# Moving Object Detection via Robust Low-Rank and Sparse Separating with High-Order Structural Constraint*

Aihua Zheng
*Anhui University*
*Computer Science and Technology*
Hefei, Anhui Province, China
ahzheng214@ahu.edu.cn

Yumiao Zhao
*Anhui University*
*Computer Science and Technology*
Hefei, Anhui Province, China
ymiaozhao@foxmail.com

Chenglong Li
*Anhui University*
*Computer Science and Technology*
Hefei, Anhui Province, China
lichenglong@ahu.edu.cn

Jin Tang
*Anhui University*
*Computer Science and Technology*
Hefei, Anhui Province, China
tangjin@ahu.edu.cn

Bin Luo
*Anhui University*
*Computer Science and Technology*
Hefei, Anhui Province, China
luobin@ahu.edu.cn

***Abstract*—Low-rank representation has been successfully applied for moving object detection by assuming the background images are linearly correlated while the moving foreground are sparse. Further, extensive works propose to incorporate the spatial pairwise smoothness of pixels to improve the robustness. In this paper, we investigate the long-range spatiotemporal relationships among pixels, and propose a novel approach to pursue the high-order consistency for moving object detection in the low-rank and sparse separation framework. In particular, we integrate the sparse unary penalty, the spatial pairwise smoothness, and the supervoxel-based high-order consistency into a unified structural constraints on the foreground. Moreover, we propose a single optimization algorithm to learn the background model and the foreground mask at a same time. Extensive experiments on the benchmark datasets GTFD and CDnet suggest that our approach achieves superior performance over several state-of-the-art algorithms.**

***Index Terms*—Moving object detection; Supervoxel; Low-Rank and Sparse Separating; High-Order Structural Constraint**

## I. Introduction

Moving object detection, aiming to locate and segment the moving objects in the video, plays a crucial role in computer vision and pattern recognition, such as target recognition [20], tracking [6], behavior analysis [4]. There are extensive studies on moving object detection over the past decades.

Recently, the low-rank and sparse separation methods have drawn much attention. The basic idea is to recover the low-rank background and the sparse outliers as foreground objects. The pioneer works include Robust Principal Component Analysis (RPCA) [2], [26] and its variants [7], [8], [15]. In order to enforce the spatial structure of the foreground, DECOLOR [29] and later on methods [3], [10], [23], [28] introduce a contiguous prior of the neighborhood pixels. Furthermore, besides the pixel-level spatial smoothness, BS-SMOD [13] presents a online matrix decomposition via max-norm constraints on each superpixel segment. SLMS [14] constructs the spatial and temporal graphs on the dense optical flow based motion-compensation. However, it removes the frames with less motion information, which may result in missing detection of sudden pause of the moving objects. Importantly, the above methods only construct the structural constraints of foreground objects on the spatial pairwise smoothness of pixels, while ignoring the long-range spatiotemporal relationships among pixels, which are usually important to the robustness of moving object detection.

In this paper, we investigate the long-range spatiotemporal relationships among pixels, and propose a novel approach of moving object detection to pursue high-order structural consistency in the low-rank and sparse separation framework. Given the accumulated sequential frames from the input video, we first form a data matrix by stacking each frame as a vector, and decompose it into the low-rank and sparse components, corresponding to the background and foreground, respectively. Second, we model the structural constraints of foreground objects by a Markov Random Filed (MRF) [17], which consists of three potential terms: i) the sparse unary term, ii) the spatial pairwise smoothness term, and iii) the high-order consistency term. The first two terms have been extensively employed in recent works [10], [23], [28], [29]. However, the long-range consistency of pixels has not been well investigated in background modeling and foreground detection, but plays a critical role in the robustness for detecting moving objects, as demonstrated in our experiments. To this end, we employ the robust $P^n$ [16] to model the supervoxel-based high-order consistency, and integrate it into MRF to model the structural

This work was partially supported by the National Natural Science Foundation of China (61502006, 61702002 and 61671018) and the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2017A017).

978-1-5386-5321-0/18/$31.00 ©2018 IEEE

Corresponding author: Chenglong Li

constraints of foreground objects. Finally, we design a single unified optimization algorithm to learn the background model and foreground mask simultaneously by iteratively employing the SOFT-IMPUTE algorithm [21] and the graph cut algorithm [22].

## II. Related Work

Moving object detection via low-rank and sparse separation boasts an extensive literature. The most representative problem formulation is the Robust Principal Component Analysis (RPCA) [26] which decomposes a given matrix/frames into a low-rank background matrix and sparse foreground matrix. [2] proposed Principal Component Pursuit (PCP) to recover the low-rank model form unknown corruption patterns. [30] extended PCP as Stable Principle Component Pursuit (SPCP), to handle sparse gross errors and small entrywise noises. DECOLOR [29] introduced the contiguous prior on foreground mask to preserve the spatial structure of the foreground. CLASS [28] proposed a collaborative framework to leverage the various size of the moving objects via introducing the global appearance consistency. COROLA [23] presented an online sequential framework via solving sequential low-rank approximation and contiguous outlier representation problem.

In order to preserve the spatial compactness, BSSMOD [13] presented an online matrix decomposition using max-norm constraint on each superpixel segment. TVRPCA [3], [10] introduced the 3-D total variance along the temporal axis to separate the moving foreground and the even sparser dynamic background. TLSFSD [11] further designed saliently fused-sparse regularizer to the tensor total variation. SLMS [14] proposed to construct the spatial and temporal graphs based on the motion-compensated binary mask generated by the dense optical flow prior.

However, these methods only consider the structural constraints of foreground objects on the spatial pairwise smoothness of pixels, while ignoring the long-range spatiotemporal relationships among pixels, which are usually important to the robustness of moving object detection. In this paper, we investigate the long-range spatiotemporal relationships among pixels and propose a novel approach to pursue the high-order consistency among the supervoxels.

## III. The Proposed Approach

The key of our method is to introduce a supervoxel-based high-order consistency into the low-rank and sparse separation framework to consider long-range spatio-temporal relationships among pixels. Given a video sequence $\mathbf{D} = \left[\mathbf{I_1}, \cdots, \mathbf{I_n}\right] \in \mathbb{R}^{m \times n}$ consists of $n$ frames with $m$ pixels per frame, we employ the video segmentation method [5], [9] to generate the supervoxel prior at the first place. We shall elaborate the proposed approach followed by the optimization in the following part of this section.

### *A. Problem Formulation*

Given the video sequence $\mathbf{D}$, our main task is to estimate the background $\mathbf{B} \in \mathbb{R}^{m \times n}$ and the binary foreground support $\mathbf{S} \in \{0,1\}^{m \times n}$ with:

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

We assume that the background frames are linearly correlated while the foregrounds are sparse [2], [26]. Furthermore, for the background region where $S_{ij} = 0$, we assume that $D_{ij} = B_{ij} + \epsilon_{ij}$, where $\epsilon_{ij}$ denotes i.i.d. Gaussian noise. Based on the above assumptions, the energy function can be written as:

$$\begin{aligned} &\min_{B_{ij}, S_{ij} \in \{0,1\}} \beta \parallel vec(\mathbf{S}) \parallel_0, \\ &s.t.\ \mathbf{S}_{\perp} \circ \mathbf{D} = \mathbf{S}_{\perp} \circ (\mathbf{B} + \epsilon),\ rank(\mathbf{B}) \leq r. \end{aligned} \tag{2}$$

where $\beta$ is the penalized factor, and $||\mathbf{X}||_0$ indicates the $l_0$ norm of a vector. $vec(\mathbf{S})$ is a vectorized operator on matrix $\mathbf{S}$. The operator "$\circ$" denotes element-wise multiplication of two matrices, $\mathbf{S}_{\perp}$ denotes the region of $S_{ij} = 0$, and $r$ is a constant that suppresses the complexity of the background model.

In order to enforce the spatial smoothness structure of the foreground, a common strategy is to penalize the neighboring pixels with diverse labels [23], [29]. Therefore the additional pairwise potential is defined as:

$$\|\mathbf{A}_1 vec(\mathbf{S})\|_1 = \sum_{(ij,kl) \in \mathcal{E}_1} |S_{ij} - S_{kl}|. \tag{3}$$

here, $\mathbf{A}_1$ is the node-edge incidence matrix denoting the connecting relationship among pixels. $\mathcal{E}_1$ is the edge set of the connected pixel pairs.

However, despite of the spatial smoothness structure, the foreground structure is also strongly related to the long-range spatiotemporal relationships among pixels. Therefore, we propose supervoxel-based high-order consistency into a unified structural constraints on the foreground.

**Supervoxel-Based High-Order Consistency:** We observe that, the structures of the moving objects are generally consistent along the temporal shifting. Similar to superpixels on 2-dimensional spatial images, the supervoxel refers to the high-order/3-dimensional voxels with both long-range spatially and temporally neighboring pixels of similar appearances. To consider the long-range spatiotemporal relationship among pixels, we enforce the pixels from the same supervoxel prior generated via [12], [27] to have the same pattern which tends to capture the global appearance compactness and temporal consistency of the foreground strucutre. Specifically, we first construct the graph between the pairs of pixels in the same superpixel projected from the supervoxel prior and define the smoothness as:

$$\|\mathbf{A}_2 vec(\mathbf{S})\|_1 = \sum_{(ij,mn) \in \mathcal{E}_2} |S_{ij} - S_{mn}|. \tag{4}$$

analogously, $\mathbf{A_2}$ denotes the connecting relationship among pixels. $\mathcal{E}_2$ is the edge set of the pixel pairs within the same superpixel projected by the supervoxel prior. Second, in order to consider the relationship in the long-range temporal domain, we further introduce a high-order consistency into the detection model. As illustrated in Fig. 1, we enforce the pixels inside the same supervoxel to possess the same pattern. Inspired by the robust $P^n$ [12], [16], we define the high-order potential among the supervoxel as:

$$\Phi\left(\mathbf{S}_{\mathcal{V}\in\mathcal{C}}\right)=\begin{cases} N\left(\mathbf{S}_{\mathcal{V}}\right)\frac{1}{Q}\tau_{max}\left(\mathcal{V}\right) & if\ N\left(\mathbf{S}_{\mathcal{V}}\right)\leqslant Q, \\ \tau_{max}\left(\mathcal{V}\right) & otherwise. \end{cases} \tag{5}$$

where $\mathcal{V}$ is one of the supervoxel clique in the supervoxel set $\mathcal{C}$, $|\mathbf{S}_{\mathcal{V}}|$ denotes the number of nodes/pixels in supervoxel clique $\mathcal{V}$, $N\left(\mathbf{S}_{\mathcal{V}}\right)=\min\left(|\mathbf{S}_{\mathcal{V}}=1|,|\mathbf{S}_{\mathcal{V}}=0|\right)$ denotes the number of nodes with nondominant label in supervoxel $\mathcal{V}$, and $Q$ is the truncation parameter controlling the rigidity within the supervoxels. $\tau_{max}=|\mathbf{S}_{\mathcal{V}}|\,exp\left(-\sigma_{\mathcal{V}}\right)$, $\sigma_{\mathcal{V}}$ is the total RGB variance in supervoxel clique $\mathcal{V}$.

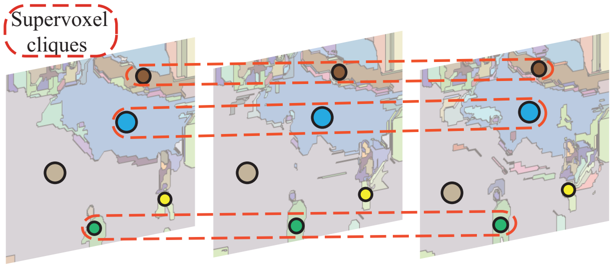

Fig. 1. Supervoxel clique in videos. The colorful circles indicate the superpixels on each frame which projected from the supervoxel prior while the circles with the same color are derived from the same supervoxel.

Based on above discussion, we can summarize the energy function as:

$$\begin{aligned}\min_{B_{ij},S_{ij}\in\{0,1\}} & \beta\parallel vec(\mathbf{S})\parallel_0+\gamma||\mathbf{A_1}\ vec(\mathbf{S})||_1 \\ + & \ \eta||\mathbf{A_2}\ vec(\mathbf{S})||_1+\lambda\Phi\left(\mathbf{S}_{\mathcal{V}}\right), \\ & s.t.\ \mathbf{S}_\perp\circ\mathbf{D}=\mathbf{S}_\perp\circ(\mathbf{B}+\epsilon),\ \ rank(\mathbf{B})\leq r.\end{aligned} \tag{6}$$

where $\gamma$, $\eta$ and $\lambda$ are the balance parameters to leverage the smoothness between the pixel pairs in the adjacent neighborhood, superpixels and the high-order supervoxels.

### B. Model Optimization

To make Eq. (6) tractable, the common strategy is to relax the rank constraint by nuclear norm. Therefore, the formulation can be rewritten as:

$$\begin{aligned}&\min_{B_{ij},S_{ij}\in\{0,1\}}\frac{1}{2}\left\|\mathbf{S}_\perp\circ(\mathbf{D}-\mathbf{B})\right\|_F^2+\alpha\left\|\mathbf{B}\right\|_*+\beta\parallel vec(\mathbf{S})\parallel_0 \\ &+\gamma\left\|\mathbf{A}_1\ vec\left(\mathbf{S}\right)\right\|_1+\eta\left\|\mathbf{A}_2\ vec\left(\mathbf{S}\right)\right\|_1+\lambda\Phi\left(\mathbf{S}_{\mathcal{V}}\right).\end{aligned} \tag{7}$$

where $\alpha$ is the balance parameter to control the complexity of the background. $\left\|\mathbf{X}\right\|_*$ and $\left\|\mathbf{X}\right\|_F$ indicate the nuclear norm and the Frobenius norm of a matrix respectively.

The objective function Eq. (7) is non-convex and not trivial to be solved due to both continuous and discrete variables. Therefore, we design a two step alternating algorithm by separating the energy minimization over $\mathbf{B}$ and $\mathbf{S}$.

**Solving-B:** In order to estimate $\mathbf{B}$ with the currently given $\hat{\mathbf{S}}$, Eq. (7) turns to be the minimization problem:

$$\min_{\mathbf{B}}\frac{1}{2}\left\|\mathbf{S}_\perp\circ(\mathbf{D}-\mathbf{B})\right\|_F^2+\alpha\left\|\mathbf{B}\right\|_*. \tag{8}$$

Eq. (8) can be solved by SOFT-IMPUTE [21] algorithm with the updating prorogation:

$$\hat{\mathbf{B}}\ \leftarrow\Theta_\alpha\left(\hat{\mathbf{S}}_\perp\circ\mathbf{D}+\hat{\mathbf{S}}_\perp\circ\mathbf{B}\right). \tag{9}$$

where $\Theta_\alpha(Z)=\mathbf{U}\Sigma_\alpha\mathbf{V}^T$ means the singular value thresholding, $\Sigma_\alpha=diag[(d_1-\alpha)_+,\cdots,(d_k-\alpha)_+]$, $\alpha\Sigma_\alpha\mathbf{V}^T$ is the SVD of $\mathbf{Z}$, $\Sigma=diag[d_1,\cdots,d_k]$ and $t_+=max(t,0)$.

**Solving-S:** Given the estimated low-rank background $\hat{\mathbf{B}}$, Eq. (7) can be rewritten as:

$$\begin{aligned}&\sum_{ij}\left(\beta-\frac{1}{2}\left(D_{ij}-\hat{B}_{ij}\right)^2\right)S_{ij}+\gamma\left\|\mathbf{A}_1\ vec\left(\mathbf{S}\right)\right\|_1+\lambda\Phi\left(\mathbf{S}_{\mathcal{V}}\right) \\ &+\eta\left\|\mathbf{A}_2\ vec\left(\mathbf{S}\right)\right\|_1+\frac{1}{2}\sum_{ij}\left(D_{ij}-B_{ij}\right)^2.\end{aligned} \tag{10}$$

where $(D_{ij}-B_{ij})^2$ and $S_{ij}$ are constants with fixed $\hat{\mathbf{B}}$. Above energy function is a standard first-order MRFs [17] with unary term, pairwise term and high-order term, which can be optimized through the graph cut [22] algorithm.

## IV. Experiment

We evaluate our method on the benchmark datasets GTFD [18] and CDnet14 [25] comparing with six state-of-the-art moving object detection algorithms including PCP [2], VIBE [1], GMM [24], TTD [19], DECOLOR [29] and COROLA [23]. We choose the default parameters of these methods.

### A. Datasets

GTFD [18] consists of 25 video sequences which contains various challenges such as intermittent motion, low illumination, bad weather, intense shadow, dynamic scene and background clutter. It consists of both visible video and infrared video for each scene. We only evaluate the visible videos in our experiments.

CDnet14 [25] is a large scale dataset which includes 11 different categories and contains 55 video sequences. To better present our algorithm have good spatial smoothness and robustness, we evaluate our method on 10 videos from 5 challenging categories including *DynamicBackground* (Boats, Fountain02), *IntermittentObjectMotion* (WinterDriveway, StreetLight), *PTZ* (TwoPositionPTZCam (PTZCam)), ZoomInZoomOut (ZoomInOut)), *Shadow* (Cubicle, CopyMachine) and *Thermal* (Corridor, Park).

### B. Evaluation Settings

**Parameters.** There are six parameters in our method, we adjust one parameter while fixing other parameters and then obtain better performance for our approach. $\alpha$ is first roughly

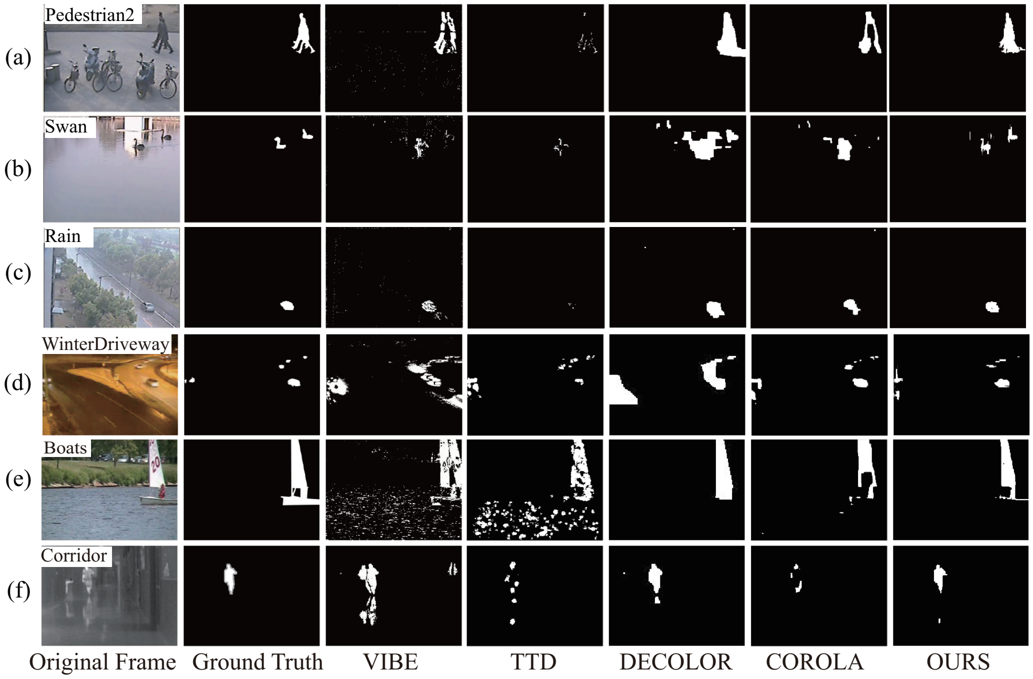

Fig. 2. Comparison results of a certain frame from six example sequences from GTFD dataset ((a) - (c)) and CDnet14 dataset ((d) - (f)). The text rectangles on the left top of the original frames indicate the name of the video sequences.

estimated as the rank of the background model and further adjusted by SOFT-IMPUTE [21] algorithm until $rank\left(\hat{\mathbf{B}}\right) > r$. $\beta = 2.5\sigma^2$ where $\sigma$ is estimated online by the mean variance of $\left\{D_{ij} - \hat{B}_{ij}\right\}$. Note that the pixel-level smoothness will result in high computational burden. Therefore, we set $\eta$ to be a large value to severely penalize the spatial inconsistency, which can speed up 3 times while increasing about $2\%$ in F-measure. The final parameter settings are $(\alpha, \beta, \gamma, \lambda, Q) = \left(0.707, 2.5\sigma^2, 2\beta, 2, 0.01\left|\mathbf{S}_{\mathcal{V}}\right|\right)$. Furthermore, we construct the spatiotemporal consistency in every 10 frames in our algorithm.

**Evaluation Metrics.** For the quantitative evaluation, we employ the Precision $(P)$, Recall $(R)$ and F-measure $(F)$ as evaluation metrics.

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F = \frac{2PR}{P + R}. \tag{11}$$

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

## C. Qualitative Results

We first demonstrate some qualitative comparison results on GTFD [18] and CDnet14 [25] in Fig. 2. From which we can see that, VIBE works on the original pixel space therefore they are quite sensitive to the noise and introduce ”ghost”. TTD is robust to the noise but fails to detect the small foreground in complex background. By introducing spatial relationship, DECOLOR and COROLA obtain more coherent foreground masks than the ones with single sparse constraint such as TTD. However, DECOLOR fails to sketch the contours of the objects, since it only consider the spatial smoothness among the neighboring pixels. COROLA, as the state-of-the-art online detection model, tends to produce the cavities on the objects due to the lack of long-range sequential information. After enforcing the long-range spatiotemporal relationship among pixels, our method can better preserve the contours of the moving objects and achieve robust detection results under the various challenging scenarios.

## D. Quantitative Results

Table I reports the average precision, recall and F-measure on the 25 videos from GTFD dataset [18] while Table II details these quantitative results of each testing video from CDnet14 [25]. The dashes (-) in the Table II denote that the corresponding methods failed to detect objects in this video. From Table I and Table II, it is clear to see that: 1) Our method achieves superior result than the state-of-the-arts in precision in most of the cases since we integrate the sparse unary potential, the spatial pairwise potential and the high-order potential to punish the foregrounds. 2) As for recall, it works worse than DECOLOR since DECOLOR tends to generate coarse contours which always leads to higher recalls. 3) The more comprehensive measurement between precision and recall, F-measure demonstrates the best trade-off performance of our method with $5\%$ and $4\%$ higher than the second best method on GTFD and CDnet respectively.

## E. Component Analysis

We evaluate the components of the spatial smoothness within the superpixel ($S_{superpixel}$) which is projected from the supervoxel prior and the high-order consistency encoded

TABLE II

COMPARISON OF PRECISION (P), RECALL (R), AND F-MEASURE (F) SCORES ON TEN TESTING VIDEOS FROM CDNET14 DATASET.

| Methods | | PCP | ViBe | GMM | TTD | DECOLOR | COROLA | OURS |
|---|---|---|---|---|---|---|---|---|
| Boats | P | 0.17 | 0.25 | 0.47 | 0.40 | 0.85 | 0.76 | **0.97** |
| | R | 0.47 | 0.39 | 0.20 | 0.36 | **0.74** | 0.64 | 0.67 |
| | F | 0.25 | 0.31 | 0.27 | 0.34 | **0.79** | 0.70 | **0.79** |
| Fountain02 | P | 0.02 | 0.16 | 0.74 | 0.44 | 0.66 | **0.87** | 0.81 |
| | R | 0.27 | 0.40 | 0.55 | 0.15 | **0.90** | 0.62 | 0.67 |
| | F | 0.04 | 0.23 | 0.63 | 0.22 | **0.76** | 0.71 | 0.74 |
| WinterStreet | P | 0.10 | 0.12 | 0.26 | 0.37 | 0.22 | 0.41 | **0.50** |
| | R | 0.41 | 0.44 | 0.62 | 0.45 | **0.96** | 0.74 | 0.62 |
| | F | 0.15 | 0.19 | 0.34 | 0.39 | 0.35 | 0.51 | **0.53** |
| StreetLight | P | 0.01 | - | - | 0.27 | 0.41 | 0.39 | **0.59** |
| | R | 0.52 | - | - | 0.44 | **0.91** | 0.56 | 0.58 |
| | F | 0.01 | - | - | 0.32 | 0.56 | 0.46 | **0.58** |
| TwoPositionPTZCam | P | 0.15 | 0.10 | 0.58 | **0.62** | 0.47 | 0.57 | 0.57 |
| | R | 0.31 | 0.30 | 0.47 | 0.53 | **0.98** | 0.86 | 0.78 |
| | F | 0.16 | 0.13 | 0.48 | 0.50 | 0.61 | **0.65** | 0.62 |
| ZoomInZoomOut | P | 0.01 | 0.02 | 0.11 | 0.05 | 0.17 | 0.04 | **0.20** |
| | R | 0.38 | **0.58** | 0.38 | 0.43 | 0.31 | 0.81 | 0.39 |
| | F | 0.03 | 0.04 | 0.15 | 0.08 | 0.16 | 0.07 | **0.24** |
| Cubicle | P | 0.10 | 0.43 | 0.88 | 0.90 | **0.92** | 0.78 | 0.86 |
| | R | 0.33 | 0.55 | 0.24 | 0.48 | 0.58 | 0.52 | **0.69** |
| | F | 0.15 | 0.48 | 0.36 | 0.62 | 0.66 | 0.58 | **0.75** |
| CopyMachine | P | 0.21 | 0.43 | 0.79 | **0.86** | 0.80 | 0.74 | **0.86** |
| | R | 0.14 | 0.32 | 0.61 | 0.85 | **0.98** | 0.70 | 0.92 |
| | F | 0.17 | 0.37 | 0.69 | 0.83 | 0.88 | 0.69 | **0.89** |
| Corridor | P | 0.04 | 0.41 | 0.52 | 0.54 | 0.59 | 0.77 | **0.94** |
| | R | 0.53 | 0.66 | 0.18 | 0.25 | **0.96** | 0.28 | 0.62 |
| | F | 0.07 | 0.48 | 0.23 | 0.32 | 0.71 | 0.40 | **0.73** |
| Park | P | 0.15 | 0.52 | **0.95** | 0.83 | 0.68 | 0.87 | 0.87 |
| | R | 0.50 | 0.31 | 0.37 | 0.47 | **0.98** | 0.54 | 0.69 |
| | F | 0.23 | 0.38 | 0.53 | 0.58 | **0.80** | 0.65 | 0.78 |
| Average | P | 0.09 | 0.24 | 0.53 | 0.53 | 0.58 | 0.62 | **0.72** |
| | R | 0.39 | 0.28 | 0.36 | 0.44 | **0.83** | 0.63 | 0.67 |
| | F | 0.12 | 0.27 | 0.36 | 0.42 | 0.63 | 0.55 | **0.67** |

TABLE I

AVERAGE PRECISION (P), RECALL (R) AND F-MEASURE (F) OF OUR METHOD AGAINST THE STATE-OF-THE-ART ALGORITHMS ON GTFD DATASET.

| **Algorithm** | $P$ | $R$ | $F$ |
|---|---|---|---|
| PCP | 0.28 | 0.18 | 0.21 |
| ViBe | 0.41 | 0.49 | 0.41 |
| GMM | 0.48 | 0.65 | 0.52 |
| TTD | 0.59 | 0.29 | 0.32 |
| DECOLOR | 0.54 | **0.83** | 0.58 |
| COROLA | 0.59 | 0.67 | 0.56 |
| OURS | **0.64** | 0.70 | **0.63** |

in the supervoxel clique ($H_{supervoxel}$) on GTFD dataset in this section. We report the results in Table III where OURS-I denotes our model without $H_{supervoxel}$ and $S_{superpixel}$ by setting $\eta = \lambda = 0$ and OURS-II indicates our model with only $H_{supervoxel}$ by setting $\eta = 0$. From which we can see, both high-order consistency and the spatial smoothness play important roles for moving object detection. Noted that the higher recall in OURS-I results from the coarse contours of the detected foregrounds. After introducing supervoxel-based high-order consistency (comparing OURS-II to OURS-I), the average F-measure value increases 3%. Furthermore, after introducing spatial smoothness within the superpixel (comparing OURS to OURS-II), the average F-measure value improves 2%.

TABLE III

COMPONENT ANALYSIS OF THE HIGH-ORDER CONSISTENCY AND THE SPATIAL SMOOTHNESS WITHIN THE SUPERPIXEL.

| **Algorithm** | $P$ | $R$ | $F$ |
|---|---|---|---|
| OURS-I(without $H_{supervoxel}$ or $S_{superpixel}$) | 0.54 | **0.83** | 0.58 |
| OURS-II (with $H_{supervoxel}$ ) | 0.61 | 0.69 | 0.61 |
| OURS (with $H_{supervoxel}$ and $S_{superpixel}$) | **0.64** | 0.70 | **0.63** |

### *F. Computational Complexity*

Our algorithm is implemented on the mixed platform of MATLAB and C++ for the background and foreground decomposition via on the Linux system for the supervoxel segmentation. All experiments are carried out on a desktop with an Intel i7 3.4GHz CPU and 16GB RAM. The total computation cost of our algorithm consists of the cost of supervoxel segmentation, background updating cost via SOFT-IMPUTE and the foreground updating cost via graph cut. Table IV reports the the our computational cost comparing with the state-of-the-arts on GTFD dataset with resolution of $320 \times 240$. Our method works slightly slower than DECOLOR and COROLA, but it can generate more robust foregrounds. Though PCP, GMM and VIBE work much faster than ours, they perform greatly worse. Therefore, our method keeps a

TABLE IV
COMPUTATIONAL COMPLEXITY COMPARISON OF OUR METHOD AGAINST THE STATE-OF-THE-ARTS (IN FPS).

| | PCP | VIBE | GMM | TTD | DECOLOR | COROLA | OURS |
|---|---|---|---|---|---|---|---|
| Code Type | Matlab | C++ | C++ | Matlab | Matlab & C++ | Matlab & C++ | Matlab & C++ |
| FPS | 15.50 | 166 | 76.78 | 0.11 | 0.90 | 3.89 | 0.65 |

good balance between the efficiency and accuracy. We believe that we can achieve better efficiency by code optimization.

## V. CONCLUSION

This paper have proposed a moving object detection method to pursue high-order structural consistency in the low-rank and sparse separation framework. By introducing the high-order potential over the supervoxel clique together with the spatial smoothness within the superpixels, our method can capture fine appearance and perform robust against the challenging scenarios. In the future work, we will focus on the video segmentation method which can provide more accurate prior for the detection model.

## REFERENCES

[1] O. Barnich and M. Van Droogenbroeck, “Vibe: A universal background subtraction algorithm for video sequences,” *IEEE Transactions on Image processing*, pp. 1709–1724, 2011.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM*, pp. 11:1–11:37, 2011.

[3] X. Cao, L. Yang, and X. Guo, “Total variation regularized rpca for irregularly moving object detection under dynamic background,” *IEEE Transactions on Cybernetics*, pp. 1014–1027, 2016.

[4] S.-H. Cho and H.-B. Kang, “Abnormal behavior detection using hybrid agents in crowded scenes,” *Pattern Recognition Letters*, pp. 64 – 70, 2014.

[5] J. J. Corso and C. Xu, “Evaluation of super-voxel methods for early video processing,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1202–1209, 2012.

[6] I. Elafi, M. Jedra, and N. Zahid, “Unsupervised detection and tracking of moving objects for video surveillance applications,” *Pattern Recognition Letters*, pp. 70–77, 2016.

[7] J. Feng, H. Xu, and S. Yan, “Online robust pca via stochastic optimization,” *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 404–412, 2013.

[8] J. Goes, T. Zhang, R. Arora, and G. Lerman, “Robust Stochastic Principal Component Analysis,” *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 266–274, 2014.

[9] M. Grundmann, V. Kwatra, M. Han, and I. Essa, “Efficient hierarchical graph-based video segmentation,” *Georgia Institute of Technology*, pp. 2141–2148, 2010.

[10] X. Guo, X. Wang, L. Yang, X. Cao, and Y. Ma, “Robust foreground detection using smoothness and arbitrariness constraints,” *European Conference on Computer Vision*, pp. 535–550, 2014.

[11] W. Hu, Y. Yang, W. Zhang, and Y. Xie, “Moving object detection using tensor-based low-rank and saliently fused-sparse decomposition,” *IEEE Transactions on Image Processing*, pp. 724–737, 2017.

[12] S. D. Jain and K. Grauman, “Supervoxel-consistent foreground propagation in video,” *Springer International Publishing*, pp. 656–671, 2014.

[13] S. Javed, S. Ho Oh, A. Sobral, T. Bouwmans, and S. Ki Jung, “Background subtraction via superpixel-based online matrix decomposition with structured foreground constraints,” *International Conference on Computer Vision Workshop*, pp. 90–98, 2015.

[14] S. Javed, A. Mahmood, T. Bouwmans, and S. K. Jung, “Spatiotemporal low-rank modeling for complex scene background initialization,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[15] S. Javed, S. H. Oh, J. Heo, and S. K. Jung, “Robust background subtraction via online robust pca using image decomposition,” *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, pp. 105–110, 2014.

[16] P. Kohli, L. Ladicky, and P. H. S. Torr, “Robust higher order potentials for enforcing label consistency,” *International Journal of Computer Vision*, pp. 302–324, 2009.

[17] Li and Z. Stan, *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., 2009.

[18] C. Li, X. Wang, L. Zhang, and J. Tang, “Weld: Weighted low-rank decomposition for robust grayscale-thermal foreground detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 725–738, 2017.

[19] O. Oreifej, X. Li, and M. Shah, “Simultaneous video stabilization and moving object detection in turbulence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 450–62, 2013.

[20] W. Phillips Iii, M. Shah, and N. da Vitoria Lobo, “Flame recognition in video,” *Pattern recognition letters*, pp. 319–327, 2002.

[21] R. T. Rahul Mazumder, Trevor Hastie, “Spectral regularization algorithms for learning large incomplete matrices,” *Journal of Machine Learning Research*, pp. 2287–2322, 2010.

[22] C. Rother, V. Kolmogorov, and A. Blake, “”grabcut”: interactive foreground extraction using iterated graph cuts,” *ACM SIGGRAPH*, pp. 309–314, 2004.

[23] M. Shakeri and H. Zhang, “Corola: A sequential solution to moving object detection using low-rank approximation,” *Computer Vision and Image Understanding*, pp. 27–39, 2015.

[24] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

[25] Y. Wang, P. M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, “Cdnet 2014: An expanded change detection benchmark dataset,” *Computer Vision and Pattern Recognition Workshops*, pp. 393–400, 2014.

[26] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, “Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization,” *Advances in neural information processing systems*, pp. 2080–2088, 2009.

[27] C. Xu, C. Xiong, and J. J. Corso, “Streaming hierarchical video segmentation,” *European Conference on Computer Vision*, pp. 626–639, 2012.

[28] A. Zheng, M. Xu, B. Luo, Z. Zhou, and C. Li, “Class: Collaborative low-rank and sparse separation for moving object detection,” *Cognitive Computation*, pp. 180–193, 2017.

[29] X. Zhou, C. Yang, and W. Yu, “Moving object detection by detecting contiguous outliers in the low-rank representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 597–610, 2013.

[30] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, “Stable principal component pursuit,” *IEEE International Symposium on Information Theory*, pp. 1518–1522, 2010.