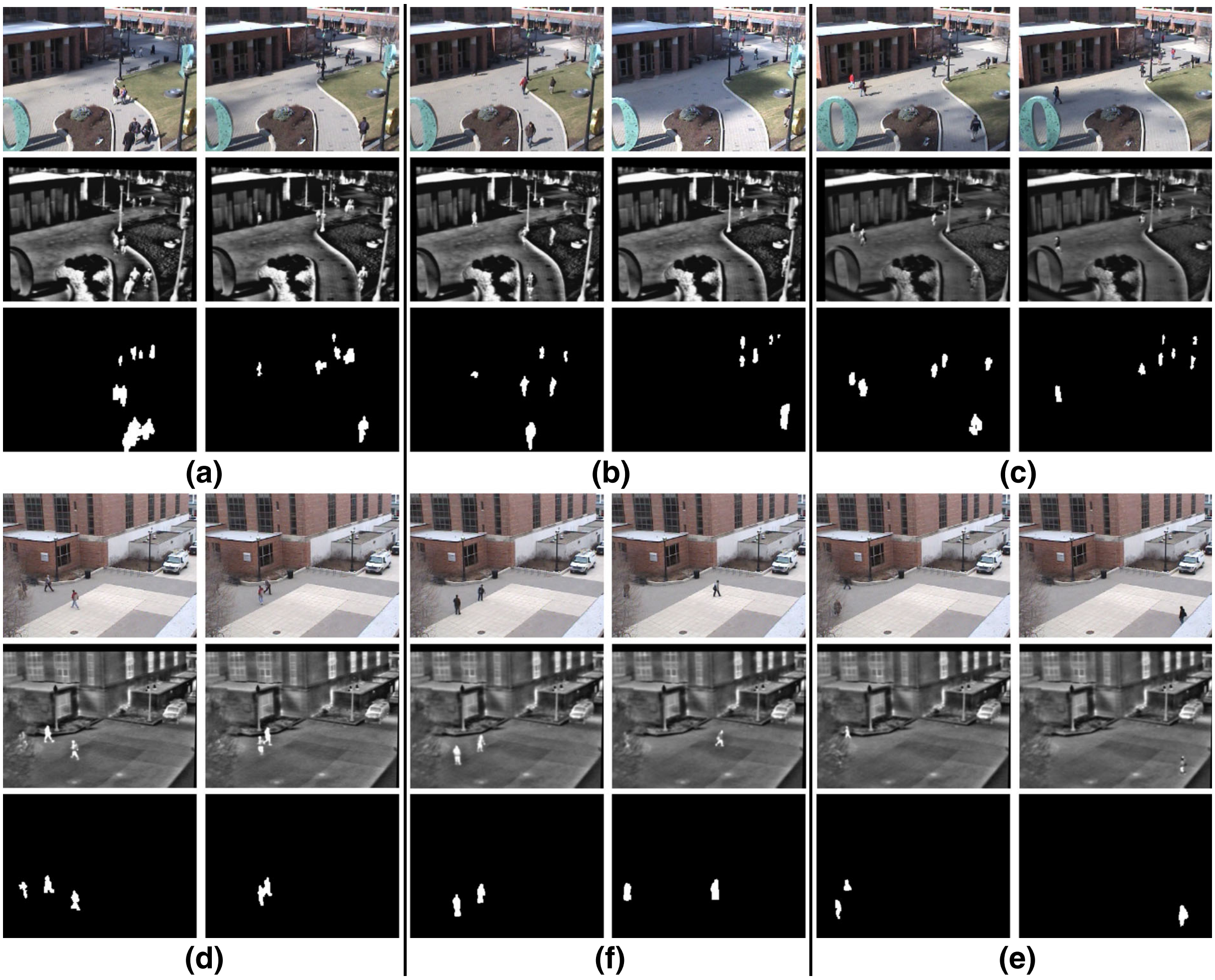

Fig. 5 Visual results of proposed foreground detection algorithm. **a** Sequence-1. **b** Sequence-2. **c** Sequence-3. **d** Sequence-4. **e** Sequence-5. **f** Sequence-6

detected foreground pixels (false alarms), and incorrectly detected background pixels (or missing pixels from the object). Specifically, these three numbers can be calculated by comparing the binary masks of the detected image and the ground truth. Furthermore, since the database does not have ground truth, we obtain a manual segmentation of the pedestrian regions in 53 frames from Sequence-1. The F measure is defined by $F_{measure} = \frac{2 \cdot P \cdot R}{P+R}$.

Key Parameter Selection

In this paper, we carefully choose the key parameter by investigating their changes on the performance. The effect of several key parameters in the proposed approach is discussed as follows. For adaptive Gaussian mixture model, the key parameters are the learning rate α, the threshold of background portion T, and the Gaussian distribution number K. Tables 1, 2, and 3 summarize the performance by changing these three parameters, respectively. From Table 1, we can see that the precision will slightly increase with the rising learning rate while the recall shows the inverse trend against the precision. As the learning rate decides how many recent frames are used for training, the larger the learning rate is, the less the recent number of frames is used. Generally, with less recent frames used to predict the background, the local information will be more detailed. On the contrary, more recent frames will make the background to have more global property and robust to local inconsistency. To this end, the learning rate can be neither too large nor too small, which is set to 0.002 (500 recent frames) in this paper based on our practical measurement [37]. From Table 2, we can find that the precision grows with the increasing T, yet the recall has the opposite tendency against the precision again. The reason for this is when the portion of background is increased some foreground regions or noise may be considered as background. Although it somehow makes the precision increase, the recall will reduce sharply. However, if the portion of background is too small, many noised regions can be considered as foreground. As a result, it will significantly degrade the precision rate due to the growth of the false alarm while the recall will not increase much. Though T was set to 0.6 in [37], we empirically choose

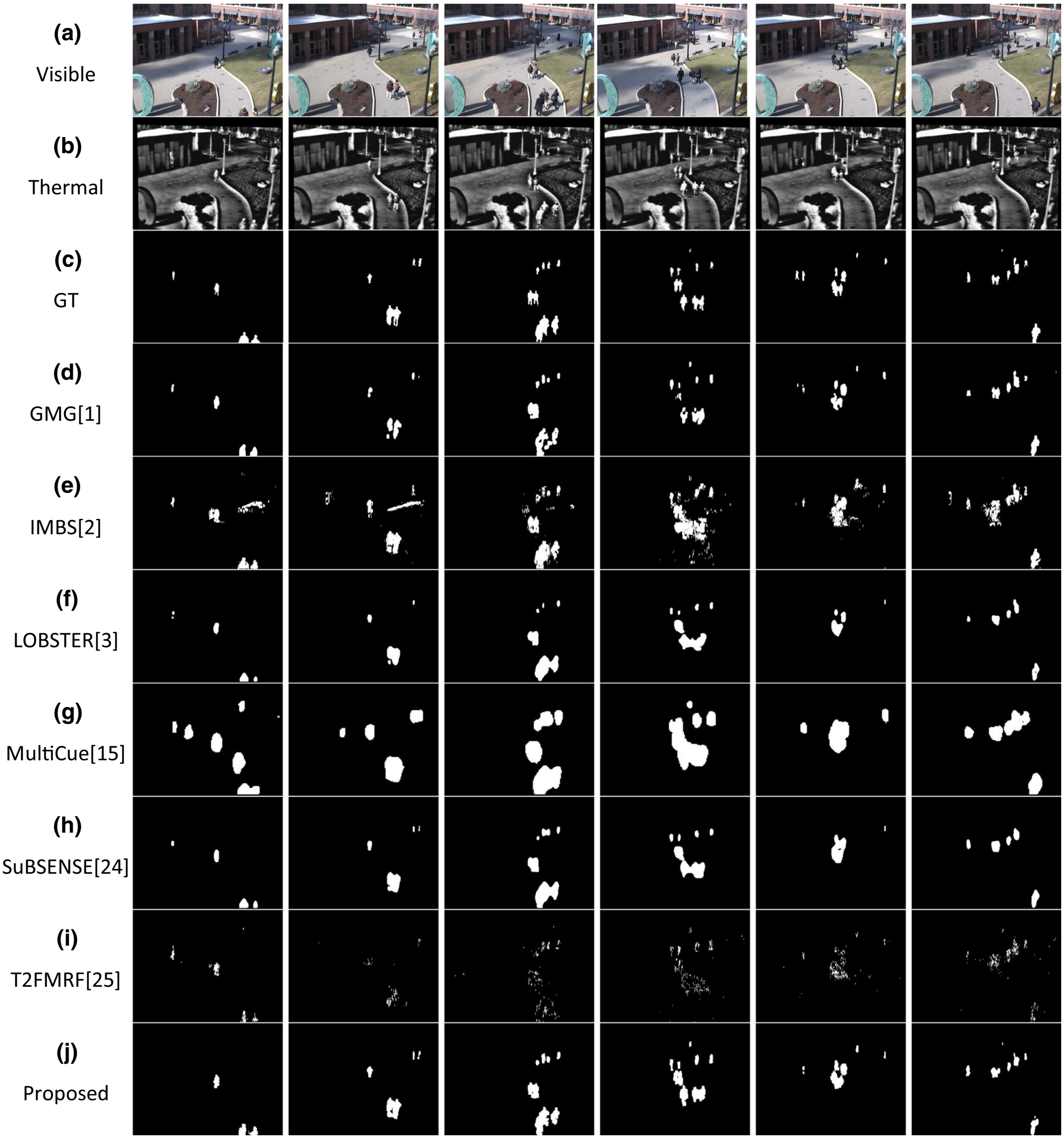

Fig. 6 Visual comparison. **a** Original images. **b** Ground truth. **c**–**j** Saliency maps generated by different methods

$T = 0.7$ for its better performance. Table 3 shows that the number of Gaussian distributions does not affect the performance much as long as it is larger than 2. Therefore, we set $K = 5$ as suggested in [38].

Assessment of Foreground Detection Method

To evaluate the quality of the extracted foreground map, we compare our proposed method with six state-of-the-art methods in terms of precision, recall, and F measure as the performance metrics with the results shown in Table 4. For fair comparison, instead of just comparing our fusion result with others' results on visible images, we do the same fusion strategy for each method where I_{vis} and I_{thm} are generated by those methods on visible and thermal images, respectively. From Table 1, we can see the precision of proposed foreground detection is comparable with GMG [21] and LOBSTER [27], and both recall and F measure of our method outperform other methods. IMBS, MultiCue, and T2FMRF yield bad performance due to their algorithms does not take too much account of the scene change. Although their methods work well in some indoor and outdoor data, those data do not have too much light change. However, in the 03 OSU Color-Thermal Database from OCTBVS, the clouds make the big shadow on the ground and the light of the scene changes as time goes by. GMG, LOBSTER, and SuBSENSE almost have similar performance and very comparable with our proposed method.

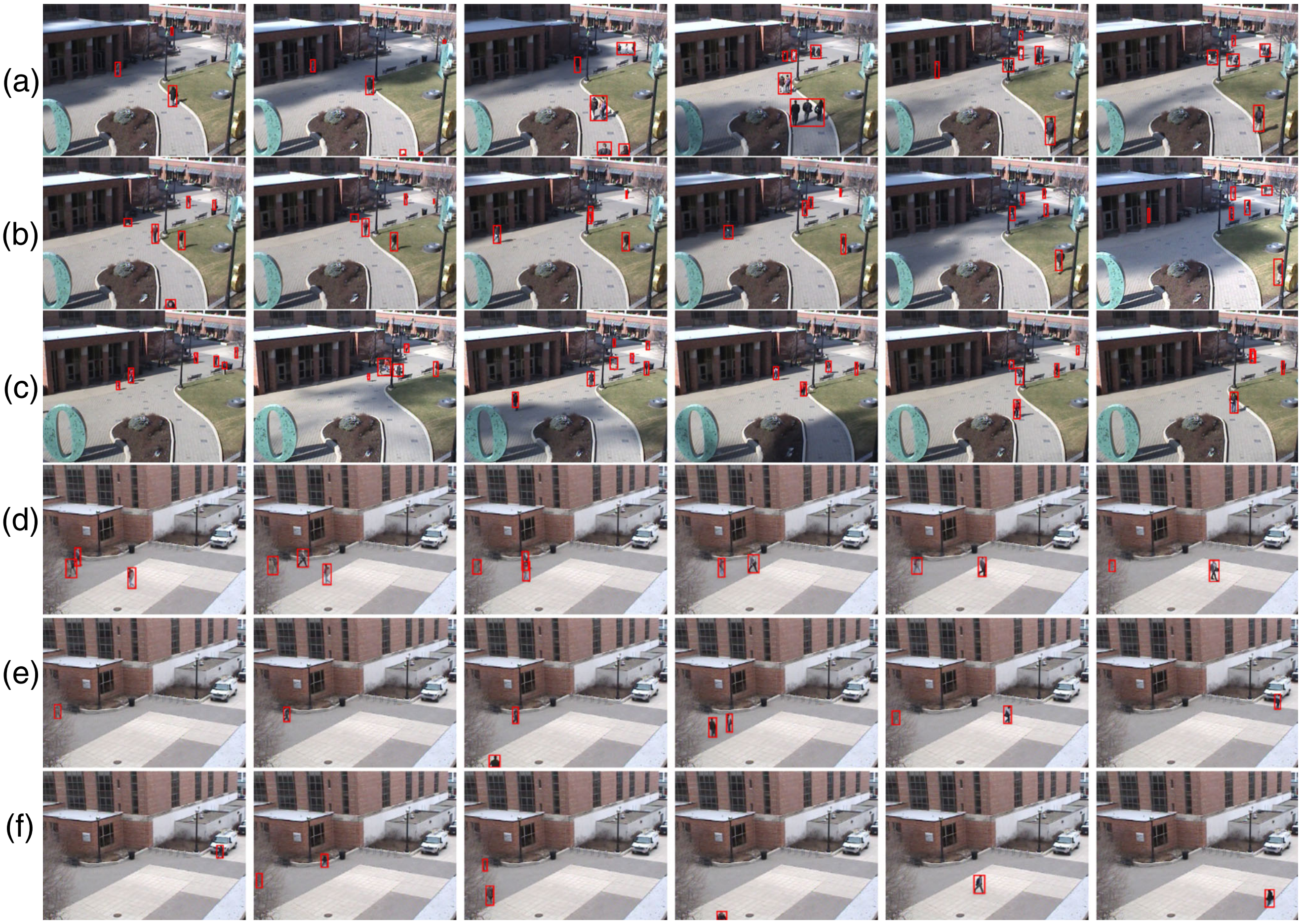

Fig. 7 Visual tracking results of the proposed approach across different images and scenarios. **a**–**f** Sequence-1–6

However, these methods are mainly designed for the object detection in the small scene. And the objects in the small scene usually have large size than the pedestrians in a surveillance system. Therefore, these methods can detect the pedestrians within close or middle range but not long range from the camera. In addition, affected by light change and weather condition, some details have been lost. As can be seen in the visible image in Fig. 6, some pedestrians' shapes in GMG are not integrated; some pedestrians' shapes in GMG are fractured, e.g., left person in the first image is split into two regions; some pedestrians that are far away from the camera cannot be detected in SuBSENSE, e.g., the fifth and sixth images. Hence, these methods have good quantitative results but their qualitative results do not fit human's cognition. However, our foreground detection result is generated by a two-stage background subtraction procedure and fusion strategy where cognition-based knowledge is applied in to refine the procedure and guide the fusion strategy.

Although our proposed method yields the best performance in terms of F measure, there are still rooms for further improvements. As seen, our proposed method has produced high recall value but relative low precision value just like other methods. There are two main reasons, i.e., missing detection and inaccurate ground truth mapping. For the cases of missing detection, this is mainly due to the failure in detecting objects dressing in similar

Table 1 Key parameter α analysis

Learning rate	Precision	Recall	F measure
0.001	69.59	88.72	78
0.002	*70.16*	*87.97*	*78.06*
0.003	71.06	84.97	77.39
0.004	71.03	81.09	75.73

The best results in term of F measure is highlighted in italic

Table 2 Key parameter T analysis

Threshold T	Precision	Recall	F measure
0.4	68.8	89.21	77.69
0.5	68.83	89.2	77.7
0.6	69.06	89.05	77.79
0.7	*70.16*	*87.97*	*78.06*
0.8	72.32	82.49	77.07
0.9	76.26	59.33	66.74

The best results in term of F measure is highlighted in italic

Table 3 Key parameter K analysis

Number of K	Precision	Recall	F measure
2	68.82	89.21	77.7
3	69.56	88.86	78.03
4	70.14	88.04	78.08
5	*70.16*	*87.97*	*78.06*
6	70.16	87.97	78.06
7	70.16	87.97	78.06

The best results in term of F measure is highlighted in italic

color to the background and behind obstacles. This can be possibly improved by introducing certain post-processing such as back-tracking. However, it can still be challenging in dealing with small objects which are frequently grouped together. This also explains the low accuracy of ground truth as in some cases and the silhouettes of the pedestrians can be hardly defined accurately even in a manual way.

Assessment of Object Tracking Method

To validate the performance of the proposed object tracking approach, all video sequences are used in our experiments. In Fig. 7, detection and tracking results from these sequences are given to illustrate the extracted/tracked objects using their bounding boxes. As can be seen, the proposed method can give reliable pedestrian detection and tracking results under various conditions, including occlusion and light changes in terms of illumination and scale. When the pedestrians are independent, we can detect them very well with proper scale bounding box. We can also identify the people even they are overlapped such as the first and third images in Fig. 7d. In addition, when there are some occlusions appear like tree or wall such as the second and third images in Fig. 7d, first image in Fig. 7e, second and third images in Fig. 7f, sixth image in Fig. 7b, and sixth image in Fig. 7c, etc. For the object is getting out of the screen, such as the third image in Fig. 7e, fourth and sixth images in Fig. 7f, we can still locate the objects and track their motion.

Table 4 Comparison of precision, recall, and F measure values

Methods	Precision	Recall	F measure
GMG [21]	70.45	70.17	70.31
IMBS [22]	37.03	74.44	49.46
LOBSTER [27]	72.95	72.19	72.57
MultiCue [26]	26.02	88.78	40.25
SuBSENSE [28]	69.31	76.87	72.89
T2FMRF [23]	50.78	29.93	37.66
Proposed	*70.16*	*87.97*	*78.06*

The best results in term of F measure is highlighted in italic

However, some failure cases, such as third and fourth images in Fig. 7**a**, the second image in Fig. 7**c** still exist in our tracking results. There are two main reasons, and the first one is that some pedestrians always walk together as a group from the beginning to the end in the sequence; therefore, our tracking system always consider the pedestrian group as a single object. The second reason is that if one pedestrian leaves a group of pedestrians and join in another group, the tracking system cannot extract its own color, texture, and spatial features. As a result, the mean-shift method may fail to track in such a context.

Conclusion

In this paper, we proposed a cognitive model by fusing visible and thermal images for pedestrian detection, along with an improved mean-shift method proposed for tracking the pedestrians in the videos. There are three key components in this model, i.e., foreground detection, fusion based object refinement, and object tracking. By estimating the background model followed by a two-stage background subtraction, foreground objects can be successfully detected. Shape-constrained morphological filtering-based fusion strategy helps to further refine the detected foreground objects. Finally, prediction-based forward and backward tracking is found particularly useful to separate overlapped or occluded objects, and robust to the scale change. However, if one certain pedestrian in a group cannot be detected individually from the beginning to the end, the tracking system will fail to estimate its own track and just estimate the track of its group instead. In future work, we will put deep learning model to further enhance the foreground detection performance and improve the tracking procedure in order to precisely estimate the objects' track even with some challenging situations.

Funding This work was partly supported by the National Natural Science Foundation of China (61672008, 61772360, and 61572351), Guangdong Provincial Application-oriented Technical Research and Development Special fund project (2016B010127006, 2015B010131017), the Natural Science Foundation of Guangdong Province (2016A030311013), and the Scientific and Technological Projects of Guangdong Province (2017A050501039).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Li M, Wei C, Yuan Y, Cai Z. A survey of video object tracking. Int J Control Autom. 2015;8(9):303–12. 10.14257/ijca.2015.8.9.29.
2. Yilmaz A, Javed O, Shah M. Object tracking: a survey. Acm Comput Surv (CSUR). 2006;38(4):13–es. https://doi.org/10.1145/1177352.1177355.
3. Yan Y, Ren J, Zhao H, Zheng J, Zaihidee EM, Soraghan J. Fusion of thermal and visible imagery for effective detection and tracking of salient objects in videos. In: Pacific Rim Conference on Multimedia; 2016. pp. 697–704.
4. Benfold B, Reid I. Stable multi-target tracking in real-time surveillance video. In: Computer Vision and Pattern Recognition (CVPR), 2011 I.E. Conference on; 2011. pp. 3457–3464.
5. Sidla O, Lypetskyy Y, Brandle N, Seer S. Pedestrian detection and tracking for counting applications in crowded situations. In: Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on; 2006. pp. 70–70.
6. Ren J, Orwell J, Jones GA, Xu M. Tracking the soccer ball using multiple fixed cameras. Comput Vis Image Underst. 2009;113(5): 633–42. https://doi.org/10.1016/j.cviu.2008.01.007.
7. Ren J, Orwell J, Jones GA, Xu M. Real-time modeling of 3-D soccer ball trajectories from multiple fixed cameras. IEEE Trans Circuits Syst Video Technol. 2008;18:350–62.
8. Ren J, Xu M, Orwell J, Jones GA. Multi-camera video surveillance for real-time analysis and reconstruction of soccer games. Mach Vis Appl. 2010;21(6):855–63. https://doi.org/10.1007/s00138-009-0212-0.
9. Ge J, Luo Y, Tei G. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. Intell Transp Syst, IEEE Trans. 2009;10:283–98.
10. Geronimo D, Lopez AM, Sappa AD, Graf T. Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans Pattern Anal Mach Intell. 2009. pp. 1239–1258.
11. Czubenko M, Kowalczuk Z, Ordys A. Autonomous driver based on an intelligent system of decision-making. Cogn Comput. 2015;7(5): 569–81. https://doi.org/10.1007/s12559-015-9320-5.
12. Poppe R. A survey on vision-based human action recognition. Image Vis Comput. 2010;28(6):976–90. https://doi.org/10.1016/j.imavis.2009.11.014.
13. Bodor R, Jackson B, Papanikolopoulos N. Vision-based human tracking and activity recognition. In: Proc. of the 11th Mediterranean Conf. on Control and Automation; 2003.
14. Castillo JC, Castro-González Á, Fernández-Caballero A, Latorre JM, Pastor JM, Fernández-Sotos A, et al. Software architecture for smart emotion recognition and regulation of the ageing adult. Cogn Comput. 2016;8(2):357–67. https://doi.org/10.1007/s12559-016-9383-y.
15. Han J, Zhang D, Cheng G, Guo L, Ren J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans Geosci Remote Sens. 2015;53(6):3325–37. https://doi.org/10.1109/TGRS.2014.2374218.
16. Han J, Zhang D, Hu X, Guo L, Ren J, Wu F. Background prior-based salient object detection via deep reconstruction residual. IEEE Trans Circuits Syst Video Technol. 2015;25:1309–21.
17. Davis JW, Keck MA. A two-stage template approach to person detection in thermal imagery. In: null; 2005. pp. 364–369.
18. Davis JW, Sharma V. Robust background-subtraction for person detection in thermal imagery. IEEE Int. Wkshp. on Object Tracking and Classification Beyond the Visible Spectrum; 2004.
19. Davis JW, Sharma V. Robust detection of people in thermal imagery. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on; 2004; pp. 713–716.
20. Kim D-E, Kwon D-S. Pedestrian detection and tracking in thermal images using shape features. In: Ubiquitous Robots and Ambient Intelligence (URAI), 2015 12th International Conference on; 2015. pp. 22–25.
21. Godbehere AB, Matsukawa A, Goldberg K. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In: American Control Conference (ACC), 2012; 2012. pp. 4305–4312.
22. Bloisi D, Iocchi L. Independent multimodal background subtraction. In: CompIMAGE; 2012. pp. 39–44.
23. Zhao Z, Bouwmans T, Zhang X, Fang Y. A fuzzy background modeling approach for motion detection in dynamic backgrounds. In: Multimedia and signal processing, ed: Springer; 2012. pp. 177–185.
24. Zeng J, Xie L, Liu Z-Q. Type-2 fuzzy Gaussian mixture models. Pattern Recogn. 2008;41(12):3636–43. https://doi.org/10.1016/j.patcog.2008.06.006.
25. Li S. Markov random field models in computer vision. Computer Vision—ECCV'94; 1994. pp. 361–370.
26. Noh S, Jeon M. A new framework for background subtraction using multiple cues. In: Asian Conference on Computer Vision; 2012. pp. 493–506.
27. St-Charles P-L, Bilodeau G-A. Improving background subtraction using local binary similarity patterns. In: Applications of Computer Vision (WACV), 2014 I.E. Winter Conference on; 2014. pp. 509–515.
28. St-Charles P-L, Bilodeau G-A, Bergevin R. Flexible background subtraction with self-balanced local sensitivity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2014; pp. 408–413.
29. Tu Z, Zheng A, Yang E, Luo B, Hussain A. A biologically inspired vision-based approach for detecting multiple moving objects in complex outdoor scenes. Cogn Comput. 2015;7(5):539–51. https://doi.org/10.1007/s12559-015-9318-z.
30. Zheng A, Xu M, Luo B, Zhou Z, Li C. CLASS: Collaborative Low-Rank and Sparse Separation for moving object detection. Cogn Comput. 2017;9(2):180–93. https://doi.org/10.1007/s12559-017-9449-5.
31. Wang Y, Zhao Q, Wang B, Wang S, Zhang Y, Guo W, et al. A real-time active pedestrian tracking system inspired by the human visual system. Cogn Comput. 2016;8(1):39–51. https://doi.org/10.1007/s12559-015-9334-z.
32. Han J, Pauwels EJ, de Zeeuw P. Fast saliency-aware multi-modality image fusion. Neurocomputing. 2013;111:70–80. https://doi.org/10.1016/j.neucom.2012.12.015.
33. Han J, Pauwels EJ, De Zeeuw P. Visible and infrared image registration in man-made environments employing hybrid visual features. Pattern Recogn Lett. 2013;34(1):42–51. https://doi.org/10.1016/j.patrec.2012.03.022.
34. De Zeeuw PM, Pauwels EJEM, Han J. Multimodality and multiresolution image fusion. In: VISAPP 2012-Proceedings of the International Conference on Computer Vision Theory and Applications; 2012. pp. 151–157.
35. Otsu N. A threshold selection method from gray-level histograms. Automatica. 1975;11:23–7.
36. Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on; 2000. pp. 142–149.
37. KaewTraKulPong P, Bowden R. An improved adaptive background mixture model for real-time tracking with shadow detection. In: Video-based surveillance systems, ed: Springer; 2002. pp. 135–144.
38. Stauffer C, Grimson WEL. Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on; 1999. pp. 246–252.