A Tensor Foreground-Background Separation Algorithm Based on Dynamic Dictionary Update and Active Contour Detection

SHENGLI PEI¹, LINHAO LI², LIJUAN YE³, AND YONGFENG DONG²

¹School of Physics and Electronics Information Engineering, Qinghai Nationalities University, Xining 810007, China
²Hebei Provincial Key Laboratory of Big Data Computing, School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
³School of Mathematics and Statistics, Qinghai Nationalities University, Xining 810007, China

Corresponding author: Linhao Li (lilinhao@hebut.edu.cn)

This work was supported in part by the Applied Basic Research Program of Qinghai under Grant 2019-ZJ-7017, in part by the National Natural Science Foundation of China under Grant 61902106, and in part by the High-Level Talent (Ph.D.) Program of Qinghai University for Nationalities under Grant 2020XJG13.

ABSTRACT Foreground-background separation of surveillance video, that models static background and extracts moving foreground simultaneously, attracts increasing attentions in building a smart city. Conventional techniques towards this always consider the background as primary target and tend to adopt low-rank constraint as its estimator, which provides finite (equal to the value of rank) alternatives when constructing the background. However, in practical missions, although general sketch of background is stable, some details change constantly. Aimed at this, we propose to represent the general background by a linear combination of some atoms and record the detailed background by spatiotemporal clustered patches. Then, the moving foreground is considered as a mixture of active contours and continuous contents. Eventually, joint optimization is conducted under a unified framework, i.e., alternating direction multipliers method (ADMM), and produces our tensor model for hierarchical background and hierarchical foreground separation (THHS). The employed tensor space, which agrees with the instinct structure of video data, benefits all the spatiotemporal designs in both background modular and foreground part. Experimental results show that THHS is more adaptive to the dynamic background and produces more accurate foreground when compared against current state-of-the-art techniques.

INDEX TERMS Dictionary learning, foreground-background separation, active contour model, quantization, hierarchical modeling.

I. INTRODUCTION

Background modeling and foreground extraction is a fundamental task in the area of artificial intelligence. It is not only the basic problem of advanced artificial intelligence tasks like smart city [1] and semantic understanding [2], but also widely used in various practical tasks such as object tracking [3] and behavior recognition [4]. A sound illustration about the potential applications is given in [5], where current solutions and serious challenges are summarized alongside. The task is actually established on the notable differences in data distribution of foreground and background from the same video, namely, the background content is relatively stable, while the foreground objects usually take up a small space and keep moving in the entire video.

At the foundation of the task is the technique of measuring the differences between foreground and background. A straightforward way is accomplished by employing extensive kinds of statistical models, i.e., Mixture of Gaussian (MOG) [6], clustering codebook [7], Support Vector Machine (SVM) [8], etc. To produce more effective inputs for these models, various pixel-wise [9], [10] and region-wise [11]–[13] features of data distribution are extracted from raw video data. Besides, some network based attempts are proposed to extract deep features and are constructed by considering the produced results of some state-of-the-art algorithms as groundtruth [14], [15]. In addition, by considering the provided groundtruth in some datasets (e.g. CDnet [16]), researchers have employed
FIGURE 1. Framework of the proposed THHS. 1. Low-frequency background is extracted by the proposed sparse dictionary learning way. 2. Group of frames (GOF) is composed by several residual frames that obtained by subtracting the low-frequency background from the input video frame. 3. Spatiotemporal quantization clustering model produces high-frequency background. 4. Background subtraction is conducted after entire background is computed by summing low-frequency background and high-frequency background. 5. A rough foreground contour is detected by the proposed level set segmentation. 6. The eventual detected foreground is produced after adding a continuity constraint. At last, above processes will iterate until convergence.

supervised approaches to construct deep learned features and thus produce more effective classifiers for background/foreground [17], [18].

Another way treats video data globally, i.e., extracting background by projecting the high-dimensional video data onto certain sparsity-base low-dimension space. These sparsity based background modeling ways have attracted more attentions in the past few years, as the sparsity constraint usually results in algorithms with stable performance and high generalization ability. Common constraints for background includes low-rank [19], rank-1 [20] and sparse coding [21]. Comparisons of the above two unsupervised ways can be queried in [22].

As a low-rank space is actually spanned by a few basis vectors, intensive selection and dynamic training of the basic vectors (or dictionary atoms) are considered able to produce more effective representation of the background [23]–[25]. However, attempts towards this have not yet produced satisfying background modeling performance. Because, the process of dictionary learning is too sensitive to noise, e.g., changing details of background and various video noises. These outliers prevent dictionary learning from converging to the optimum and thus influence the accuracy of the sparse represented background model.

In this paper, a dictionary learning based tensor hierarchical background and hierarchical foreground separation (THHS) model is proposed, where dynamic dictionary updating model for background and active contour based model for foreground are established. Better estimations of various interferences produces more robust training phase of dictionary learning. The framework of THHS are shown in Fig. 1. Firstly, motivated by the hierarchical background in [20], we regenerate the background as general background (low frequency BG in Fig. 1) and detail background (high frequency BG in Fig. 1). Dynamical learning of the basis of general BG is conducted and a spatiotemporal clustering is arranged for detail BG. Then, unlike most conventional foreground models that mainly emphasise the saliency and continuity of the content, we propose to estimate the characteristics of both content and contour simultaneously. Eventually, alternating optimization of background and foreground results in more accurate estimation of each video component. Besides, all above operations are conducted in tensor space as spatiotemporal distribution features of video data are essential in both our background model and foreground model. Contributions can be summarized as follows.

- As to produce more effective dictionary atoms of background, we establish specialized models for all the potential components of video, so that the data for training the dictionary of general background is clean.
- Detail background is estimated by a spatiotemporal clustering manner.
• A two-layer foreground model is proposed by considering both spatiotemporal continuous content and level set based contour simultaneously.
• A tensor based modeling framework enables us to explore high order relationships of video pixels.

The remainder of this paper is organized as follows. After reviewing basic works about background-foreground separation in Section 2, we show details about the proposed THHS in Section 3. Related optimizations of THHS is given in Section 4. We show the experimental results in Section 5 and then conclude after that.

II. RELATED WORK
To extract foreground from background, difference of data distributions is a most powerful separator. Most conventional estimations of the difference are conducted statistically or sparsity-based.

A. STATISTICAL MODELS
Statistical models explore the laws of pixel-wise/region-wise data distribution changing ways, which are then employed as the criterion of classifying the foreground data and background data. In 1999, pixel series from background are modelled independently by a Mixture of Gaussian (MOG) distribution [6]. The pixels do not follow the distributions are from the foreground accordingly. Starting from this framework, different hypopaper and information [10] are further considered and have produced improving performance. Then, similarly, other statistical measurements, e.g., Kernel Density Estimation (KDE) [26] and clustering codebook [7], are utilized to model the distribution of the pixel series from background. Besides, some advanced tools, e.g., Support Vector Machine (SVM) [8], Restricted Boltzmann Machine (RBM) [27] and Weightless Neural Networks (WNN) [28], are also employed to build up a convincing background distribution estimator.

Secondly, pixels can be treated regionally as they are always related to each other in most practical tasks. This kind of relationship is usually robust enough to the noises in the video. Thus, features, e.g., Local Binary Patterns (LBP) [29] and motion trajectories [12], are extracted and classified to indicate the labels of the corresponding pixels. Then, some nonparametric estimators, e.g., Visual Background Extractor (ViBe) [11] and Pixel-Based Adaptive Segmenter (PBAS) [9], are proposed to capture the characters of the background pixel patches automatically. In addition, some recent attempts propose to model the pixel relationships by regional receptive of convolutional networks. Hand-crafted groundtruth or pre-manufactured groundtruth are required to produce the effective encoder-decoder spatialtemporal network [15] and multiscale fully convolutional network [14].

B. SPARSITY-BASED MODELS
Sparsity-based models are established based on the idea of projecting high-dimensional data onto a specifical lower dimensional subspace. By keeping the eigenbackgrounds associated with the few largest eigenvalues, the first sparsity based model is proposed by Oliver et al. in 2000, which utilizes principal component analysis (PCA) as the background estimator. Then, in 2011, Hu et al. [38] conduct an in-depth analysis of the noise robustness of the PCA model, and then propose a principal component pursuit model (PCP) for the low-rank background. Furthermore, a local PCP considers the case that data distribute in multiple subspaces, in which total variation (TV) is employed to deal with complex and variable noise interferences in practical data when extracting the foreground [30]. Following this idea, a total variation regularized robust principal component analysis (TVRPCA) is built up in 2017 [31]; a tensor based total variation regularization is proposed in 2018 [32]; a 3-D total variation measure is proposed in 2019 [33]. Another common constraint for foreground is Markov Random Field (MRF) that is first combined with the low-rank background in 2013 [34]. In 2015, Liu et al. propose a new norm that derived from structured sparsity of foreground objects and establish a low rank and structured sparsity decomposition (LSD) algorithm. Structured sparsity is further discussed to form a spatiotemporal structured-sparse RPCA for complex scenes in 2019 [35]. In addition, a Gaussian scale mixture model is proposed for jointly estimating the variances of the foreground sparse coefficients and the unknown sparse coefficients in [36].

The low-rank based models provide a few options (equal to the value of rank) for the background decision. An improving way of this is to represent the background by the linear combination of these options, which are called dictionary atoms in this paper. In 2009, Huang et al. first propose to accomplish foreground-background separation by sparse representation, where the dictionary are obtained by preprocessing of some selected frames [23]. Then, the model is improved by a more precise foreground model and the dictionary atoms are just some randomly selected frames [24]. Learning of the dictionary is first discussed in 2011 by Zhao and the dictionary is updated in an iterative way [37]. Then, in 2016, a more effective foreground model is proposed for reducing the influence of noise on the dictionary learning processing, which results in more satisfying separation performance [25].

C. SPATIOTEMPORAL SOLUTIONS
In addition to the frameworks introduced in above sections, spatiotemporal characteristics of practical video data are frequently employed as additional constraints or principles when constructing effective foreground model and background model.

Formulation of low-rank constraint in the tensor space is discussed by amounts of researchers: Hu et al. arrange a high-order kernel function for the background in 2017 [38]; Liu et al. come up with a new way to extract the low-rank component in tensor data whose low-rank entries are from the diagonal elements of the core tensor in 2018 [39] and
Wang et al. find a computable strategy on calculating the rank of a given video tensor [40]; Sobral et al. propose a tensor subspace learning scheme to solve high order low-rank model [41] while Li et al. come up with an online solution for the low-rank tensor data [42]; Javed et al. use a stochastic optimization way to improve the efficiency of tensor decomposition; in 2019, both $l_1/2$ regularization and total variation regularization for tensor data are discussed in [43].

Then, structural features of pixel-wise or region-wise spatiotemporal relationships of adjacent pixels are generally considered robust enough to video noises: temporal information that contained in object motion are extracted to assist background subtraction in [44] and the framework is then improved in [45]; in [46], spatial descriptors produced by saliency detectors are integrated with shape constraint to constrain the moving objects; a spatiotemporal slice-based SVD is proposed by Kajo et al. to produce more effective tensor completion in 2018 [47]; in 2019, [20] proposes a pixel-wise short term temporal quantization way to extract detail background patches.

In addition, other structural information are also explored, e.g., manifold learning of the background linear or nonlinear space is conducted to conform the spatiotemporal similarity of the background [48]. Meanwhile, super-pixel is also considered an effective regional pixel similarity estimating way that is robust enough to outliers. A coarse-to-fine segmentation strategy is proposed in [49] to extract the low-rank component using matrix decomposition with max-norm of super-pixels. Then, in [50], a super-pixel based tree structure is arranged for the foreground objects detection. In 2020, super-pixel segmentation technology is allocated for the foreground objects while the background is modelled by a non-convex approximation [51].

Based on above works, a tensor dictionary learning framework and a high order quantization clustering way are adopted in this paper. To improve the temporal quantization in [20], a spatiotemporal manner is proposed. Then, to estimate the regional foreground pixel similarity, a level set based segmentation way is introduced. Eventually, these structural constraints result in effective foreground-background separation performance of THHS.

III. FORMULATION OF THHS

A. HIERARCHICAL BACKGROUND MODEL

For a given video $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_N] \in \mathbb{R}^{I \times J \times 3 \times N}$, background is denoted by $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_N]$ and foreground is $\mathbf{F} = [\mathbf{F}_1, \ldots, \mathbf{F}_N]$. The binary foreground mask is $\Omega = [\Omega_1, \ldots, \Omega_N] \in \mathbb{R}^{I \times J \times N}$ where $\Omega_{i,j,n} = 1$ if pixel $(i, j, n)$ is from foreground and $\Omega_{i,j,n} = 0$ if pixel $(i, j, n)$ is from background.

The hierarchical background model constructed in this paper convinces that the background of the video contains low-frequency components and high-frequency components, which is

$$\mathbf{B} = \mathbf{B}^h + \mathbf{B}^l.$$  

The two components are modelled by sparse combination of some dictionary atoms and spatiotemporal quantization clustering.

1) SPARSE DICTIONARY BACKGROUND MODEL

Here, it is considered that the low-frequency components in the video are obtained from a sparse linear combination of multiple dictionary atoms, that is

$$\mathbf{B}^l = \mathbf{S} \times_4 \mathbf{Z}. \quad (2)$$

where $\mathbf{S} \in \mathbb{R}^{I \times J \times 3 \times N}$ is a dictionary of dictionary atoms, $N_S$ is the size of the dictionary, $\mathbf{Z} \in \mathbb{R}^{N \times N_S}$ is the representation coefficient. In the Eq. (2), $\mathbf{S} \in \mathbb{R}^{I \times J \times 3 \times N_S}$ is composed of $N_S$ low-frequency background representation factors, the size of which is equivalent to $N_S$ frames for the image, $\mathbf{Z} \in \mathbb{R}^{N \times N_S}$ is the selection matrix of the dictionary atom, which is sparse.

Conventional low-rank based models do not emphasise the continuity of the frame image, and decomposes each frame into a weighted summation of multiple representation factors pixel-wisely. Here we consider that the changing ways of different pixels on the same frame image are related. Besides, this formulation enables us to approximate the actual background by infinite attempts, instead of the finite alternatives in low-rank based works.

Then, we add a sparsity constraint to the linear weighted summation representation and give the following objective function

$$\begin{align*}
\min_{\mathbf{S}, \mathbf{Z}} & \quad \|\mathbf{B}^l - \mathbf{S} \times_4 \mathbf{Z}\|_F^2 \\
\text{s.t.} & \quad \|\mathbf{Z}\|_0 \leq K.
\end{align*}$$

where $K$ refers to the maximum number of dictionary atoms in the sparse representation.

2) SPATIOTEMPORAL QUANTIZATION CLUSTERING MODEL

The high-frequency component $\mathbf{B}^h$ in the background reflects the detail information in the background and the irregularly changing modules. Here, we use the method of quantization clustering to analyze the residual part of each group of frames (GOF) after removing the low-frequency components. In [20], a pixel-wise quantization way is proposed to support the pixel-wise changing tendency. In this paper, in addition to temporal clustering, we still conduct quantization spatially.

As is shown in Fig. 2(c), a tensor based spatiotemporal quantization way is proposed. We get the residual of the low-frequency background $\mathbf{E}^h$ by

$$\mathbf{E}^h = \mathbf{B} - \mathbf{B}^l. \quad (4)$$

Here, the residual corresponds to the high-frequency background is recorded as $\mathbf{E}^h$. The residual for the non-foreground region is $P_{\mathbf{E}^h}$.

The entire video sequence is divided into $K$ GOFs (each GOF contains $f$ frame images), that is $K = N / f$. For the $k$th GOF, as shown in Figure 2(c), it is divided into $[I] \times [J]$
The codebook for quantization clustering can be obtained by training the following mode

\[
\min_{Q} \sum_{k=1}^{K} \sum_{i=1}^{[I]} \sum_{j=1}^{[J]} \sum_{t=1}^{[T]} \|E_{k[i],[j],[t]} - q_c\|,
\]

\[\text{s.t. } q_c \in Q, (k[i], j, t) \in \Omega. \quad (6)\]

Furthermore, the high-frequency background of the video foreground and background regions are constructed separately as follows

\[
B_{k[i],[j],[t]}^{h} = \begin{cases} 
\arg\min_{q_c \in Q} \|E_{k[i],[j],[t]} - q_c\|, & \text{if } E_{k[i],[j],[t]} \in \Omega, \\
\arg\min_{q_c \in Q} \sum_{E_{k[i],[j],[t]} \in \Omega} \|E_{k[i],[j],[t]} - q_c\|, & \text{otherwise}.
\end{cases}
\]

(7)

Among them, \(E_{k[i],[j],[t]}^{h}\) represents any GOF in a non-foreground area. Here, after quantifying each region block \([i] = 1, \ldots, [I]; [j] = 1, \ldots, [J]; k = 1, \ldots, K\), we get the detailed background of the video \(B^{h}\).

Considering Eq. (2) and Eq. (7), the background model is

\[
B = S \times _4 Z + B^{h}. 
\]

(8)

In addition, in the non-foreground region \((\Omega)\), the background is actually the video. That is, by defining \(W_{i,j,n} = 1\) if pixel \((i, j, n)\) is in foreground region and \(W_{i,j,n} = 0\) otherwise, we have \(W_{\Omega} \otimes D = W_{\Omega} \otimes B\). Here, \(\otimes\) indicates that the elements at the same position is correspondingly multiplied. Thus, as to minimize the noise in the non-foreground region, the hierarchical background model is given by

\[
min_{S,Z,B^h} \|W \otimes (D - B)\|_F^2,
\]

\[s.t. \ B = S \times _4 Z + B^h, \|Z\|_0 \leq K. \quad (9)\]

B. HIERARCHICAL FOREGROUND MODEL

Most foreground extraction models require consistent distribution of neighboring pixels, which improves their robustness to discretely distributed noise. However, when solving the boundary of the foreground area, this way is far from optimum as most concave curves are recovered as straight lines or zigzag lines.

Therefore, in this section, an active contour model is introduced to better characterize the boundary of the foreground region, and then the active contour model is combined with the spatial continuity assumption to form a more accurate foreground model.

1) CONTINUITY CONSTRAINT IN THE FOREGROUND REGION

We construct the constraint on the continuity of the foreground region at first. For each pixel, we consider the consistency of the categories of four adjacent pixels (up, down, left, right) in the same frame and two pixels in adjacent frames, i.e.,

\[
\min_{\Omega} \sum_{G} \|\Omega_{i,j,n} - \Omega_{x,y,z}\|,
\]

\[s.t. \ G : |i-x| + |j-y| + |n-z| \leq 1. \quad (10)\]

Then, to introduce weighting for different pixels, we have

\[
\min_{\Omega} \sum_{G} f(\alpha)\|\Omega_{i,j,n} - \Omega_{x,y,z}\|.
\]

(11)

where \(f(\alpha)\) equals \(exp(\alpha - D_{i,j,n} - D_{x,y,z})\). In addition, the size of the foreground area should also be limited to a certain
range, that is, the foreground area is restricted by a norm of \( l_0 \). The continuity constraint model constructed is

\[
\min_{\Omega} \sum_{G} f(\phi) \| \Omega_{i,j,n} \ominus \Omega_{x,y,z} \| + \beta \| \Omega \|_0
\] (12)

From the perspective of the graph model, Eq. (12) can be organized into the following form

\[
\min \left\{ \sum_{i,j,n} \beta \Omega_{i,j,n} + \sum_{G} f(\phi) \| \Omega_{i,j,n} \ominus \Omega_{x,y,z} \| \right\}.
\] (13)

That is, each node \( \Omega_{i,j,n} \) in the graph is assigned with a weight of \( \beta \), and each edge (connecting nodes \( \Omega_{i,j,n} \) and \( \Omega_{x,y,z} \)) is assigned with a weight of \( f(\phi) \).

2) ACTIVE FOREGROUND CONTOUR MODEL

The active contour model [52] is first proposed by Kass et al. in 1988 and is widely used for various image segmentation-related tasks. In 1995, Caselles et al. [53] further propose a more optimized Geodesic Active Contour (GAC) model. This model is widely used in various computer vision tasks [54].

For any frame \( D_n, n = 1, \ldots, N \), we record the result of the background subtraction as \( X_n \), that is \( X_n = D_n - B_n \). The operations on any frame are the same here, so for the convenience of presentation, we do not introduce subscripts to distinguish, that is, use \( X \) to represent the background subtraction result of any frame. As for \( X \), assuming that the foreground area is \( \Omega \), the subscript is omitted here as well. In fact, the entire foreground region consists of multiple closed subsets (assuming \( T \), that is, \( \Omega = \Omega_1 \cup \ldots \cup \Omega_T \).

In our model, like most conventional works, two kinds of energy are employed to constrain the foreground boundary, i.e., the internal energy to maintain the contour smoothness and the topology; the external energy to restrict the contour to the edge with the largest gradient. The energy function is given by

\[
\min_{\Omega, c_1, c_2, \ldots, c_T, \lambda} E(\Omega, c_1, c_2, \ldots, c_T, \lambda) = f(\Omega, X) + \lambda \int_{\Omega} \| \nabla x \| dx
\]

\[
+ \lambda \int_{\Omega_1} (c_1 - x)^2 dx + \lambda \int_{\Omega_2} (c_2 - x)^2 dx + \cdots + \lambda \int_{\Omega_T} (c_T - x)^2 dx,
\] (14)

where \( \lambda \) is a parameter that weighs the two energies, \( x \in X \). \( f(\Omega, X) \) is the first energy to constrain the contour of the foreground region, which is similar to the literature [31]. Here, the full variational norm is used to constrain the smoothness of the foreground region, that is

\[
f(\Omega, X) = \int_{\Omega} \| \nabla x \| dx.
\] (15)

Formulation (14) considers that the pixel values of the background subtraction residuals in each foreground region block (\( \Omega_i \)) are approximately the same and are distributed around \( (c_i) \). In addition, in the background area \( (\Omega) \), the pixel values after the background subtraction should all be 0.

Then, we can define a piecewise function on \( \Omega \)

\[
P_\Omega(x) =
\begin{cases}
(c_1 - x)^2, & x \in \Omega_1, \\
(c_2 - x)^2, & x \in \Omega_2, \\
\cdots, & \\
(c_T - x)^2, & x \in \Omega_T.
\end{cases}
\] (16)

Defining \( C = \{c_1, c_2, \ldots, c_T\} \), we reformulate the energy function as

\[
E(\Omega, C, \lambda) = f(\Omega, X) + \lambda \int_{\Omega} \| \nabla x \| + \lambda \int_{\Omega} P_\Omega(x) dx.
\] (17)

A level set function [55] \( \phi \) can be constructed accordingly, and its energy function is

\[
E(\Omega, C, \lambda, \phi) = \int_{\Omega} \| \nabla \phi \| + \lambda \int_{\Omega} Q_\Omega(x, C) \phi dx,
\]

s.t., \( Q_\Omega(x, C) = P_\Omega(x) - (x)^2 \). (18)

When solving \( \phi \), the horizontal plane used in our experiments is \( \Psi = 0.5 \).

By assuming that the foreground area obtained by the active contour model is denoted by \( \Omega_\phi \), a constraint for foreground is given by

\[
\min_{\Omega} \| \Omega - \Omega_\phi \|.
\] (19)

This requires that the foreground segmentation results obtained by the overall model are consistent with those obtained by the active contour model. In fact, \( \Omega_\phi \) can more accurately characterize the boundaries of the foreground region, but performs weakly when effectively detecting the complete foreground region. So, only the foreground area of \( \Omega_\phi \) is used as a constraint in this paper, which is

\[
\min_{\Omega} \| \Omega - \Omega_\phi \|_{\Omega_\phi=1}.
\] (20)

Introducing the constraints (20) into the model (12), we have

\[
\min_{\Omega} \sum_{G} f(\phi) \| \Omega_{i,j,n} \ominus \Omega_{x,y,z} \| + \beta \| \Omega \|_0 + \gamma \| \Omega - \Omega_\phi \|_{\Omega_\phi=1},
\] (21)

or

\[
\min \left\{ \sum_{i,j,n} \beta \Omega_{i,j,n} + \sum_{G} f(\phi) \| \Omega_{i,j,n} \ominus \Omega_{x,y,z} \| \right\}
\]

\[
+ \sum_{\Omega_\phi=1} \gamma \Omega_{i,j,n}.
\] (22)

Therefore, the active contour model established in this section adds a weight \( \gamma \) of size \( \phi \) to the nodes located in the foreground region.
IV. OPTIMIZATION OF THHS

The overall model of video foreground and background separation based on the previous discussion is

\[
\min_{B, \Omega_s, \Omega} \|W \otimes (D - B)\|_F^2 + \sum_G f(\alpha)\|\Omega_{i,j,n} - \Omega_{t,y,z}\| \\
+ \beta\|\Omega_0 + y\| - \Omega_\phi = 0, \nonumber \\
s.t. \quad B = S \times 4 Z + B^h, \|Z\|_0 \leq K. \quad (23)
\]

The model is non-convex, so it is difficult to give explicit solutions directly. So, the method of iteratively solving each component is used to obtain the foreground and background of the video.

A. SOLVING BACKGROUND PHASE

Assuming the foreground \(F\) in the video is known and the foreground area \(\Omega\) is known, the objective function is the same as (9). The model can be solved by augmented Lagrangian method. Without considering \(\|Z\|_0 \leq K\), the augmented Lagrangian function is

\[
\mathcal{L}(S, Z, B^h, \Lambda, \rho) = \|W \otimes (D - B)\|_F^2 \\
+ \langle \Lambda, B - S \times 4 Z - B^h \rangle \\
+ \frac{\rho}{2}\|B - S \times 4 Z - B^h\|_F^2. \quad (24)
\]

Then the solution of the objective function can be completed in a fixed format, that is, the formula for the iteration at step \(t + 1\) is

\[\begin{align*}
(B^{t+1}, Z^{t+1}, B^h^{t+1}) &= \arg\min_{B, Z, B^h} \mathcal{L}(S, Z, B^h, \Lambda^{t}, \rho^{t}), \\
\Lambda^{t+1} &= \Lambda^{t} + \rho^{t}\langle B^{t+1} - S \times 4 Z - B^h^{t+1} \rangle, \\
\rho^{t+1} &= \rho^{t} + \alpha (\rho^{t}).
\end{align*}\]

where \(\alpha\) is a positive parameter. The first sub-problem is

\[
\min_{S, Z, B^h} \|W \otimes (D - B)\|_F^2 + \langle \Lambda^{t}, B - S \times 4 Z - B^h \rangle \\
+ \frac{\rho^{t}}{2}\|B - S \times 4 Z - B^h\|_F^2. \quad (26)
\]

It can be solved by alternating iterations. First, when \(S, Z\) and \(B^h\) are known, solving the problem of \(B\) can be organized into an one-variable quadratic equation, and the solution is

\[B = \frac{2W \otimes D + \rho^{t}(S \times 4 Z + B^h) - \Lambda^{t}}{2W + \rho^{t}}. \quad (27)\]

After fixing \(B\), the tasks of solving \(S, Z\) and \(B^h\) variables are

\[
\min_{S, Z, B^h} \langle \Lambda^{t}, B - S \times 4 Z - B^h \rangle + \frac{\rho^{t}}{2}\|B - S \times 4 Z - B^h\|_F^2. \quad (28)
\]

Then, by considering the constraint \(\|Z\|_0 \leq K\) and assuming \(B = B + \frac{\Lambda^{t}}{\rho^{t}}\), the model (8) is same as the formulation in model (8). Thus dictionary learning can be done by the K-SVD algorithm [56], and sparse representation be done by the OMP algorithm [57]. Finally, the solution of \(B^h\) is given by formula (7). Another iterative optimization of the three parts results in better performance.

B. SOLVING LEVEL SET FUNCTIONS

The objective function of the level set function is

\[E(\Omega, C, \lambda, \phi) = \int_{\Omega} \|\nabla \phi\| + \lambda \int_{\Omega} Q_{\Omega}(x, C)\phi dx. \quad (29)\]

With the background \(B\) of the video known, \(x \in X\) becomes constant. At this time, the objective function can be solved by fixed iteration [58], that is

\[\begin{align*}
\phi(t+1) &= \nu - \theta \text{div} p^{t}, \\
v^{t+1} &= \min \{\max \{\phi(t+1) - \delta \lambda Q_{\Omega}(x, C), 0\}, 1\}, \\
p^{t+1} &= \frac{\rho^{t} + \delta \text{div} (p^{t} - v^{t})}{1 + \delta \text{div} (p^{t} - v^{t})}. 
\end{align*}\]

Among them, \(\delta\) is a parameter that controls the step size of the iteration, \(\theta > 0\), and \(p^{0} = 0\)

C. SOLVING FOREGROUND MODELS

With the background \(B\) and \(\Omega_s\) of the video known, the objective function of the foreground region solution becomes

\[
\min_{\Omega_s} \sum_{i,j,n} f(\alpha)\|\Omega_{i,j,n} - \Omega_{t,y,z}\| + \beta\|\Omega_0\| \\
+ \gamma\|\Omega_\phi\|_{\Omega_s=1} - \sum_{(i,j,c,n) \in \Omega_s} \|B_{i,j,c,n} - B_{i,j,c,n}\|^2_F, \quad (31)
\]

which is equal to

\[
\min_{i,j,n} \{\beta\Omega_{i,j,n} + \sum_{i,j,n} f(\alpha)\|\Omega_{i,j,n} - \Omega_{t,y,z}\| \\
+ \sum_{\Omega_s=1} \gamma\|\Omega_{i,j,n} - \Omega_{i,j,n}\|^2_F \Omega_{i,j,n} \}. \quad (32)
\]

where \(c = 1, 2, 3\). From the perspective of graph theory, the model (32) gives constraints for each node \(\Omega_{i,j,n}\) and constraints for any edge \(\Omega_{i,j,n}, \Omega_{s,y,z}\), so the model can be solved by the graph segmentation algorithm [59].

D. MODEL ALGORITHM

The algorithm (THHS) proposed in this paper is shown in algorithm 1.

V. COMPARATIVE TEST

In this section, the proposed algorithm will be shown and compared with the latest algorithms of six publicly released codes, including HDMI [20], TVRPCA [31], GFL [60], DECOLOR [34], LSD [61], PCP [62], OMoGMF [10] and SOIR. The parameters of the algorithm are optimized according to the default settings, or according to the suggestions discussed in the corresponding papers. The benchmark
dataset includes the I2R dataset [63] and the ChangeDetection dataset 2014 (CDnet2014) [16]. The computing environment for all experiments in this paper is as follows: Matlab R2016a; 3.40GHz Intel (R) Core (TM) CPU; 16-bit operating system; 64G memory.

A. TRAINING DICTIONARY DISPLAY
In this paper, the main components in the video background are obtained from a sparse linear combination of multiple background atoms. Examples of dictionaries trained on different video data are shown here. Training results for the videos Watersurface, Campus, and Escalator are shown in Figures 3, 4 and 5, respectively. This paper uses 220 frames as a video block. The dictionary of each video block contains 10 background dictionary atoms (which means, the dictionary size is 10). The background of each frame image is linearly represented by 3 dictionary atoms.

It can be seen from the figures that, on one hand, the main components of different dictionary atoms are similar, which are very close to the real background of the video. On the other hand, different atoms contain different details, among which some details originate from foreground or noise in the
video. For example, partially blurred grass on Watersurface and vague ghosting on platforms in Escalator. Here, each dictionary atom can be approximated as a state of the video background, so the dictionary atoms shown in the figure are very similar to the potential background result obtained by the low-rank representation theory (PCP, RPCA) models.

In addition, here are statistics of the frequency with which the 10 dictionary atoms of the Watersurface videos are used to represent all the backgrounds of the video. The entire video contains a total of 220 frames, and the background of each frame is represented by 3 dictionary atoms. The distribution of different dictionary atoms for a total of 660 usage frequencies is shown in Figure 6.

As can be seen from the figure, the frequencies of usage of different dictionary atoms are significantly different. The atom 4 in the figure is used in representing the background of almost all the frames (165/220). The number 10 atom is only used in the 10 times background representation of the total 220 frames.

B. DISPLAY OF COMPREHENSIVE FOREGROUND EFFECT

This section shows the effects of the active contour model without continuity constraints, and the effects of a comprehensive model combining Markov random fields and active contour constraints. The examples use the fall video and fountain02 video in the CDnet dataset, and the results are shown in Figure 7 and Figure 8 respectively. Among them,
Algorithm 1 Tensor Model for Hierarchical Background and Hierarchical Foreground Separation (THHS)

Input: original video \(D \in \mathbb{R}^{I \times J \times 3 \times N}\).
Output: areas of video background \(B\) and foreground objects \(\Omega\).

Compute:
When not converge (outer loop)
Step 1
When not converge (mid loop)
1) Lagrangian parameter \((\lambda, \rho)\): the update formula is shown in Eq. (25);
2) Hierarchical background:
   When not converge (inner loop)
   2.1) Overall background \((B)\): as Eq. (27);
   2.2) Dictionary background \((S, Z)\): K-SVD algorithm [56] solves dictionary \(S\), OMP algorithm [57] solves sparse coefficients \(Z\);
   2.3) Detail background \((B^d)\): Based on training Eq. (6) and resulting Eq. (7);
End when (inner loop)
End when (mid loop)
Step 2
Active contour foreground: based on iterative formulation \((\Omega_\psi)\): Eq. (30);
Step 3
Overall foreground: solving model(32): through graph segmentation algorithm [59];
End when (outer loop).

The two datasets in the figure show the two situations that are more difficult in practical applications. The shaking leaves and the constantly changing fountains, where the changing background textures occupy an area that is much larger than the moving foreground objects. It can be found from the figure that the active contour model can detect most of the foreground areas in the background subtraction result. But the detection results of the active contour model are often distributed in discrete regions. Therefore, this article uses prudent parameter selection way to ensure the effectiveness of the active contour model to detect the foreground region. So the continuity constraint is included to connect the detection results of multiple regional distributions in the third column of the figure to obtain a more accurate foreground region test result.

C. NUMERICAL COMPARISON EXPERIMENT

In this section, we compare the proposed model (THHS) with the results of all comparison algorithms for video foreground and background separation tasks. The data set contains all videos of the dynamic background task in the I2R data set and CDnet data set. The numerical comparison indicators are precision, recall, and F-measure indicators.

1) COMPARISON OF I2R DATASETS

First, experimental comparisons on the I2R dataset are performed. The results of randomly selected test pictures on each video are shown in Figure 9.

It can be seen from the figure that foreground extraction task is challenging when the background textures keeps changing and the noises are saliency enough. Changing background textures such as changing light, dancing leaves, surging fountains, and moving elevators interfere with almost all algorithms. In comparison, the algorithms based on spatial continuity constraints (THHS, HMAO, and DECOLOR) are more robust to noise, that less affected by discretely distributed noise but affected by regional noise, for example, HMAO incorrectly detects the elevator as the foreground, and DECOLOR often connects the foreground areas that are closer to each other. Models based on gradient information (THHS and TVRPCA) are robust to various noise, but multiple foreground areas are mis-detected as background areas inside. After comparison, the model of integrated spatial continuity constraints and gradient information (THHS) proposed in this paper achieved the best results.

The index calculations for the test results of all test images in the entire data set are given in Table 1.

As can be found from the table, THHS performs well on most videos. Especially, on the Escalator and Lobby datasets, THHS shows a significant improvement over most algorithms. This is because THHS combines the active contour model and the continuous constraint, that is, the gradient information is considered when dividing the foreground region, so it is more robust
to various types of noise interference. Finally, THHS achieved the best performance in terms of average accuracy.

2) CDnet DATASET COMPARISON
This section will compare the algorithm in this paper with other algorithms on the CDnet dataset. For the comparison
of 220 frames of each video, a representative one is selected and shown in Figure 10.

It can be found from the figure that the model THHS is still robust to noise interference when the interference is increasing in the background. At the same time, because the parameter selection is cautious, so the ability to remove noise is enhanced and some foreground areas that are not significant are judged as the background as expense. Then, similar to the experimental results in the previous section, the model based on spatial continuity constraints is more robust to discrete noise, but weaker in discriminating regionally distributed noise. The model based on gradient information is more robust to various noise, but difficult to accurately judge and classify the interior of the foreground area. Further, numerical measurements of all test results are recorded in Table 2.

In both videos, there is a slight disturbance of the large-scale background texture, that is, the specific value of each pixel changes within a certain interval around the real background value. THHS has achieved the best foreground and background separations on two datasets with large local noise interference in fall and fountain02. Finally, in terms of average accuracy, THHS performs a bit better than HMAO, and is significantly better than other algorithms.

3) ALGORITHM SPEED COMPARISON
This section gives a comparison of the running time of THHS with other methods, that is shown in Table 3. Here, the running time of 9 videos in the I2R dataset is averaged and the results are given. It can be seen that the complexity of THHS and TVRPCA is slightly higher than the traditional

| TABLE 1. Comparison of numerical indicators of foreground detection results on the I2R dataset. |
|-----------------------------------------------|---|---|---|---|---|---|---|---|---|---|
| video | THHS | HMAO | GPL | LSD | OMoGMP | PCP | DECOLOR | TVRPCA | SOIR |
| R | P | F | R | P | F | R | P | F | R | P | F | R | P | F | R | P | F |
|------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bootstrap | 0.86 | 0.68 | 0.76 | 0.72 | 0.81 | 0.65 | 0.69 | 0.65 | 0.65 | 0.64 | 0.58 | 0.75 | 0.53 | 0.62 | 0.86 | 0.50 | 0.63 |
| Campus | 0.82 | 0.83 | 0.75 | 0.90 | 0.81 | 0.65 | 0.62 | 0.65 | 0.65 | 0.64 | 0.58 | 0.75 | 0.53 | 0.62 | 0.86 | 0.50 | 0.63 |
| Curtain | 0.92 | 0.90 | 0.87 | 0.90 | 0.87 | 0.65 | 0.59 | 0.65 | 0.65 | 0.64 | 0.58 | 0.75 | 0.53 | 0.62 | 0.86 | 0.50 | 0.63 |
| Escalator | 0.80 | 0.66 | 0.77 | 0.58 | 0.62 | 0.66 | 0.63 | 0.65 | 0.64 | 0.58 | 0.75 | 0.53 | 0.62 | 0.86 | 0.50 | 0.63 |
| Fountain | 0.88 | 0.84 | 0.71 | 0.84 | 0.87 | 0.74 | 0.77 | 0.77 | 0.77 | 0.65 | 0.67 | 0.75 | 0.53 | 0.64 | 0.83 | 0.71 | 0.83 |
| Hall | 0.83 | 0.64 | 0.75 | 0.63 | 0.61 | 0.63 | 0.63 | 0.72 | 0.72 | 0.66 | 0.67 | 0.75 | 0.53 | 0.64 | 0.83 | 0.71 | 0.83 |
| ShopMall | 0.85 | 0.73 | 0.66 | 0.75 | 0.71 | 0.82 | 0.71 | 0.80 | 0.67 | 0.76 | 0.70 | 0.74 | 0.52 | 0.64 | 0.83 | 0.71 | 0.83 |
| Wattsface | 0.84 | 0.76 | 0.70 | 0.82 | 0.65 | 0.63 | 0.57 | 0.65 | 0.65 | 0.64 | 0.59 | 0.73 | 0.52 | 0.64 | 0.83 | 0.71 | 0.83 |
| Lobby | 0.85 | 0.78 | 0.74 | 0.79 | 0.95 | 0.50 | 0.65 | 0.65 | 0.64 | 0.59 | 0.76 | 0.56 | 0.64 | 0.83 | 0.71 | 0.83 |
| Average | 0.79 | 0.78 | 0.67 | 0.67 | 0.69 | 0.62 | 0.71 | 0.70 | 0.70 | 0.66 | 0.67 | 0.75 | 0.53 | 0.64 | 0.83 | 0.71 | 0.83 |

| TABLE 2. Comparison of foreground detection results on the CDnet2014 dataset. |
|-----------------------------------------------|---|---|---|---|---|---|---|---|---|---|
| video | THHS | HMAO | GPL | LSD | OMoGMP | PCP | DECOLOR | TVRPCA | SOIR |
| R | P | F | R | P | F | R | P | F | R | P | F | R | P | F | R | P | F |
|------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| boats | 0.53 | 0.64 | 0.62 | 0.34 | 0.33 | 0.35 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |
| canoo | 0.78 | 0.71 | 0.63 | 0.54 | 0.38 | 0.44 | 0.38 | 0.56 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 |
| fall | 0.52 | 0.54 | 0.56 | 0.47 | 0.43 | 0.46 | 0.39 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 |
| fountain | 0.08 | 0.10 | 0.12 | 0.09 | 0.09 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| fountain2 | 0.85 | 0.77 | 0.68 | 0.78 | 0.71 | 0.67 | 0.72 | 0.70 | 0.70 | 0.67 | 0.72 | 0.67 | 0.72 | 0.67 | 0.72 | 0.67 | 0.72 |
| overpass | 0.71 | 0.72 | 0.73 | 0.76 | 0.85 | 0.83 | 0.84 | 0.50 | 0.65 | 0.65 | 0.64 | 0.65 | 0.65 | 0.64 | 0.65 | 0.65 | 0.64 |
| Average | 0.61 | 0.60 | 0.60 | 0.46 | 0.46 | 0.50 | 0.43 | 0.39 | 0.49 | 0.49 | 0.51 | 0.46 | 0.51 | 0.46 | 0.51 | 0.46 | 0.51 |

| TABLE 3. Comparison of running time of various algorithms on the I2R dataset (unit: second). |
|-----------------------------------------------|---|---|---|---|---|---|---|---|
| THHS | HMAO | GPL | LSD | OMoGMP | PCP | DECOLOR | TVRPCA |
| 1909 | 65470 | 122897 | 52643 | 19.4 | 303.4 | 164.9 | 1898.4 | 60.7 |
PCA-based algorithms (PCP and DECOLOR) and the online algorithm OMoGMF.

VI. CONCLUSION
This paper proposes a joint optimization model for hierarchical foreground and hierarchical background separation. In the proposed model, a background model based on dynamical dictionary updating and spatiotemporal quantization clustering, and a foreground model based on active contours and Markov random fields, are introduced to improve the robustness of the model to noise. The experimental results show that the model THHS is able to deal with more drastic changes in the background than most conventional algorithms, and is more robust to regional noise interference.

REFERENCES
[1] L. Tian, H. Wang, Y. Zhou, and C. Peng, “Video big data in smart city: Background construction and optimization for surveillance video processing,” Future Gener. Comput. Syst., vol. 66, pp. 1371–1382, Sep. 2018.
[2] Y. Chen, J. Wang, Y. Bai, G. Castanon, and V. Saligrama, “Probabilistic semantic retrieval for surveillance videos with activity graphs,” IEEE Trans. Multimedia, vol. 21, no. 3, pp. 704–716, Mar. 2019.
[3] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, “Robust visual tracking via consistent low-rank sparse learning,” Int. J. Comput. Vis., vol. 111, no. 2, pp. 171–190, Jan. 2015.
[4] P. V. K. Borges, N. Conci, and A. Cavallaro, “Video-based human behavior understanding: A survey,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 11, pp. 1993–2008, Nov. 2013.
[5] B. Garcia-Garcia, T. Bouwmans, and A. J. Rosales Silva, “Background subtraction in real applications: Challenges, current models and future directions,” Comput. Sci. Rev., vol. 35, Feb. 2020, Art. no. 100204.
[6] 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., Ft. Collins, CO, USA, Dec. 1999, pp. 2246–2252.
[7] J.-M. Guo, C.-H. Hsia, Y.-F. Liu, M.-H. Shih, C.-H. Chang, and J.-Y. Wu, “Fast background subtraction based on a multilayer codebook model for moving object detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 10, pp. 1809–1821, Oct. 2013.
[8] B. Han and L. S. Davis, “Density-based multifeature background subtraction with support vector machine,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 5, pp. 1017–1023, May 2012.
[9] M. Hofmann, P. Tiefenbacher, and G. Rigoll, “Background segmentation with feedback: The pixel-based-adaptive segmenter,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Providence, RI, USA, Jun. 2012, pp. 38–43.
[10] H. Yong, D. Meng, W. Zhou, and L. Zhang, “Robust online matrix factorization for dynamic background subtraction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 7, pp. 1726–1740, Jul. 2018.
[11] M. Van Droogenbroeck and O. Paquot, “Background subtraction: Experiments and improvements for ViBe,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Providence, RI, USA, Jun. 2012, pp. 32–37.
[12] J. Stücker and S. Behnke, “Hierarchical object discovery and dense modelling from motion cues in RGB-D video,” in Proc. 23rd Int. Joint Conf. Artif. Intell. (IJCAI), Beijing, China, Aug. 2013, pp. 2502–2509.
[13] W. He, Y. Kim, H. Ko, J. Wu, W. Li, and B. Tu, “Local compact binary count based nonparametric background modeling for foreground detection in dynamic scenes,” IEEE Access, vol. 7, pp. 9229–92340, 2019.
[14] D. Zeng and M. Zhu, “Background subtraction using multiscale fully convolutional network,” IEEE Access, vol. 6, pp. 16010–16021, 2018.
[15] M. Qiu and X. Li, “A fully convolutional encoder-decoder spatial-temporal network for real-time background subtraction,” IEEE Access, vol. 7, pp. 85949–85958, 2019.
[16] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, “Changectection: Net: A new change detection benchmark dataset,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Providence, RI, USA, Jun. 2012, pp. 1–8.
[40] X. Wang and C. Navasca, “Low-rank approximation of tensors via sparse optimization,” *Numer. Linear Algebra with Appl.*, vol. 25, no. 2, Mar. 2018, Art. no. e2136.

[41] A. Sobral, C. G. Baker, T. Bouwmans, and E. Zahzah, “Incremental and multi-feature tensor subspace learning applied for background modeling and subtraction,” in *Proc.ICIAR*, Vilamoura, Portugal, Oct. 2014, pp. 94–103.

[42] P. Li, J. Feng, X. Jin, L. Zhang, X. Xu, and S. Yan, “Online robust low-rank tensor modeling for streaming data analysis,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1061–1075, Apr. 2019.

[43] A. J. Tom and S. N. George, “A three-way optimization technique for noise robust moving object detection using tensor low-rank approximation, Li/2, and TTV regularizations,” *IEEE Trans. Cybern.*, early access, Jun. 19, 2019, doi: 10.1109/TCYB.2019.2921827.

[44] J. Yang, X. Sun, X. Ye, and K. Li, “Background extraction from video sequences via motion-assisted matrix completion,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 2437–2441.

[45] 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.

[46] A. Sobral, T. Bouwmans, and E.-H. Zahzah, “Double-constrained RPCA based on saliency maps for foreground detection in automated maritime surveillance,” in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS)*, Karlsruhe, Germany, Aug. 2015, pp. 1–6.

[47] I. Kajo, N. Kamel, Y. Ruichek, and A. S. Malik, “SVD-based tensor completion technique for background initialization,” *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3114–3126, Jun. 2018.

[48] S. Javed, A. Mahmood, T. Bouwmans, and S. K. Jung, “Background–foreground modeling based on spatiotemporal sparse subspace clustering,” *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5840–5854, Dec. 2017.

[49] S. Javed, S. H. Oh, A. Sobral, T. Bouwmans, and S. K. Jung, “Background subtraction via superpixel-based online matrix decomposition with structured foreground constraints,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 930–938.

[50] S. E. Ebadi and E. Izquierdo, “Foreground segmentation via dynamic tree-structured sparse RPCA,” in *Proc. 14th Eur. Conf.*, Amsterdam, the Netherlands, Oct. 2016, pp. 314–329.

[51] Z. Hu, Y. Wang, R. Su, X. Bia, H. Wei, and G. He, “Moving object detection based on non-convex RPCA with segmentation constraint,” *IEEE Access*, vol. 8, pp. 41026–41036, 2020.

[52] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[53] V. Caselles, R. Kimmel, and G. Sapiro, “Geodesic active contours,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Cambridge, MA, USA, Jun. 1995, pp. 690–699.

[54] A. Badoual, D. Schmitter, V. Uhlmann, and M. Unser, “Multiresolution subdivision snakes,” *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1188–1201, Mar. 2017.

[55] S. Osher and J. A. Sethian, “Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations,” *J. Comput. Phys.*, vol. 79, no. 1, pp. 12–49, Nov. 1988.

[56] M. Aharon, M. Elad, and A. Bruckstein, “rmK-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[57] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[58] X. Bresson, S. Esedoḡlu, P. Vandergheynst, J.-P. Thiran, and S. Osher, “Fast global minimization of the active Contour/Snake model,” *J. Math. Imag. Vis.*, vol. 28, no. 2, pp. 151–167, Aug. 2007.

[59] 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, Dec. 2001.

[60] B. Xin, Y. Tian, Y. Wang, and W. Gao, “Background subtraction via generalized fused lasso foreground modeling,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4676–4684.

[61] X. Liu, G. Zhao, J. Yao, and C. Qi, “Background subtraction based on low-rank and structured sparse decomposition,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.

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

[63] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, “Statistical modeling of complex backgrounds for foreground object detection,” *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.