Neurocomputing 371 (2020) 27–38

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-modal foreground detection via inter- and intra-modality-consistent low-rank separation

Aihua Zheng[a,b], Naipeng Ye[a], Chenglong Li[a,b,c,*], Xiao Wang[a], Jin Tang[a,b]

[a] School of Computer Science and Technology, Anhui University, Hefei, China
[b] Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, Hefei, China
[c] Institute of Physical Science and Information Technology, Anhui University, Hefei, China

ARTICLE INFO

Article history:
Received 24 January 2019
Revised 30 July 2019
Accepted 5 August 2019
Available online 6 September 2019

Communicated by Dr C. Chen

Keywords:
Soft cross-modal consistency
Foreground detection
Low-rank separation
Local-global appearance consistency

ABSTRACT

Multi-modal foreground detection, which integrates multiple complementary data like visible and thermal infrared sources for moving object detection, has received more and more attention recently. In this paper, we propose a novel **M**ulti-modal **F**oreground **D**etection approach that pursues the inter- and intra-modality consistencies in a unified **L**ow-rank and **S**parse separation model called MFDLS. In particular, we first introduce a soft cross-modal constraint to pursue the inter-modal consistency among different modalities, while allowing sparse inconsistency to model their heterogeneous properties. Second, in addition to the conventional local appearance consistency within each modality, we further propose to preserve the global appearance consistency via Gaussian Mixture Model as the intra-modality consistency. Extensive experimental results yield to a new state-of-the-art comparing to the prevalent multi-modal foreground detection methods.

1. Introduction

Foreground detection aims to extract moving objects from background in a video segment. It plays an important and fundamental role in computer vision due to its potential applications, such as behavior analysis, video surveillance, visual object tracking [1], pedestrian detection and other application scenarios in [2,3]. Representative works on foreground detection include Gaussian Mixture Models (GMM) [4], ViBe [5], and multiple features based methods [6], *etc*. More development on foreground detection (background subtraction) can refer to comprehensive surveys [7–9]. Despite of decades of efforts [10,11], it still suffers from many challenges, such as complicated background, low illumination, bad weather, *etc*.

Low-rank and sparse separation, which decomposes a video sequence matrix into low-rank background matrix and sparse foreground matrix, has attracted increasing attention for foreground detection [12,13]. In the past years, many progress has been made for single-mode foreground detection based on low-rank and sparse separation [14–16]. However, single-model foreground detection still faces the above challenging issues.

Recently, some works introduced the heterogeneous thermal infrared data as a supplementary source to boost the robustness of foreground detection. Different from the visible images where each pixel represents the color of the object appearance captured by the electromagnetic specturm that can be perceived by human eyes, the pixels on the infrared images mainly depends on the emissivity and temperature distribution of the object. They are heterogeneous with distinctive properties in context, texture and color space, which can complement each other. Li et al. [17] proposed to learn the share foreground mask in multi-modal foreground detection based on a weighted low-rank and sparse separation approach to adaptively fuse the data from different modalities. Yang et al. [18] designed a fast foreground detection method by collaboratively separating and integrating the foregrounds in thermal and visible modalities. However, the hard consistency of sharing a foreground mask between different modalities [17] might be overstrict, since different modalities are heterogeneous and thus with distinctive properties. As demonstrated in Fig. 1(a), the pedestrian obscured in visible modality will be compensated by the source/image from the thermal modality. And the visible one, which was greatly disturbed by lamplight as demonstrated in Fig. 1(b), was less affected in thermal modality.

Moreover, some works successfully explored the intra-modal appearance consistency in foreground detection [19,20], in which the local appearance consistency was introduced to improve detection performance based on the assumption that neighboring pixels

* Corresponding author at: School of Computer Science and Technology, Anhui University, Hefei, China.
E-mail address: lcl1314@foxmail.com (C. Li).

https://doi.org/10.1016/j.neucom.2019.08.087
0925-2312/

Fig. 1. Illustration of the heterogeneous foregrounds with different sizes in different modalities. The first and the second rows indicate the visible and thermal image pairs from GTFD dataset.

tended to have a similar appearance. However, such local consistency lacked of global information and thus might be easily disturbed by noises. In addition, existing low-rank and sparse separation methods assume the moving objects as sparse outliers, which might limit their performance when moving objects are relatively in large size such as in Fig. 1(c).

According to the above observations, we project an innovative multi-modal foreground detection method to simultaneously capture cross-modal consistency among the heterogeneous modalities and local-global appearance consistency within each modality in a single low-rank separation framework. Given the two accumulative matrices of the videos from both visible and thermal modalities, we first assume that the potential background along the sequential frames in each modality are linearly correlated while the foregrounds or outliers are generally sparse. Within each modality, we first incorporate the local appearance consistency constraint among the neighborhood pixels and further enforce the global appearance consistency to improve the robustness to noises and sparse assumption on foreground objects. In different modalities, to take both collaboration and heterogeneity into account, we propose the soft cross-modal consistency to make foreground mask consistent while allowing the sparse inconsistency. The all consistent constraints are integrated into a unified low-rank separation model (as shown in Fig. 2). Finally, we jointly optimize the proposed multi-modal foreground detection model via an efficient solver to simultaneously separate the background models and the foreground masks which are heterogeneous in distinctive modalities.

The major contributions of this work can be summarized as:

- We propose an effective method to integrate complementary yet heterogeneous source data in multi-modal foreground detection. More experiments demonstrate the effectiveness of the proposed model against the state-of-the-art multi-modal foreground detection methods.
- We propose to introduce a soft consistency to capture both collaboration and heterogeneity among different modalities for multi-modal foreground detection.
- We integrate the global appearance consistency model by the GMM within each modality to improve the robustness to noises and sparse assumption on foreground objects.

The preliminary version of this work can be referred to [21], where we proposed to integrate soft cross-modal consistency and local appearance consistency into the low-rank separation model to capture both collaboration and heterogeneity among different modalities. Based on the observation that existing low-rank and sparse separation methods present limited performance on larger size foregrounds due to the assumption that the foregrounds are generally sparse, in this work, we further consider the global appearance consistency in each modality for detecting moving objects with diverse sizes and with higher robustness to the noises. Moreover, more experiments have been implemented to verify the effectiveness of our model compared to the preliminary version.

The rest parts of this paper are organized as follows. In Section 2, the related work to our MFDLS is introduced. We detail the MFDLS and the proposed constraints in Sections 3. The experimental results and analysis betweening our method and the state-of-the-art methods are shown in Sections 4. Section 5 concludes this paper.

2. Related works

Background subtraction is the most commonly used method for foreground detection, while low-rank and sparse separation is one of the representative framework in background subtraction due to its robust to noises. We focus on multi-modal foregrund detection in the low-rank and sparse separation framework. Therefore, we briefly discuss the state-of-the-art literatures on background subtraction, low-rank and sparse separation and multi-modal foreground detection as the related works.

Background subtraction. The key task of background subtraction is to establish a solid background model [22]. Background subtraction has been extensively usaged since the 1990s and primarily for moving object detection. Representative approaches consist of mixture of Gaussian [4], the variations on Gaussian distribution [23,24], and other models [25,26]. Ali et al. [24] utilized the Gaussian components to model the intensity value of pixel blocks based on the dynamic learning rate. Chen et al. [23] proposed a hierarchical superpixel segmentation method based on the optical flow and spanning trees based GMM. Hofmann et al. [25] designed a method that treats decision threshold and

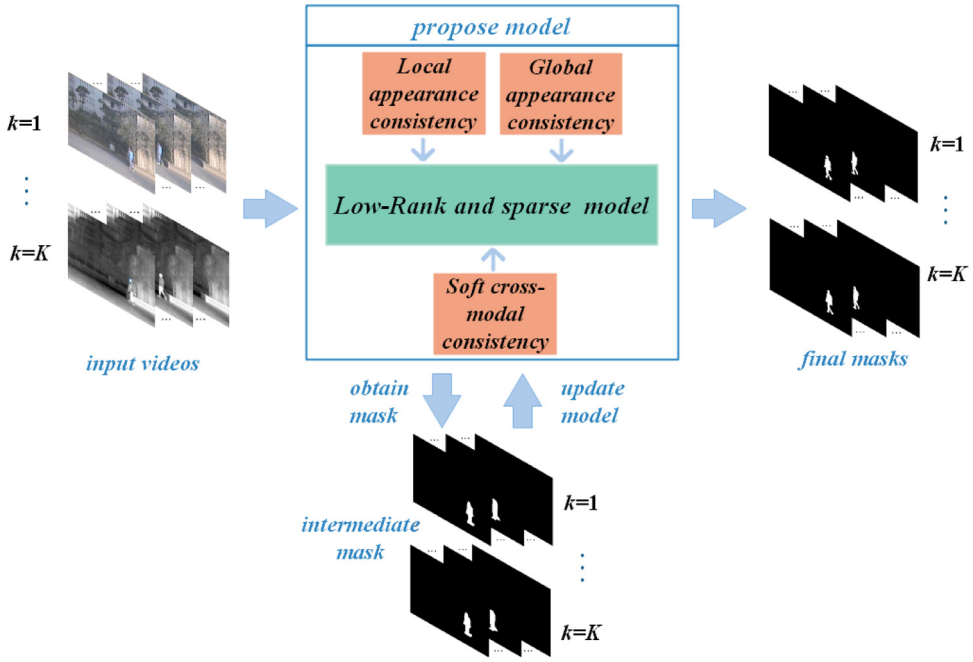


Fig. 2. The diagram of the proposed MFDLS, where k indicates the number of modality of the input video.

randomness parameters of all pixels as adaptive state variables. Zhong et al. [26] integrated based pixel level method and based object level method for foreground detection. However, these methods independently modeled the background for each pixel on each frame, while ignoring the relationship between continuous frames, therefore the sensitive to noise and occlusion.

Recently, regarding foreground detection as a classification task, deep learning methods [27] have gained increasing attention for detecting the object from video sequences. For instance, Zeng et al. [28] proposed a fully convolutional network structure that used the results of foreground detection of different background subtraction algorithms as input. Instead of the traditional background model strategy, some methods [27] calculated the probability of the foreground for each pixel. Braham and Van Droogenbroeck [27] designed a CNN model to learn the spatial features and then subtracted the background from an input image patch. Meanwhile, some deep learning based approaches [29] regarded the foreground detection as a segmentation problem, which directly predicted the category (foreground/background) of each pixel via deep learning networks (i.e., full convolution neural networks) for a single input image. Even if deep learning gives significant improvements in foreground detection, it presents the drawback to be supervised requiring hand labeled images for training. On the contrary, low-rank representation methods (like the proposed method) are unsupervised.

Low-rank and sparse separation. Low-rank and sparse background modeling aims to detect the foreground by decomposing the correlated background from the sequential frames in the low-rank subspace. Robust Principal Component Analysis (RPCA) [30] first devoted to the low-rank and sparse background modeling. Principal Component Pursuit (PCP) [31] proposed to recover the low-rank background and the sparse foregrounds individually via a convex program. Zhou et al. [10] further proposed to enforce a spatial consistency constraint into the low-rank representation framework to encourage the spatial smoothness inside the foregrounds. Xin et al. [20] proposed to regularize the foreground and background by the generalized fused lasso into the low-rank representation model.

The main challenge of applying the low-rank and sparse decomposition to foreground detection are to have online/incremental algorithms and spatio-temporal algorithms. As for the online/incremental algorithms, He et al. [32] proposed an online method based robust subspace estimation from randomly subsampled data for faster separation rate. MERoP [33] presented a provably correct algorithm to achieve the online fashion for both the static and dynamic RPCA. Guo et al. [34] indicated that the time sequences of sparse vectors and dense vectors can be recovered by their sum when the slowly changing background is separated from the moving foreground object. Rodriguez and Wohlberg [35] proposed a new incremental fashion for PCP with comparable performance to the batch PCP. pROST [36] used the ℓ_p-norm as a convex relaxation of the sparse function to identify sparse foreground objects and a framework which efficient alternating online optimization.

In order to make use of the spatial-temporal information during the videos, Sobral et al. [37] used the spatial and temporal saliency detector to build shape constraint and confidence map for constraining the spare component. Javed et al. [38] used the max-norm constraint to obtain initial foregrounds followed by efficient GFL (Generalized Fussed Lasso) to acheive the background subtraction in the online fashion on superpixel. Ebadi et al. [39] proposed to handle the camera motion method by inducing the structured sparsity incorporating the spatial prior. Javed et al. [40] investigated an online graph regularization spatiotemporal RPCA algorithm by saving low-rank spatiotemporal information in a dual spectral graphs.

Multi-modal foreground detection. Multi-modal data has attracted increasing attention in the community [41] with the rapid development of various sensors, such as thermal infrared and depth sensors. Researchers started to develop the foreground detection with multi-modal resources to compensate for the limitations of the data from single modality [17,18]. Nadimi and Bhanu [42] provided a physical model to combine the external conditions with the visible and infrared data into a globally consistent dynamic representation. Becker et al. [43] investigated the strategy of fusing infrared and visible video streams from a

vibrating camera. Bouwmans et al. [44] conducted a comprehensive study on the types and sizes of features used in background modeling and foreground detection, as well as its inherent spectrum, spatial and temporal characteristics.

Some literature [43,44] investigated the fusion of visible and infrared spectrum and analyze the role of features in background modeling. Conaire et al. [41] used infrared and visible features to remove incorrectly detected foreground regions to model a robust background. Li et al. [17] proposed a weighted low-rank decomposition (WELD) to adaptively fuse the variables from each modality with the shared foreground mask for multi-modal foreground detection. Yang et al. [18] designed a collaborative low-rank decomposition (CLoD), which efficiently detected the foreground under thermal and visible modalities by incroporating the multi-modal separation. However, most existing methods ignore the global appearance consistency.

3. Proposed methodology

Our Multi-modal Foreground Detection via inter- and intra-modality consistent Low-rank Separation MFDLS is in a batch processing manner.

3.1. Problem formulation

As for the k-th modal video sequences $k = 1, \ldots, K$, we first reshape each frame into a column of vectors, and then accumulate the video sequence into a $m \times n$ matrix, *i.e.*, $\mathbf{D}^k = [\mathbf{d}_1^k, \mathbf{d}_2^k, \ldots, \mathbf{d}_n^k] \in R^{m \times n}$, where m and n indicate the total number of pixels on each frame and the number of frames of the video respectively. Herein, we use grayscale-thermal data, therefore $K = 2$. One can easily extend our model to larger K since on one hand, we formulate our model in a general fashion by using an arbitrary number K as the number of modalities. On the other hand, when the inputs are more than 2 modalities, the proposed soft cross-modal consistency could indeed handle such scenarios to achieve collaborative fusion of all modalities. We assume that the foregrounds are contiguous and sparse and the background along the sequential frames in each modality are linearly correlated, which has been widely employed in existing works [10,45]. Therefore, our ultimate task is to separate the input multi-modal video sequences/matrices $\mathbf{D}^k, k = 1, 2, \ldots, K$ into the contiguous and sparse foreground masks $\mathbf{S}^k$ and the low-rank backgrounds $\mathbf{B}^k \in R^{m \times n}$ for each modality as:

$$\begin{aligned} &\min_{\mathbf{B}^k, \mathbf{S}^k} \frac{1}{2} ||f_{\mathbf{S}_\perp^k}(\mathbf{D}^k - \mathbf{B}^k)||_F^2 + \beta ||vec(\mathbf{S}^k)||_0, \\ &s.t.\ rank(\mathbf{B}^k) \le r^k,\ \ k = 1, 2, \ldots, K, \end{aligned} \tag{1}$$

where $vec(\cdot)$ converts a matrix into a vector. β is a tuning parameter. $||\cdot||_0$ and $||\cdot||_F$ represent the l_0 norm of a vector and the Frobenius norm of a matrix, respectively. r^k is a constant denoting the rank of a matrix, which controls the correlation of the background matrix in the k-th modality. Herein, $\mathbf{D}^k = \mathbf{B}^k + \mathbf{S}^k$. $f_{\mathbf{S}}(\mathbf{X}^k)$ orthogonally project the matrix $\mathbf{X}^k$ to the linear space based on the support matrix $\mathbf{S}^k$:

$$f_{\mathbf{S}}^k(\mathbf{X}^k)(i,j) = \begin{cases} 0, & \mathbf{S}_{ij} = 0, \\ \mathbf{X}_{ij}^k, & \mathbf{S}_{ij} = 1. \end{cases} \tag{2}$$

$f_{\mathbf{S}_\perp^k}(\mathbf{X}^k)$ is the complementary projection, *i.e.*, $f_{\mathbf{S}}^k(\mathbf{X}^k) + f_{\mathbf{S}_\perp^k}(\mathbf{X}^k) = \mathbf{X}^k$. $\mathbf{S}_{i,j}^k$ implies the binary foreground indications:

$$\mathbf{S}_{i,j}^k = \begin{cases} 1, & \text{if } ij \text{ is foreground}, \\ 0, & \text{if } ij \text{ is background}. \end{cases} \tag{3}$$

Local appearance consistency. In order to maintain the spatial smoothness of foregrounds, it is necessary to encourage adjacent pixels to have similar appearance, which is called intra-modal appearance consistency in this paper. $||vec(\mathbf{S}^k)||_0$ and the intra-modal appearance consistency on $\mathbf{S}$ can be regarded as the unary term and pairwise term of Markov Random Field (MRF) which can be solved by graph cuts algorithm [46] in the same manner as [10]. Followed by [19,20], we construct the adaptive weights $w^k{}_{ij,\,pq}$ into the smoothness to enforce the local appearance consistency:

$$\begin{aligned} &||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1 = \sum_{(ij,kl)\in\varepsilon^k} w^k{}_{ij,pq}\ |\mathbf{S}_{ij}^k - \mathbf{S}_{pq}^k|; \\ &w^k{}_{ij,pq} = \exp\frac{-||d^k{}_{ij} - d^k{}_{pq}||_2^2}{2\theta^2}. \end{aligned} \tag{4}$$

where $||\mathbf{X}||_1 = \sum_{ij} |\mathbf{X}_{ij}|$ indicates the l_1-norm. $\mathbf{C}^k$ denotes an adjacency matrix, which represents the degree of relationship between pixel nodes in the k-th modality and ε^k denotes the edge connection among the spatial neighborhood pixels in the k-th modality, θ is a tunning parameter and $d^k{}_{ij}$ and $d^k{}_{pq}$ represent respectively the values of pixel ij and pq in the k-th modality. Then, our model incorporates this appearance consistency can be expressed as follows:

$$\begin{aligned} &\min_{\mathbf{B}^k, \mathbf{S}^k} \frac{1}{2} ||f_{\mathbf{S}_\perp^k}(\mathbf{D}^k - \mathbf{B}^k)||_F^2 + \beta ||vec(\mathbf{S}^k)||_0 + \mu ||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1, \\ &s.t.\ rank(\mathbf{B}^k) \le r^k,\ \ k = 1, 2, \ldots, K, \end{aligned} \tag{5}$$

where μ is a balance parameter for local appearance consistency.

Global appearance consistency. Due to the assumption of sparseness on moving objects, Eq. (1) is theoretically insufficient to detect the foregrounds with lage size. From the perspective of statistics, we further assume that the backgrounds and foregrounds in each modality are with Gaussian distribution, which has been successfully and widely applied in foreground detection [14]. Therefore, we further enforce a global appearance consistency to compete the sparse term and thus to improve the capability of detecting the foreground with large sizes. The global interaction between the appearance model of the foreground and the background can be written as:

$$\sum_{k=1}^{K} \sum_{l=0}^{1} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta(l, \mathbf{S}_{ij}^k) \mathbf{A}_l^k(i,j) = \sum_{k=1}^{K} \sum_{l,i,j} \delta(l, \mathbf{S}_{ij}^k) \mathbf{A}_l^k(i,j) \tag{6}$$

where $\mathbf{A}_0^k$ and $\mathbf{A}_1^k$ are the Gaussian Mixture Models of background and foreground respectively, which are employed to simulate the distribution of background and foreground pixels in the k-th modality. $\delta(l, \mathbf{S}_{ij}^k)$ is the Dirac delta function that represents the $\mathbf{S}_{ij}^k$ related with the value of l. The $\mathbf{A}_l^k(i,j)$ is a unary potential that assesses the possibility of pixel i becoming a foreground or background based on the appearance model of frame j.

Soft cross-modal consistency. Unlike existing multi-modal foreground detection methods [17,18], we further propose to enforce the soft cross-modal consistency among the multi-spectral data, which models the interdependency between two modalities while allowing the sparse inconsistency for their heterogeneous properties. The soft cross-modal constraint is formulated as:

$$\sum_{k=2,(ij)\in\mathcal{F}}^{K} |\mathbf{S}_{ij}^k - \mathbf{V}_{ij}^{k-1}|, \tag{7}$$

where $\mathcal{F}$ is the set of all the pixels as vertices in each modality (as shown in Fig. 2), and $\mathbf{V}_{ij}$ denotes the set of four neighbors corresponding to the pixel $\mathbf{S}_{ij}^k$ in another modality. Thus $\mathbf{V}_{ij}^{k-1}$ is defined as $[\ \mathbf{S}_{ij}^{k-1}, \mathbf{S}_{(i+1)j}^{k-1}, \mathbf{S}_{(i)(j+1)}^{k-1}, \mathbf{S}_{(i+1)(j+1)}^{k-1}]$. Eq. (7) urges the pixel $\mathbf{S}_{ij}^k$ and its cross-modal neighbors $\mathbf{V}_{ij}^{k-1}$ possessing to the similar property, which is helpful to improving robustness of cross-modal foreground detection. Note that we allow sparse inconsistency in different modalities to account for their heterogeneous properties,

so we use the absolute value instead of square value to employ the sparse properties.

Therefore, taking all above considerations together, our model can be summarized as:

$$\min_{\mathbf{B}^k,\mathbf{S}^k} \frac{1}{2}||f_{\mathbf{S}_\perp^k}(\mathbf{D}^k-\mathbf{B}^k)||_F^2+\ \beta||vec(\mathbf{S}^k)||_0+\mu||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1 + \gamma\sum_{k=2,(ij)\in\mathcal{F}}^{K}|\mathbf{S}_{ij}^k-\mathbf{V}_{ij}^{k-1}|+\rho\sum_{k=1}^{K}\sum_{l,i,j}\delta(l,S_{ij}^k)\mathbf{A}_l^k(i,j) \quad (8)$$

$$s.t.\ rank(\mathbf{B}^k)\le r^k,\ \ k=1,2,\ldots,K.$$

where γ is a tuning parameter for soft cross-modal consistency.

Eq. (8) is a NP-hard problem. We utilize the nuclear norm to make Eq. (8) more convenient to calculate, which has demonstrated to be an effective convex surrogate of the rank operator [47] to relax the constraints of rank operator on $\mathbf{B}^k$. Meanwhile, we spliced matrices of different modalities together for collaborative optimization. The final model of our propose is:

$$\min_{\mathbf{B},\mathbf{S}^k}\sum_{k=1}^{K}\frac{1}{2}||f_{\mathbf{S}_\perp^k}(\mathbf{D}^k-\mathbf{B}^k)||_F^2+\ \beta||vec(\mathbf{S}^k)||_0+\lambda||\mathbf{B}^k||_* +\mu||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1+\gamma\sum_{k=2,(ij)\in\mathcal{F}}^{K}|\mathbf{S}_{ij}^k-\mathbf{V}_{ij}^{k-1}| 1+\rho\sum_{k=1}^{K}\sum_{l,i,j}\delta(l,S_{ij}^k)\mathbf{A}_l^k(i,j) \quad (9)$$

where $||\cdot||_*$ denotes the nuclear norm of a matrix and λ is a balance constant.

3.2. Optimization

The objective function defined in Eq. (9) is a non-convex function, including continuous and discrete variables. The joint optimization of B and S is extremely difficult. Hence, we usaged the alternating algorithm to decompose the minimization of B and S into two subproblems. B-subproblem is a convex optimization problem and S-subproblem is a combinatorial optimization problem. Experiments show that the optimal solutions of B-subproblem and S-subproblem can be computed efficiently.

Problem of Eq. (9) can be transformed into the following two subproblems:

B – *subproblem*. Fixing the current estimation of the foreground mask $\hat{\mathbf{S}}^k$, we first minimize Eq. (9) to estimate $\mathbf{B}^k$:

$$\min_{\mathbf{B}^k}\sum_{k=1}^{K}\frac{1}{2}||f_{\hat{\mathbf{S}}_\perp^k}(\mathbf{D}^k-\mathbf{B}^k)||_F^2+\lambda\parallel\mathbf{B}^k\parallel_* \quad (10)$$

Here we learn a low-rank background matrix, and utilize SOFT-IMPUTE [48] algorithm to estimate $\mathbf{B}^k$ by iteratively employing Eq. (11) and Θ_λ means the singular value thresholding.:

$$\hat{\mathbf{B}}^k \longleftarrow \Theta_\lambda\left(f_{\hat{\mathbf{S}}_\perp^k}(\mathbf{D}^k)+f_{\hat{\mathbf{S}}^k}(\hat{\mathbf{B}}^k)\right) \quad (11)$$

S – *subproblem*. Fixing the current estimation of the background position matrix $\hat{\mathbf{B}}^k$, Eq. (9) can be transferred into following optimization function:

$$\min_{\mathbf{S}^k}\sum_{k=1}^{K}\frac{1}{2}\ ||f_{\mathbf{S}_\perp^k}(\mathbf{D}^k-\hat{\mathbf{B}}^{\mathbf{k}})||_F^2+\ \beta||vec(\mathbf{S}^k)||_0 +\mu||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1+\gamma\sum_{k=2,(ij)\in\mathcal{F}}^{K}|\mathbf{S}_{ij}^k-\mathbf{V}_{ij}^{k-1}| +\rho\sum_{k=1}^{K}\sum_{l,i,j}\delta(l,S_{ij}^k)\mathbf{A}_l^k(i,j) \quad (12)$$

Algorithm 1 Optimization Process to Eq. (9).

Input: $\mathbf{D}^k$, $(k=1,\ldots,K)$.
Set $\mathbf{S}^k=\mathbf{0}$, $\mathbf{B}^k=\mathbf{D}^k$, $maxIter=28$, $\epsilon=1e-4$.
Output: $\mathbf{S}^k$, $\mathbf{B}^k$.

```
for i = 1 : maxIter do
    Update B^k according to Eq. (10);
    if rank(B̂^k) ≤ r^k then
        adjust parameter λ, go to step 2.
    end if
    Update {S^k} according to Eq. (12);
    Convergence condition: if the number of the interation
    reaches maxIter or the absolute difference betweentwo
    continuous iterations is smaller than ϵ, then break the loop.
end for
```

The Eq. (12) can be calculated in a standard format of a first-order Markov Random Fields [49] as:

$$\begin{aligned}\min_{\mathbf{S}^k}&\frac{1}{2}\sum_{k=1}^{K}\sum_{ij}\left(\mathbf{D}_{ij}^k-\hat{\mathbf{B}}_{ij}^k\right)^2\left(1-\mathbf{S}^k{}_{ij}\right)+\beta\sum_{ij}\mathbf{S}_{ij}^k\\&+\mu||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1+\gamma\sum_{k=2,(ij)\in\mathcal{F}}^{K}|\mathbf{S}_{ij}^k-\mathbf{V}_{ij}^{k-1}|\\&+\rho\sum_{k=1}^{K}\sum_{l,i,j}\delta(l,S_{i,j}^k)\mathbf{A}_l^k(i,j)\\=\min_{\mathbf{S}^k}&\sum_{k=1}^{K}\sum_{lij}\left[\beta-\frac{1}{2}\left(\mathbf{D}_{ij}^k-\hat{\mathbf{B}}_{ij}^k\right)^2+\rho\delta(l,S_{i,j}^k)\mathbf{A}_l^k(i,j)\right]\\&+\mu||\mathbf{C}^k\ vec(\mathbf{S}^k)||_1+\gamma\sum_{k=2,(ij)\in\mathcal{F}}^{K}|\mathbf{S}_{ij}^k-\mathbf{V}_{ij}^{k-1}|+\mathscr{S}\end{aligned} \quad (13)$$

where $\mathscr{S}=\frac{1}{2}\sum_{k=1}^{K}\sum_{ij}(\mathbf{D}_{ij}^k-\hat{\mathbf{B}}_{ij}^k)^2$ is a constant as a result of fixed $\hat{\mathbf{B}}^k$. We utilize the graph cuts algorithm [46] to solve Eq. (13) effectively.

In Algorithm 1, we summarize the whole process of alternating optimizate $\{\mathbf{B}^k\}$ and $\{\mathbf{S}^k\}$ detailedly. With the optimal results of each sub-problems, our model achieves the best convergence state.

4. Experiments

Our method is implemented on the mixing platform of C++ and MATLAB without any code optimization on a desktop computer of the Intel i5-7500 3.4 GHz CPU and 16GB RAM.

4.1. Experimental settings

Datasets. We evaluate our method on the public challenging datasets OSU03 [50], GTFD [17] and Cropped GTFD [1] and and comparing to the prevalent multi-modal foreground detection algorithms.

OSU03 dataset consists of 6 thermal and visible sequences pairs of outdoor pedestrains. We extract 100 consecutive frames from each sequence and manually labeled one frame in each 10 consecutive frames with groundtruth for evaluation purpose.

GTFD dataset consists of 25 video sequence pairs captured from both visible and thermal modalities in challenging scenariors, including TC (thermal crossover), IM (intermittent motion), IL (low illumination), BW (bad weather), IS (intense shadow), DS (dynamic scene) and BC (background clutter).

[1] URL: https://github.com/yenaipeng/cropped-GTFD.git.

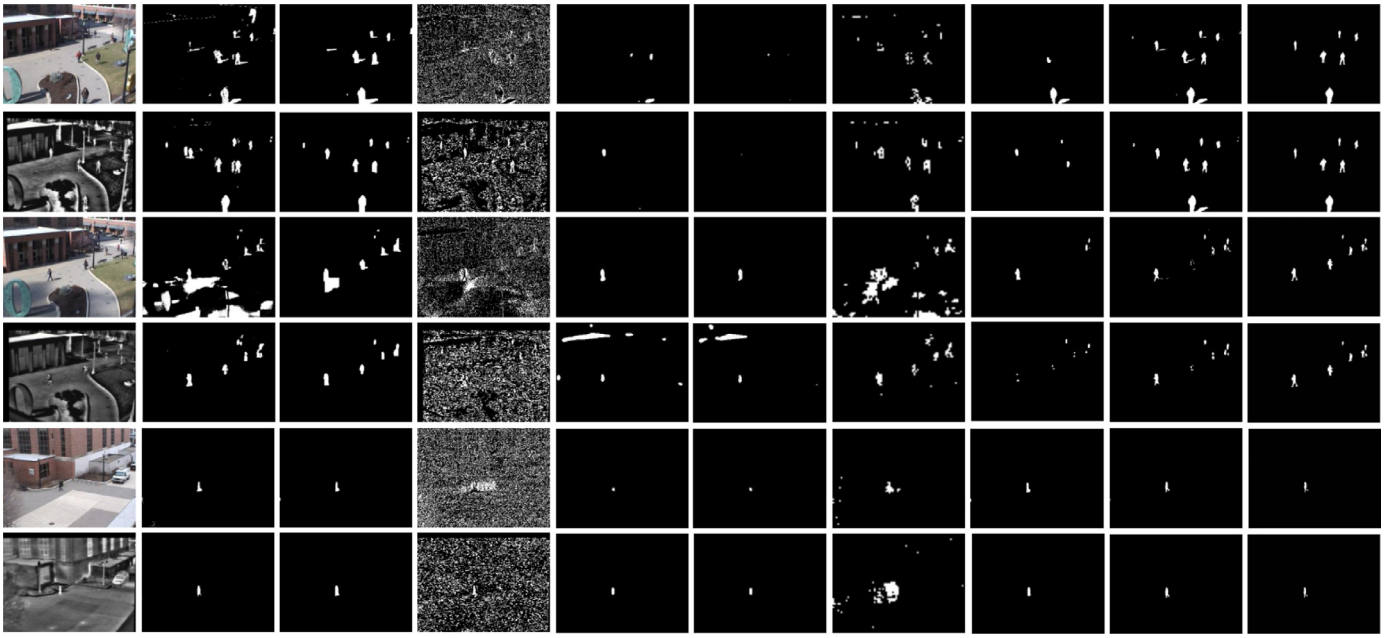


Fig. 3. Qualitative examples of our method against the prevalent methods on the OSU03 dataset. The odd and the even rows illustrate the original frames (the first column) along with the corresponding detected foregrounds (the rest columns) in grayscale modality and thermal modality, respectively.

Cropped GTFD. To justify the effectiveness of our approach on larger size foreground detection, we obtain the Cropped dataset by cropping the active area of 11 video sequence pairs from GTFD dataset. Therefore the foreground objects occupy a larger proportion of entire frames (averagely 10% with largest and least proportion as 32% and 5%) than the original GTFD (approximately 2%). There are no no duplicate video sequences in GTFD dataset the Cropped GTFD dataset.

Evaluation metrics. We evaluate the detecting results using the following three metrics, precision, recall and F-measure (denoting P, R, F, respectively):

$$\mathbf{P} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}, \mathbf{R} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}, \mathbf{F} = \frac{\mathbf{2PR}}{\mathbf{P}+\mathbf{R}} \tag{14}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively.

4.2. Comparison results

To demonstrate the superiority of our model, we conduct the comparison to the prevalent foreground detection algorithms, including grayscale, thermal and thermal-grayscale detection methods. Following the protocols in [17,18], the results on single Grayscale or Thermal modality of the multi-modal methods like WELD [17] or CLoD [18] are achieved by considering the two duplications of each modality as the multi-modal input. Note that the foreground masks from each modality will complement each other during the optimization and we directly adopt the masks from grayscale modality as the final result. Therefore, the grayscale and the thermal sources can mutually promote and complement the detecting during optimization, which is the main advantage of employing both sources on grayscale and thermal data. While WELD [17] or CLoD [18] utilizes a shared mask for each modality and generate the final result by a fusing scheme.

4.2.1. Results on OSU03 dataset

Qualitative results. Fig. 3 illustrates several examples of detecting results on the OSU03 dataset. From which we can see that our method is clearly superior to other state-of-the-art methods with finer details such as the body's posture and less influence from the background.

Quantitative results Table 1 reports the quantitative results of our method against the state-of-the-art methods on precision, recall and F-measure on the OSU03 dataset. We can see that: (1) Our method substantially surpasses the state-of-the-art methods, which verifying the contribution of the proposed local and global appearance consistency. (2) DECOLOR [10] and SAC [45] achieve higher in recall which result from the coarse contour as shown in Fig. 3. Our method achieves much higher F-measure, which claims the comprehensive performance between the precision and recall. (3) Our method is not satisfactory in terms of time consumption due to the more complex graphs which brings more burden to graph cuts. Furthermore, introduing of GMM will also increase the computation time. However, our method is superior in accuracy and each compoment plays important role in our model as shown in Section 4.4. We shall also discuss this limitation in the Section 4.6.

4.2.2. Results on GTFD dataset

Qualitative results. Fig. 4 illustrates several examples of detecting results on GTFD dataset, from which we can see, our method can better detect the foreground mask from both modalities benefited from the soft cross-modal consistency. Furthermore, our method can produce more smooth foreground objects due to the local and global appearance consistency.

Quantitative results. We compare the results of our method with prevalent methods in precision (P), recall (R) and F-measure (F), together with the computational complexity on GTFD dataset in Table 2. It is clear that: (1) Our method substantially surpasses other thermal-grayscale methods in recall, precision and F-measure, proving the validity of the proposed constraints of soft cross-modal consistency. (2) Although CLoD [18] and WELD [17] achieve satisfying performance after fusing the results of grayscale and thermal, they still perform worse than ours. (3) From Table 2, we can see that our method works slightly overshadowed than WELD [17] and comparatively to by DECOLOR [10] in

Table 1
Quantitative comparison of the average Precision (P), Recall (R) and F-measure (F) on the OSU03 dataset. The top three results are highlighted in **red**, *blue* and ***green***, respectively. [6,10,17,18,31,45,51–53]

Algorithm	Grayscale			Thermal			Grayscale-Thermal			FPS
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
PCP [31]	0.02	0.20	0.02	0.03	0.68	0.06	-	-	-	16.45
SAC [45]	0.42	0.91	0.51	0.56	0.88	0.67	-	-	-	1.03
DECOLOR [10]	0.62	0.91	0.72	0.63	0.90	0.73	-	-	-	0.64
PAWCS [51]	0.62	0.25	0.34	0.50	0.15	0.20	-	-	-	1.89
SuBSENSE [6]	0.71	0.27	0.34	0.56	0.31	0.35	-	-	-	1.5
WeSamBE [52]	0.67	0.29	0.36	0.48	0.21	0.35	-	-	-	1.16
LSD [53]	0.30	0.81	0.42	0.30	0.86	0.43	-	-	-	0.24
CLoD [18]	0.18	0.75	0.28	0.25	0.80	0.35	0.19	0.85	0.29	6.76
WELD [17]	0.74	0.65	0.63	0.82	0.50	0.55	0.79	0.61	0.64	0.64
MFDLS(ours)	0.79	0.80	0.79	0.75	0.80	0.76	0.82	0.83	0.81	0.1

Table 2
Quantitative comparison of the average Precision (P), Recall (R) and F-measure (F) on GTFD dataset. The top three results are highlighted in **red**, *blue* and ***green***, respectively. [4–6,10,17,18,31,45,51–59]

Algorithm	Grayscale			Thermal			Grayscale-Thermal			FPS
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
ASOM [54]	0.18	0.07	0.06	0.16	0.07	0.08	-	-	-	111.11
FCFT [55]	0.39	0.20	0.22	0.25	0.22	0.20	-	-	-	38.46
APKV [56]	0.38	0.42	0.36	0.42	0.20	0.24	-	-	-	0.03
ViBe [5]	0.41	0.49	0.41	0.41	0.47	0.39	-	-	-	318.48
TTD [57]	0.59	0.29	0.32	0.58	0.38	0.40	-	-	-	0.07
PCP [31]	0.28	0.18	0.21	0.49	0.40	0.43	-	-	-	20.42
GMM [4]	0.48	0.65	0.52	0.48	0.65	0.50	-	-	-	93.37
SAC [45]	0.42	0.74	0.41	0.47	0.71	0.53	-	-	-	1.15
DECOLOR [10]	0.54	0.84	0.59	0.52	0.82	0.59	-	-	-	1.98
MAMR [58]	0.57	0.67	0.60	0.59	0.63	0.59	-	-	-	3.37
GMM-GT [4]	-	-	-	-	-	-	0.53	0.60	0.53	34.04
JSC [59]	-	-	-	-	-	-	0.17	0.43	0.18	10.21
PAWCS [51]	0.55	0.30	0.33	0.29	0.12	0.15	-	-	-	1.31
SuBSENSE [6]	0.61	0.46	0.46	0.53	0.36	0.37	-	-	-	1.13
WeSamBE [52]	0.63	0.47	0.47	0.50	0.30	0.32	-	-	-	0.83
LSD [53]	0.42	0.79	0.51	0.44	0.78	0.53	-	-	-	1.62
CLoD [18]	0.53	0.71	0.55	0.63	0.62	0.57	0.62	0.80	0.66	45.66
WELD [17]	0.58	0.80	0.64	0.50	0.63	0.50	0.64	0.81	0.67	2.43
MFDLS (ours)	0.57	0.80	0.59	0.57	0.69	0.58	0.66	0.81	0.68	0.5

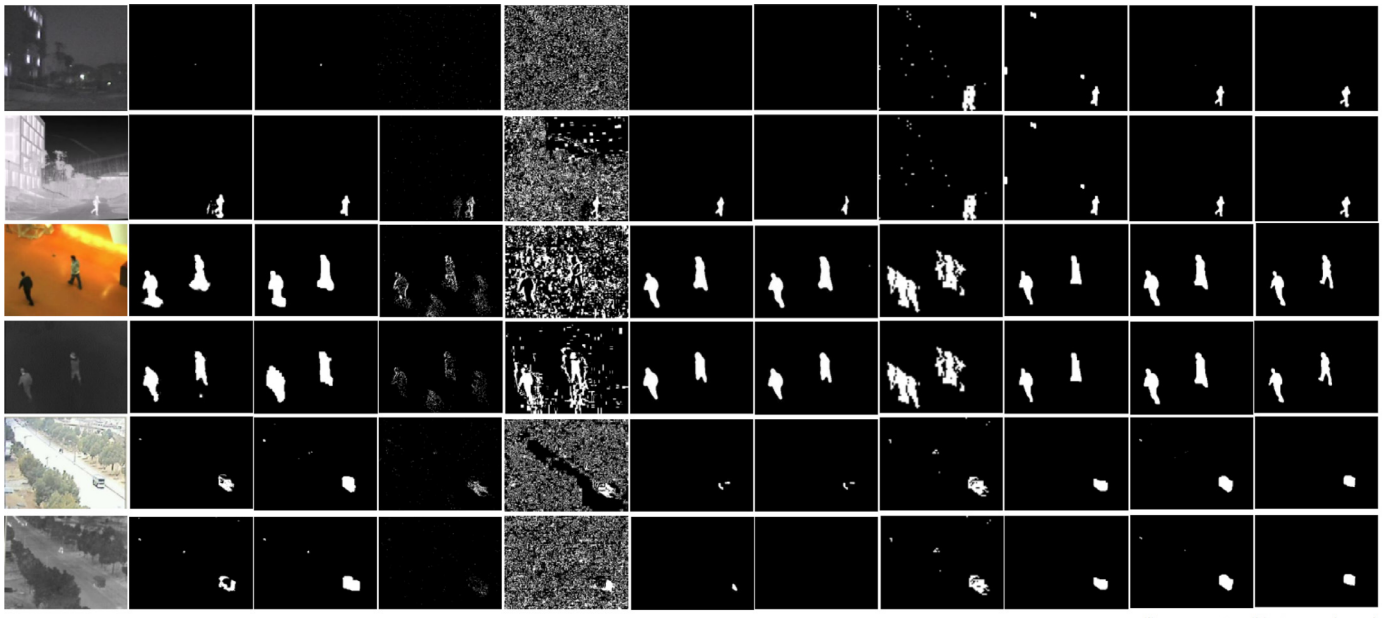


Fig. 4. Qualitative examples of our method against the prevalent methods on GTFD dataset. The odd and the even rows illustrate the original frames (the first column) along with the corresponding detected foregrounds (the rest columns) in grayscale modality and thermal modality, respectively.

Table 3
Quantitative comparison of the average Precision (P), Recall (R) and F-measure (F) on cropped GTFD dataset. The top three results are highlighted in **red**, *blue* and ***green***, respectively. [6,10,17,18,31,45,51–53]

Algorithm	Grayscale			Thermal			Grayscale-Thermal			FPS
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
PCP [31]	0.28	0.18	0.21	0.49	0.40	0.43	-	-	-	20.42
SAC [45]	0.49	0.61	0.44	0.51	0.48	0.41	-	-	-	1.15
DECOLOR [10]	0.48	0.34	0.39	0.46	0.47	0.36	-	-	-	1.98
PAWCS [51]	0.58	0.29	0.33	0.37	0.12	0.15	-	-	-	8.42
SuBSENSE [6]	0.55	0.37	0.38	0.54	0.35	0.37	-	-	-	7.07
WeSamBE [52]	0.57	0.37	0.38	0.57	0.28	0.32	-	-	-	5.96
LSD [53]	0.54	0.32	0.36	0.52	0.38	0.41	-	-	-	1.61
CLoD [18]	0.61	0.35	0.38	0.55	0.33	0.36	0.65	0.35	0.38	16.3
WELD [17]	0.62	0.85	0.68	0.53	0.87	0.63	0.58	0.90	0.68	3.45
MFDLS(ours)	0.60	0.82	0.65	0.62	0.77	0.65	0.66	0.88	0.74	0.3

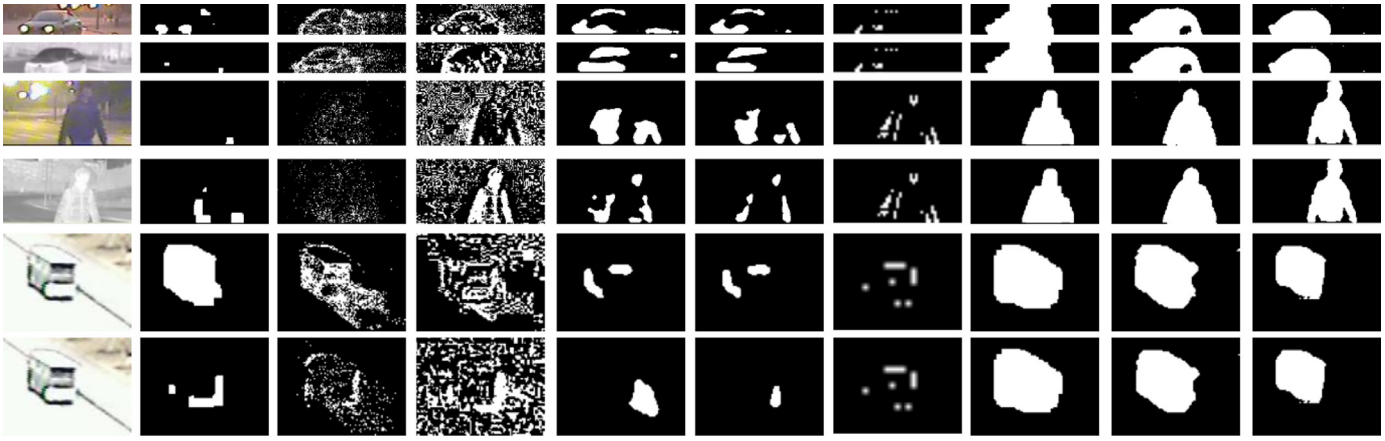


Fig. 5. Qualitative examples of our method against the prevalent methods on cropped GTFD dataset. The odd and the even rows illustrate the original frames (the first column) along with the corresponding detected foregrounds (the rest columns) in grayscale modality and thermal modality, respectively.

F-measure on either Grayscale or Thermal modality, the reason may be the wrong detection on the single modality will suppress the contribution of the proposed soft consistency. However, our MFDLS significantly beat all the state-of-the-art methods on multi-modal (grayscale and thermal) case, which verifies the necessity and effectiveness of multi-modal foreground detection and the proposed model respectively.

4.2.3. Results on cropped GTFD dataset

Qualitative results. Fig. 5 illustrates several examples of detecting results on Cropped GTFD dataset under the challenging of larger size foregrounds. The detection results are consistent to that on GTFD dataset. Our method can better preserve the compact foreground structures and the fine details from both modalities, especially when the other state-of-the-art methods fail to capture the large foregrounds.

Quantitative results. Table 3 reports the quantitative results of our method against the state-of-the-art methods on precision, recall and F-measure on the cropped GTFD dataset. It is clear that: (1) Our method substantially surpasses the state-of-the-art methods. (2) Together with Table 2 we can see that, the improvement of our method is more prominent on the cropped GTFD dataset with larger foregrounds, verifying the contribution of the proposed global appearance consistency. (3) WELD [17] generate coarse a contour of the foregrounds, which results in a higher recall, while our method achieves much higher F-measure, which claims the comprehensive performance between the precision and recall.

4.3. Challenge-based performance

We further evaluate our approach (MFDLS) against the state-of-the-arts on variouse challenges [17] in GTFD dataset to analyze the challenge-sensitive performance as shown in Fig. 6. It is clearly to see that our MFDLS (GT) significantly improves F-measure in the challenging of BW (bad weather), IS (intense shadow), DS (dynamic scene) and BC (background clutter), which verifies the effectiveness of our method while hanlding the challenging senarios. Alghouth our MFDLS (GT) works overshadowed than WELD (GT) [17] in TC (thermal crossover), IM (intermittent motion) and IL (low illumination) senarios, both MFDLS (G) and MFDLS (T) outperforms WELD (G) and WELD (T), which in turn means the better performance of WELD (GT) [17] results from the fusing of WELD (T) and WELD (T). Furthermore, our MFDLS beats the state-of-the-art methods on the whole GTFD dataset with all seven challenges as shown in Table 2, which indicates the generality and robustness of the proposed method.

4.4. Ablation study

We implement the ablation study on our model with three variants on cropped GTFD dataset and report the results in Table 4 and Fig. 7. Specifically, Ours-I, Ours-II and Ours-III indicate the variants of our model of eliminating the soft cross-modal consistency, global appearance consistency and local appearance consistency, by setting γ to 0, ρ to 0 and $w^k_{ij,\,kl}$ to 1 in Eq. (9), respectively. By comparing Ours-I, Ours-II and Ours-III to Ours, respectively, from Table 4, we observe that the significance of the

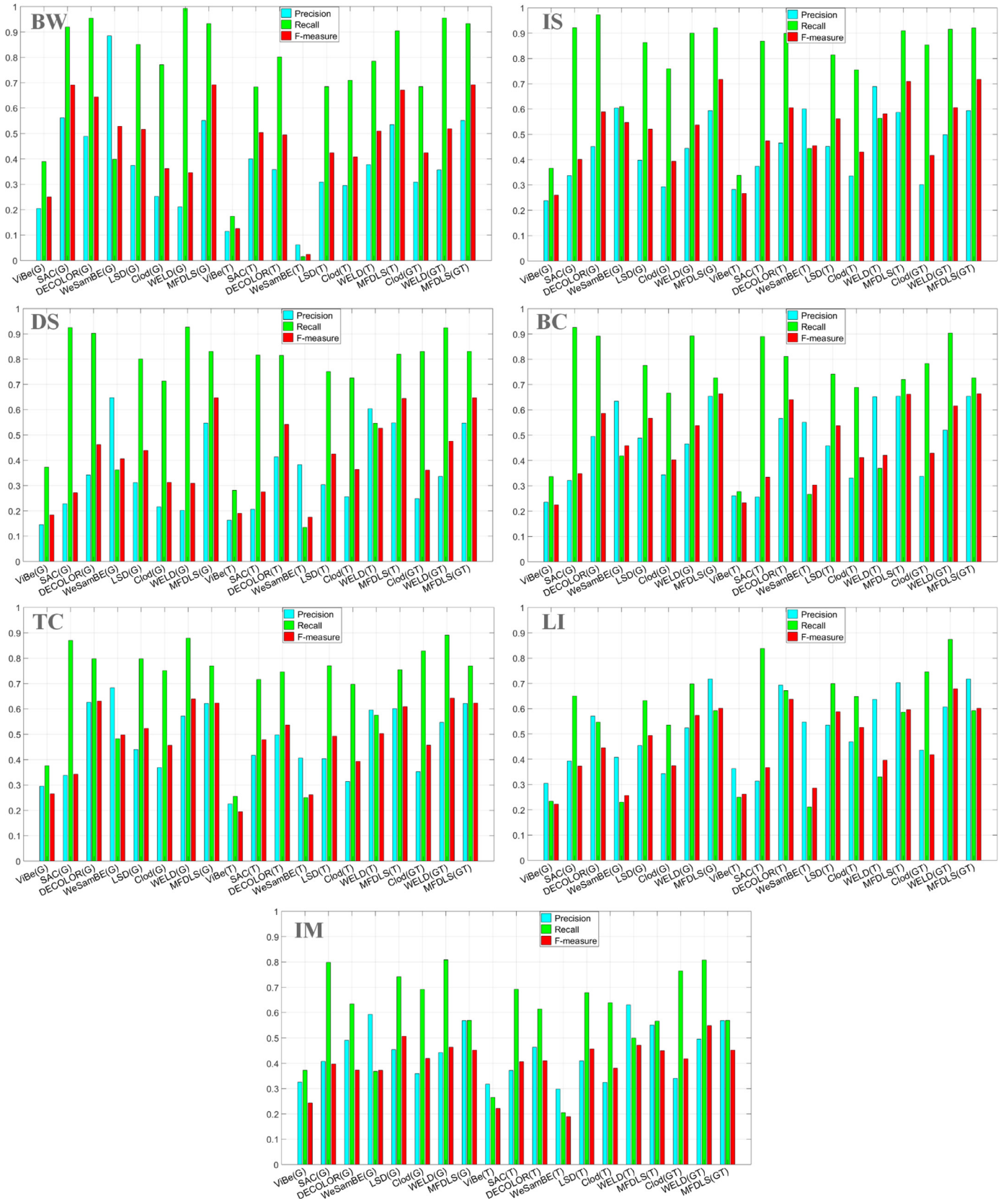


Fig. 6. The comparison of our MFDLS against the state-of-the art methods on the seven challenging scenarios on GTFD dataset, where G, T, and GT denote Grayscale, Thermal and Grayscale-thermal, respectively.

Table 4
Ablation study with different variants by progressively introducing three components including (a) soft cross-modal consistency (CM), (b) global appearance consistency (GA), and (c) local appearance consistency (LA). The top three results are highlighted in **red**, *blue* and ***green***, respectively.

Component		Ours	Ours-I	Ours-II	Ours-III
(a) CM		✓	×	✓	✓
(b) GA		✓	✓	×	✓
(c) LA		✓	✓	✓	×
Metrics					
Grayscale	P	0.659	0.584	0.601	0.657
	R	0.887	0.771	0.713	0.873
	F	0.738	0.626	0.618	0.729
Thermal	P	0.655	0.558	0.602	0.652
	R	0.877	0.648	0.704	0.859
	F	0.732	0.555	0.614	0.720

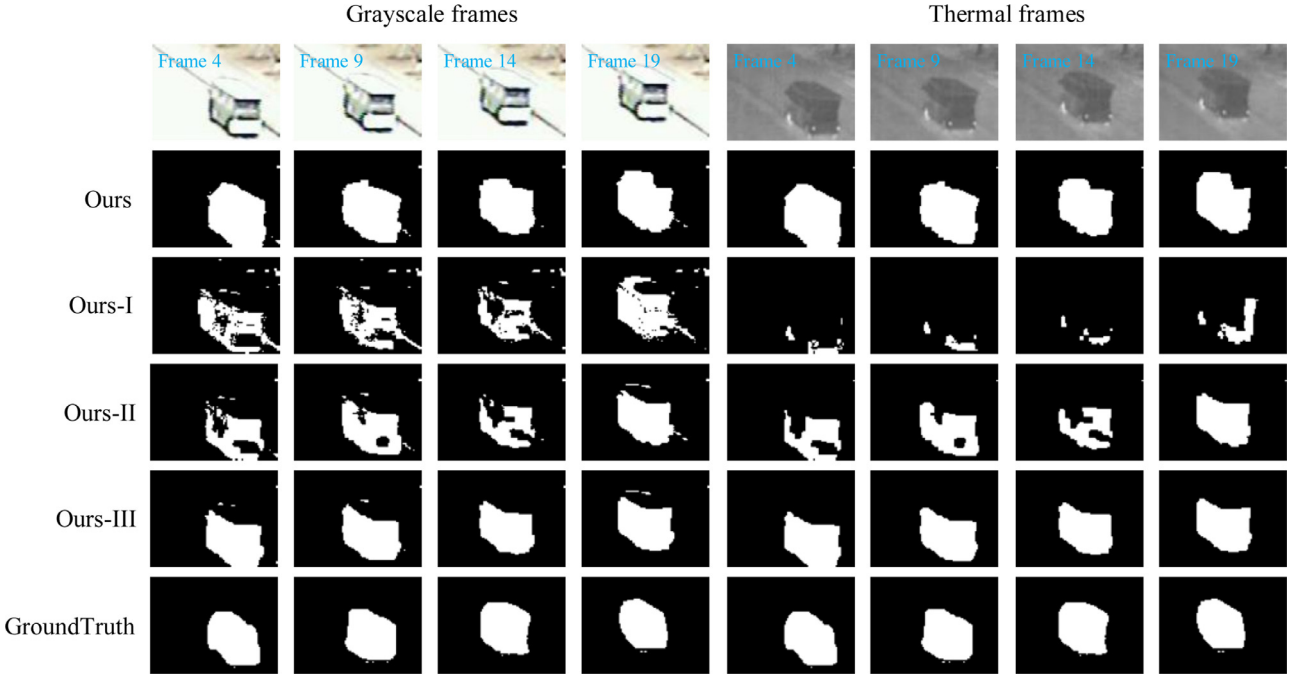


Fig. 7. The detecting results of the continuous frames of our method with its variants. The first four columns denote the frames taken in every 5 frames in grayscale modality (the first row), followed by the detecting results of our method (the second row) together with the three variants (the third to the fifth rows) and the ground truth (the sixth row). The last four columns indicate the corresponding frames and results in thermal modality.

component contribution can be descendingly ordered as soft cross-modal consistency (CM), global appearance consistency (GA) and local appearance consistency (LA).

From Fig. 7 we can observe that: (1) By introducing the soft cross-modal consistency (comparing Ours-I to Ours), our method can benefit from complementary information from the thermal modality especially when the grayscale modality encounters illumination clutter. (2) By introducing the global appearance consistency (comparing Ours-II to Ours), our method produce more compact structures of the foregrounds with much less "cavities". (3) By introducing the local appearance consistency (comparing Ours-III to Ours), our method can better detect the fine details along the contours of the foregrounds.

4.5. Parameter analysis

Our model contains seven important parameters, $\{\beta, \mu, \gamma, \rho, \theta, r, \lambda\}$. In particular, β controls the sparsity of the foreground masks, μ controls the spatial smoothness to punish the neighboring pixels with different labels, γ balances the soft cross-modal consistency of the foreground masks to promote the pixels with same label from different modality, ρ and θ adjust the global appearance consistency and the local appearance consistency. r is the rough estimation of the $rank(B)$, λ controls the complexity of the background model. We empirically set $\beta = 0.045\sigma^2$, where σ denotes the mean variance of the difference matrix $\{\mathbf{D}^k - \mathbf{B}^k\}$. The initial β of our algorithm is very large, and then decreases by 0.08 until it reaches $0.045\sigma^2$. We adjusted each parameter while fixing the other five to achieve the best performance of our model. We initialize λ to be the mean of the second largest singular values of $\mathbf{D}^k$ in our implementation, and run the SOFT-IMPUTE algorithm. If the existing k is subject to $\text{rank}(\mathbf{B}^k) \leq r$, we reduce λ by a factor ($\frac{1}{\sqrt{2}}$ in our implementation) and repeat the SOFT-IMPUTE algorithm until $\text{rank}(\mathbf{B}^k) > r$ for all $k = 1, 2, \ldots, K$. We empirically set the other parameters as $\{\beta, \mu, \gamma, \rho, \theta, r\} = \{0.045\sigma^2, 0.5\beta, 0.2\beta, 0.15\mu, 10, \sqrt{n}\}$, where n is the total number of pixels.

Table 5 reports the parameter analysis on the cropped GTFD dataset. Since β, μ, γ and ρ are interrelated and jointly adjusted during optimization, we only analyze γ instead of four of them. In addition, we further analyze r and θ in our model. From Table 5 we

Table 5
Parameter analysis of our method on cropped GTFD dataset.

Param	Setting	P	R	F
r	$2\sqrt{n}$	0.653	0.639	0.603
	$\sqrt{n}$	0.659	0.887	0.738
	$\frac{1}{2}\sqrt{n}$	0.609	0.913	0.706
γ	0.3β	0.656	0.885	0.734
	0.2β	0.659	0.887	0.738
	0.1β	0.664	0.861	0.731
θ	5	0.657	0.875	0.733
	10	0.659	0.887	0.738
	15	0.659	0.880	0.736

observe that: (1) Our model is insensitive to the parameters γ and θ which demonstrate the generality of our model. (2) Although the results are slightly unstable with different r, it is noted that the best result generated from $r = \sqrt{n}$ consistent to [10], where n is the total number of pixels, which is a dynamic value for different videos. Therefore, it also verifies the generality of our model.

4.6. Limitation

From the above experiments, we can conclude that our method has a significant advantage over many existing methods on public challenging datasets GTFD, OSU03 and GTFD's extended dataset, but our method is not satisfactory in terms of time consumption. The main reason is that our algorithm adds the calculation of GMM and the adjacency matrix constructed between modalities will increase the burden of graph cuts algorithm. In the future, we will optimize the code or replace the graph cuts algorithm with some other algorithm or choose a higher computing platform to reduce the time consumption.

5. Conclusion

We have proposed a novel approach for multi-modal foreground detection via pursuing the inter- and intra-modality consistencies in a united low rank and sparse decomposition framework. First, we have explored the intra-modality consistency via preserving the global appearance consistency in addition to the conventional local appearance consistency. Second, we have explored the inter-modal consistency among different modalities by enforcing a soft cross-modal constraint. At last, the comprehensive experimental results on benchmark multi-modal motion detection datasets validate the promising performance of our method comparing to the state-of-the-art methods. Our future work will focus on investigating the online version of algorithm and extending our method with more modalities like RGB-T-D (RGB-Themal-Depth) case.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Aihua Zheng: Funding acquisition, Project administration, Writing - review & editing. **Naipeng Ye:** Validation, Investigation, Writing - original draft, Data curation. **Chenglong Li:** Funding acquisition, Project administration, Writing - review & editing. **Xiao Wang:** Methodology. **Jin Tang:** Methodology.

Acknowledgment

This research is supported in part by the Open Project Program of the National Natural Science Foundation of China (Nos. 61976002, 61976003, 61702002, 61872005 and 61671018), the National Laboratory of Pattern Recognition (NLPR) (201900046), the Natural Science Foundation of Anhui Province (1808085QF187) and the Open fund for Discipline Construction, Institute of Physical Science and Information Technology.

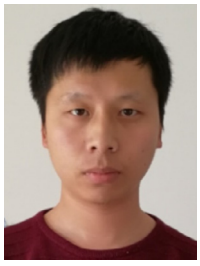
References

[1] C. Li, X. Wu, N. Zhao, X. Cao, J. Tang, Fusing two-stream convolutional neural networks for RGB-T object tracking, Neurocomputing 281 (2018) 78–85.
[2] T. Bouwmans, B. Garcia-Garcia, Background Subtraction in Real Applications: Challenges, Current Models and Future Directions, arXiv:1901.03577 (2019).
[3] Z. Tu, Z. Guo, W. Xie, M. Yan, R.C. Veltkamp, B. Li, J. Yuan, Fusing disparate object signatures for salient object detection in video, Pattern Recognit. 72 (2017) 285–299.
[4] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings of IEEE International Conference on Computer Vision, 1999.
[5] B. Olivier, V.D. Marc, Vibe: a universal background subtraction algorithm for video sequences, IEEE Trans. Image Process. 20 (6) (2011) 1709–1724.
[6] P.L. Stcharles, G.A. Bilodeau, R. Bergevin, Subsense: a universal change detection method with local adaptive sensitivity., IEEE Trans. Image Process. 24 (1) (2014) 359–373.
[7] L. Maddalena, A. Petrosino, Background subtraction for moving object detection in RGB-D data: a survey, J. Imaging 4 (5) (2018) 71.
[8] T. Bouwmans, Traditional and recent approaches in background modeling for foreground detection: an overview, Comput. Sci. Rev. 11 (2014) 31–66.
[9] T. Bouwmans, S. Javed, M. Sultana, S.K. Jung, Deep neural network concepts for background subtraction: a systematic review and comparative evaluation, Neural Netw. 117 (2019) 8–66.
[10] X. Zhou, C. Yang, W. Yu, Moving object detection by detecting contiguous outliers in the low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 597–610.
[11] A. Zheng, T. Zou, Y. Zhao, B. Jiang, J. Tang, C. Li, Background subtraction with multi-scale structured low-rank and sparse factorization, Neurocomputing 328 (2018) 113–121.
[12] T. Bouwmans, A. Sobral, S. Javed, S.K. Jung, E.-H. Zahzah, Decomposition into low-rank plus additive matrices for background/foreground separation: a review for a comparative evaluation with a large-scale dataset, Comput. Sci. Rev. 23 (2017) 1–71.
[13] N. Vaswani, T. Bouwmans, S. Javed, P. Narayanamurthy, Robust subspace learning: robust PCA, robust subspace tracking, and robust subspace recovery, IEEE Signal Process. Mag. 35 (4) (2018) 32–55.
[14] A. Zheng, M. Xu, B. Luo, Z. Zhou, C. Li, Class: collaborative low-rank and sparse separation for moving object detection, Cogn. Comput. 9 (2) (2017) 1–14.
[15] M. Wu, Y. Sun, R. Hang, Q. Liu, G. Liu, Multi-component group sparse RPCA model for motion object detection under complex dynamic background, Neurocomputing 314 (2018) 120–131.
[16] J. Dou, J. Li, Q. Qin, Z. Tu, Moving object detection based on incremental learning low rank representation and spatial constraint, Neurocomputing 168 (2015) 382–400.
[17] C. Li, W. Xiao, Z. Lei, T. Jin, L. Liang, Weld: weighted low-rank decomposition for robust grayscale-thermal foreground detection, IEEE Trans. Circuits Syst. Video Technol. 27 (4) (2016) 725–738.
[18] S. Yang, B. Luo, C. Li, G. Wang, J. Tang, Fast grayscale-thermal foreground detection with collaborative low-rank decomposition, IEEE Trans. Circuits Syst. Video Technol. 28 (2017) 2574–2585.
[19] M. Xu, C. Li, H. Shi, J. Tang, A. Zheng, Moving object detection via integrating spatial compactness and appearance consistency in the low-rank representation, in: Proceedings of the Chinese Conference on Computer Vision, 773, 2017, pp. 50–60.
[20] B. Xin, Y. Tian, Y. Wang, W. Gao, Background Subtraction via Generalized Fused Lasso Foreground Modeling, arXiv preprint (2015) 4676–4684.
[21] A. Zheng, Y. Zhao, C. Li, J. Tang, B. Luo, Multispectral foreground detection via robust cross-modal low-rank decomposition, in: Proceedings of Pacific Rim Conference on Multimedia, 2018.
[22] A. Sobral, A. Vacavant, A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos, Comput. Vis. Image Underst. 122 (2014) 4–21.
[23] M. Chen, X. Wei, Q. Yang, Q. Li, G. Wang, M.H. Yang, Spatiotemporal GMM for background subtraction with superpixel hierarchy, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1518–1525.
[24] S.T. Ali, K. Goyal, J. Singhai, Moving object detection using self adaptive gaussian mixture model for real time applications, in: Proceedings of IEEE International Conference on Recent Innovations in Signal Processing and Embedded Systems, 2017.
[25] M. Hofmann, P. Tiefenbacher, G. Rigoll, Background segmentation with feedback: the pixel-based adaptive segmenter, in: Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012.
[26] Z. Zhong, B. Zhang, G. Lu, Y. Zhao, Y. Xu, An adaptive background modeling method for foreground segmentation, IEEE Transactions on Intelligent Transportation Systems 18 (5) (2017) 1109–1121.

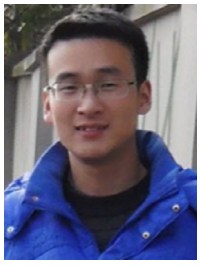
[27] M. Braham, M. Van Droogenbroeck, Deep background subtraction with scene-specific convolutional neural networks, in: Proceedings of IEEE International Conference on Systems, Signals and Image Processing, 2016.
[28] D. Zeng, M. Zhu, A. Kuijper, Combining background subtraction algorithms with convolutional neural network, J. Electron. Imaging 28 (1) (2019) 013011.
[29] Y. Wang, Z. Luo, P.-M. Jodoin, Interactive deep learning method for segmenting moving objects, Pattern Recognit. Lett. 96 (2017) 66–75.
[30] F. De La Torre, M.J. Black, A framework for robust subspace learning, Int. J. Comput. Vis. 54 (1–3) (2003) 117–142.
[31] E. Candes, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (2011) 175–181.
[32] J. He, L. Balzano, A. Szlam, Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2012.
[33] P. Narayanamurthy, N. Vaswani, A fast and memory-efficient algorithm for robust PCA (merop), in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.
[34] H. Guo, C. Qiu, N. Vaswani, Practical reprocs for separating sparse and low-dimensional signal sequences from their sumpart 1, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.
[35] P. Rodriguez, B. Wohlberg, Incremental principal component pursuit for video background modeling, J. Math. Imaging Vis. 55 (1) (2016) 1–18.
[36] M. Kleinsteuber, F. Seidel, C. Hage, Prost: a smoothed ip-norm robust online subspace tracking method for realtime background subtraction in video, Mach. Vis. Appl. 25 (5) (2013) 1227–1240.
[37] A. Sobral, T. Bouwmans, E.-h. ZahZah, Double-constrained RPCA based on saliency maps for foreground detection in automated maritime surveillance, in: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance, 2015.
[38] S. Javed, S. Ho Oh, A. Sobral, T. Bouwmans, S. Ki Jung, Background subtraction via superpixel-based online matrix decomposition with structured foreground constraints, in: Proceedings of IEEE International Conference on Computer Vision Workshops, 2015.
[39] S.E. Ebadi, V.G. Ones, E. Izquierdo, Dynamic tree-structured sparse rpca via column subset selection for background modeling and foreground detection, in: Proceedings of IEEE International Conference on Image Processing, 2016.
[40] S. Javed, T. Bouwmans, M. Sultana, S.K. Jung, Moving object detection on rgb-d videos using graph regularized spatiotemporal rpca, in: Proceedings of International Conference on Image Analysis and Processing, 2017.
[41] C. Conaire, E. Cooke, N. O'Connor, N. Murphy, A. Smearson, Background modelling in infrared and visible spectrum video for people tracking, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2005.
[42] S. Nadimi, B. Bhanu, Physics-based models of color and ir video for sensor fusion, in: Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2003.
[43] S. Becker, N. Scherer-Negenborn, P. Thakkar, W. Hübner, M. Arens, The effects of camera jitter for background subtraction algorithms on fused infrared-visible video streams, Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII, 2016.
[44] T. Bouwmans, C. Silva, C. Marghes, M.S. Zitouni, H. Bhaskar, C. Frelicot, On the role and the importance of features for background modeling and foreground detection, Comput. Sci. Rev. 28 (2018) 26–91.
[45] X. Guo, X. Wang, L. Yang, X. Cao, Y. Ma, Robust foreground detection using smoothness and arbitrariness constraints, in: Proceedings of IEEE International Conference European Conference on Computer Vision, 2014.
[46] V. Kolmogorov, R. Zabin, What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 147–159.
[47] B. Recht, M. Fazel, P.A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Rev. 52 (3) (2010) 471–501.
[48] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (11) (2009) 2287.
[49] S.Z. Li, Markov Random Field Modeling in Image Analysis, 2009.
[50] J.W. Davis, V. Sharma, Background-subtraction using contour-based fusion of thermal and visible imagery, Comput. Vis. Image Underst. 106 (2) (2007) 162–182.
[51] P.-L. St-Charles, G.-A. Bilodeau, R. Bergevin, A self-adjusting approach to change detection based on background word consensus, in: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2015.
[52] S. Jiang, X. Lu, Wesambe: a weight-sample-based method for background subtraction, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2105–2115.
[53] X. Liu, G. Zhao, J. Yao, C. Qi, Background subtraction based on low-rank and structured sparse decomposition, IEEE Trans. Image Process. 24 (8) (2015) 2502–2514.
[54] L. Maddalena, A. Petrosino, A self-organizing approach to background subtraction for visual surveillance applications, IEEE Trans. Image Process. 17 (7) (2008) 1168–1177.
[55] H. Zhang, D. Xu, Fusing color and texture features for background model, in: Proceedings of Third International Conference on Fuzzy Systems and Knowledge Discovery, 2006.
[56] M. Narayana, A. Hanson, E. Learned-Miller, Background modeling using adaptive pixelwise kernel variances in a hybrid feature space, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[57] O. Omar, L. Xin, S. Mubarak, Simultaneous video stabilization and moving object detection in turbulence, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 450–462.
[58] X. Ye, J. Yang, X. Sun, K. Li, C. Hou, Y. Wang, Foreground-background separation from video clips via motion-assisted matrix restoration, IEEE Trans. Circuits Syst. Video Technol. 25 (11) (2015) 1721–1734.
[59] G. Han, X. Cai, J. Wang, Object detection based on combination of visible and thermal videos using a joint sample consensus background model, J. Softw. 8 (2013) 987–994.

Aihua Zheng received B.Eng. degrees and finished Master-Docter combined program in computer science and technology from Anhui University of China in 2006 and 2008, respectively. And received Ph.D. degree in computer science from University of Greenwich of UK in 2012. She visited University of Stirling during June to September in 2013. She is currently an Associate Professor in Anhui University. Her main research interests include vision based artificial intelligence and pattern recognition. Especially on Person/Vehicle Re-identification, Audio Visual Computing, Motion Detection and Tracking.

Naipeng Ye received the B.S. degree from Anhui University of Technology, Ma'anshan, China, in 2018, where he is currently working toward the M.S. degree in computer science. His research interests include computer vision, machine learning, and pattern recognition.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he worked as a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a postdoctoral research fellow at the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an Associate Professor at the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.



Xiao Wang received the B.S. degree in Western Anhui University, Luan, China, in 2013. He is currently pursuing the Ph.D. degree in computer science in Anhui University. From 2015 and 2016, he was a visiting student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He is now having a visiting at UBTECH Sydney Artificial Intelligence Centre, the Faculty of Engineering, the University of Sydney, in 2019. His current research interests mainly about computer vision, machine learning, pattern recognition and deep learning. He also serves as a reviewer for a number of journals and conferences such as TCSVT, CVPR, ICCV, and so on.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is a Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.