Weighted Low-Rank Decomposition for Robust Grayscale-Thermal Foreground Detection

Chenglong Li, Xiao Wang, Lei Zhang, Jin Tang, Hejun Wu, and Liang Lin

***Abstract*— This paper investigates how to fuse grayscale and thermal video data for detecting foreground objects in challenging scenarios. To this end, we propose an intuitive yet effective method called weighted low-rank decomposition (WELD), which adaptively pursues the cross-modality low-rank representation. Specifically, we form two data matrices by accumulating sequential frames from the grayscale and the thermal videos, respectively. Within these two observing matrices, WELD detects moving foreground pixels as sparse outliers against the low-rank structure background and incorporates the weight variables to make the models of two modalities complementary to each other. The smoothness constraints of object motion are also introduced in WELD to further improve the robustness to noises. For optimization, we propose an iterative algorithm to efficiently solve the low-rank models with three subproblems. Moreover, we utilize an edge-preserving filtering-based method to substantially speed up WELD while preserving its accuracy. To provide a comprehensive evaluation benchmark of grayscale-thermal foreground detection, we create a new data set including 25 aligned grayscale-thermal video pairs with high diversity. Our extensive experiments on both the newly created data set and the public data set OSU3 suggest that WELD achieves superior performance and comparable efficiency against other state-of-the-art approaches.**

***Index Terms*— Adaptive fusion, foreground detection, grayscale-thermal processing, low-rank representation, video surveillance.**

I. Introduction

FOREGROUND detection (sometimes called moving object detection) is a fundamental problem in computer vision, and plays a critical role in numerous vision applications, such as object tracking, activity recognition, and video indexing. Although much progress has been made in recent years, it is still a challenging problem in complex and challenging scenarios, like low illumination (LI), background clutter (BC), as well as bad weather (BW).

Fortunately, thermal infrared sensors can provide complementary information for visible spectrum sensors to alleviate the effects of the above factors [1]. Thermal sensors are a kind of passive sensors that capture the infrared radiation emitted by all objects with a temperature above absolute zero, and thus the imaging procedure is not sensitive to light conditions. In addition, this type of sensor, originally developed for military use (e.g., surveillance during night), has recently opened up a broader field of applications due to decrease in its price [1], [2]. Meanwhile, visible spectrum sensors may be more effective to separate objects from the background when they have similar temperatures [also called thermal crossover (TC)]. Therefore, grayscale and thermal data can complement information to each other to achieve more robust moving object detection in challenging scenarios.

This paper will address the following issues in grayscale-thermal foreground detection through existing works.

1) How to collaboratively employ grayscale-thermal information to achieve robust foreground detection. In many scenarios, grayscale and thermal data can complement each other, and thus the effective fusion of these two modalities is important to detect foreground objects robustly. Existing approaches [3]–[6] employed some cues, such as contour and saliency, to integrate grayscale and thermal data for detecting moving objects. These methods were difficult to handle challenging scenarios. The others preferred to exploit thermal information to assist in grayscale detection and ignored the complementary benefit of the grayscale source when the thermal information was crossover.
2) How to efficiently and robustly detect moving objects in challenging scenarios. Foreground detection is always cast as a prerequisite step for subsequent applications, which demands an efficient and robust solution. Some fast methods, such as Gaussian mixture model (GMM) [7], [8], ViBe [9], and nonparametric models [10], [11], have been applied to many practical systems due to their efficient solutions. However, they are easy to produce poor performance in challenging scenarios. Other complex approaches [12]–[15] can obtain robust foreground detection results in various challenging scenarios, but usually have long latencies.
3) How to create a comprehensive grayscale-thermal benchmark for moving object detection. Given the potentials of grayscale-thermal data, however, the related research is limited by the lack of a comprehensive video

Manuscript received August 27, 2015; revised December 18, 2015 and March 9, 2016; accepted April 14, 2016. Date of publication April 20, 2016; date of current version April 3, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61472002 and Grant 61370185, in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2015A110, in part by the Guangdong Natural Science Foundation under Grant S2013010013432 and Grant S2013010015940, in part by the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067, and in part by the Fundamental Research Funds for the Central Universities. The work of L. Lin was supported by the Guangdong National Science Foundation for Distinguished Young Scholars. This paper was recommended by Associate Editor Y. Liu. *(Chenglong Li and Xiao Wang contributed equally to this work.) (Corresponding author: Liang Lin.)*

C. Li, X. Wang, L. Zhang, and J. Tang are with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com; xwang0717@foxmail.com; lzhang15@foxmail.com; tj@ahu.edu.cn).

H. Wu and L. Lin are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: wuhejun@mail.sysu.edu.cn; linliang@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2556586

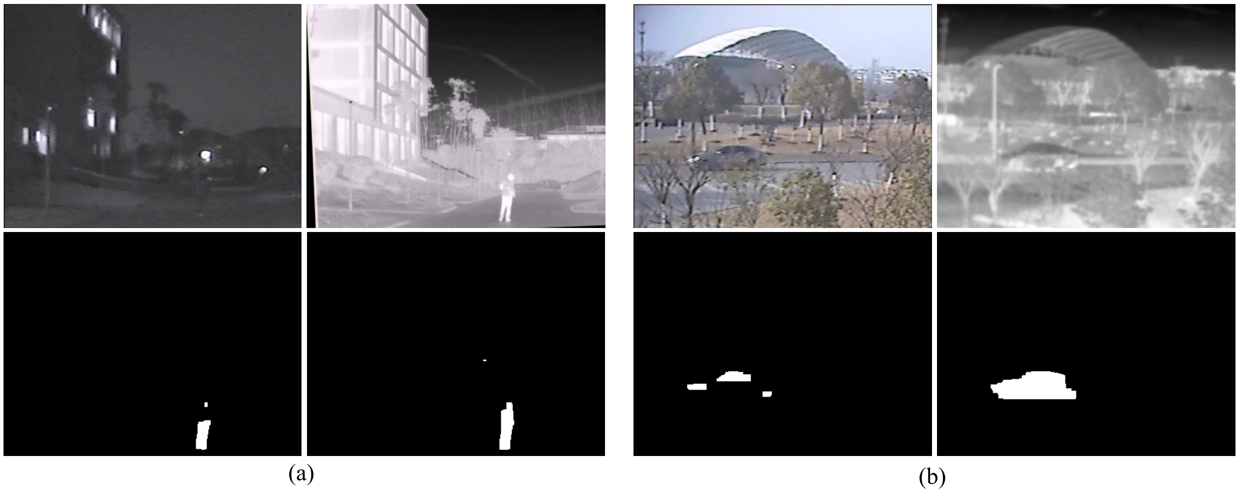


Fig. 1. Typical results generated by WELD. (a) and (b) First row: grayscale and thermal frames, respectively, and second row: WELD results with equal quality weights and WELD results, respectively. The optimized quality weights of grayscale and thermal videos in (a) are 0.32 and 1.81, while in (b) are 0.43 and 0.36, respectively.

benchmark. On the one hand, existing grayscale-thermal data sets, like OSU Color-Thermal [16] and LITIV [17], [18], contain a small number of videos with less challenge and induce a significant bias. In addition, these data sets are used for different computer vision tasks, such as registration, fusion, and tracking, and thus are not suitable for addressing different challenges in moving object detection. On the other hand, the lack of a grayscale-thermal evaluation platform also limits the research progress of grayscale-thermal foreground detection.

Aiming at addressing the above issues and motivated by the effectiveness of low-rank representation on background modeling, we propose a general algorithm, called weighted low-rank decomposition (WELD), which integrates grayscale and thermal data for robust moving object detection. Given two data matrices by accumulating sequential frames from the grayscale and thermal videos, WELD adaptively pursues a cross-modality low-rank representation. More specifically, the foreground object pixels are jointly detected as sparse outliers in grayscale-thermal videos against the low-rank structure backgrounds. In addition, we also integrate the contiguous constraints, which are usually employed for modeling the spatial distribution of moving objects into WELD to improve its robustness to noise.

For adaptively integrating information from different modalities, we develop a quality-based fusion scheme to detect moving objects robustly. In particular, we incorporate the weight variables into WELD to make the models of two modalities complementary to each other. The weight variable of one modality represents its reliability, and thus WELD achieves an adaptive fusion based on the reliabilities of different modalities. The effectiveness of introducing quality weights in WELD is shown in Fig. 1. For optimization, we efficiently solve WELD by iteratively optimizing: 1) the background matrices by SOFT-IMPUTE algorithm [19]; 2) the foreground matrix by graph cut [20], [21]; and 3) the weight variables with closed-form solutions.

To further improve the practicality of WELD, we substantially speed up WELD by an edge-preserving-based method while preserving its accuracy. Specifically, we first perform WELD on a gradient-driven-based downsampled video, and then employ an edge-preserving upsampling [22]–[24] to recover the full-resolution detection results. In this paper, we call the fast version of WELD as F-WELD.

We also create a grayscale-thermal data set with 25 video pairs (3 video pairs from [16], [17] for high diversity) recorded under different challenging scenes. The ground truths of all the frames are aligned and annotated by one person to ensure high consistency.[1] The annotated data set, implemented baselines, and evaluation metrics provide a comprehensive benchmark for grayscale-thermal foreground detection.

This paper makes the following contributions to video processing and related applications.

1) It proposes a general framework for detecting moving objects in multimodal videos. To the best of our knowledge, we are the first to address the problem of grayscale-thermal foreground detection in an adaptive low-rank framework. Our framework is able to deal with challenging scenarios by adaptively leveraging complementary information about different modalities, and substantially outperforms other moving object detection approaches on both the newly created data set and the public data set OSU3 [16].
2) It presents a fast algorithm to greatly speed up WELD. In particular, this fast algorithm preserves the accuracy by the advantage of utilizing the frame structural information.
3) It creates a standard grayscale-thermal benchmark for grayscale-thermal foreground detection. This benchmark will be beneficial for further study of foreground detection in grayscale-thermal videos. We will release the evaluation platform online for free academic usage.

[1] Grayscale-thermal data set webpage: http://vision.sysu.edu.cn/projects/grayscale_thermal_detection/.

The rest of this paper is organized as follows. In Section II, the methods relevant to our WELD are introduced. We describe the details of WELD and F-WELD in Sections III and IV, respectively. The creation of a grayscale-thermal benchmark and the experiment results on this benchmark are shown in Sections V and VI, respectively. Section VII concludes this paper.

II. Related Work

This paper is closely related to the advances in research streams and the development of a grayscale-thermal data set for moving object detection.

Single-modality foreground detection has been extensively studied over the past decades. The representative methods include GMMs [7], nonparameter algorithms [9], fuzzy-based methods [25], multiple features-based methods [26], low-rank representation models [13], [27]–[30], and neural and neuro-fuzzy methods [31]. These works focus on a single-modality video sequence (thermal information can be viewed as the gray value of image), and thus will suffer from the aforementioned challenges, i.e., LI, BW, etc.

Multimodality foreground detection has drawn more and more attention in the community [3], [4], [6], [16] with the popularity of various sensors, such as depth sensors and thermal infrared sensors [32]. Davis and Sharma [3] provided a framework for effectively combining information from thermal and visible videos. They first identified the initial regions-of-interest in the thermal domain, and then propagated it into the grayscale domain. The contour saliency map was obtained by combining both thermal and grayscale information on the detected regions-of-interest, and further flood-filled to produce silhouettes. Han and Bhanu [4] proposed a hierarchical scheme to automatically align synchronous grayscale and thermal frames, and probabilistically combined cues from registered grayscale-thermal frames for improving human silhouette detection. Davis and Sharma [16] proposed a new background-subtraction technique fusing contours from thermal and grayscale videos for persistent object detection in urban settings. Using the region saliency detection method in [6], infrared and visible images were integrated with different strategies applied for salient and nonsalient regions. The background subtraction method is then employed using GMM.

Low-rank representation has been recently applied to background modeling and achieved impressive results in challenging scenarios. The framework of robust principal component analysis (RPCA) has drawn a lot of attention in computer vision. The seminal work in [12] showed that the low-rank model can be recovered from unknown corruption patterns by principal component pursuit (PCP), a convex program. The examples in [12] justify the superior performance of PCP against the previous methods of RPCA and its promising potential for background subtraction. Zhou *et al.* [33] proposed stable PCP (SPCP), an extension of PCP, to handle both sparse gross errors and small entrywise noises. PCP and SPCP relaxed l_0-penalty to l_1-penalty for convex optimization. However, the l_0-penalty works effectively for sparse noise detection in regression, while l_1-penalty does not [34]. The work, DEtecting Contiguous Outliers in the LOw-rank Representation (DECOLOR), [13] keeps l_0-penalty to preserve the robustness to outliers, and also modeled the continuity earlier on foreground masks to improve the accuracy of detecting contiguous outliers.

TABLE I

Challenges of Our Data Set Against Other Public Grayscale-Thermal Data Sets. Herein, the Challenges Include IM, LI, BW, IS, DS, BC, and TC. For More Details, Please Refer to Table II

Dataset	Challenge
Our dataset	IM, LI, BW, IS, DS, BC, TC
OSU Color-Thermal [16]	IM, IS, DS
Torabi [17]	IM, LI, IS
Torabi [18]	IM, IS

There have been several *grayscale-thermal video data sets* for various vision tasks. For example, the OSU color-thermal data set [16] contains six thermal/color video sequence pairs recorded from two different locations with only people moving. Two other grayscale-thermal data sets are collected in [17] and [18]. Most of them, however, suffer from their limited size, low diversity, and high bias. This paper addresses this issue and creates a reasonable size grayscale-thermal video data set that provides comprehensive evaluations. Table I presents the challenges of our data set against other public grayscale-thermal data sets.

III. WELD Algorithm

Given a grayscale-thermal video pair, we solve the moving object detection in a batch way and adaptively incorporate the information from different modalities by their respective video quality.

A. Model Formulation

We formulate the problem of grayscale-thermal foreground detection in a low-rank representation framework due to its robustness to noise. For the kth modal video, we stack each frame as column vectors into a matrix, i.e., $\mathbf{D}^k = [\mathbf{d}_1^k, \mathbf{d}_2^k, \ldots, \mathbf{d}_n^k] \in R^{g\times n}$, where g is the number of pixels in one frame and n is the number of frames in one modal video. Herein, we consider the K modalities for general formulation and grayscale-thermal data in this paper is a special case with $K = 2$. Our goal is to discover the object mask $\mathbf{S}$ from data matrices $\mathbf{D}^{[1,\ldots,K]}$, which is an abbreviation of the matrix set $\{\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^K\}$. $\mathbf{S} \in \{0, 1\}^{g\times n}$ is a binary matrix denoting the foreground mask as

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

To this end, we assume that single-modal underlying background images are linearly correlated and foregrounds are sparse. This assumption has been successfully applied in background modeling [13], [27]. Thus, based on this assumption, the low-rank representation model can be formulated as

$$\min_{\mathbf{B}^k,\mathbf{S}^k} \frac{1}{2}\|f_{\bar{\mathbf{S}}^k}(\mathbf{D}^k-\mathbf{B}^k)\|_F^2+\beta\|\text{vec}(\mathbf{S}^k)\|_0$$
$$\text{s.t. } \text{rank}(\mathbf{B}^k)\le r^k,\quad k=1,2,\ldots,K \tag{2}$$

where $\mathbf{B}^k \in R^{g\times n}$ denotes the underlying background images and β is a balance parameter. vec$(\cdot)$ is a vectorize operator on a matrix. $\|\cdot\|_F$ and $\|\cdot\|_0$ indicate the Frobenius norm of a matrix and the l_0 norm of a vector, respectively. $f_{\mathbf{S}}(\mathbf{X})$ represents the orthogonal projection of a matrix $\mathbf{X}$ onto the linear space of matrices supported by $\mathbf{S}$

$$f_{\mathbf{S}}(\mathbf{X})(i,j)=\begin{cases}0, & \mathbf{S}_{ij}=0\\ \mathbf{X}_{ij}, & \mathbf{S}_{ij}=1\end{cases} \tag{3}$$

and $f_{\bar{\mathbf{S}}}(\mathbf{X})$ is its complementary projection, i.e., $f_{\mathbf{S}}(\mathbf{X})+f_{\bar{\mathbf{S}}}(\mathbf{X})=\mathbf{X}$.

To make the rank constraints tractable, we relax it with the nuclear norm, which has been proved to be an effective convex surrogate of the rank operator [35]. Thus, we can reformulate (2) as

$$\min_{\mathbf{B}^k,\mathbf{S}^k} \frac{1}{2}\|f_{\bar{\mathbf{S}}^k}(\mathbf{D}^k-\mathbf{B}^k)\|_F^2+\lambda\|\mathbf{B}^k\|_*+\beta\|\text{vec}(\mathbf{S}^k)\|_0$$
$$k=1,2,\ldots,K \tag{4}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix and λ is a balance parameter. Though minimizing l_0 norm of the foreground mask $\mathbf{S}^k$ is also nonconvex, we can optimize it through introducing the contiguous constraint on $\mathbf{S}^k$, which is a prior that foreground objects should be contiguous pieces, and naturally model it by a Markov Random Field (MRF) [13], [36]. In this way, l_0 norm of $\mathbf{S}^k$ can be converted into a unary term in the MRF minimization function (see Section III-B for more details). We denote E^k as the edge set connecting spatially neighboring pixels in the kth modality. It is worth mentioning that the edges in the temporally neighboring pixels are ignored for reducing the computational complexity. This contiguous constraint is formulated as

$$\sum_{(ij,kl)\in E^k}\left|\mathbf{S}_{ij}^k-\mathbf{S}_{kl}^k\right|=\|\mathbf{A}^k\ \text{vec}(\mathbf{S}^k)\|_1 \tag{5}$$

where $\mathbf{A}^k$ is the node-edge incidence matrix denoting the connecting relationship among pixels in kth modality. Therefore, we enforce contiguous constraint into (4) as

$$\min_{\mathbf{B}^k,\mathbf{S}^k} \frac{1}{2}\|f_{\bar{\mathbf{S}}^k}(\mathbf{D}^k-\mathbf{B}^k)\|_F^2+\lambda\|\mathbf{B}^k\|_*+\beta\|\text{vec}(\mathbf{S}^k)\|_0$$
$$+\gamma\,\|\mathbf{A}^k\ \text{vec}(\mathbf{S}^k)\|_1,\quad k=1,2,\ldots,K \tag{6}$$

where γ is a balance parameter.

In (6), it is inherently indicated that the available modalities are independent and contribute equally. This may significantly limit the performance in dealing with occasional perturbation or malfunction of individual sources. Therefore, we propose a novel collaborative model for robustly detecting moving objects that: 1) adaptively recovers the low-rank backgrounds based on their respective modal qualities; 2) collaboratively computes one sparse foreground mask shared by all modalities; and 3) efficiently optimizes the quality weights of all modalities with closed-form solutions. In this way, we can detect moving objects by adaptively leveraging the information about different modalities based on their reliabilities. The formulation of the WELD algorithm is proposed as

$$\min_{\{\mathbf{B}^k\},\mathbf{S},\{\delta^k\}} \sum_{k=1}^{K}\frac{(\delta^k)^m}{2}\|f_{\bar{\mathbf{S}}}(\mathbf{D}^k-\mathbf{B}^k)\|_F^2+\lambda\|\mathbf{B}^k\|_*$$
$$+\beta\|\text{vec}(\mathbf{S})\|_0+\gamma\,\|\mathbf{A}\ \text{vec}(\mathbf{S})\|_1+\sum_{k=1}^{K}\phi_{\delta^k}(1-\delta^k)^m \tag{7}$$

with the constraints $\delta^k > 0$ $(k=1,\ldots,K)$, where δ^k is the quality weight of the kth modality and $m\in(1,\infty)$ is a fuzzifier parameter, similar to the formulation of fuzzy c-means clustering [37]. $\mathbf{S}$ is the shared foreground mask matrix by all the modalities. ϕ_{δ^k} is determined by the reconstruction error of the kth modality after the first iteration, as shown in Algorithm 1. The last term in (7) is a possibility-like constraint to avoid degenerate solutions of $\{\delta^k\}$, similar to the possibilistic fuzzy c-means clustering [38], allowing the weights of different modalities to be specified independently.

Algorithm 1 Optimization Procedure to (7)

```
**Input:** $\mathbf{D}^k$, $(k=1,\ldots,K)$.
    Set $\mathbf{B}^k=\mathbf{D}^k$ $(k=1,2,\ldots,K)$, $\mathbf{S}=\mathbf{0}$, $\lambda^k=\frac{1}{K}$, $maxIter=20$, $\epsilon=1e-4$.
**Output:** $\mathbf{S}$, $\mathbf{B}^k$, $\delta_k$, $(k=1,2,\ldots,K)$.
for $i=1:maxIter$ do
    if $i==1$ then
        $\phi_{\delta^k}=\|f_{\bar{\mathbf{S}}}(\mathbf{D}^k-\mathbf{B}^k)\|_F^2,\ k=1,2,\ldots,K.$
    end if
    Parallelly update $\{\mathbf{B}^k\}$ by Eq. (8);
    Update $\mathbf{S}$ by Eq. (9);
    Update $\{\delta^k\}$ by Eq. (10);
    Check the convergence condition: if the maximum objective change between two consecutive iterations is less than $\epsilon$, then terminate the loop.
end for
```

B. Optimization

Although (7) seems complex, we can efficiently solve it by the alternating optimization algorithm. Given $\{\delta^k\}$ and $\mathbf{S}$, the minimization of $\{\mathbf{B}^k\}$ in (7) can be transformed to be the matrix completion problem [19]

$$\min_{\{\mathbf{B}^k\}}\sum_{k=1}^{K}\frac{(\delta^k)^m}{2}\|f_{\bar{\mathbf{S}}}(\mathbf{D}^k-\mathbf{B}^k)\|_F^2+\lambda\|\mathbf{B}^k\|_*. \tag{8}$$

This is to learn a low-rank background matrix from partial observations. The optimal $\mathbf{B}^k$ in (8) can be efficiently computed by the SOFT-IMPUTE algorithm whose convergence property has been proved in [19]. Note that the computations of $\{\mathbf{B}^k\}$ are independent in (8). Thus, we can optimize them in a parallel way for efficiency.

Given $\{\delta^k\}$ and $\{\mathbf{B}^k\}$, (7) can be rewritten as

$$\begin{aligned}
&\min_{\mathbf{S}} \sum_{k=1}^{K} \frac{(\delta^k)^m}{2} \| f_{\bar{\mathbf{S}}}(\mathbf{D}^k - \mathbf{B}^k)\|_F^2 + \beta \|\text{vec}(\mathbf{S})\|_0 + \gamma \,\|\mathbf{A}\ \text{vec}(\mathbf{S})\|_1 \\
&= \min_{\mathbf{S}} \sum_{k=1}^{K} \frac{(\delta^k)^m}{2} \sum_{ij} \left(\mathbf{D}_{ij}^k - \mathbf{B}_{ij}^k\right)^2 (1 - \mathbf{S}_{ij}) + \beta \sum_{ij} \mathbf{S}_{ij} \\
&\quad + \gamma \,\|\mathbf{A}\ \text{vec}(\mathbf{S})\|_1 \\
&= \min_{\mathbf{S}} \sum_{ij} \left[\beta - \frac{1}{2} \sum_{k=1}^{K} (\delta^k)^2 \left(\mathbf{D}_{ij}^k - \mathbf{B}_{ij}^k\right)^2 \right] \mathbf{S}_{ij} \\
&\quad + \gamma \,\|\mathbf{A}\ \text{vec}(\mathbf{S})\|_1 + \mathscr{C} \\
&= \min_{\mathbf{S}} \sum_{ij} \left[\beta - \frac{1}{2} \sum_{k=1}^{K} (\delta^k)^2 \left(\mathbf{D}_{ij}^k - \mathbf{B}_{ij}^k\right)^2 \right] \mathbf{S}_{ij} \\
&\quad + \gamma \,\|\mathbf{A}\ \text{vec}(\mathbf{S})\|_1 \qquad (9)
\end{aligned}$$

where $\mathscr{C} = (1/2) \sum_{k=1}^{K} (\delta^k)^2 \sum_{ij} (\mathbf{D}_{ij}^k - \mathbf{B}_{ij}^k)^2$ is a constant with respect to $\mathbf{S}$. Therefore, the minimization of $\mathbf{S}$ can be converted to the first-order MRFs problem, which can be efficiently solved by graph cut algorithm [20], [21].

Given $\{\mathbf{B}^k\}$ and $\mathbf{S}$, the quality weights in (7) can be written as

$$\min_{\{\delta^k\}} \sum_{k=1}^{K} \left\{ \frac{\| f_{\bar{\mathbf{S}}}(\mathbf{D}^k - \mathbf{B}^k)\|_F^2}{2} (\delta^k)^m + \phi_{\delta^k} (1 - \delta^k)^m \right\}$$
$$\delta^k > 0 \ (k = 1, \ldots, K) \qquad (10)$$

which has a closed-form solution

$$\delta^k = \frac{1}{1 + \left(\frac{\| f_{\bar{\mathbf{S}}}(\mathbf{D}^k - \mathbf{B}^k)\|_F^2}{\phi_{\delta^k}} \right)^{\frac{1}{m-1}}}, \quad k = 1, 2, \ldots, K. \qquad (11)$$

A suboptimal solution can be obtained by alternating the optimization to $\{\mathbf{B}^k\}$, $\mathbf{S}$, and $\{\delta^k\}$; the algorithm is summarized in Algorithm 1. The convergence of WELD can be guaranteed obviously, as each subproblem converges to an optimal solution.

IV. F-WELD: Fast Implementation

In this section, we will present an edge-preserving filtering-based method to improve the efficiency of WELD while preserving its accuracy, using F-WELD.

Instead of processing the full-resolution videos, we perform WELD algorithm on the low-resolution videos subsampled from the original videos in a gradient-driven way. More specifically, we pick the pixel with the largest gradient magnitude from a 3 × 3 patch on every frame to form the low-resolution videos. In this way, we can obtain the low-resolution detection maps. For each low-resolution detection map, we employ it to recover the full-resolution one with the edge-preserving upsampling technique. Herein, we regard the reliable modal frame, which has a higher quality weight, as the guidance image. It can produce a smoothly varying dense detection map without blurring the edges of objects. In this paper, the edge-preserving upsampling method consists of two steps:

Algorithm 2 Summarization of Our System

Input: One grayscale-thermal video pair.
Output: Full-resolution foreground detection maps.

```
1: Run gradient-driven downsampling on video pair to obtain
   the low-resolution video pair;
2: Run Alg. 1 on the low-resolution video pair to obtain the
   low-resolution foreground detection maps;
3: For each foreground detection map, run edge-preserving
   upsampling to recover the full-resolution one.
```

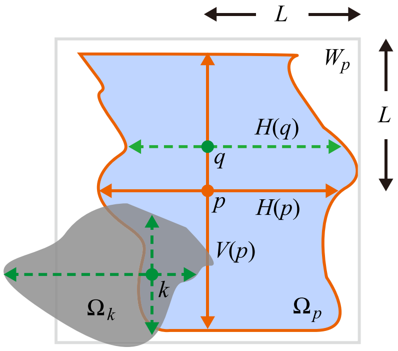


Fig. 2. Generating a pixelwise shape-adaptive region of CLMF [23] (see text for more details).

1) *Shape-Adaptive Region Generation*: We first obtain an arbitrary-shaped observation region for each pixel in an image. Specifically, for a pixel p centered at a square window W_p, the color similarity criterion for a pixel q is defined as

$$|I_c(q) - I_c(p)| \le \tau, \quad c \in \{R, G, B\}, \quad q \in W_p \qquad (12)$$

where I_c is the intensity of the color band c of the 3×3 median smoothed guidance image I and L denotes the preset maximum arm length of the observation window W_p centered at pixel p of size $(2L + 1) \times (2L + 1)$. τ controls the confidence level of the color similarity. The details of generating the shape-adaptive region $\mathbf{\Omega}_p$ are presented in cross-based local multipoint filtering (CLMF) [23] and we briefly review the main idea for clarity. CLMF decides a pixelwise adaptive cross with four arms (left, right, top, bottom) for every pixel p. These arms record the largest left/right horizontal and top/bottom vertical span of the anchor pixel p, where all the pixels covered by the arms satisfies 12. Let $H(p)$ and $V(p)$ denote all the pixels covered by the horizontal and vertical arms of p, respectively, as shown in Fig. 2. Let q denote any pixel covered by the vertical arms of p, i.e., $q \in V(p)$ and we can construct the arbitrary-shaped region of p by integrating multiple $H(q)$ sliding along $V(p)$: $\mathbf{\Omega}_p = \bigcup_{q \in V(p)} H(q)$.

2) *Edge-Preserving Upsampling*: Given the low-resolution input image J^l, we can upsample it to the full-resolution image J without perturbing the object edges by the edge-preserving filtering. For the pixel $p \in J$, similar to the joint bilateral upsampling [22], its value is

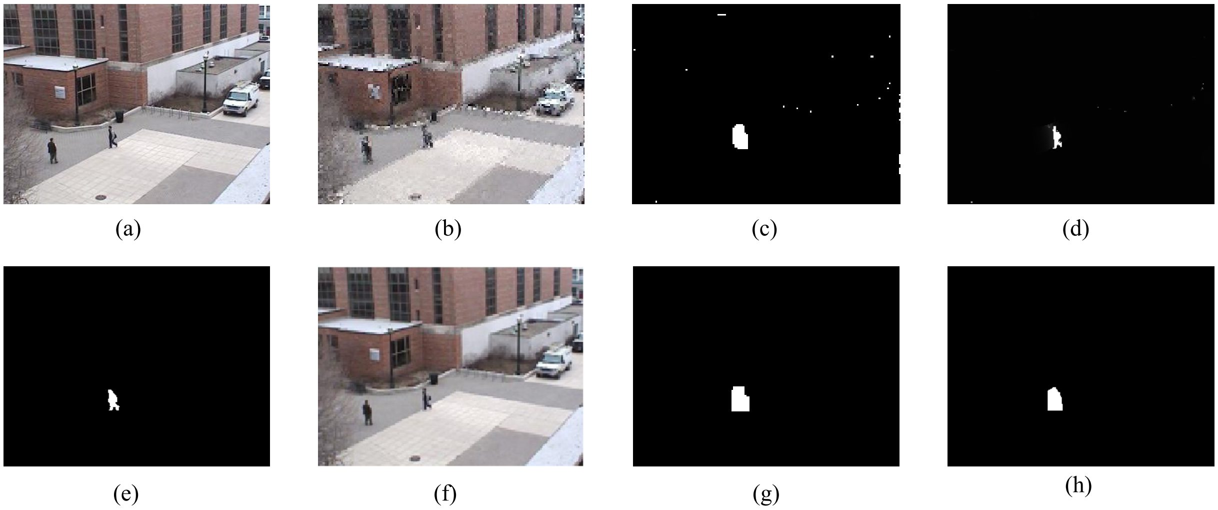


Fig. 3. Edge-preserving filtering for F-WELD. Herein, one modality is shown for clarity. (a) Original frame. (b) Gradient-driven downsampling. (c) WELD result. (d) Edge-preserving upsampling. (e) Ground truth. (f) Bilinear downsampling. (g) WELD result. (h) Bilinear upsampling. (c) and (g) Results of WELD running on (b) and (f), respectively. One can see that the edge-preserving filtering-based F-WELD achieves better accuracy than the general resize operation.

estimated by

$$J(p) = \frac{1}{|\Omega_k|} \sum_{k \in \Omega_k} \omega_{p,k} J^l(k) \tag{13}$$

where Ω_k is the shape-adaptive region generated by the guidance image I of pixel $k \in J^l$, as shown in Fig. 2. $|\Omega_k|$ denotes the number of pixels in Ω_k and $\omega_{p,k} = \exp(-\|\mathbf{x}_p, \mathbf{x}_k\|/\sigma)$, where $\mathbf{x}_p$ indicates the position of p. In this way, the pixel values are weighted average by the spatial distance in the homogeneous regions of the guidance image to form the full-resolution one without perturbing the edges of objects. Overall, our complete system of grayscale-thermal foreground detection is summarized in Algorithm 2.

There are several commonly used downsampling and upsampling techniques. On the one hand, the interpolation-based methods usually generated bad performance, as it employed less information. In particular, the bilinear interpolation is one of the commonly used interpolation methods, and thus we demonstrate the effectiveness of the proposed method against the bilinear interpolation in Fig. 3. On the other hand, the filtering-based methods regarded the original-resolution images as guidance and had the advantage of utilizing the structural information about the original-resolution images for detection recovery. For the filters using piecewise constant modeling [39], [40], they usually cannot preserve the image gradient information well. Although using piecewise linear local modeling in the guided filter (GF) [41], our used filter is a more generalized form, which performs local averaging over a shape-adaptive support region, rather than within a fixed-sized square window in the GF.

V. Grayscale-Thermal Benchmark

This section introduces a new grayscale-thermal video benchmark for moving object detection and presents some essential analyses.

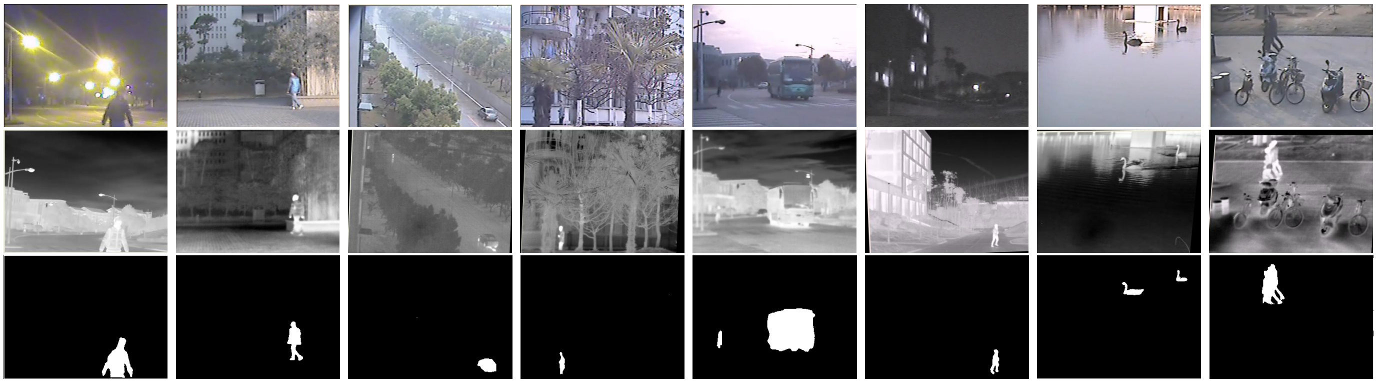
A. Data Set

Our recording system consists of an online thermal image (MAG32) and a charge-coupled device camera (SONY TD-2073). We mount these two cameras on tripods and make their views overlapped as much as possible for convenient alignment.

Unlike the industry registration in RGBD sensor, which consists of one RGB sensor and one depth sensor, which consists of one RGB sensor and one depth sensor, we manually construct the recording system and develop an annotation tool to align grayscale-thermal videos in the following way. We uniformly select a number of point correspondences in the keyframe of the video pair and compute the homography matrix by the least-square method. Then, the video pair can be aligned by applying the computed homography matrix to transform the remaining frame pairs. This registration method can accurately align video pairs due to two main reasons. First, we carefully choose the planar and nonplanar scenes to make the homography assumption effective. Second, since two camera views are almost coincident as we made, the transformation between two views is simple.

We annotate the ground truths of the data set using the more distinguishable modality. In addition, all the frames are manually annotated by one person to keep a high consistency. When occlusion occurs, the ground truth is annotated by the visible portion of the target. Fig. 4 presents some typical frame samples of our data set. The following main aspects are taken into account in creating the grayscale-thermal video.

1) *Scene Category:* We captured video pairs in 15 scenes, including laboratory rooms, campus roads, play grounds, water pools, etc.
2) *Object Category:* Our grayscale-thermal data set includes rigid and nonrigid objects, such as vehicles, pedestrians, and animals.
3) *Intermittent Motion (IM):* When the objects move slowly or stop at one or more frames, many detection methods easily tend to classify them as background.



Pedestrian4
LI, BC

Pedestrian5
BC, TC

Rain
BW, DS, TC

Pedestrian8
IM, DS, BC

Bus
IM, TC

Night1
LI

Swan
IM, IS, DS, BC

Pedestrian2
IS, TC

Fig. 4. Sample pairs and the corresponding annotated challenges of our data set. The grayscale frames, thermal frames, and their corresponding ground truths are shown in the first row, second row, and third row, respectively. We only present one frame pair of each video pair for clarity.

TABLE II
CHALLENGES ANNOTATED TO OUR GRAYSCALE-THERMAL DATA SET

Challenge	Description
IM	Intermittent Motion - the objects stop at one or more frames in their movement.
LI	Low Illumination - the illumination in the object regions is low.
BW	Bad Weather - the videos are captured in bad weathers, such as rainy, snowy and cloudy.
IS	Intense Shadow - the obvious shadows are cast by moving objects or backgrounds with varying ambient lighting.
DS	Dynamic Scene - the non-static background in the scene.
BC	Background Clutter - the background is complex.
TC	Thermal Crossover - the temperature of moving objects is similar with the temperature of their surrounding background in some frames.

TABLE III
ANNOTATED CHALLENGES OF ALL VIDEO SEQUENCES, WHERE ✓ INDICATES THE VIDEO PAIR AND INCLUDES THE CORRESPONDING CHALLENGE

Sequence Name	IM	LI	BW	IS	DS	BC	TC
Pedestrian1				✓			
Pedestrian2				✓			✓
Pedestrian3	✓			✓			
Car1				✓	✓		
Pedestrian4		✓				✓	
Car2							✓
Pedestrian5						✓	✓
Car3				✓			✓
Car4				✓			✓
Swan	✓			✓	✓	✓	
Night1		✓					
Car5	✓	✓					✓
Bus	✓						✓
Pedestrian6	✓						✓
Pedestrian7					✓	✓	
Rain			✓		✓		✓
Pedestrian8	✓				✓	✓	
Pedestrian9							✓
Pedestrian10				✓			
Pedestrian11				✓			
Car6							✓
Car7		✓		✓			
Car8						✓	
Car9		✓				✓	
Car10							✓

4) *Shadow Effect:* The shadows cast by foregrounds or backgrounds with varying ambient lighting usually cause fake motions in the background. They increase the difficulty to detect the intact foregrounds.
5) *Illumination Condition:* The video sequences are captured under different light conditions (sunny, rainy, and nighttime, etc). The LI and illumination variation caused by different light conditions usually bring big challenges in grayscale videos.
6) *Background Factor:* First, a similar background to the moving objects in appearance or temperature will introduce ambiguous information, and bring big challenges to detection methods. Second, it is difficult to separate objects accurately from a cluttered background. Third, the dynamic scene (DS) is also a challenge in moving object detection, such as swaying leaves and waving water surfaces.

Table II summarizes the challenges of the newly built video data sets. We present the challenge distribution in Fig. 5 and the challenges of all the video sequences in Table III.

B. Baseline Approaches

To evaluate the proposed approach and providing a comprehensive evaluation platform, we add some popular methods as baselines as a part of our benchmarks. Specifically, we implement three kinds of baselines, including grayscale, thermal, and grayscale-thermal detection methods.

1) Eleven grayscale baselines are included in the benchmark, including DEtecting Contiguous Outliers in the LOw-rank Representation (DECOLOR) [13], Motion-Assisted Matrix Restoration (MAMR) [29], Local Adaptive Sensitivity (LAS) [26], Adaptive Self-Organizing Model (ASOM) [42], Fusing Color and Texture Features (FCTF) [43], Smoothness and Arbitrariness Constraints (SAC) [14], Principle Component Pursuit (PCP) [12], Three Term Decomposition (TTD) [45], Adaptive Pixelwise Kernel Variances (APKV) [44], Gaussian Mixture

TABLE IV

AVERAGE PRECISION, RECALL, AND F-MEASURE OF OUR METHOD AGAINST DIFFERENT KINDS OF BASELINE METHODS ON THE NEWLY CREATED DATA SET. CODE TYPE AND FRAMES PER SECOND ARE ALSO PRESENTED. BOLD FONTS OF THE RESULTS INDICATE THE BEST PERFORMANCE

Algorithm	Grayscale			Thermal			Grayscale-Thermal			Code Type	FPS
	P	R	F	P	R	F	P	R	F		
ASOM [42]	0.18	0.07	0.06	0.16	0.07	0.08	-	-	-	C++	111.11
FCFT [43]	0.39	0.20	0.22	0.25	0.22	0.20	-	-	-	C++	38.46
APKV [44]	0.38	0.42	0.36	0.42	0.20	0.24	-	-	-	MATLAB & C++	0.03
ViBe [9]	0.41	0.49	0.41	0.41	0.47	0.39	-	-	-	C++	318.47
LAS [26]	0.58	0.49	0.46	0.50	0.40	0.39	-	-	-	C++	12.66
TTD [45]	**0.59**	0.29	0.32	0.58	0.38	0.40	-	-	-	MATLAB	0.07
PCP [12]	0.28	0.18	0.21	0.49	0.40	0.43	-	-	-	MATLAB	20.42
GMM [7]	0.48	0.65	0.52	0.48	0.65	0.50	-	-	-	C++	93.37
SAC [14]	0.42	0.74	0.41	0.47	0.71	0.53	-	-	-	MATLAB	1.15
DECOLOR [13]	0.54	**0.84**	0.59	0.52	**0.82**	0.59	-	-	-	MATLAB & C++	1.98
MAMR [29]	0.57	0.67	0.60	0.59	0.63	0.59	-	-	-	MATLAB & C++	3.77
GMM-GT [7]	-	-	-	-	-	-	0.53	0.60	0.53	C++	34.04
JSC [5]	-	-	-	-	-	-	0.17	0.43	0.18	MATLAB	10.21
WELD	0.58	0.80	**0.64**	0.50	0.63	0.50	0.64	**0.81**	0.67	MATLAB & C++	2.43
F-WELD	**0.59**	0.80	**0.64**	**0.65**	0.78	**0.68**	**0.70**	0.80	**0.73**	MATLAB & C++	9.54

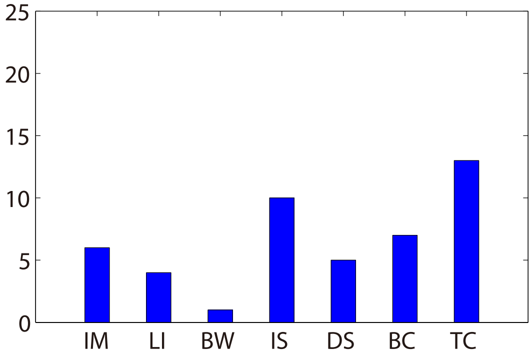


Fig. 5. Challenge distribution on the entire data set. The vertical ordinate indicates the number of video pairs.

Model (GMM) [7], and Visual Background Extractor (ViBe) [9].

2) Regarding thermal modality as input in the above methods, we obtain 11 thermal baselines.

3) To the best of our knowledge, none of the grayscale-thermal methods release the codes online. Thus, we implement two grayscale-thermal baselines for comprehensive evaluation, including GMM-GT and JSC [5]. In particular, GMM-GT fuses the detection results of GMM [7] with grayscale and thermal inputs.

It is worth mentioning that two variants of the proposed approach can also be regarded as additional baseline approaches in the benchmark (see Section VI-C for more details).

C. Evaluation Metrics

For the quantitative evaluation, we employ precision, recall, and F-measure as evaluation metrics, denoting P, R, and F, respectively. Their calculations are

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

where TP, FP, and FN denote true positive, false positive, and false negative, respectively.

VI. EXPERIMENTS

We present empirical evaluation and analysis of the proposed method against several baseline methods on both the newly created data set and the public OSU3 data set [16]. We further analyze the component contributions and efficiency of the proposed method. Finally, we discuss our limitations through failure cases.

A. Evaluation Settings

1) Datasets: We evaluate the proposed method on two data sets, the newly created one and the public one [16]. Our (OSU3) data set includes 25 (6) video sequence pairs and 1067 (8544) frame pairs in total, where the shortest and the longest video lengths are 24 (601) and 131 (2031), respectively. The video sequence pairs in the newly created data set are with three resolutions: 1) 320 × 240; 2) 384 × 288; and 3) 400 × 296, while the video sequence pairs in OSU3 are with the resolution 320 × 240. More details of the newly created data set are presented in Section V. For OSU3, its main challenges are shadow, IM, and DS. Since the original data set reported in [16] does not include the detection ground truths, we manually annotate the ground truths with 30 frame intervals for facilitating the evaluations.

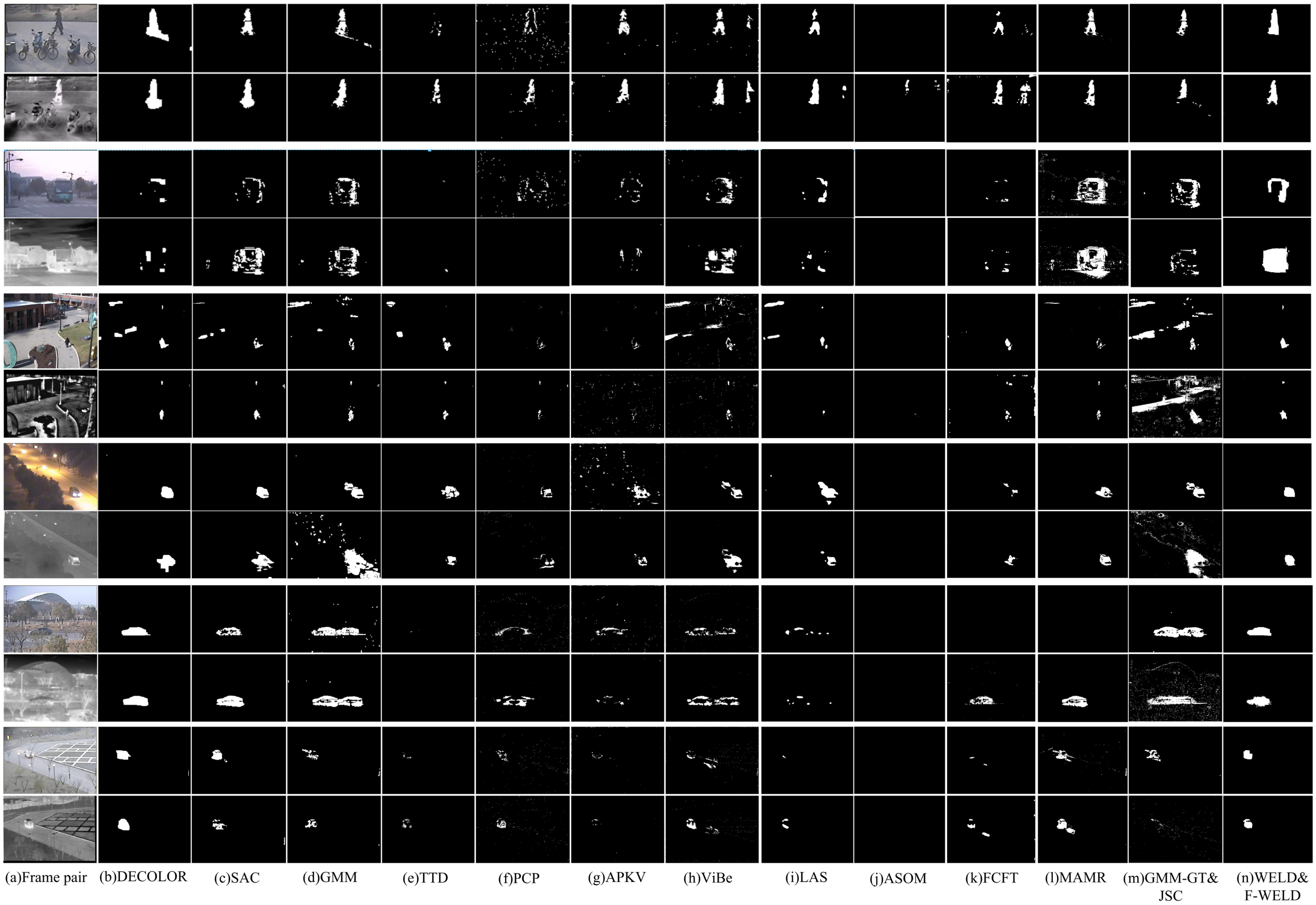


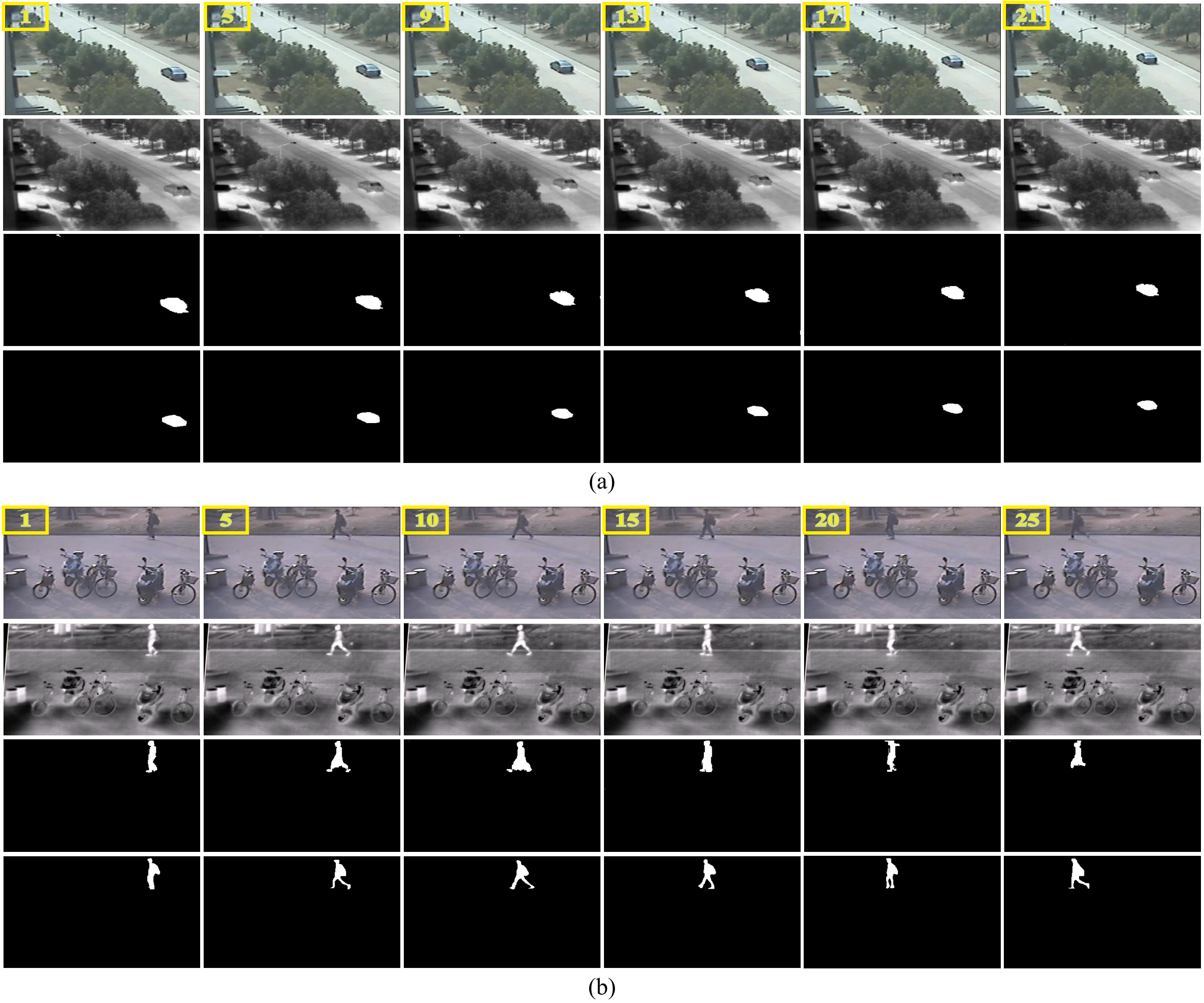
Fig. 6. Sample results of our method against other methods. The odd rows indicate the grayscale frames and the corresponding detection results generated by grayscale methods, and the even rows denote the thermal frames and the corresponding detection results generated by thermal methods, except for the last two columns. Herein, (i) and (j) indicate the detection results of grayscale-thermal methods (GMM-GT and JSC) and our methods (WELD and F-WELD), respectively.

2) Parameters: Although the proposed method has many parameters, we only need to set some of them, and fix them on both the newly created data set and the OSU3 data set for fair comparisons.

1) The parameter λ controls the complexity of the background model. We first give a rough estimate to the rank of the background model, denoting r for all modalities. Then, we initialize λ with a large value, 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 $(1/\sqrt{2}$ in our implementation) and repeat the SOFT-IMPUTE algorithm until $\text{rank}(\mathbf{B}^k) > r$ for all $k = 1, 2, \ldots, K$. Thus, we just need to set r to control the complexity of the background model in the experiments. Note that the shortest video length of two data sets is 24, and the longest video length of two data sets is 2031. r should be adjusted by the video length to adapt its variation. Therefore, we empirically set it to be $\sqrt{n}$, where n denotes the video length.
2) The parameter β controls the sparsity of the foreground masks. We typically set $\beta = 4.5\sigma^2$, where σ is estimated online by the mean variance of $\{\mathbf{D}^k - \mathbf{B}^k\}$. Since the estimation of $\{\mathbf{B}^k\}$ and σ is biased at the beginning iterations, similar to λ, we start our algorithm with a large β and then reduce it by a factor (0.5 in our implementation) after each iteration until it reaches $4.5\sigma^2$.
3) The parameter γ controls the spatial smoothness of the foreground and background, and can be adaptively adjusted by β. We empirically set it to be 6β.
4) The parameters $\{\phi_{\delta^k}\}$ are determined by the reconstruction error after the first iteration, as shown in Algorithm 1.
5) The above parameter settings are for WELD. In F-WELD, we empirically set $\{r, \gamma, \beta\} = \{(\sqrt{n}/3), 4\beta, 2.7\sigma^2\}$, and set the window size L and the similarity threshold τ to be 20 and 16, respectively.

B. Comparison Results

1) Overall Performance: We first report the overall performance of both the newly created and the public data sets. The quantitative and qualitative comparison results of the proposed approach on the newly created data set with three other kinds of baselines are presented in Table IV and Fig. 6. Here, we input the same modal data into WELD and F-WELD



(a)

(b)

Fig. 7. (a) and (b) Two sequential results of our method. The left-top number indicates the frame index. The grayscale frame, thermal frame, F-WELD results, and ground truth are presented in the first, second, third, and fourth rows, respectively.

to obtain their respective results on the single modality. From Table IV and Fig. 6, we can see that our F-WELD achieves a superior performance over other methods with different inputs, demonstrating the effectiveness of the proposed approach. We can further make the following main conclusions through the comparison results.

1) F-WELD and WELD achieve a superior performance than the grayscale and thermal methods. This observation not only demonstrates that fusing grayscale and thermal data can obtain more robust results than single modality, but it also identifies the importance of thermal information in foreground detection under challenging scenarios.
2) F-WELD and WELD substantially outperform other grayscale-thermal methods, justifying that the proposed methods can adaptively integrate grayscale and thermal information to achieve robust foreground detection.
3) F-WELD, WELD, and DECOLOR obtain a big superiority in recall. It validates the effectiveness of introducing the contiguous constraints of moving objects.
4) The superior performance of F-WELD on a single modality against other state-of-the-art approaches further demonstrates the effectiveness of the proposed approach.
5) The low-rank methods usually achieve promising detection results in challenging scenarios, such as F-WELD, WELD, MAMR, and DECOLOR. This may be attributed to the robustness to noises of the low-rank model. In addition, we also present the sequential results generated by our F-WELD in Fig. 7.

To further validate the effectiveness of the proposed methods, we evaluate them with other baseline methods on the public data set OSU3 [16]. The comparison results are shown in Fig. 8, which mainly accord with the observations on the newly created data set. In particular, the performance of F-WELD is slightly worse than WELD and significantly outperforms three other kinds of baseline approaches. Comparing the results on two data sets, we can find that many methods have a

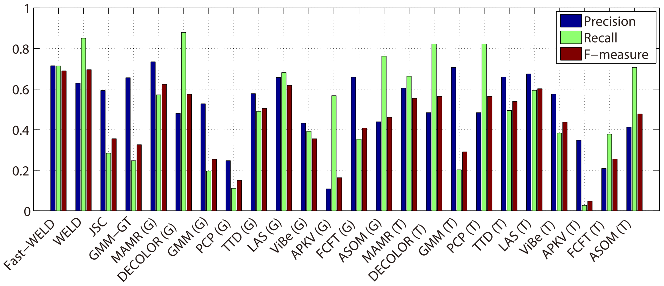


Fig. 8. Evaluation results of the proposed methods against other baseline methods on the public data set OSU3. G and T in the brackets indicate the grayscale data and the thermal data, respectively.

TABLE V

CHALLENGE-BASED RESULTS OF F-WELD AGAINST ITS RESULTS ON A SINGLE MODALITY, WHERE G INDICATES THAT WE ONLY INPUT THE GRAYSCALE DATA INTO F-WELD, WHILE T INDICATES THAT WE ONLY INPUT THE THERMAL DATA INTO F-WELD

Algorithm	LI			TC		
	P	R	F	P	R	F
F-WELD	0.73	0.72	0.71	0.73	0.75	0.72
F-WELD (G)	0.53	0.58	0.52	0.73	0.71	0.69
F-WELD (T)	0.83	0.64	0.68	0.65	0.55	0.55

bad performance in the newly created data set, but obtain good results in the OSU3, such as ASOM, FCFT, LAS, TTD, and PCP. It suggests that the proposed data set has bigger challenges than the public one, and will be helpful to promote the progress in grayscale-thermal foreground detection.

2) Challenge-Based Performance: To quantify the complementarity between the two modalities, we present the challenge-based comparison results in Fig. 9, including LI, intense shadow (IS), BC, and TC. For LI and IS, grayscale data are affected by the lighting condition, while thermal data are not. The results show that most of the compared methods have a higher performance on thermal data, as shown in Fig. 9(a) and (b). In BC, the moving objects are easier to be influenced by clutter backgrounds in grayscale data than in thermal data. Fig. 9(c) shows this point, i.e., most compared methods obtain a superior performance on thermal data than on grayscale data. In contrast, thermal videos with TC usually make detection algorithms have a weaker performance. In such a circumstance, grayscale information may be effective to distinguish the moving objects from the background, as demonstrated in Fig. 9(d). Note that our methods achieve high performance in all the four challenges. In addition, we also present the challenge-based results of F-WELD against its results on a single modality, as shown in Table V. The comparison results further justify the complementarity of these two modalities. Therefore, we can conclude that the proposed methods can complement each other to achieve robust detection under challenging scenarios.

3) Other Discussions: We first discuss the influences of the frame numbers on the performance and computational efficiency. The F-measure and the runtime of different frame numbers are shown in Fig. 10. Here, a different frame number indicates the different partition of input videos. For example, given a certain frame number n_0, we partition the input video into several video clips, in which each clip has n_0 adjacent frames. The detection results with the frame number n_0 are obtained by performing F-WELD on all video clips. From Fig. 10, on the one hand, we can see that the performance changes a little when the frame number is between 10 and 100 and the runtime increases slightly. Although F-WELD obtain much gain in efficiency, its F-measure decreases greatly when the frame number is 5. Therefore, we can conclude that at least 10 frames are required in our algorithm to keep a high performance. On the other hand, if we input all the frames of a video at one time, the length that can be processed depends on the memory of the computer used. However, if we temporally partition the video into some video clips, in which each clip is computable for the computer used, our algorithm can process an arbitrarily long video.

TABLE VI

PERFORMANCE AND RUNTIME OF DIFFERENT DOWNSAMPLING SCALARS ON THE NEWLY CREATED DATA SET

Scalar	P	R	F	**FPS**
2	0.72	0.81	0.74	3.50
3	0.70	0.80	0.73	9.54
4	0.67	0.78	0.71	9.82
5	0.64	0.76	0.68	19.61

TABLE VII

AVERAGE PRECISION, RECALL, AND F-MEASURE OF OUR METHOD AND ITS VARIANTS ON THE ENTIRE DATA SET. BOLD FONTS OF RESULTS INDICATE THE BEST PERFORMANCE

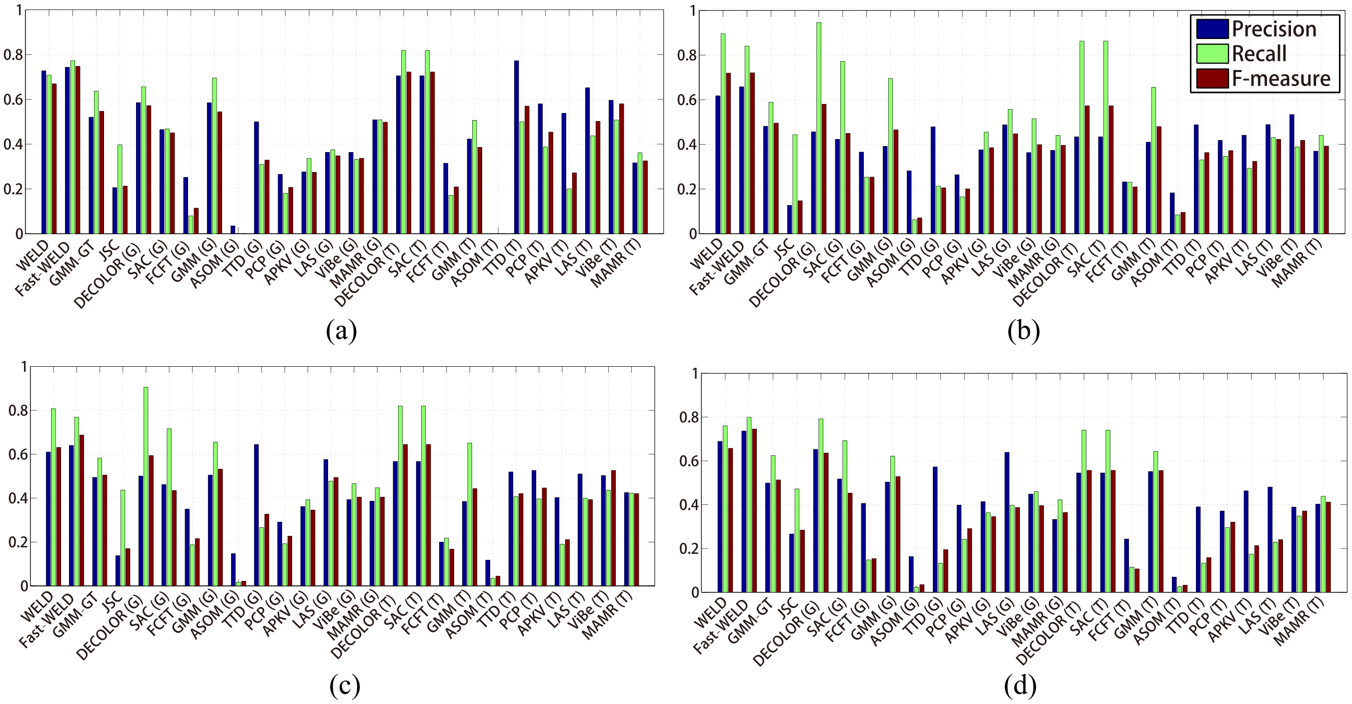
Algorithm	P	R	F
WELD	0.64	0.81	0.67
WELD-I	0.67	0.78	0.61
F-WELD	**0.70**	0.80	**0.73**
F-WELD-I	0.57	**0.90**	0.66

Then, we present the influences of the downsampling scalars on the performance and computational efficiency in Table VI. From the results, we can see that the performance decreases slightly, while the runtime decreases greatly when the downsampling scalar becomes 3 from 2. When the downsampling scalar is between 3 and 5, the results change a little in both the performance and computational efficiency. Therefore, we set the downsampling scalar to 3 in this paper to balance the accuracy-efficiency tradeoff.

C. Component Analysis

To justify the component contributions of the proposed approach, we further implement the following two variants.

1) WELD-I, which sets $\delta^k = 1/K$, $k = 1, 2, \ldots, K$, making each modality equal in reliable weights.
2) F-WELD-I, which replaces the gradient-driven downsampling and edge-preserving upsampling with the bilinear downsampling and bilinear upsampling, respectively.



(a) (b)

(c) (d)

Fig. 9. Challenge-based performance of the proposed methods against other baseline methods on the newly created data set. G and T in the brackets indicate the grayscale data and the thermal data, respectively. (a) Low illumination. (b) Intense shadow. (c) Background clutter. (d) Thermal crossover.

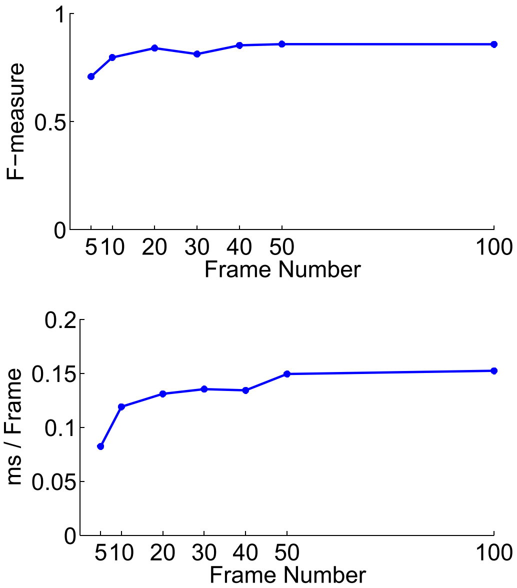


Fig. 10. F-measure and runtime of F-WELD of different number of frames, which are computed from three videos in the newly created data set. Each video has a length of 100 frames.

Table VII shows the quantitative results on the newly created data set and we can make the following observations.

1) F-WELD outperforms WELD in precision and F-measure. It justifies the contribution of the proposed accelerate algorithm. In F-WELD, we regard the original-resolution frames as guidance and recover the original-resolution results from the low-resolution ones by employing edge-preserving filtering. This process has the advantage of utilizing the structural information about the original-resolution frames for detection recovery.
2) WELD significantly outperforms WELD-I in precision and F-measure, and is slightly worse in recall. It demonstrates the importance of weights in the proposed model to achieve adaptive fusion of different modalities.
3) F-WELD substantially outperforms F-WELD-I in precision and F-measure, showing that the edge-preserving filtering-based accelerate algorithm can greatly improve the detection accuracy.

D. Efficiency Analysis

The experiments are carried out on a desktop with an Intel i7 4.0-GHz CPU and 32-GB RAM, and implemented on the mixing platform of C++ and MATLAB without any code optimization. The runtime of our method against other methods is presented in Table IV and all the frames are with a 320 × 240 resolution. From Table IV, we can see that F-WELD can speed up WELD by about 4 times (achieving about 10 frames/s) while improving its accuracy by 6%. Although ASOM, ViBe, GMM, FCFT, GMM-GT, and PCP are much faster than ours, these methods are much worse than F-WELD in precision, recall, and F-measure. LAS and JSC are comparable with F-WELD in efficiency and also have much worse accuracy than F-WELD. For MAMR and DECOLOR, they have heavy computational burdens in spite of obtaining good accuracies. These demonstrate that F-WELD obtains a good balance in efficiency and accuracy. Note that our F-WELD is near-real-time, and we can further reduce the

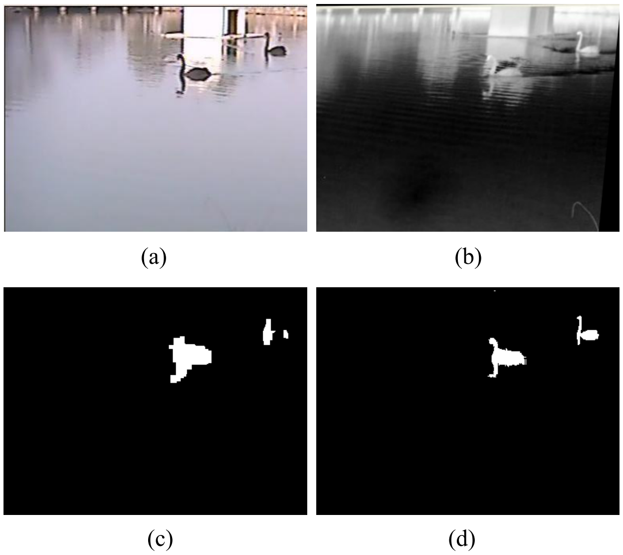


Fig. 11. Unsatisfying results generated by our method. (a) Grayscale frame. (b) Thermal frame. (c) WELD. (d) F-WELD.

computational burden by increasing the downsampling scalar without losing much accuracy, as shown in Table VI.

E. Limitation

We also present the unsatisfying results generated by our method, as shown in Fig. 11. When severe noises occur in one modality, such as moving shadows, another modal information can alleviate their effects. The detection results of the third frame pair in Fig. 6 show that our method can handle this situation well. However, when such noises occur in both the modalities [see Fig. 11(a) and (b)], our method will fail [see Fig. 11(c) and (d)]. This problem could be tackled by incorporating shape or other high-level knowledge to remove these noises.

VII. Conclusion

In this paper, we proposed a general algorithm for robust foreground detection in challenging scenarios by adaptively leveraging grayscale-thermal information and also substantially sped up our algorithm while preserving its accuracy by the edge-preserving filtering. Extensive experiments on the newly created and public grayscale-thermal data sets suggest that our approach achieved superior performance against other state-of-the-art approaches. In the future work, we will develop other priors on the foreground or background into our framework to further improve the robustness and extend our algorithm in a streaming or an online fashion for processing arbitrarily long videos with limited computational sources and spaces.

References

[1] R. Gade and T. B. Moeslund, “Thermal cameras and applications: A survey,” *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, 2013.

[2] C. Li, S. Hu, S. Gao, and J. Tang, “Real-time grayscale-thermal tracking via Laplacian sparse representation,” in *Proc. Int. Conf. MultiMedia Modeling*, 2016, pp. 54–65.

[3] J. W. Davis and V. Sharma, “Fusion-based background-subtraction using contour saliency,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, p. 11.

[4] J. Han and B. Bhanu, “Fusion of color and infrared video for moving human detection,” *Pattern Recognit.*, vol. 40, no. 6, pp. 1771–1784, 2007.

[5] G. Han, X. Cai, and J. Wang, “Object detection based on combination of visible and thermal videos using a joint sample consensus background model,” *J. Softw.*, vol. 8, no. 4, pp. 987–994, 2013.

[6] B. Zhao, Z. Li, M. Liu, W. Cao, and H. Liu, “Infrared and visible imagery fusion based on region saliency detection for 24-hour-surveillance systems,” in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2013, pp. 1083–1088.

[7] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 246–252.

[8] T. H. Tsai, C.-Y. Lin, and S.-Y. Li, “Algorithm and architecture design of human–machine interaction in foreground object detection with dynamic scene,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 1, pp. 15–29, Jan. 2013.

[9] O. Barnich and M. Van Droogenbroeck, “ViBe: A universal background subtraction algorithm for video sequences,” *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[10] C. Cuevas and N. García, “Efficient moving object detection for lightweight applications on smart cameras,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 1, pp. 1–14, Jan. 2013.

[11] M.-H. Yang, C.-R. Huang, W.-C. Liu, S.-Z. Lin, and K.-T. Chuang, “Binary descriptor based nonparametric background modeling for foreground extraction by using detection theory,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 595–608, Apr. 2015.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *J. ACM*, vol. 58, no. 3, 2011, Art. no. 11.

[13] X. Zhou, C. Yang, and W. Yu, “Moving object detection by detecting contiguous outliers in the low-rank representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[14] X. Guo, X. Wang, L. Yang, X. Cao, and Y. Ma, “Robust foreground detection using smoothness and arbitrariness constraints,” in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 535–550.

[15] X. You, Q. Li, D. Tao, W. Ou, and M. Gong, “Local metric learning for exemplar-based object detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1265–1276, Aug. 2014.

[16] J. W. Davis and V. Sharma, “Background-subtraction using contour-based fusion of thermal and visible imagery,” *Comput. Vis. Image Understand.*, vol. 106, no. 2, pp. 162–182, 2007.

[17] A. Torabi, G. Massé, and G.-A. Bilodeau, “An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications,” *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 210–221, 2012.

[18] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, “Thermal–visible registration of human silhouettes: A similarity measure performance evaluation,” *Infr. Phys. Technol.*, vol. 64, pp. 79–86, May 2014.

[19] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Aug. 2010.

[20] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[21] V. Kolmogorov and R. Zabin, “What energy functions can be minimized via graph cuts?” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[22] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. no. 96.

[23] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, “Cross-based local multipoint filtering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 430–437.

[24] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, “PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence,” *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.

[25] F. El Baf, T. Bouwmans, and B. Vachon, “Foreground detection using the Choquet integral,” in *Proc. 9th Int. Workshop Image Anal. Multimedia Interact. Services*, 2008, pp. 187–190.

[26] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, “SuBSENSE: A universal change detection method with local adaptive sensitivity,” *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[27] X. Cui, J. Huang, S. Zhang, and D. N. Metaxas, “Background subtraction using low rank and group sparsity constraints,” in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 612–625.

[28] C. Li, L. Lin, W. Zuo, S. Yan, and J. Tang, “SOLD: Sub-optimal low-rank decomposition for efficient video segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5519–5527.

[29] X. Ye, J. Yang, X. Sun, K. Li, C. Hou, and Y. Wang, "Foreground–background separation from video clips via motion-assisted matrix restoration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1721–1734, Nov. 2015.

[30] C. Li, L. Lin, W. Zuo, W. Wang, and J. Tang, "An approach to streaming video segmentation with sub-optimal low-rank decomposition," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1947–1960, May 2016.

[31] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 179–186, 2010.

[32] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*, vol. 4338. Madurai, India: Springer, 2006, pp. 528–539.

[33] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, "Stable principal component pursuit," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1518–1522.

[34] Y. She and A. B. Owen, "Outlier detection using nonconvex penalized regression," *J. Amer. Statist. Assoc.*, vol. 106, no. 494, pp. 626–639, 2011.

[35] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[36] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[37] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy *c*-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.

[38] S. Bahrampour, B. Moshiri, and K. Salahshoor, "Weighted and constrained possibilistic C-means clustering for online fault detection and isolation," *Appl. Intell.*, vol. 35, no. 2, pp. 269–284, 2011.

[39] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 839–846.

[40] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 1073–1079, Jul. 2009.

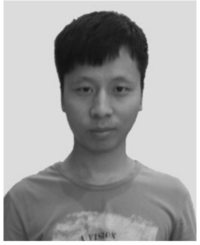
[41] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[42] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.

[43] H. Zhang and D. Xu, "Fusing color and texture features for background model," in *Proc. 3rd Int. Conf. Fuzzy Syst. Knowl. Discovery*, 2006, pp. 887–893.

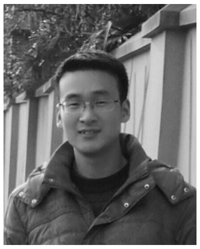
[44] M. Narayana, A. Hanson, and E. Learned-Miller, "Background modeling using adaptive pixelwise kernel variances in a hybrid feature space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2104–2111.

[45] O. Oreifej, X. Li, and M. Shah, "Simultaneous video stabilization and moving object detection in turbulence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 450–462, Feb. 2013.



Chenglong Li received the B.S. degree in applied mathematics and the M.S. degree in computer science from Anhui University, Hefei, China, in 2010 and 2013, respectively, where he is currently working toward the Ph.D. degree in computer science.

His research interests include computer vision, machine learning, and intelligent media technology.

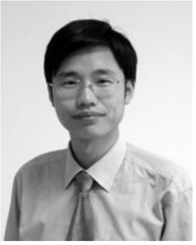


Xiao Wang received the B.S. degree from Western Anhui University, Luan, China, in 2013 and the M.S. degree from Anhui University, Hefei, China, in 2016, where he is currently working toward the Ph.D. degree in computer science.

His research interests include computer vision, machine learning, and pattern recognition.

Lei Zhang received the B.S. degree from Western Anhui University, Luan, China, in 2013 and the M.S. degree from Anhui University, Hefei, China, in 2016.

His research interests include computer vision and machine learning.



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.

Hejun Wu received the Ph.D. degree in computer science and engineering from Hong Kong University of Science and Technology, Hong Kong.

He joined Sun Yat-sen University, Guangzhou, China, in 2009. He served as a Post-Doctoral Research Fellow with Hong Kong University of Science and Technology and Hong Kong Ploytechnic University, Hong Kong, in 2008 and 2009. He is currently an Associate Professor with the Department of Computer Science, Sun Yat-sen University, and the Assistant Dean of the School of Information Science and Technology. His research interests include wireless sensor network, mobile cloud computing, distributed databases, embedded systems, and pervasive computing.

Liang Lin received the B.S. and Ph.D. degrees from Beijing Institute of Technology, Beijing, China, in 1999 and 2008, respectively.

He was a Joint Ph.D. Student with the Department of Statistics, University of California at Los Angeles (UCLA), Los Angeles, CA, USA, from 2006 to 2007. He was a Post-Doctoral Research Fellow with the Center for Vision, Cognition, Learning, and Art, UCLA. He is currently a Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has authored over 80 papers in top-tier academic journals and conferences. His research interests include new models, algorithms, and systems for intelligent processing and understanding of visual data, such as images and videos.

Prof. Lin was a recipient of the Best Paper Runners-Up Award in NPAR 2010, the Google Faculty Award in 2012, the Hong Kong Scholars Award in 2014, and the Best Student Paper Award in the IEEE ICME in 2014. He serves as an Associate Editor of *Neurocomputing* and *The Visual Computer*.