Background subtraction with multi-scale structured low-rank and sparse factorization

Aihua Zheng, Tian Zou, Yumiao Zhao, Bo Jiang, Jin Tang, Chenglong Li*

School of Computer Science and Technology, Anhui University, Hefei, China

Abstract

Low-rank and sparse factorization, which models the background as a low-rank matrix and the foreground as the contiguously corrupted outliers, exhibits excellent performance in background subtraction, in which the structured constraints of the foreground usually play a very essential role. In this paper, we propose a novel approach with multi-scale structured low-rank and sparse factorization for background subtraction. Different from the conventional methods that only enforce the smoothness between the spatial neighbors, we propose to explore the structured smoothness with both appearance consistency and spatial compactness in the low-rank and sparse factorization framework. Moreover, we integrate structural information at different scales into the formulation for robustness. We also design a low-rank decomposition scheme to improve the computational efficiency of the optimization algorithm. Extensive experiments on benchmark datasets GTFD and CDnet suggest that our approach achieves big superior performance against the state-of-the-art methods.

1. Introduction

Moving object detection is a fundamental problem in computer vision, and plays a critical role in numerous vision applications, such as intelligent transportation [1], vehicle navigation [2] and scene understanding [3]. In the past decades, extensive algorithms made remarkable efforts to moving object detection, as referred in the interesting surveys [4–7], while background subtraction has been recognized as one of the most competitive approaches. Conventional background subtraction methods include GMM [8], VIBE [9] and their variants. Qin et al. [10] further introduced the Gabor filter into VIBE to reduce the “ghost” and then accelerated the updating procedure via GPU parallel computing. Zhang et al. [11] revealed the foreground and background imbalance problem and designed a two-stage compensation framework on the data and the algorithm level sequentially. Kim and Jung [7] comprehensively discussed the issue of illumination changes which is ubiquitous in the outdoor scenarios from the methodology and performance prospectives. However, it is still a challenging task due to undesired environment such as dynamic scene, illumination changes, severe weather, occlusions and so on.

Meanwhile, deep learning based methods have been widely used in computer vision and pattern recognition. Some state-of-the-art networks have been designed for semantic segmentation, including FCN [12], DeepLab [13], MSCNN [14] and so on. Specifically, moving object detection tries to segment the moving foreground out of the background. Zhang et al. [15] designed a block-wise deep learning module to encode the image representation and then accelerated the detection via binary scene modeling in Hamming space. Heo et al. [16] combined the appearance network (a pre-trained VGG-16 network) and the motion network (a shallow network) for moving object detection. Chen et al. [17] designed an end-to-end architecture which first extracted high-level CNN features via an encoder–decoder network then modeled pixel-wise temporal changes via an Attention ConvLSTM. Kim and Kelles [18] developed a encoder–decoder neural network embedded the image into multi-scale feature space under a triplet framework during the encoder while learning the feature to image mapping via a transposed convolutional network during decoder.

Recently, the low-rank and sparse factorization framework has attracted a considerable attention. The basic idea is to factorize the given matrix of the accumulated frames into the low-rank background and sparse foreground as outliers. The pioneering work is Robust Principal Component Analysis (RPCA) [19,20] and its variants [21–23]. However, RPCA-based methods are sensitive to the outlier noise and tend to produce cavities due to the laciness of structural contiguous constraints. In order to enforce structured constraint on the foreground, DECOLOR [24] proposed to model the spatial contiguity among the neighboring pixels to preserve the spatial homogeneity of the foreground, Xin et al. [25] exploited...
the continuous structural information via efficient GFL [26] to strengthen the fusion among the adjacent pixels. GOSUS [27] proposed to impose the group sparsity on the pre-processed sliding windows or superpixels to encourage the perceptually meaningful groups. Yang et al. [28] proposed a subspace-based motion segmentation method by integrating the global sparse subspace optimization via PCA projection and local refinement via a simple error estimation. Javed et al. [29] presented an online matrix decomposition with max-norm regularization and structured sparsity constraints on each superpixel segment. However, they construct the structured constraint only between the spatial neighbors while ignoring the high-level spatial compactness of the foreground. Moreover, none of them take into account the multi-scale cues which can further promote the appearance consistency and spatial compactness on varying scales.

In this paper, we propose a novel multi-scale structured low-rank and sparse factorization for background subtraction. Comparing to the deep-based methods, we do not require laborious pre-training or the large training set. In addition, we have no need of saving a large pre-trained deep model. We enforce our structured constraints in terms of integrating the multi-scale appearance consistency and spatial compactness of the foreground into a unified model. First, we observe that, the foregrounds are generally consistent in appearance. As demonstrated in Fig. 1(b), we should penalize the neighboring pixels 1 and 3 with inconsistent appearance while enhancing the structure between neighboring pixels 1 and 2 with consistent appearances. Second, we observe that the foregrounds are homogeneous in the same concept of spatial region, such as the same superpixel as shown in Fig. 1(c) and (d). Therefore, we further encourage this structure of spatial compactness on the foregrounds in this paper.

In addition, we have observed that structure of the foreground presents diversely on varying scales. As compared in Fig. 1(c) and (d), Coarse scale (with smaller number of superpixels as Fig. 1(c)) imposes global structure of the pattern (such as the whole body imposed in a single superpixel), while fine scale captures local structure (such as the head imposed in a single superpixel in Fig. 1(d)). In this paper, we propose to integrate the multi-scale cues into an unified structured low-rank and sparse factorization model to capture the diverse structure of the foregrounds.

As for the optimization, the low-rank model always suffers from heavy computational burden due to singular value decomposition procedure in most of the existing batch mannered models. To relieve this issue, we introduce an alternative definition of the nuclear-norm with an efficient decomposition strategy in this paper.

As summarized, we make the following contributions for moving object detection and related applications. We propose an effective approach for background subtraction. Our approach takes both appearance consistency and spatial compactness in a multi-scale manner for the separation of background and moving objects. In this way, it is capable of capturing rich information among pixels and thus improves the detection performance significantly. We design an efficient algorithm to optimize the proposed model. In particular, we decompose the low-rank background matrix into two sub-matrices to avoid SVD operations in each iteration, and thus deliver an efficient solver to the proposed model. Extensive experiments on two benchmark datasets GTFD and CDnet demonstrate that, our approach can better preserve the boundary of the foregrounds and achieves big superior performance against the state-of-the-arts.

The rest of paper is organized as follow: Section 2 briefly reviews the related work on low-rank and sparse factorization for moving object detection. We elaborate our approach and the efficient optimization in Section 3. Section 4 demonstrates the experimental results on two benchmark datasets while Section 5 concludes our paper.

2. Related work

In the past years, extensive work have developed on low-rank and sparse factorization for moving object detection. One pioneering work is Robust Principal Component Analysis (RPCA) [19,20], which decomposes a given matrix/frames into a low-rank background matrix and sparse foreground matrix. Candès et al. [21] proposed to recover the low-rank and sparse components individually by solving a convenient convex program called Principal Component Pursuit (PCP). Zhou et al. [22] proposed to handle both small entrywise noises and gross sparse errors. Dou et al. [23] proposed a incremental learning model using K-SVD for dictionary learning, Ebadi et al. [31] constructed the image sequence into the low-rank background and a dynamic tree-structured sparse foreground. Javed et al. [32] introduced a motion-aware prior obtained by optical flow as the regularization of graphs into the low-rank component. To enforce the spatial smoothness, Zhou et al. [24] proposed to relax the requirement of sparse and random distribution of corruption by preserving $l_0$-penalty and modeling the spatial contiguity of the sequence. CLASS [33] proposed a collaborative framework to leverage the various size of the moving objects via introducing the global appearance consistency. However, they considered the spatial smoothness only according to the coordinates of the pixels while ignoring the appearance similarity. Xin et al. [25] introduced the intensity similarities to the neighboring pixels via regularization terms to enforce the appearance smoothness. Shakeri et al. [34] presented an online sequential framework via solving sequential low-rank approximation and contiguous outlier representation problem. However, these methods constructed the graph only based on pixel level which ignored the spatial compactness. Recently, Javed et al. [29] presented an online matrix decomposition using max-norm constraints on each superpixel segment. However, they only enforce the structured constraint between the spatial neighbors while ignoring the more informative structures of the foreground including the spatial compactness and the multi-scale cues.

3. Our approach

In this section, we will elaborate our multi-scale structured low-rank and sparse factorization model, followed by the alternating optimization algorithm.
3.1. Problem formulation

We formulate the problem of background subtraction as structured low-rank and sparse factorization model. A video sequence $D = \{f_1, f_2, ..., f_t\} \in \mathbb{R}^{m \times n}$ is composed of $n$ frames by of $m$ pixels per frame. $B \in \mathbb{R}^{m \times n}$ is a background matrix, which denotes the underlying background images. Our goal is to discover the object mask $S$ from data matrices $D$, where $S_{ij}$ is a binary matrix:

$$S_{ij} = \begin{cases} 0, & \text{if } ij \text{ is background,} \\ 1, & \text{if } ij \text{ is foreground.} \end{cases}$$

We assume that the underlying background images with low-rank structure and sparse and contiguous structure, which has been successfully applied in structured background modeling [24,35]. Furthermore, for the background region where $S_{ij} = 0$, we assume that $D_{ij} = B_{ij} + e_{ij}$, where $e_{ij}$ denotes i.i.d. Gaussian noise. Based on the above assumptions, we have:

$$\min_{S_{ij} \in \{0,1\}} \alpha \| \text{vec}(S) \|_0$$

subject to $S_{ij} \circ D = S_{ij} \circ (B + \epsilon)$, rank $(B) \leq r$.

where $\alpha$ is a penalized factor, $\|X\|_0$ indicates the $l_0$ norm of a vector. "$\circ$" denotes element-wise multiplication of two matrices, $S_{ij}$ denotes complementary matrix of $S$, i.e., $S + S = 1$. $r$ is a constant that suppresses the structure complexity of the background model.

3.1.1. Appearance consistency

In order to preserve the spatial structure of the objects, [24,35] constructed the graph based on the neighboring pixels. However, as we discussed in Fig. 1(b), it is essential to penalise the neighbors with inconsistent appearances while enhancing the ones with consistent appearances [25,29]. Therefore, we enforce the appearance consistency into the structure of the informative graphs (as shown in Fig. 2(a)) by:

$$\min_{S_{ij} \in \{0,1\}} \alpha \| \text{vec}(S) \|_0$$

subject to $S_{ij} \circ D = S_{ij} \circ (B + \epsilon)$, rank $(B) \leq r$.

where $\alpha$ is a penalized factor, $\|X\|_0$ indicates the $l_0$ norm of a vector. "$\circ$" denotes element-wise multiplication of two matrices, $S_{ij}$ denotes complementary matrix of $S$, i.e., $S + S = 1$. $r$ is a constant that suppresses the structure complexity of the background model.

3.1.2. Spatial compactness

It is observed that, the pixels from the same superpixel, which is a perceptually compact unit with consistent color and texture, are basically derived from the same pattern (background/foreground) as shown in Fig. 1(c) and (d). In order to enforce the structure of spatial compactness, we first generate the superpixel segmentation via SLIC [30], which is simple and efficient comparing to LRW [36] and GWT [37]. Then, we construct the fully connected graph between the pixels within each superpixel as illustrated in Fig. 2(b). We introduce this structured constraint into the model via:

$$\min_{B^k, S_{ij} \in \{0,1\}} \sum_{k=1}^{K} \sum_{ij \in N_k} \alpha \| \text{vec}(S_{ij}) \|_0 + \| \text{vec}(S^k) \|_1 + \| \text{vec}(S^k) \|_1$$

subject to $S_{ij} \circ D = S_{ij} \circ (B^k + \epsilon)$, rank $(B^k) \leq r$.

where $K$ indices edge set connecting all the pixel pairs within each superpixel and $A$ is the node-edge incidence matrix denoting the connecting relationship among pixels. Noted that we don't enforce additional appearance similarities onto the graph within a superpixel since the superpixel is a conceptual group with similar appearances. In a sense, the compact structure in a superpixel can promote the appearance consistency simultaneously.

3.1.3. Multi-scale integration

Noted that the scale of superpixel segmentation controls the diversity of the foreground structure. As shown in Fig. 1(c) and (d), the coarse/fine scale with smaller/larger number of superpixels represents the global/local structure of the foreground. In order to capture the diverse foreground structures on varying scales, we further integrate above structured constraints into a multi-scale fashion. Considering the $S_{ij}$, $k = 1, \ldots, K$ indicates the foreground support matrix on the $k$th scale, we encourage the foreground supports from varying scales close to the true foreground $S$. As concluded, we formulate the multi-scale structured foreground integration as:

$$\min_{B^k, S_{ij} \in \{0,1\}} \sum_{k=1}^{K} \sum_{ij \in N_k} \alpha \| \text{vec}(S_{ij}) \|_0 + \| \text{vec}(S^k) \|_1 + \| \text{vec}(S^k) \|_1$$

subject to $S_{ij} \circ D = S_{ij} \circ (B^k + \epsilon)$, rank $(B^k) \leq r$.

with:

$$\| \text{vec}(S^k) \|_1 = \beta \| \text{vec}(S^k) \|_1 + \gamma \| \text{vec}(S^k) \|_1,$$

where $\beta$, $\gamma$ and $\eta$ are the tuning parameters leveraging the contribution of appearance consistency, spatial compactness and the multi-scale integration respectively.

3.2. Model optimization

Eq. (6) is a NP-hard problem, to make it tractable, we relax the rank operator on $B$ with the nuclear norm which has been proven as an effective convex surrogate of the rank operator [38]. Therefore, Eq. (6) can be reformulated as:

$$\min_{B^k, S_{ij} \in \{0,1\}} \frac{1}{2} \| \text{vec}(S_{ij}) \|_0^2 + \alpha \| \text{vec}(S) \|_0$$

subject to $S_{ij} \circ D = S_{ij} \circ (B^k + \epsilon)$, rank $(B^k) \leq r$.

where $\alpha$ is a balancing parameter. $\| \cdot \|$ and $\| \cdot \|$ indicate the nuclear norm and the Frobenius norm of a matrix, respectively.

We adopt an alternating algorithm by separating Eq. (8) over $B^k$, $S^k$ and $S$ via solving the following three subproblems.
3.2.1. Solving $B^k$

Due to the high computational complexity of singular value decomposition via optimizing the nuclear norm, we decompose the background $B^k$ into two sub-matrices as $B^k = M^kN^k$ inspired by Mazumder et al. [39] and reformat Eq. (8) as:

$$
\min_{\{S_k, s_k\}_k=1} \sum_{k=1}^{K} \left( \frac{1}{2} ||S_k \odot (D - M^kN^k) ||_F^2 + \alpha \| \text{vec}(S^k) \|_0 \right.
$$

$$
+ \frac{1}{2} \|E^k \text{vec}(S^k)\|_1 + \eta \| S^k - S \|_F^2 + \frac{\lambda}{2} \| M^k \|_F^2 + \frac{\lambda}{2} \| N^k \|_F^2 \right). 
$$

(9)

Given an current foreground mask $S^k$, estimating $B^k$ by minimizing Eq. (8) turns out to be the matrix completion problem. This is to learn a low-rank background matrix from partial observations:

$$
\arg \min_{M^k, N^k} || P^k - M^kN^k ||_F^2 + \lambda (\| M^k \|_F^2 + \| N^k \|_F^2),
$$

(10)

where $P^k = S^k \odot D + S^k \odot B^k$, which can be minimized along one coordinate direction at each iteration. We expand this procedure as follows:

First, fix the other variables and update $M^k$ by solving the problem:

$$
\arg \min_{M^k} || P^k - M^kN^k ||_F^2 + \lambda (\| M^k \|_F^2 + \| N^k \|_F^2),
$$

(11)

which has a closed-form solution given as:

$$
M^k = P^k N^k (N^k N^k + \lambda I)^{-1}.
$$

(12)

Then, $N^k$ plays a symmetric role to $M^k$ which can be updated by solving:

$$
\arg \min_{N^k} || P^k - M^kN^k ||_F^2 + \lambda (\| M^k \|_F^2 + \| N^k \|_F^2),
$$

(13)

with a closed-form solution given as:

$$
N^k = (M^k + \lambda I)^{-1} M^k P^k.
$$

(14)

Finally, $B^k$ can be achieved via:

$$
B^k = M^kN^k.
$$

(15)

3.2.2. Solving $S^k$

Given a current estimate of the background position matrix $B^k$, Eq. (8) can be transferred into following optimization function:

$$
\min_{S^k} \frac{1}{2} || S_k \odot (D - B^k) ||_F^2 + \| \text{vec}(S^k) \|_0 + \| E^k \text{vec}(S^k)\|_1
$$

$$
+ \eta \| S^k - S \|_F^2.
$$

(16)

The energy function Eq. (16) can be rewritten in line with the standard form of a first-order Markov Random Fields [40] as:

$$
\frac{1}{2} || S_k \odot (D - B^k) ||_F^2 + \alpha \| \text{vec}(S^k) \|_0 + || E^k \text{vec}(S^k)\|_1
$$

$$
+ \eta \| S^k - S \|_F^2
$$

$$
= \frac{1}{2} \sum_{i,j} (D_{ij} - B^k_{ij})^2 (1 - S^k_{ij}) + \alpha \sum_{i,j} S^k_{ij} \| E^k \text{vec}(S^k)\|_1
$$

$$
+ \eta \sum_{i,j} (1 - 2S^k_{ij}) S^k_{ij}
$$

$$
= \sum_{i,j} \left[ (\alpha - \frac{1}{2} (D_{ij} - B^k_{ij})^2 + \eta (1 - 2S^k_{ij}) S^k_{ij} + \| E^k \text{vec}(S^k)\|_1
$$

$$
+ \frac{1}{2} \sum_{i,j} (D_{ij} - B^k_{ij})^2.
$$

(17)

$S^k_{ij}$ is a constant, $\frac{1}{2} \sum_{i,j} (D_{ij} - B^k_{ij})^2$ is also a constant with fixed $B^k$. Meanwhile, $S^k_{ij}$ which can be achieved via Eq. (19) is also a constant. Known Markov unary term and pairwise smoothing term, one can easily obtain the optimal foreground matrix though graph cuts method [41,42] since $S^k_{ij} \in [0,1]$ is discrete.

3.2.3. Solving $S$

Once attained the current $S^k$ and $B^k$ for each scale, Updating $S$ turns to be:

$$
\min_{S^k} \frac{1}{K} \sum_{k=1}^{K} || S^k - S \|_F^2,
$$

(18)

which has a closed-form solution given as:

$$
S = \frac{1}{K} \sum_{k=1}^{K} S^k.
$$

(19)

A sub-optimal solution can be obtained by alternating optimizing $B^k, S^k$ and $S$ and the algorithm is summarised in Algorithm 1.

Algorithm 1 Optimization Algorithm to Eq. (8).

Input: $D = [I_1, I_2, ..., I_L] \in \mathbb{R}^{M \times N}$.

Initialize $B^k = D, S = S^k = 0, \tau = 1 + e - 4, maxIter = 20$.

Output: $S, B^k, S^k$.

1. For $k = 1$ to $K$ do

2. Updating $B^k$: optimizing energy function Eq. (10) via:

   (1) updating $M^k$: $M^k \leftarrow P^k N^k (N^k N^k + \lambda I)^{-1}$.

   (2) updating $N^k$: $N^k \leftarrow (M^k + \lambda I)^{-1} M^k P^k$.

   (3) updating $B^k$: $B^k \leftarrow M^k N^k$.

3. Updating $S^k$: using graph cuts to optimize energy function Eq. (16) by:

   $S^k \leftarrow \text{argmin}_{S} \sum_{i,j} \left[ (\alpha - \frac{1}{2} (D_{ij} - B^k_{ij})^2 + \eta (1 - 2S_{ij}) S^k_{ij} + \| E^k \text{vec}(S^k)\|_1
$$

   $+ \frac{1}{2} \sum_{i,j} (D_{ij} - B^k_{ij})^2)^2.

4. End For.

5. Updating $S$: $S \leftarrow \frac{1}{K} \sum_{k=1}^{K} S^k$.

6. Check the convergence condition: if the maximum objective change between two consecutive iterations is less than $\tau$ or the maximum number of iterations reaches maxIter, then terminate the loop.

4. Experiments

We evaluate our approach on GTFD [35] and CDnet14 [43] against five state-of-the-arts including COROLA [34], DECOLOR [24], GMM [8], VIBE [9] and PCP [21]. To keep things fair, we choose the default parameters released by the authors for corresponding methods and resize the resolution of all the frames into $160 \times 120$.

4.1. Evaluation settings

4.1.1. Datasets

GTFD [35] dataset consists of 25 video sequence pairs in both visual and thermal modality with various challenges including intermittent motion, low illumination, bad weather, intense shadow, dynamic scene and background clutter etc. We evaluate the proposed method on visual modality videos.

CDnet14 [43] is a large scale dataset consisting of 11 different categories with 55 video sequences. We evaluate our
Table 1
Evaluated parameters on GTFD dataset.

<table>
<thead>
<tr>
<th>Param</th>
<th>Setting</th>
<th>F-measure</th>
<th>Param</th>
<th>Setting</th>
<th>F-measure</th>
<th>Param</th>
<th>Setting</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>16.2σ²</td>
<td>0.69</td>
<td>β</td>
<td>0.14</td>
<td>0.63</td>
<td>γ</td>
<td>0.014</td>
<td>0.64</td>
</tr>
<tr>
<td>η</td>
<td>0.006</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Sample results of our method against the state-of-the-arts on six videos from GTFD dataset.

method on 10 videos from 5 challenging categories including Base-
line (Office, PETS2006), IntermittentObjectMotion (Sofa, winterDrive-
way), nightVideos (TramStation, WinterStreet), Shadow (CopyMachine, BusStation) and Thermal (Park, Corridor).

4.1.2. Evaluation criterion
The Precision, Recall, F-measure are first comprehensively evalu-
ated, which are defined as following:

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

(20)

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

Furthermore, the Mean Absolute Error (MAE) is evaluated to 
measure the disagreement between the detected results and the 
groundtruth:

\[
\text{MAE} = \frac{1}{N \times F} \sum_{i=1}^{F} \sum_{p \in \text{DR}, \hat{p} \in \text{GT}} \text{XOR}(p, \hat{p})
\]

(21)

where N denotes the resolution and frame and F denotes the 
number of the frames in the video clip. DR and GT indicate the 
“Detection Result” and the “Ground Truth” respectively. XOR(·) 
denotes the logic operator “exclusive OR”. p, \( \hat{p} \in \{0, 1\} \) denotes the background/foreground pixels.

4.1.3. Parameters
There are six important parameters in our model. We adjust 
one parameter while fixing other parameters and then obtain best 
performance for our approach. In our model of Eq. (8), the pa-
parameter α which controls the sparsity structure of the foreground 
masks is set as \( \alpha = 16.2\sigma^2 \), where \( \sigma^2 \) is estimated online by the 
mean variance of \( |\mathbf{D}_0 - \mathbf{B}_0| \). The parameter \( \beta, \gamma, \eta \) and \( \lambda \) control the relative contribution of each term in Eqs. (7) and (8), re-
spectively. We empirically set as \( \{\beta, \gamma, \eta, \lambda\} = \{0.21, 0.028, 0.007, 8.5\} \). The number of superpixels on two different scales are \( \{A_1, A_2\} = \{650, 950\} \) with \( K = 2 \). We evaluate the parameter variations on 
GTFD dataset and report the corresponding F-measure in Table 1. It 
is worth to note that our method is insensitive to the parameters.

4.2. Qualitative results
Figs. 3 and 4 demonstrate the detected results on a certain 
frame of video clips from GTFD [35] and CDnet14 [43] respec-
tively. From which we can see, our method can produce more pre-
cise boundary information and better suppress the influence of the 
noise. PCP is not able to capture contiguous structure of the fore-
ground since only enforcing the sparse structure of the foreground. 
GMM and VIBE work on the original pixel space therefore they 
are quite sensitive to the noise and introduce “ghost”. DECOLOR is 
much more robust but fails to sketch the contours of the objects. 
COROLA, as the state-of-the-art online detection model, tends to 
produce the cavities on the objects due to the lack of structure 
constraint of spatial compactness.

4.3. Quantitative results
Tables 2 and 3 report Precision, Recall, F-measure, and MAE on 
public GTFD dataset [35] and CDnet14 [43] respectively. It is clear 
to see that our method significantly outperforms the state-of-the-
arts in precision. Although the recall of our method looks lower
than DECOLOR [24], from Fig. 3 we can see, DECOLOR [24] tends to produce coarse boundary which always leads to high recall. The more balanced criteria between precision and recall, F-measure, together with the MAE verify the promising performance of our method.

4.4. Component analysis

In order to validate the contribution of our multi-scale structure integration of appearance consistency and spatial compactness, we evaluate several variants of our model on GTFD and report the results on Table 4, where OURS: the proposed model; OURS-I: our model without spatial compactness structure by setting γ = 0; OURS-II: our model without appearance consistency structure by setting all $w_{ij,m} = 1$; OURS-III: our model on the single scale (without multi-scale structure) by setting $\eta = 0$ with the number of superpixels $A = 950$. From Table 4 we can see that: 1) Both appearance consistency and spatial compactness structures play important roles for moving object detection. 2) The multi-scale integration can further boost the performance. Together with the visualized examples on Fig. 5 we can see that: 1) After introducing the spatial compactness structure (comparing OURS to OURS-I), the model can better capture the global spatial coherence and suppress the influence of the noises. 2) After introducing the appearance consistency structure (comparing OURS to OURS-II), the model can better preserve the local contour of the objects. 3)

### Table 2

The average Precision, Recall, F-measure and MAE values on GTFD dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.29</td>
<td>0.40</td>
<td>0.51</td>
<td>0.54</td>
<td>0.53</td>
<td><strong>0.65</strong></td>
</tr>
<tr>
<td>Recall</td>
<td>0.18</td>
<td>0.47</td>
<td>0.64</td>
<td><strong>0.82</strong></td>
<td>0.62</td>
<td>0.80</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.22</td>
<td>0.39</td>
<td>0.51</td>
<td>0.59</td>
<td>0.52</td>
<td><strong>0.69</strong></td>
</tr>
<tr>
<td>MAE</td>
<td>0.0155</td>
<td>0.0169</td>
<td>0.0123</td>
<td>0.0061</td>
<td>0.0135</td>
<td><strong>0.0048</strong></td>
</tr>
</tbody>
</table>

Fig. 4. Sample results of our method against the state-of-the-arts on ten videos from CDnet dataset.
Table 3
Comparison of Precision (P), Recall (R), F-measure (F) and MAE on videos from CDnet14 dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Office</td>
<td>0.65</td>
<td>0.75</td>
<td>0.81</td>
<td>0.87</td>
<td>0.81</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.73</td>
<td>0.54</td>
<td><strong>0.83</strong></td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.73</td>
<td>0.63</td>
<td>0.82</td>
<td>0.76</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>0.69</td>
<td>0.74</td>
<td>0.88</td>
<td>0.64</td>
<td>0.59</td>
<td>0.89</td>
</tr>
<tr>
<td>PETS2006</td>
<td>0.13</td>
<td>0.22</td>
<td>0.27</td>
<td>0.23</td>
<td>0.15</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.33</td>
<td>0.40</td>
<td>0.30</td>
<td>0.23</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>0.48</td>
<td>0.47</td>
<td>0.85</td>
<td>0.87</td>
<td>0.82</td>
<td>0.87</td>
</tr>
<tr>
<td>Sofa</td>
<td>0.05</td>
<td>0.36</td>
<td>0.30</td>
<td>0.44</td>
<td>0.36</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.40</td>
<td>0.43</td>
<td>0.56</td>
<td>0.47</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.27</td>
<td>0.62</td>
<td>0.66</td>
<td>0.64</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>WinterDriveway</td>
<td>0.05</td>
<td>0.32</td>
<td>0.17</td>
<td>0.48</td>
<td>0.35</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.41</td>
<td>0.26</td>
<td>0.52</td>
<td>0.37</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td><strong>0.50</strong></td>
<td>0.27</td>
<td>0.32</td>
<td>0.41</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>TransStation</td>
<td>0.19</td>
<td>0.65</td>
<td>0.54</td>
<td><strong>0.98</strong></td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>0.27</td>
<td>0.37</td>
<td>0.39</td>
<td>0.57</td>
<td>0.60</td>
<td><strong>0.61</strong></td>
</tr>
<tr>
<td></td>
<td>0.51</td>
<td>0.21</td>
<td>0.22</td>
<td>0.20</td>
<td>0.45</td>
<td>0.61</td>
</tr>
<tr>
<td>WinterStreet</td>
<td>0.21</td>
<td>0.69</td>
<td>0.68</td>
<td><strong>0.97</strong></td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>0.29</td>
<td>0.31</td>
<td>0.31</td>
<td>0.34</td>
<td>0.50</td>
<td><strong>0.56</strong></td>
</tr>
<tr>
<td></td>
<td>0.28</td>
<td>0.48</td>
<td>0.69</td>
<td>0.79</td>
<td>0.77</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>CopyMachine</td>
<td>0.06</td>
<td>0.80</td>
<td>0.68</td>
<td><strong>0.99</strong></td>
<td>0.76</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.57</td>
<td>0.68</td>
<td>0.87</td>
<td>0.75</td>
<td><strong>0.93</strong></td>
</tr>
<tr>
<td></td>
<td><strong>0.92</strong></td>
<td>0.78</td>
<td>0.87</td>
<td>0.86</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>BusStation</td>
<td>0.12</td>
<td>0.34</td>
<td>0.36</td>
<td><strong>0.78</strong></td>
<td>0.28</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.46</td>
<td>0.50</td>
<td>0.79</td>
<td>0.33</td>
<td><strong>0.80</strong></td>
</tr>
<tr>
<td></td>
<td><strong>0.99</strong></td>
<td>0.67</td>
<td>0.92</td>
<td>0.65</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>Park</td>
<td>0.27</td>
<td>0.37</td>
<td>0.45</td>
<td><strong>0.99</strong></td>
<td>0.37</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>0.47</td>
<td>0.60</td>
<td><strong>0.78</strong></td>
<td>0.48</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.46</td>
<td>0.48</td>
<td>0.57</td>
<td>0.66</td>
<td><strong>0.89</strong></td>
</tr>
<tr>
<td>Corridor</td>
<td>0.20</td>
<td>0.68</td>
<td>0.25</td>
<td>0.99</td>
<td>0.16</td>
<td><strong>0.87</strong></td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>0.52</td>
<td>0.30</td>
<td>0.70</td>
<td>0.24</td>
<td><strong>0.88</strong></td>
</tr>
<tr>
<td></td>
<td>0.61</td>
<td>0.55</td>
<td>0.67</td>
<td>0.65</td>
<td>0.70</td>
<td><strong>0.82</strong></td>
</tr>
<tr>
<td>Average</td>
<td>0.14</td>
<td>0.52</td>
<td>0.42</td>
<td><strong>0.77</strong></td>
<td>0.46</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.46</td>
<td>0.45</td>
<td>0.63</td>
<td>0.47</td>
<td><strong>0.74</strong></td>
</tr>
<tr>
<td>MAE</td>
<td>0.0368</td>
<td>0.0404</td>
<td>0.0349</td>
<td>0.0284</td>
<td>0.0259</td>
<td><strong>0.0069</strong></td>
</tr>
</tbody>
</table>

Fig. 5. Example results of our method and its variants on four video sequences from GTFD dataset.

Table 4
Component analysis of our method and its variants on the GTFD dataset.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>OURS</td>
<td><strong>0.65</strong></td>
<td>0.80</td>
<td><strong>0.69</strong></td>
</tr>
<tr>
<td>OURS-I</td>
<td>0.58</td>
<td>0.81</td>
<td>0.64</td>
</tr>
<tr>
<td>OURS-II</td>
<td>0.62</td>
<td>0.78</td>
<td>0.66</td>
</tr>
<tr>
<td>OURS-III</td>
<td>0.60</td>
<td><strong>0.82</strong></td>
<td>0.66</td>
</tr>
</tbody>
</table>

By multi-scale integration, our model yields to better global and local consistency in both spatial and appearance aspects.

4.5. Computational complexity

Our method are implemented on the mixing platform of C++ and MATLAB without any code optimization on a desktop with an
Intel i7 3.4 GHz CPU and 32 GB RAM. Runtime together with the F-measure of our method against other methods on GTFD dataset is presented in Table 5, where OURS-SS: our model excludes the time for superpixel segmentation, OURS-SS+S1: OURS-SS model solving B via conventional SOFT-INPUT [44] instead of our decomposition strategy. From which we can see: 1) Though GMM, VIBE and PCP run much faster than ours, these methods perform significantly worse. 2) Our method performs slightly slower than DECOLOR and COROLA. The main reason is the time for superpixel segmentation and more connections on the informative graphs. 3) The matrix decomposition strategy can speed up the optimization comparing to the conventional SOFT-INPUT [44] without losing accuracy. Our method achieves significantly superior accuracy. We believe that it can be accelerated by further code optimization.

4.6. Limitation

We also encounter unsatisfactory results and detection failure as shown in Fig. 6. (1) The unfaithful superpixel segmentation of the car in Fig. 6(b1) and the pedestrian in Fig. 6(d1) will introduce false contour of the moving object as shown in Fig. 6(b2) and (d2). (2) The tiny object (the bicycle man indicated in Fig. 6(a1)) moving in the dark environment fails to be detected as shown in Fig. 6(d1). The first limitation could be refined by more robust superpixel segmentation method. The second failure could be significantly alleviated via multi-modal resources like thermal images.

5. Conclusion

In this paper, we have proposed a novel method for moving object detection via multi-scale structured low-rank and sparse factorization. We first encourage the structure of the neighboring pixels with close appearance. Then we explore the spatial compactness structure among the pixels within the same superpixel. Finally we integrate the multi-scale coherence with different number of superpixels into a unified framework. We optimize the proposed model via an efficient alternating algorithm. Extensive experiments against state-of-the-arts on the public datasets suggest that, the proposed method can better preserve the boundary of the objects and is robust to the noise. In future work, we will focus on extending our model to online or streaming fashion for real-life applications.

Acknowledgment

This study was partially funded by the National Nature Science Foundation of China (61502006, 61702002, 61602001, 61671018), the Natural Science Foundation of Anhui Province (1508085QF127), and Co-Innovation Center for Information Supply & Assurance Technology, Anhui University.

References
