Moving Object Detection in Video via Hierarchical Modeling and Alternating Optimization

Linhao Li, Qinghua Hu, *Senior Member, IEEE*, and Xin Li, *Fellow, IEEE*

***Abstract*— In conventional wisdom of video modeling, the background is often treated as the primary target and foreground is derived using the technique of background subtraction. Based on the observation that foreground and background are two sides of the same coin, we propose to treat them as peer unknown variables and formulate a joint estimation problem, called *Hierarchical modeling* and *Alternating Optimization* (HMAO). The motivation behind our *hierarchical* extensions of background and foreground models is to better incorporate a priori knowledge about the *disparity* between background and foreground. For background, we decompose it into temporally low-frequency and high-frequency components for the purpose of better characterizing the class of video with dynamic background; for foreground, we construct a Markov random field prior at a spatially low resolution as the pivot to facilitate the noise-resilient refinement at higher resolutions. Built on hierarchical extensions of both models, we show how to successively refine their joint estimates under a unified framework known as alternating direction multipliers method. Experimental results have shown that our approach produces more discriminative background and demonstrates better robustness to noise than other competing methods. When compared against current state-of-the-art techniques, HMAO achieves at least comparable and often superior performance in terms of F-measure scores, especially for video containing dynamic and complex background.**

***Index Terms*— Hierarchical modeling, dictionary learning, joint estimation, alternating direction multipliers method (ADMM).**

I. Introduction

SEPARATING foreground (moving objects) from background is a fundamental problem in various computer vision and video processing applications including object tracking [1], [2], video surveillance [3], [4], behavior recognition [5], category prediction [6] and so on. Historically, background (BG) modeling has received more attention than foreground (FG) modeling partially because it is relatively easier to model the BG especially in the absence of camera motion. Based on a good estimation of BG, it is relatively easy to solve the problem of FG segmentation by subtracting BG from the videos (e.g., so-called background subtraction [33]). To enforce the smoothness constraint of object boundaries during FG segmentation, different models have been constructed in the literature - e.g., regional continuity [25], [26], total variation norm [34] and spatio-temporal sparsity [31].

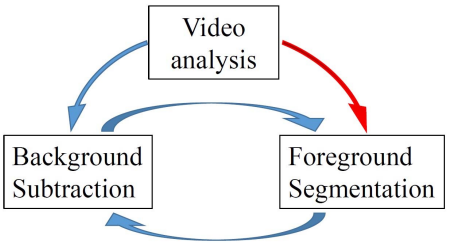


Fig. 1. The proposed Hierarchical Modeling and Alternating Optimization (HMAO) framework for FG-BG separation (red color highlights the difference between ours and previous approaches).

However, such a *background subtraction* framework can be challenged from several perspectives. **First**, a background-first approach would introduce unnecessary bias in background model to foreground segmentation. In fact, given the binary nature of video segmentation, resolving uncertainty with one immediately resolves the other. During each round of iteration, background estimate is successively refined by video analysis as shown by the blue color in Fig. 1, while foreground estimate is updated by prior constraints without resorting to input video at all [25], [26], [31], [34]. As highlighted by the red color in Fig. 1, a more principled way is to segment the foreground based on both background subtraction and video analysis results; in other words, BG and FG are treated as peer unknown variables (both successively refined at each iteration). **Second**, even if one acknowledges the priority of BG (e.g., it usually contain a lot more pixels than FG), the complexity of accurately modeling BG is high. Irregular motion (e.g., rippling water, waving leaves, fluttering flags) and textures (e.g., meadowland with varying depths and illuminations) in the physical world are two primary interfering factors, which make BG modeling a long-standing open problem. In view of various limitations with BG modeling, obtaining FG by BG subtraction is arguably ad-hoc and far from being optimized.

In this paper, we propose to take a *hierarchical* approach toward modeling BG/FG and formulate FG-BG separation as

Manuscript received March 28, 2018; revised August 20, 2018 and September 28, 2018; accepted November 10, 2018. Date of publication November 22, 2018; date of current version December 12, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61732011, 61432011, U1435212, and 61502332, and in part by the Applied Foundamental Research Program of Qinghai Province under Grant 2019-ZJ-7017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chun-Shien Lu. *(Corresponding author: Qinghua Hu.)*

L. Li and Q. Hu are with the Tianjin Key Lab of Machine Learning, School of Computer Science and Technology, Tianjin University, Tianjin 300350, China (e-mail: huqinghua@tju.edu.cn).

X. Li is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506-6109 USA.

Digital Object Identifier 10.1109/TIP.2018.2882926

a *joint* optimization problem. In our *temporally* hierarchical model, BG consists of two components: *averaging* and *detail* targeting at characterizing low-frequency and high-frequency components respectively; our two-component model can be interpreted as the combination of previous work on dynamic background models [23] and texture background model [35] in order to more accurately characterize complex BG in the physical world. As an example, Fig. 2 shows that regularly changing patterns in the BG (water scene) correspond to the averaging component; while irregularly changing patterns (e.g., intensity variations arising from ripples and reflections) the detail component. Furthermore, while building a Markov random field model for the FG, we have taken a *spatially* hierarchical approach of starting from a low-resolution and propagating the class label from low-resolution to high-resolution in a supervised manner. This way we can improve the robustness of FG modeling to noise (including the errors caused by BG estimation). By treating BG and FG as a pair of peer variables, we formulate a joint optimization problem and solve it by the Alternating direction multipliers method (ADMM) [36]. The main contributions of this paper are summarized as follows:

- Hierarchical modeling of BG. To better characterize dynamic structures in natural scenes, we sequentially estimate temporally low-frequency and high-frequency components of BG which respectively model the averaging and detail patterns. We argue that modeling detail patterns of BG (instead of treating them as outliers) improves the accuracy especially for the class of video containing dynamic background and self-repeating textures.
- Hierarchical modeling of FG. To improve robustness to noise (including potential errors in BG estimate), we propose to first detect FG at a low resolution (LR) and hierarchically refine such estimation at spatially higher resolutions. To propagate label information from LR to HR, rank-1 constraint of the BG and l_1-norm constraint of the FG are jointly enforced by graph cut techniques.
- Treating FG and BG as peer variables. Despite the existing joint optimization framework for FG-BG separation (e.g., DECOLOR [25]), BG is often viewed as the primary and carries more weight than FG. We propose to treat FG and BG as peer unknown variables and update their estimates by alternating optimization. Unlike conventional approaches, FG is also refined by exploiting additional information from video (e.g., label information) as highlighted by the red color in Fig. 1.

Our approach based on Hierarchical Modeling of BG/FG and Alternating Optimization (HMAO) has been experimentally verified on two popular video datasets for moving object detection: *I2R* and *CDNet 2014*. The proposed HMAO has been compared against seven leading algorithms whose codes are publicly available. It has been found that HMAO has achieved at least comparable and often superior performance to other competing approaches in terms of F-measure performance. Especially for those video containing dynamic background, HMAO demonstrates improved robustness to complex background and accuracy for moving object detection.

The remainder of this paper is organized as follows. Section II briefly reviews existing works on statistical and sparsity-based BG models. Section III provides the formulation of joint optimization problem and the derivation of the solution algorithm. Section IV reports our experimental results including the comparison between this work and other competing approaches. Finally, Section V provides some concluding remarks and outlines the direction for future research.

II. Related Works on Background Modeling

Existing works on modeling/estimating BG from video can be classified into two categories: *statistical models* and *sparsity-based models*.

A. Statistical Models

Statistical background models in the literature often employ individual pixel values or pixels within a region as input features. For example, individual pixel values were modeled by Gaussian distributions in 1997 [11] and by Mixture of Gaussian (MOG) in [7]; in the following years, other Gaussian-based algorithms [8], [9] have also reported good performance. Along this line of research, Kernel density estimation (KDE) was proposed to model the local pixel value variations in [16]; a uniform kernel with variable size was developed in [18] and density estimation was combined with support vector machine (SVM) in [20]. When separating BG from FG, the codebook of clustered pixel value series robust to environmental changes was considered in [14]; its multi-scale and multilayer extensions appeared in [21] and [15] respectively. In [23], radial basis function neural network was used to model pixel value series; a universal algorithm named Visual Background Extractor (ViBe) was proposed in [19] and later improved in [37]. Similar strategy also appeared in Pixel-Based Adaptive Segmenter (PBAS) [38]. The consensus of sample was employed in SAmple CONsensus (SACON) algorithm in [39], which later became the consensus of word [17] and the consensus of lightness [40].

Region-based approaches are mostly based on the observation that neighboring pixels are not independent from each other in video. Local binary patterns (LBP), which is insensitive to illumination changes, has been widely used to capture textured BG [35]; Local difference patterns (LDP) was later introduced to tackle the characterization of dynamic background [41]; Markov random field (MRF) was employed to estimate similarities between regions in [22]. More recently, Self Organized Maps (SOM) was developed for adapting dynamic background in [42]; region cues were introduced into Gaussian Mixture Model (GMM) to produce a regional spatially-consistent background model in [10]. Last but not the least, proper combination of different features or statistics often achieves improved performance - e.g., an efficient background model integrating six kinds of local features demonstrated superior performance in [43] when compared with conventional local models.

B. Sparsity-Based Models

Sparsity-based BF models are often related to the idea of projecting high-dimensional data onto a lower dimensional subspace. Among early attempts, Principal component

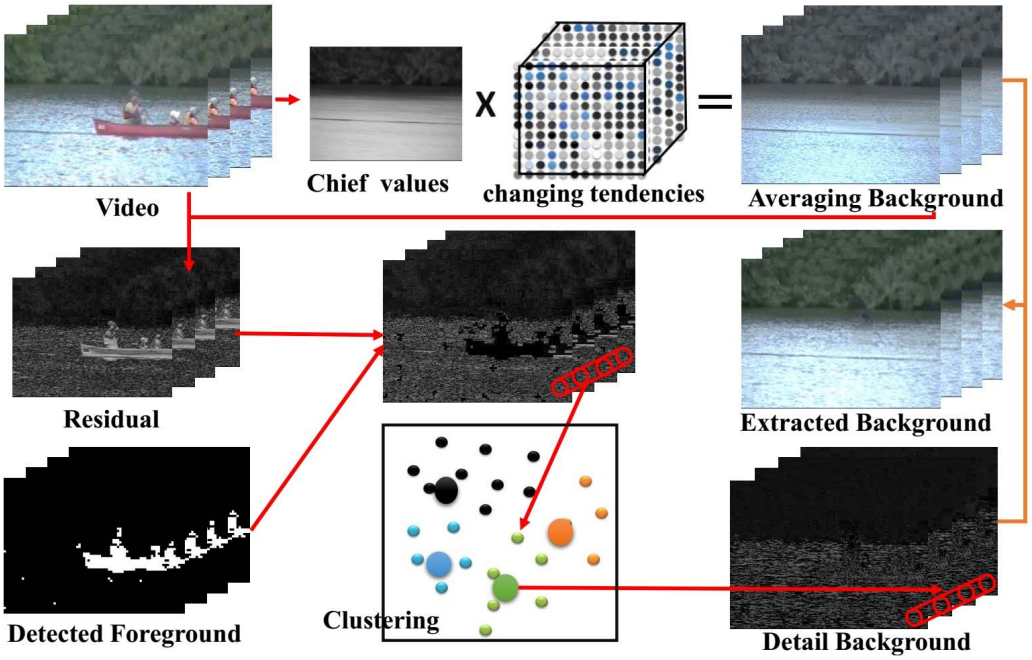


Fig. 2. The framework of the hierarchical background (input: video sequence and detected foreground; output: averaging component and detail component). Note that regularly changing patterns in BG are modeled by the averaging component; irregularly changing patterns in BG, caused by the variations in the background, are modeled by the detail component.

analysis (PCA) was proposed for modeling the BG in [24]- i.e., keep only the eigenbackgrounds associated with the few largest eigenvalues. Later a video frame was decomposed into the combination of a low-rank matrix and a sparse one in [44]; rank-1 constraint was used to derive an efficient BG estimation algorithm in [26]. More recently, tensor-Based low-rank framework was analyzed in [29]; the low-rank model was integrated with sparse subspace clustering in [32] based on the assumption that BG be spanned across multiple manifolds. Based on a similar low-rank hypothesis, other researchers have worked on the constraints for the FG matrix - e.g., Total Variation (TV) penalty on sparse deviations was employed to better handle noisy FG data in [45]; this framework was later improved in [34]. In [25], Markov Random Field (MRF) prior was introduced for better suppressing noise components and small background motion; in [28], structured sparsity-inducing norm was proposed to model the FG component.

In sparsity-based models, robustness is an issue that has attracted increasingly more attention in recent years. In [46], the robustness of BG modeling was improved by sparse signal recovery - i.e., a new frame can be represented by the sparse linear combination of a few preceding frames plus a sparse outlier term; in [30], BG was modeled by robust dictionary learning. This framework was further improved by maintaining historical pixels in [47] and by incorporating a spatio-temporal group sparsity constraint in [31]. Besides the above local and spatial models, other works concentrate on exploring extra information from video to improve the robustness - e.g., in [48], superpixel was proposed as the prior information in the background substraction framework; in [49], extra information such as the Gaussian and Laplacian images of raw video data have also proven effective.

III. Formulation of Hierarchical Modeling and Alternating Optimization Model

We introduce some necessary notations first. For a given video $[\mathbf{D}_1, \ldots, \mathbf{D}_N] \in \mathbb{R}^{I\times J\times 3\times N}$ (N is the number of frames), background is denoted by $[\mathbf{B}_1, \ldots, \mathbf{B}_N]$ and foreground is denoted by $[\mathbf{F}_1, \ldots, \mathbf{F}_N]$. The binary FG mask is $\Omega : \Omega = [\Omega_1, \ldots, \Omega_N] \in \mathbb{R}^{I\times J\times N}$ where $\Omega_{i,j,n} = 1$ if pixel (i, j, n) is in foreground and $\Omega_{i,j,n} = 0$ if pixel (i, j, n) is in background. In other words, Ω denotes the support of FG regions; the complement of Ω ($\overline{\Omega}$) denotes the support of BG regions. We assume that video is decomposed of short group of pictures (GOP) $[\mathbf{D}_{(1)}, \ldots, \mathbf{D}_{(K)}]$ each containing f frames and $K = N/f$ is the number of GOPs. Since the operations are identical for all picture groups $\mathbf{D}_{(k)}$, we drop the subscript (k) and use $\mathbf{D} \in \mathbb{R}^{I\times J\times 3\times f}$ to denote an arbitrary $\mathbf{D}_{(k)}$ $(k = 1, \ldots, K)$ for notational simplicity; similarly $\mathbf{B}$ represents an arbitrary BG group $\mathbf{B}_{(k)}$.

A. Hierarchical Background

We propose a hierarchical representation for the BG (as shown in Fig. 3) by decomposing it into low-frequency and high-frequency components - i.e.,

$$\mathbf{B} = \mathbf{B}^h + \mathbf{B}^l, \tag{1}$$

where, h, l denote high-frequency and low-frequency respectively. The motivation behind such hierarchical decomposition of BG is two-fold.

First, low-frequency component ($\mathbf{B}^l$) corresponds to constant or regularly changing patterns in the BG. For pixel

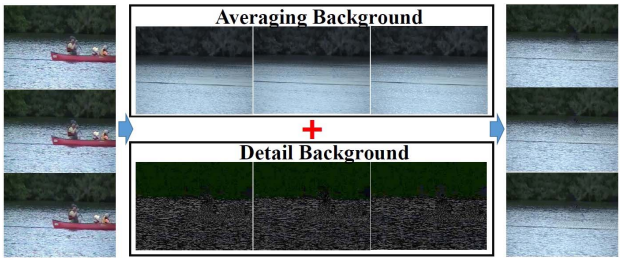


Fig. 3. The proposed hierarchical background model. The background of each frame is composed by one averaging background layer and one detail background layer.

values $\mathbf{B}^l_{i,j,:,n}(n = 1, \ldots, N)$,[1] we assume that intensity values are either constant or vary slightly. Throughout this paper, we name the center of N pixel values as "*chief values*" (conceptually similar to "historical pixels" in [47]). As shown in Fig. 2, we use the tensor product of chief value tensor ($\mathbf{B}^{l*}$) and changing tendency tensor (T) to model the low-frequency component of BG - i.e.,

$$\mathbf{B}^l = \mathbf{B}^{l*} \times_4 T. \tag{2}$$

Here, $\mathbf{B}^{l*} \in \mathbb{R}^{I \times J \times 3 \times 1}$ is a frame decomposed of the chief values from all locations, $T \in \mathbb{R}^{N \times 1}$ is a first-order matrix indicating the changing tendency for the entire frame, and "$\times_4$" is the 4−mode product that denotes multiplying a tensor by a matrix [52] (a brief introduction is given in Appendix). Note that we require $\|T\|_2 = 1$ in order to ensure: 1) changing tendency will not be influenced by averaging pixel values; 2) BG intensity is only reflected by chief values.

Since the changing tendencies of different pixels are usually independent in video, the choices of T are not unique. Here, we assume that the maximum number of different choices for T is N_T and introduce a selecting variable ($I_{i,j} \in \mathbb{R}^{N_T \times 1}$) for pixel at location (i, j) ($i = 1, \ldots, I$; $j = 1, \ldots, J$). Then Eq. (2) can be rewritten pixel-wisely - i.e.,

$$\mathbf{B}^l = \mathbf{B}^{l*}_{i,j,:} \times_4 (T \times I_{i,j}) = \mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T, \tag{3}$$

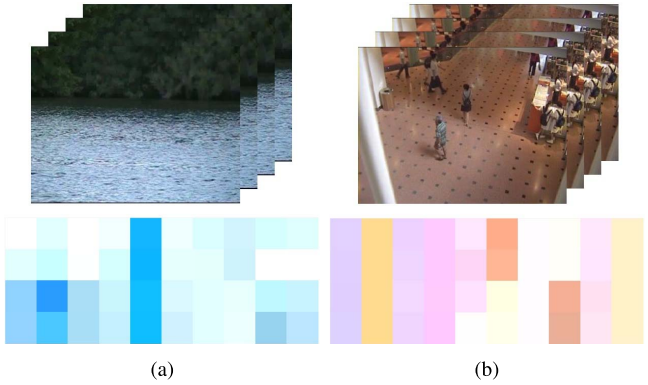
where $I_{i,j} \in \mathbb{R}^{N_T \times 1}, T = [T_1, \ldots, T_t] \in \mathbb{R}^{N \times N_T}$ is the candidate set of changing tendency matrices.

By contrast, long-term changing tendencies are usually difficult to model. For example, consider a fixed physical position of a flowing river, running water often makes the long-term changing tendency *irregular* due to complicated interaction between reflection surface and light source. To alleviate this difficulty, we propose to model each GOP *locally* - i.e., [2]

$$\mathbf{B}^l = \mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T, \quad \|T\|_2 = 1, \quad I \in \mathbb{R}^{N_T \times 1}, \; T \in \mathbb{R}^{f \times N_T}. \tag{4}$$

[1] We use (:) to denote all indexes in this dimension - e.g., $(i, j, :, n)$ means $(i, j, 1:3, n)$ or the RGB value at position (i, j) in the n-th frame.

[2] You should notice that, from here, as is illustrated in the first paragraph of Sec. III, $\mathbf{B}$ is used for representing an arbitrary BG group $\mathbf{B}_{(k)}$ and $\mathbf{D}$ denotes an arbitrary picture group $\mathbf{D}_{(k)}$. Similarly, here, T refers to the changing tendency that intended for an arbitrary group T_{GOP}. So, $\mathbf{B}$ $(\mathbf{B}^l, \mathbf{B}^h) \in \mathbb{R}^{I \times J \times 3 \times f}$, $\mathbf{D} \in \mathbb{R}^{I \times J \times 3 \times f}$ and $T \in \mathbb{R}^{f \times N_T}$ in the rest of the paper.



(a) (b)

Fig. 4. The estimated changing tendency matrices (T) of Canoe (CDnet dataset) and ShoppingMall (I2R dataset). (a) Canoe. (b) ShoppingMall.

According to (4), the changing tendency of entire video data is segmented into pixel-wise and short-term representation. In Fig. 4, we have shown the estimated latent changing tendency matrix T for some exemplar video.

For chief value tensor of each GOP ($\mathbf{B}^{l*}$), we assume that the BG should be constant (i.e., the chief values from different GOPs to be the same) except for unexpected illumination variations. It follows that most illumination changes can be characterized by the previous changing tendency matrix T. After vectorizing the chief value tensor (i.e., transforming each $\mathbf{B}^{l*}$ into a vector $\mathbf{B}^{l*}_{vec}$), we conclude that the global chief value matrix $\mathbf{B}^{l*}_{vec}$ is low-rank - i.e., $\mathbf{B}^{l*}_{vec}$ satisfies the rank constraint $rank(\mathbf{B}^{l*}_{vec}) = 1$.[3]

Second, high-frequency components ($\mathbf{B}^h$) reflect details or irregularly changing patterns in the BG. In order to model those irregular patterns, we propose to cluster pixel-wise residuals of each GOP and use the centroid of each cluster as the representative codeword. First, we obtain the residual ($\mathbf{E}^h$) representation by

$$\mathbf{E}^h = \mathbf{B} - \mathbf{B}^l. \tag{5}$$

Then we orthogonally project the residuals onto the linear space spanned by non-FG pixels; or equivalently, we consider $P_{\overline{\Omega}}(\mathbf{E}^h)$ decomposed of detail BG and noise only. Since the deviation between a FG detection result and the ground truth is inevitable, detailed features missed by FG detection can still be counted as leftover noise in the detail BG, which improves the robustness of BG estimation (as shown in Fig. 2).

It should be noted that we take the *principal* component of the residuals as the *detail* BG instead of treating them as outliers (unlike those in conventional models). We argue that most dynamic and self-repeating textures - e.g., rippling water, waving leaves and fluttering flags - are actually a part of video background (instead of moving objects in FG). To model these dynamic textures, we simply cluster short-term residuals on a pixel-by-pixel basis - i.e., self-repeating texture leads to periodic residual values; while random noise produces stochastic residual values. Therefore we can search the most

[3] We adopt the rank-1 constraint for its effectiveness and simplicity. It's possible to use a low-rank constraint instead - i.e., the corresponding solution of $\mathbf{B}^{l*}_{vec}$ will take the SVD (rather than average) of the background matrix.

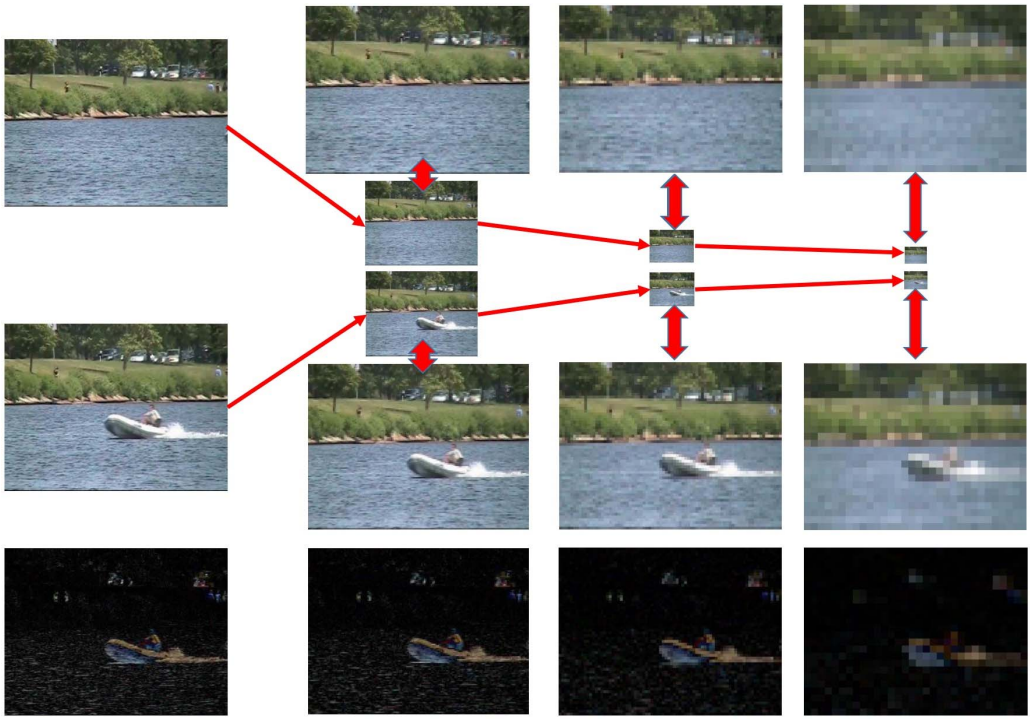


Fig. 5. Background subtraction results (bottom line) of frames in different resolutions. Here the low resolution frames are obtained by averaging every P-by-P pixels in the original frames. From left to right: P = 1 (original frames); $P = 2$; $P = 4$; $P = 8$.

frequent short-term residuals by

$$\min_{Q} \| P_{\overline{\Omega}} \mathbf{E}^h_{i,j,:,:} - q_c \|, \quad s.t.\ q_c \in Q. \tag{6}$$

where Q is the quantization codebook sized by $\mathbb{R}^{3\times f\times C}$ and C is the size of codebook.

Next we quantize high-frequency components ($\mathbf{B}^h_{i,j,:,:}$) and represent them using trained codebooks. For non-FG regions such as shown in Fig. 2, the high-frequency component of BG is represented by the corresponding center of residual sequences. However, in FG region where BG is covered by FG objects, residual representation becomes meaningless (because its theoretically impossible to recover those missing pixels in the BG). For simplicity, we empirically choose the most frequent centroid to replace them (can be interpreted as a strategy of inpainting). The high-frequency component of BG is then quantized by

$$\mathbf{B}^h_{i,j,:,:} = \begin{cases} \arg\min_{q_c} \|\mathbf{E}^h_{i,j,:,:} - q_c\|, & if\ \mathbf{E}^h_{i,j,:,:} \in \overline{\Omega}. \\ \arg\min_{q_c} \sum_{\mathbf{E}_{\overline{\Omega}}} \|(\mathbf{E}_{\overline{\Omega}})^h_{i,j,:,:} - q_c\|, & if\ \mathbf{E}^h_{i,j,:,:} \in \Omega. \end{cases} \tag{7}$$

where $\mathbf{E}_{\overline{\Omega}}$ denotes arbitrary GOP that is in the non-foreground regions, q_c is the assigned centroid recording $\mathbf{B}^h_{i,j,:,:}$ and c is the number of codebook centroids.

Putting things together, we can rewrite the complete BG model as follows

$$\mathbf{B}_{i,j,:,:} = \mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T + \mathbf{B}^h_{i,j,:,:} \quad s.t.\ rank(\mathbf{B}^{l*}_{vec}) = 1, \quad \|T\|_2 = 1. \tag{8}$$

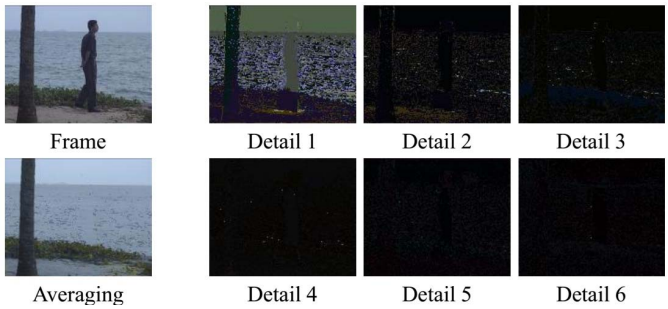


Fig. 6. HMAO with a multi-layer structure (1 averaging layer and 6 detail layers). Note that the pixel values of all detail layers are enlarged by 5 times for better visualization.

Note that the idea of decomposing video background into a combination of averaging and detail background can be extended in a multi-resolution manner. More specifically, by iteratively clustering the residual, we can build up a multi-layer decomposition of the BG - i.e., $\mathbf{B} = \mathbf{B}^h + \mathbf{B}^{l1} + \mathbf{B}^{l2} + \mathbf{B}^{l3} + \ldots$. This way, the detail in the i-th layer is obtained by clustering the residual of $\mathbf{B}^h + \mathbf{B}^{l1} + \ldots + \mathbf{B}^{l_{i-1}}$. An example is given in Fig. 6, where a 6-layer decomposition of detail BG is presented.

B. Hierarchical Foreground Detection

Conventional approaches toward FG modeling mostly focus on the enforcement of spatial continuity constraints [25], [47], which cannot effectively discover concealed FG regions misclassified as noise. Since noise can be more salient than the concealed FG objects in the residual, it is difficult to

overcome this limitation within the conventional framework of BG subtraction. As illustrated in Fig. 1, a more effective strategy is to explore the latent supervising information from raw video and use it to enforce the constraint about FG regions. To implement this strategy, we advocate a hierarchical approach of starting from a down-sampled version of the given video and propagating the labeling of FG detection from low-resolution to high-resolution in a supervised manner. There are three specific issues to be addressed in our hierarchical approach.

First, we propose to propagate FG label information in a hierarchical manner to improve the robustness to noise. As shown in Fig. 5, we obtain a low-resolution (LR) representation of video by spatial averaging (noise components are less salient than FG objects after averaging). In most videos with dynamic BG, noise typically associates with isolated pixels or small patches, which become gradually less salient as the resolution decreases; by contrast, FG objects often remain as continuous and salient regions even at low resolutions. Let us denote the low-resolution video by $\mathbf{D}^{lr}$ and low-resolution foreground by Ω^{lr} respectively. To detect the foreground from a LR video, we set up a model based on rank-1 hypothesis of BG and the l_1-norm constraint of FG as follows

$$\begin{aligned} &\min_{\mathbf{B}^{lr},\mathbf{F}^{lr}} \ \|\mathbf{D}^{lr}-\mathbf{B}^{lr}-\mathbf{F}^{lr}\|_F^2+\mu\|\mathbf{F}^{lr}\|_1. \\ &s.t.\ rank(\mathbf{B}^{lr})=1; \end{aligned} \tag{9}$$

where the l_1-norm constraint can be enforced by a soft-thresholding operator. Then we can infer the FG region Ω^{hr} from Ω^{lr} at a high resolution (HR) pixel-wisely.

Meantime, we assume the pursued FG regions should be similar - i.e.,

$$\min_{\Omega} \quad \|\Omega-\Omega^{hr}\|. \tag{10}$$

Note that the hierarchical constraint Ω^{hr} is robust to the noise, but it might not perform well in identifying the contour of FG.

Since the efficiency of the soft-thresholding operator is determined by the parameter (μ), we need to carefully select the parameter μ in Eq. (9) so that the FG regions are detected with high confidence. Since the extracted non-FG regions are often error-prone, we propose to refine constraint (10) by allocating different weights for FG and non-FG regions respectively - i.e.,

$$\mathbf{C}(\Omega)=\|\Omega-\Omega^{hr}\|_{\Omega^{hr}=1}+\nu\|\Omega-\Omega^{hr}\|_{\Omega^{hr}=0}. \tag{11}$$

where $\nu=1/3$ reflects our confidence about the detected FG regions (it is hand-crafted).

Second, we propose to construct a spatio-temporal Markov random field (MRF) model as the FG prior. For each pixel, we consider its eight surrounding neighbors within the same frame and two adjacent neighbors in the previous/next frames which should be labeled the same as the current pixel. We have adopted the following notation for neighboring pixels ($\mathbb{G}$)

$$\mathbb{G}: |i-x|+|j-y|+|n-z|\leq 1, \tag{12}$$

and the following objective function

$$\min_{\Omega}\sum_{\mathbb{G}}\|\Omega_{i,j,n}-\Omega_{x,y,z}\|. \tag{13}$$

Such spatio-temporal MRF is a good fit for pixels within moving objects; however, it requires extra attention while dealing with the boundary of different regions (i.e., discontinuities). To address this issue, we note that the difference between intensity values of neighboring pixels is typically large suggesting the existence of object boundaries. Therefore, one can leverage the intensity difference into the formulation of a weighted objective function (similar to the idea of edge-stopping in classical Perona-Malik diffusion [13])

$$\min_{\Omega}\sum_{\mathbb{G}} exp(\alpha_0-\alpha(\mathbf{D}_{i,j,:,n}-\mathbf{D}_{x,y,:,z}))\|\Omega_{i,j,n}-\Omega_{x,y,z}\|. \tag{14}$$

For notational simplification, we can use $f(\alpha)$ to denote $\exp^{\alpha_0-\alpha(\mathbf{D}_{i,j,:,n}-\mathbf{D}_{x,y,:,z})}$ and rewrite Eq. (14) into

$$\min_{\Omega}\sum_{\mathbb{G}} f(\alpha)\|\Omega_{i,j,n}-\Omega_{x,y,z}\|. \tag{15}$$

Third, a physical constraint arising from Pauli's exclusion principle dictates that BG is often *occluded* by FG objects - i.e., $\mathbf{D}=P_{\overline{\Omega}}\mathbf{B}+\mathbf{F}$ where $\mathbf{B}$ is the extracted hierarchical BG. Additionally, the size of FG region should also be limited to a certain range; in other word, the region is constrained by a l_0-norm - i.e.,

$$\min_{\Omega}\|\Omega\|_0, \quad s.t., \ \mathbf{D}=P_{\overline{\Omega}}\mathbf{B}+\mathbf{F}. \tag{16}$$

Putting things together, we can rewrite the overall objective function of FG as follows

$$\begin{aligned} &\min_{\mathbf{F},\Omega}\sum_{\mathbb{G}} f(\alpha)\|\Omega_{i,j,n}-\Omega_{x,y,z}\|+\beta\|\Omega\|_0+\gamma\,\mathbf{C}(\Omega) \\ &s.t., \ \mathbf{D}=P_{\overline{\Omega}}\mathbf{B}+\mathbf{F}. \end{aligned} \tag{17}$$

C. Formulation and Optimization of HMAO

Based on the above BG and FG models, we propose to formulate FG-BG separation as the following joint optimization problem:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{F},\Omega}\{&\sum_{\mathbb{G}} f(\alpha)\|\Omega_{i,j,n}-\Omega_{x,y,z}\|+\gamma\,\mathbf{C}(\Omega)+\beta\|\Omega\|_0 \\ &+\sum_{i,j}\|\mathbf{B}_{i,j,:,:}-\mathbf{B}^{l*}_{i,j,:}\times_4 I_{i,j}\times_4 T-\mathbf{B}^{h}_{i,j,:,:}\|\}. \\ s.t.\ &rank(\mathbf{B}^{l*}_{vec})=1, \quad \|T\|_2=1, \ \mathbf{D}=P_{\overline{\Omega}}\mathbf{B}+\mathbf{F}. \end{aligned} \tag{18}$$

In the above framework, the newly employed components are the changing tendency T, the detail background layer $\mathbf{B}^h$ and the low-resolution foreground priori knowledge $\mathbf{C}(\Omega)$. To estimate the contribution of each component, some ablation tests are shown in Table I.

In the table, H-CT stands for HMAO without changing tendency (CT) part; H-DB is HMAO without detail background (DB); H-CT-DB means HMAO without hierarchical background (CT and DB); H-LF corresponds to HMAO without hierarchical foreground priori (LF). Actually, changing tendency (CT) is the least effective component. The hierarchical foreground contributes a little more than the hierarchical background with regard to the average performance.

TABLE I
ABLATION TEST OF EACH COMPONENT IN HMAO ON I2R DATASET

Video	HMAO	H-CT	H-DB	H-CT-DB	H-LF
WaterSurface	0.9293	0.8728	0.9257	0.8716	0.9060
Fountain	0.8380	0.8017	0.8343	0.8056	0.7503
Campus	0.8050	0.7770	0.8005	0.7733	0.7372
Curtain	0.8995	0.5751	0.8931	0.6177	0.7480
Hall	0.6830	0.6801	0.6815	0.6260	0.5040
Average	0.8310	0.7413	0.8270	0.7388	0.7291

As to Eq. (18), we note that its objective function is non-convex, which is difficult to solve in general. A more tractable approach is to alternatively solve the two subproblems of estimating BG and FG using Alternating direction multipliers method (ADMM) [36].

1) Estimation of Background: Once foreground ($\mathbf{F}$) is available along with an estimated FG region (Ω), $\mathbf{B}$ can be updated by minimizing the following objective function

$$\min \sum_{i,j} \|\mathbf{B}_{i,j,:,:} - \mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T - \mathbf{B}^{h}_{i,j,:,:}\|$$
$$s.t.\ \ rank(\mathbf{B}^{l*}_{vec}) = 1, \quad \|T\|_2 = 1,\ \mathbf{D} = P_{\overline{\Omega}}\mathbf{B} + \mathbf{F}. \quad (19)$$

or equivalently,

$$\min \sum_{i,j} \|P_{\overline{\Omega}}(\mathbf{D} - \mathbf{F})_{i,j,:,:} - \mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T - \mathbf{B}^{h}_{i,j,:,:}\|$$
$$s.t.\ rank(\mathbf{B}^{l*}_{vec}) = 1, \quad \|T\|_2 = 1. \quad (20)$$

The above problem can be solved in the following three steps:

First, we consider the low-frequency BG as the primary part of the hierarchical model - i.e.,

$$\min \sum_{i,j} \|(\mathbf{D} - \mathbf{F})_{i,j,:,:} - \mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T\| \quad (21)$$

Recall that low-frequency components are decomposed of chief values and changing tendencies, which can then be solved alternatively - i.e., once chief values are estimated, changing tendencies of the averaging background can be updated by

$$\min \sum_{i,j \in \overline{\Omega}} \|\mathbf{T}_{i,j} - T \times I_{i,j}\|, \quad \mathbf{T}_{i,j} \in \mathbb{R}^{f \times 1} \quad (22)$$

where $\mathbf{T}_{i,j} = (\mathbf{D} - \mathbf{F})_{i,j,:,:} \times_3 (\mathbf{B}_{i,j,:})^{-1}$ ("$\times_3$" is the 3−mode product). Then the result is projected onto the subspace such that $\|T\|_2 = 1$. To solve T, Eq. (22) is a standard dictionary learning problem and can be solved by heuristic algorithms. Similarly, solving I is a sparse coding problem with the constraint $\|I_{i,j}\|_1 = 1$.

Second, as to estimate the chief values in low-frequency components, assuming that $V_{i,j} = T \times I_{i,j} \in \mathbb{R}^{f \times 1}$, we have

$$\mathbf{B}^{l*}_{i,j,:} \times_4 I_{i,j} \times_4 T = \mathbf{B}^{l*}_{i,j,:} \times_4 V_{i,j}. \quad (23)$$

Then we can obtain chief value matrices for all GOPs by

$$\min \sum_{i,j \in \overline{\Omega}} \|(\mathbf{D} - \mathbf{F})_{i,j,:,:} - \mathbf{B}^{l*}_{i,j,:} \times_4 V_{i,j}\|, \quad (24)$$

By enforcing the constraint that these matrices should be similar to each other, we compute the global optimal matrix by averaging the estimated BG across non-FG regions - i.e., for each $\mathbf{B}$ ($\mathbf{B}_{(k)}$, $k = 1, \ldots, K$),

$$\mathbf{B}^{l*}_{i,j,:} = \sum_{k:i,j,k \in \overline{\Omega}} \mathbf{B}^{l*}_{(k)i,j,:} / \sum_{k:i,j,k \in \overline{\Omega}} 1. \quad (25)$$

Third, residual of the low-frequency components are calculated by

$$\mathbf{E}^h = \mathbf{D} - \mathbf{F} - \mathbf{B}^l. \quad (26)$$

Then the detail background can be solved by Equation (7).

2) Estimation of Foreground: We first estimate the latent FG label (Ω^{hr}) by solving Eq. (9), which can also be obtained by alternatively solving $\mathbf{B}^{lr}$ and $\mathbf{F}^{lr}$. More specifically, once $\mathbf{B}^{lr}$ is solved, $\mathbf{F}^{lr}$ is given by $\mathbf{F}^{lr} = \mathcal{T}_\mu(\mathbf{D}^{lr}_n - \mathbf{B}^{lr}_n)$ where $\mathcal{T}_\mu$ is the soft-thresholding operator. Then the rank-1 BG can be calculated by $\mathbf{B}^{lr*} = \sum_{n=1}^{N}(\mathbf{D}^{lr}_n - \mathbf{F}^{lr}_n)/N$ and we set $\mathbf{B}^{lr}_n = \mathbf{B}^{lr*}, n = 1, \ldots, N$. Last, Ω^{hr} can be obtained by pixel-wise upsampling the FG region of LR video.

When BG ($\mathbf{B}$) and reconstructed FG (Ω^{hr}) are available, FG detection problem becomes

$$\min_{\mathbf{F},\Omega} \{\gamma\, \mathbf{C}(\Omega) + \beta\|\Omega\|_0 + \sum_{\mathbb{G}} f(\alpha)\|\Omega_{i,j,n} - \Omega_{x,y,z}\|\},$$
$$s.t.\ \mathbf{D} = P_{\overline{\Omega}}\mathbf{B} + \mathbf{F}. \quad (27)$$

The above problem can be reformulated as

$$\min\ \{const + \sum_{i,j,n}(\beta - \|(\mathbf{D}_{i,j,:,n} - \mathbf{B}_{i,j,:,n})\|)\Omega_{i,j,n} + \gamma \sum_{i,j,n} \mathbf{C}(\Omega_{i,j,n}) + \sum_{\mathbb{G}} f(\alpha)\|\Omega_{i,j,n} - \Omega_{x,y,z}\|\}. \quad (28)$$

Now it is easy to find that the objective function is decomposed of two parts - i.e., the constraints for each point $\Omega_{i,j,n}$ and the constraints for arbitrary pair ($\Omega_{i,j,n}$ and $\Omega_{x,y,z}$). Reformulating this problem as a graph function by regarding each point as a node, we can obtain an energy function for the entire FG; accordingly the optimization problem in Eq. (28) is translated into an energy minimization one and can be solved by standard graph cut techniques [53].

3) Algorithm: Putting the above two building blocks together, we obtain a moving object detection algorithm for video based on alternating the estimations of BG and FG. The complete flow-chart of the proposed HMAO algorithm (Algorithm 1) is shown below. It should be noted that unlike existing approaches in the literature, FG estimation is also refined along with BG estimation by exploiting additional information from input video at each iteration as highlighted by the red color in Fig. 1. When both BG and FG are modeled individual hierarchically, we argue that alternating optimization becomes more effective because it has the potential of jointly and successively refining the spatio-temporal estimation of BG/FG in a closed loop. In summary, improved capability of modeling complex video data (e.g., those with dynamic BG) and robustness to noise interference are the key salient features of the proposed HMAO approach (Algorithm 1).

The computational bottleneck of Algorithm 1 lies in solving Eq. (28) by graph cuts. It can be shown that energy minimization via graph cuts has the cost of

Fig. 7. Results of background extraction. From left to right: true background, HMAO, OMoGMF, TVRPCA, GFL, LSD, DECOLOR. From top to bottom: boats, overpass (overp), canoe, fountain (fount), watersurface (water) and winterDriveway (winter).

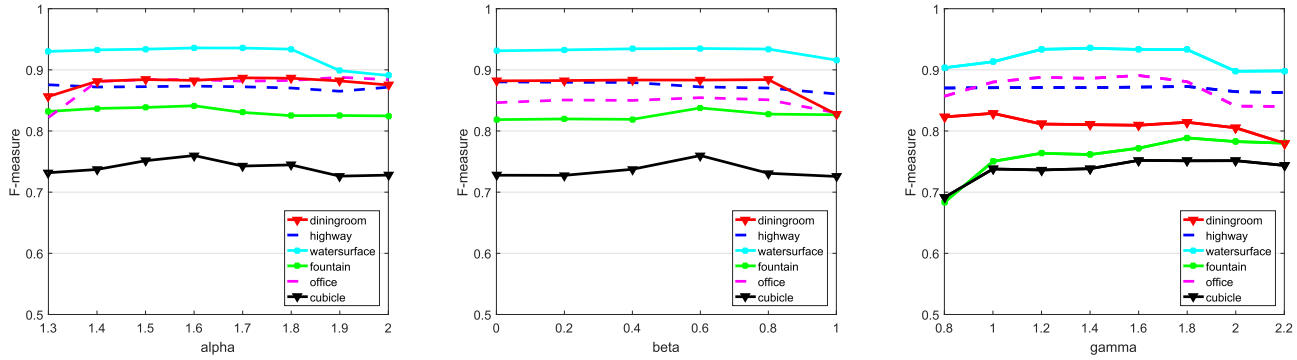


Fig. 8. Effects of the parameters from the Formula (28). (a) α. (b) β. (c) γ.

$(5 \times IJN)(IJN)^2 = O(I^3J^3N^3)$. As reported by the experimental results of the next section, the overall running time of Algorithm 1 is comparable to that of LSD [28] (lower than that of GFL [54] but higher than that of DECOLOR [25]). GPU-based acceleration techniques might lead to more efficient implementation of Algorithm 1 but it is outside the scope of this work.

IV. Experimental Results

In this section, we report our experimental results and compare the proposed algorithm against previous techniques. In our experiments, the following parameter setting has been adopted: the size of GOP is $f = 4$ and $P = 4$. The benchmark datasets include I2R dataset [50] and ChangeDetection dataset 2014 (CDnet) [51].

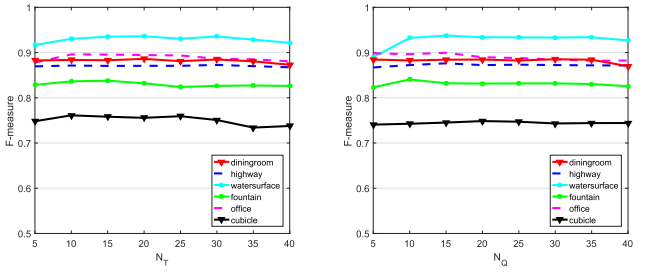
A. Comparison to Model Variants

In order to understand the relative contribution of various parameters of our model, we have conducted an empirical study as follows. The exemplar videos are taken from I2R dataset and CDnet 2014 dataset in our experiment of parameter tuning.

First, we target at the process of dictionary learning designed for encoding changing tendencies of different pixels

Algorithm 1 Algorithm for HMAO

Input: $\mathbf{D} \in \mathbb{R}^{I \times J \times 3 \times N}$.
output: $\mathbf{B}$ and $\mathbf{\Omega}$.
while not converged do $(outer\ loop)$:
Background:
1) **while not converged do** $(inner\ loop\ 1)$:
(1) **changing tendency matrix** (T):
solve T by problem (22),
and project it into the subspace $\|T\|_2 = 1$.
(2) **chief value matrix**:
solve each $\mathbf{B}^{l*}$ by problem (24)
then, by Eq. (25).
end while $(inner\ loop\ 1)$;
2) **low-frequency background**:
$\mathbf{B}^l_{i,j,:,:} = \mathbf{B}^{l*}_{i,j,:} \times_4 I^l_{i,j} \times_4 T$
3) **high-frequency background**:
build codebook (Q) by Eq. (6), then obtain $\mathbf{B}^h$ by Eq. (7).
4) **hierarchical background**:
$\mathbf{B} = \mathbf{B}^l + \mathbf{B}^h$.
Foreground:
5) **low-resolution foreground**
while not converged do $(inner\ loop\ 2)$:
(1) solve low-resolution background by
$\mathbf{B}^{lr*} = \sum_{n=1}^{N}(\mathbf{D}^{lr}_n - \mathbf{F}^{lr}_n)/N$,
and $\mathbf{B}^{lr}_n = \mathbf{B}^{lr*}, n = 1, \ldots, N$.
(2) solve low-resolution foreground by:
$\mathbf{F}^{lr} = \mathcal{T}_\mu(\mathbf{D}^{lr}_n - \mathbf{B}^{lr}_n)$.
end while $(inner\ loop\ 2)$;
6) **Entire foreground**:
solving problem (28) by graph cuts.
end while $(outer\ loop)$.

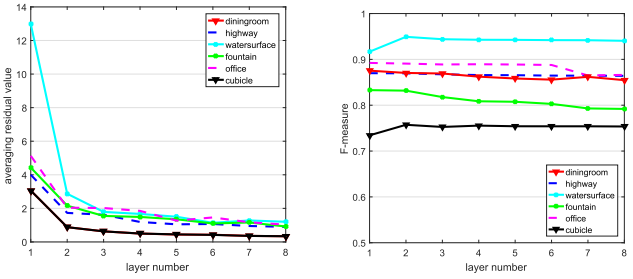


(a) (b)

Fig. 9. Performances of HMAO with different sizes of dictionary and codebook. (a) N_T. (b) N_Q.

in the averaging background (as shown in Fig. 4). In practice, the changing tendencies of different pixels are usually finite. As manifested in Fig. 9 (a), $N_T = 15$ is enough and a larger dictionary size may lead to over-fitting, where some individual changing tendencies are also recorded. Second, the detail background is modeled by clustering the residuals of low-frequency background. The performance of HMAO with different clustering numbers is shown in Fig. 9 (b). The resulting curves resemble those in Fig. 9 (a) and we have found $N_Q = 15$ is large enough.

Second, we have empirical tuned the parameters α, β and γ in formula (28). Specifically, α is the weight for the spatio-temporal Markov random field (MRF) constraint, β means the expectation of the sparsity of the foreground regions and γ reflects the confidence of the obtained low resolution background. As can be found from Fig. 8, too large or too small choices will degrade the performance of HMAO



(a) (b)

Fig. 10. Performances of multi-layers structure. (a) the averaging residual value in each layer. (b) F-measure vs. layer number.

algorithm. Although the optimal values of these parameters vary from video to video, we manage to approximate nearly optimized parameters for most videos, which are given by $\alpha = 1.65$, $\beta = 0.6$ and $\gamma = 1.6$.

Third, we have studied multi-layer structure for the detail background. By iteratively clustering the residuals, we can solve each $\mathbf{B}^{li}, i = 1, 2, 3, \ldots$. The corresponding results are given in Fig. 10. Figure (a) represents the magnitude of values in different layers. It's obvious that the data in the first layer are far more notable than those from other layers; the second layer is still noticeably higher than the rest. In Figure (b), the best performance is observed for using only 1 detail layer on 4 videos (highway, diningroom, fountain, office); for two other videos (watersurface and cubicle) whose background are dynamic, some information of the background still exist in the second detail layer. Therefore, we can opt to use at most 2 detail layers (for handling video with dynamic background) in practical tasks once we have prior knowledge about the video. In our experiments, the detail layer number is set as 1.

B. Comparison on Background Estimation

In this section, we will report the comparison result of BG extraction between this work and some recently proposed methods - i.e., OMoGMF (OMoG) [12], TVR-PCA (TVRP) [34], GFL [54], DECOLOR (DECO) [25] and LSD [28]. The parameters of the benchmark algorithms are the default values that accompany the release of their source codes. In this comparative study, we have focused on test videos are those with dynamic background in CDnet and I2R because they are more challenging.

It can be seen from Fig. 7 that most algorithms suffer from the weakness of missing a significant portion of the details in the extracted background. For example, on 'boats' and 'canoe' datasets, the ripples on the water are often treated as outliers and accordingly misclassified as FG; on 'overpass' dataset, details of the recovered waving leaves are blurry due to the limitations of background models. By contrast, HMAO effectively distinguishes regular or self-repeating details in the background from noises and therefore produces the most discriminative backgrounds. The extracted background of HMAO is the closest to the ground truth (as shown in the left column in Fig. 7) because the proposed two-layer hierarchical

Fig. 11. Results of background extraction on SBMnet dataset.

TABLE II
RMSE OF THE EXTRACTED BACKGROUNDS SHOWN IN FIG. 7

Video	HMAO	OMoG	TVRP	GFL	LSD	DECO
boats	**0.0515**	0.1042	0.0580	0.1196	0.0996	0.0813
canoe	**0.0421**	0.1228	0.0879	0.1050	0.2568	0.1978
fount	0.0571	0.0438	0.0314	0.0720	**0.0202**	0.0297
overp	**0.0289**	0.0591	0.0312	0.0350	0.0591	0.0479
water	**0.0265**	0.0766	0.0371	0.0419	0.2884	0.2577
winte	0.0416	0.0444	**0.0311**	0.0344	0.0399	0.0619
Average	**0.0413**	0.0751	0.0461	0.0680	0.1258	0.1127

modeling of background more faithfully characterizes various uncertainty sources for video containing dynamic background.

To objectively evaluate the accuracy of background extraction, we have compared the difference between the extracted background ($\mathbf{B}$) and the groundtruth ($\mathbf{B}^*$) as measured by Root Mean Square Error (RMSE) - i.e., RMSE $= \|\mathbf{B} - \mathbf{B}^*\|_F / \|\mathbf{B}^*\|_F$, as shown in Table II. It can be observed that HMAO achieves the lowest RMSE on the average (four out of six). Even though the advantages of HMAO are obvious for the class of video containing dynamic background, HMAO does still have weakness when dealing with some real datasets - e.g., 'fountain' dataset. In HMAO, we model the detail background by clustering the pixel-wise short-term residuals because we assume that the textures of the background result in certain regular residual value series. Unfortunately this assumption fails to model the pathological case of a fountain which produces irregular (more like stochastic) residual values.

Additionally, we have conducted the comparison on some complex scenes under cluttered background- i.e., the clutter category from the Scene Background Modeling.Net (SBMnet) dataset.[4] The results of BG extraction are shown in Fig. 11. We can find that the problem of BG extraction becomes more challenging when BG pixels are less visible than FG ones. In some extreme cases, all competing algorithms fail to extract the background - e.g., the board in the first column and the car in the forth column. However, we can still observe that, relatively speaking, HMAO noticeably outperform others on this difficult dataset- e.g., more revealed areas on the board in the first column, the disclosure of the car/chair in the fourth/sixth column.

[4]http://www.scenebackgroundmodeling.net/

C. Comparison on Foreground Estimation

In this section, we will show how the supervised information extracted from low-resolution video helps the robustness to noise in foreground estimation. In our experiments, by carefully choosing parameters, we can obtain the most confident regions of foreground objects from LR video (Column 4 in Fig. 12 and Fig. 13), which is fairly robust to the noise. Direct foreground estimation results (Column 3 in Fig. 12 and Fig. 13) obtained by background subtraction can find some part of foreground objects, but a significant portion of noise components are misclassified as the background. Additionally, one can observe that some noise components are caused by the inaccuracy of background models - e.g., the trunk region of the tree in Fig. 12. Therefore it is natural to improve the estimation results by combining the strengths of those two approaches (Column 3 and 4). In our hierarchical foreground model, weights for some foreground regions are strengthened based on the supervised information (passed from LR), while those for the background are weakened accordingly. Thanks to the propagation of FG estimation from LR to HR in a hierarchical fashion, almost all neighboring

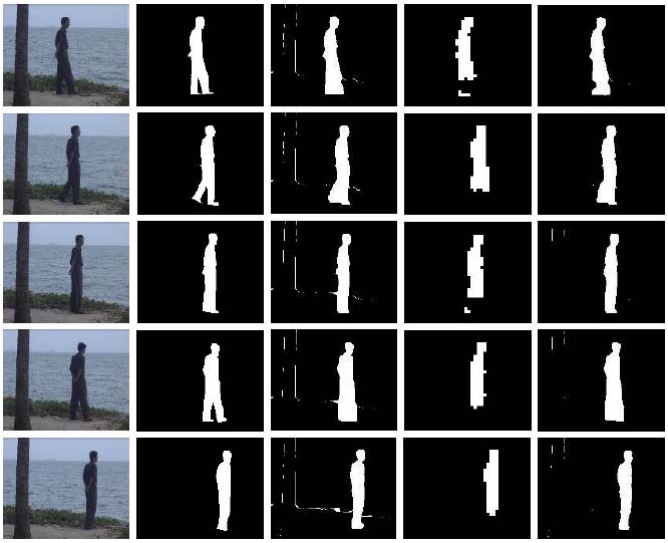


Fig. 12. Foreground estimation results on WaterSurface dataset. From left to right: input image frame, groundtruth, foreground estimation without supervised information, detected foreground in low-resolution video, final results of HMAO.

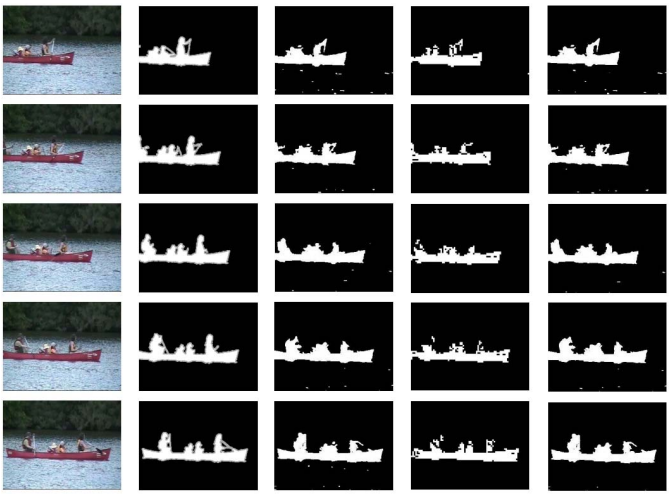


Fig. 13. Foreground estimation results on Canoe dataset. From left to right: input image frame, groundtruth, foreground estimation without supervised information, detected foreground in low-resolution video, final results of HMAO.

foreground regions are successfully detected and noise interference are suppressed almost completely.

D. Comparisons With Other Algorithms

In our study, we have compared the proposed HMAO with seven current state-of-the-art algorithms whose codes have been made publicly available - i.e., TVRPCA[5] [34], GFL[6] [54], DECOLOR[7] [25], LSD[8] [28], PCP [44], OMoGMF[9] [12] and SOIR [26]. The parameter settings of the algorithms are the default settings or are optimized following the suggestions discussed in the corresponding papers. Our extensive comparison results have been organized into the following five subsections.

1) Short-Term Moving Object Detection: Our comparisons of short-term moving object detection are conducted on the I2R dataset, which contains 9 videos. For each video, the ground-truth (manually-segmented foreground regions) of 20 frames are provided in the dataset. In our experiments, we have used these 20 frames with ground truth to test the short-term performance of the proposed moving object detection algorithm. The sequences and detection results are shown in Fig. 14. For each video, we have randomly chosen 1 from 20 test frames to compare the detection results of all competing algorithms. As can be seen from the figure, the difficulties of foreground detection are mainly caused by the interference of unwanted noise and the blurring of object contours. It can be observed that GFL, OMoGMF, LSD and PCP have misclassified some noises as FG regions. These four algorithms can find the approximate outlines of foreground objects but all miss some salient parts inside the objects. By contrast, DECOLOR, HMAO, SOIR and TVRPCA have shown better performances due to their robust and accurate BG models. However, DECOLOR can not perform well in finding detailed object contours due to its strong MRF prior for FG; while TVRPCA tends to break the object boundary (the opposite to DECOLOR). When compared with the ground truth, only HMAO produces the most satisfying results combining the strengths of DECOLOR and TVRPCA.

To quantitatively evaluate the performances of the different algorithms, we have computed the F-measure, which is derived from the precision and recall and defined by

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{29}$$

The detection results in terms of Recall (R), Precision (P) and F-measure (F) are given in Table III. In addition to the mentioned algorithms, TLSFSD [29] whose F-measure results are available in the paper is also included here. We can see from the results that HMAO clearly shows advantages on some videos, especially those with dynamic backgrounds - e.g., 'Campus', 'Curtain' and 'WaterSurface'. Although the backgrounds of 'Bootstrap' and 'ShoppingMall' datasets are not dynamic, steadily moving pedestrians in the video lead to irregularly changing illumination in the scene, which make these video resemble those with dynamic background. Not surprisingly HMAO also achieves satisfying performances on these videos. For sequences 'Hall' and 'Fountain', HMAO produces highly comparable performance to the competing ones.

2) Long-Term Moving Object Detection: Our comparisons of long-term moving object detection are performed on the 6 dynamic background videos from CDnet 2014 dataset. In this dataset, hand-segmented foreground regions of all video frames are provided. In our experiments, a video sequence

[5]http://yangliang.github.io/code/TVRPCA.rar

[6]http://idm.pku.edu.cn/staff/wangyizhou/code/code_bs_cvpr15.rar

[7]https://fling.seas.upenn.edu/∼xiaowz/dynamic/wordpress/my-uploads/codes/decolor.zip

[8]http://www.ee.oulu.fi/∼xliu/research/lsd/LSD.zip

[9]http://gr.xjtu.edu.cn/c/document_library/get_file?folderId=2456216&name=DLFE-97966.zip

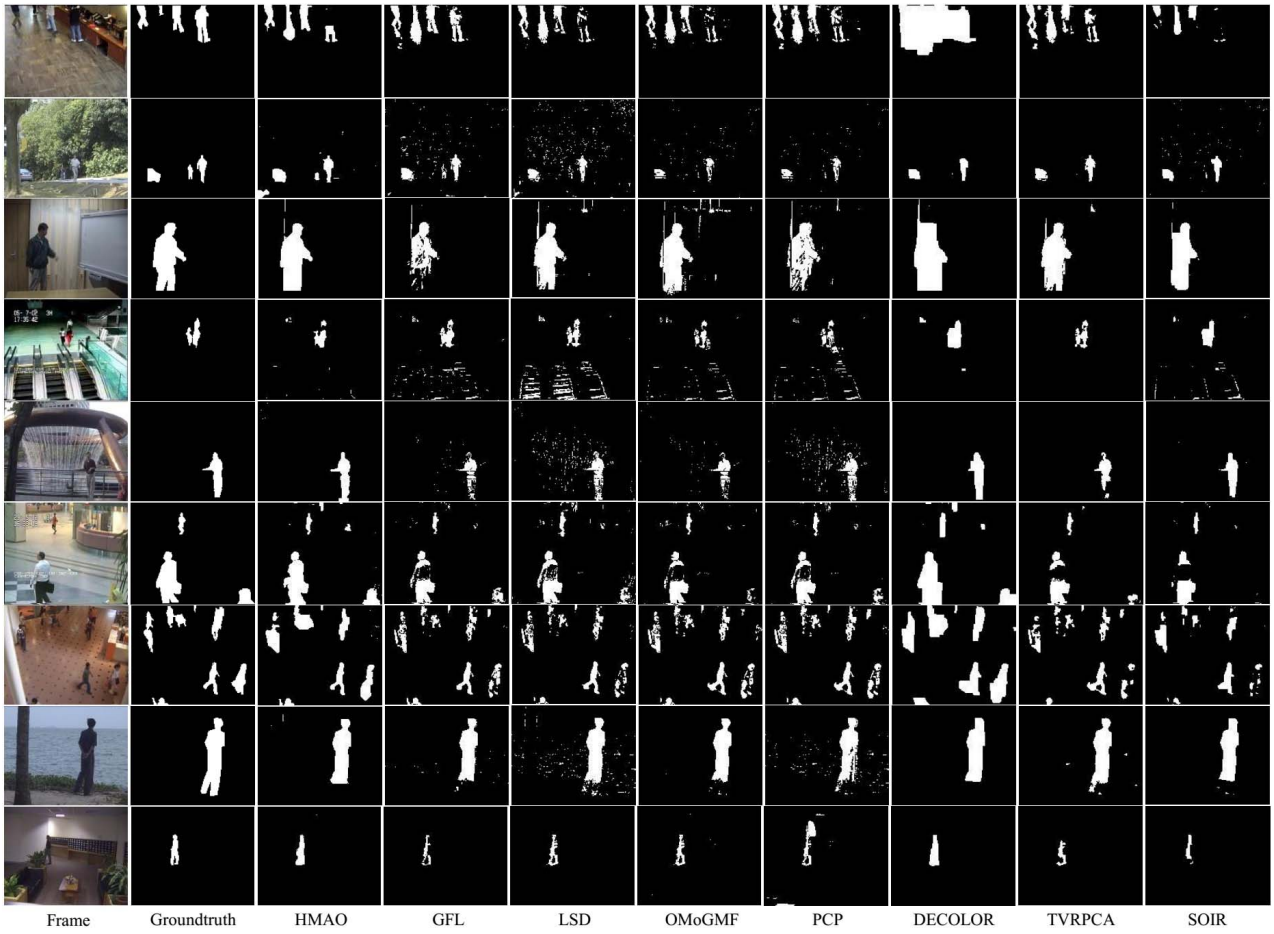


Fig. 14. Results of foreground detection on I2R dataset. The selected frames (top to bottom) are b02514 (Bootstrap), trees1831 (Campus), Curtain22847 (Curtain), Escalator3585 (Escalator), Fountain1494 (Fountain), airport2180 (Hall), ShoppingMall1606 (ShoppingMall), WaterSurface1577 (WaterSurface) and SwitchLight2019 (Lobby).

TABLE III

COMPARISON OF THE FOREGROUND DETECTION RESULTS IN TERMS OF THE F-MEASURE ON I2R DATASET

video	HMAO		GFL		LSD		OMoGMF		PCP		DECOLOR		TVRPCA		TLSFSD[10]		SOIR	
	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F
Bootstrap	0.76 0.68	**0.72**	0.81 0.65	**0.72**	0.81 0.54	0.65	0.69 0.62	0.65	0.65 0.63	0.64	0.42 0.93	0.58	0.75 0.53	0.62	0.45 0.67	0.53	0.86 0.50	0.63
Campus	0.75 0.90	**0.81**	0.53 0.75	0.62	0.45 0.58	0.51	0.54 0.36	0.43	0.54 0.37	0.44	0.91 0.68	0.77	0.91 0.56	0.70	0.87 0.73	0.79	0.33 0.84	0.47
Curtain	0.87 0.93	**0.90**	0.87 0.45	0.59	0.84 0.79	0.81	0.85 0.84	0.84	0.74 0.65	0.69	0.65 0.98	0.78	0.85 0.72	0.78	0.76 0.91	0.83	0.91 0.78	0.84
Escalator	0.58 0.66	0.62	0.54 0.76	0.63	0.41 0.66	0.51	0.62 0.54	0.57	0.56 0.59	0.57	0.60 0.93	0.73	0.82 0.68	**0.74**	0.68 0.74	0.71	0.72 0.67	0.70
Fountain	0.79 0.89	**0.84**	0.87 0.65	0.74	0.59 0.77	0.67	0.85 0.60	0.70	0.63 0.73	0.68	0.76 0.92	0.83	0.87 0.71	0.78	0.85 0.84	**0.84**	0.87 0.79	0.83
Hall	0.72 0.63	**0.68**	0.80 0.51	0.63	0.70 0.52	0.60	0.72 0.63	0.67	0.52 0.53	0.52	0.56 0.75	0.64	0.83 0.49	0.61	0.55 0.74	0.63	0.77 0.55	0.64
ShopMall	0.66 0.75	0.71	0.82 0.63	0.71	0.80 0.58	0.67	0.76 0.65	0.70	0.74 0.65	0.69	0.52 0.94	0.67	0.73 0.59	0.65	0.75 0.72	**0.74**	0.82 0.57	0.67
Watsface	0.94 0.92	**0.93**	0.96 0.76	0.85	0.86 0.89	0.88	0.98 0.78	0.86	0.80 0.76	0.78	0.95 0.76	0.84	0.93 0.86	0.89	0.81 0.99	0.89	0.95 0.79	0.86
Lobby	0.72 0.84	0.79	0.95 0.40	0.56	0.82 0.72	0.77	0.79 0.71	0.75	0.80 0.55	0.65	0.78 0.51	0.61	0.89 0.42	0.57	0.89 0.91	**0.90**	0.93 0.47	0.62
Average[11]	**0.78**		0.67		0.67		0.69		0.62		0.71		0.70		0.76		0.70	

[10] The results of TLSFSD are from [29].
[11] The average of the results in terms of F-measure.

composed of 220 continuous frames are selected for long-term detection performance evaluation. The sequences and comparison results are shown in Fig. 15. One can observe that modeling foreground objects of these videos is much more challenging. The most difficult task for foreground detection is 'fountain01', where the foreground regions are really small

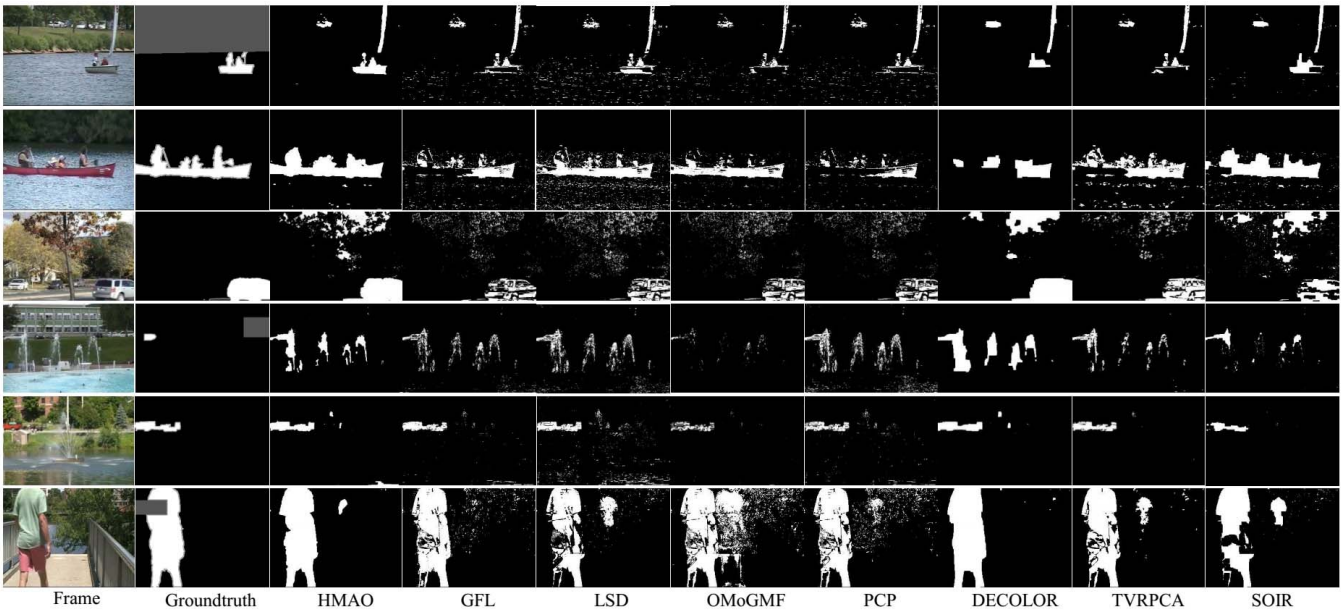


Fig. 15. Results of background extraction on CDnet dataset. The selected frames (top to bottom) are in007869 (boats), in000951 (canoe), in002067 (fall), in001164 (fountain01), in000750 (fountain02), in2371 (overpass).

TABLE IV
COMPARISON OF THE FOREGROUND DETECTION RESULTS IN TERMS OF THE F-MEASURE ON CDNET2014 DATASET

video	HMAO		GFL		LSD		OMoGMF		PCP		DECOLOR		TVRPCA		SOIR	
	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F	R P	F
boats	0.55 0.72	**0.62**	0.34 0.37	0.35	0.54 0.69	0.60	0.27 0.36	0.31	0.29 0.31	0.30	0.40 0.31	0.35	0.39 0.31	0.35	0.53 0.53	0.53
canoe	0.81 0.83	**0.82**	0.54 0.38	0.44	0.57 0.74	0.65	0.81 0.57	0.67	0.54 0.24	0.33	0.96 0.47	0.63	0.55 0.61	0.58	0.33 0.68	0.45
fall	0.33 0.80	0.47	0.61 0.35	0.43	0.36 0.66	0.46	0.30 0.53	0.38	0.37 0.48	0.41	0.26 0.93	0.41	0.40 0.72	**0.51**	0.28 0.59	0.38
fountain01	0.04 0.71	0.08	0.05 0.54	0.09	0.04 0.64	0.08	0.11 0.19	**0.14**	0.03 0.60	0.06	0.03 0.82	0.05	0.06 0.58	0.11	0.08 0.52	0.13
fountain02	0.72 0.86	**0.78**	0.67 0.68	0.67	0.51 0.77	0.62	0.89 0.37	0.52	0.51 0.57	0.54	0.71 0.76	0.73	0.89 0.58	0.71	0.73 0.45	0.55
overpass	0.96 0.76	**0.85**	0.83 0.69	0.75	0.82 0.47	0.60	0.51 0.66	0.58	0.73 0.66	0.69	0.80 0.76	0.79	0.86 0.69	0.77	0.87 0.57	0.69
Average	**0.60**		0.46		0.50		0.43		0.39		0.49		0.51		0.46	

while the area of the fountain is large. For this challenging sequence, all algorithms fail to return the correct foreground regions; for other videos, the challenge is less severe and the foreground objects can be approximately detected by most algorithms. However, heavy noises in the videos are still the major difficulties for most algorithms. Overall, HMAO performs well on most datasets, except misclassifying the fountain (fountain01) and leaves (fall) as FG objects.

Objective performance evaluation in terms of Recall (R), Precision (P) and F-measure (F) is given in Table IV. It is easy to see that the advantage of HMAO is obvious for videos containing dynamic background - e.g., 'boats', 'canoe' and 'overpass'. In those videos, the details of dynamic background are approximately self-repeating; therefore our hierarchical background model produces more detail components in BG extraction and more accurate FG regions accordingly. For 'fountain02' sequence, the performance of HMAO is at least comparable to that of other algorithms; while for 'fall' and 'fountain01' sequences, HMAO is slightly inferior to TVRPCA and OMoGMF. As discussed above, these two sequences are the cases where our hierarchical model fails (the dynamic motion in background is less regular). Nevertheless, we note that the average performance of HMAO is still noticeably better than all other competing algorithms.

3) Comparison With Online Algorithms: In this section, we provide more comparison against online methods such as GRASTA [55] and incPCP [56]. The suggested frame number for warm-start of OMoGMF, GRASTA and incPCP are 30, 100 and 1, respectively. Objective comparison results in terms of F-measure are shown in Table V.

As is shown in the table, the suggested numbers of frames for warm-start are usually effective enough. Especially in incPCP, which only requires 1 frame for warm-start and related parameters are given accordingly, more frames results in worse performances. Then, the performances of incPCP are actually influenced by the selected initialization frame.

TABLE V
COMPARISON OF THE FOREGROUND DETECTION RESULTS IN TERMS OF THE AVERAGE F-MEASURE

video	HMAO	OMoGMF				GRASTA				incPCP			
		10	30	50	70	70	100	130	160	1	3	10	20
Bootstrap	0.73	0.64	0.65	0.64	0.65	0.81	**0.82**	**0.82**	**0.82**	0.66	0.61	0.53	0.50
Campus	**0.82**	0.41	0.43	0.43	0.42	0.32	0.30	0.30	0.30	0.43	0.47	0.39	0.37
Curtain	**0.89**	0.85	0.84	0.82	0.70	0.54	0.68	0.75	0.76	0.74	0.55	0.52	0.43
Escalator	**0.67**	0.55	0.57	0.57	0.56	0.50	0.55	0.56	0.56	0.61	0.55	0.52	0.48
Fountain	**0.83**	0.69	0.70	0.69	0.71	0.53	0.68	0.69	0.68	0.71	0.60	0.41	0.30
Hall	**0.67**	**0.67**	**0.67**	**0.67**	**0.67**	0.51	0.60	0.61	0.60	0.59	0.53	0.45	0.43
ShopMall	**0.75**	0.70	0.70	0.70	0.70	0.46	0.66	0.67	0.67	**0.75**	0.70	0.54	0.38
Watsface	**0.93**	0.86	0.86	0.86	0.85	0.81	0.82	0.82	0.83	0.79	0.71	0.66	0.60
Lobby	0.80	0.76	0.75	0.73	0.72	0.28	0.50	0.49	0.49	**0.85**	0.45	0.30	0.29
boats	**0.64**	0.29	0.31	0.32	0.30	0.23	0.28	0.28	0.28	0.40	0.31	0.21	0.15
canoe	**0.85**	0.54	0.67	0.60	0.62	0.33	0.39	0.33	0.33	0.55	0.29	0.24	0.21
fall	**0.47**	0.37	0.38	0.38	0.39	0.28	0.34	0.34	0.34	0.36	0.38	0.35	0.36
fountain01	0.08	0.13	0.14	0.14	0.14	0.08	0.10	0.09	0.09	**0.15**	**0.15**	0.11	0.09
fountain02	**0.75**	0.52	0.52	0.52	0.52	0.33	0.52	0.51	0.50	0.65	0.57	0.47	0.44
overpass	**0.85**	0.58	0.58	0.56	0.57	0.51	0.66	0.67	0.66	0.65	0.57	0.58	0.55
Average	**0.71**	0.57	0.58	0.58	0.57	0.43	0.53	0.53	0.53	0.59	0.50	0.42	0.37

TABLE VI
COMPARISON OF THE FOREGROUND DETECTION RESULTS IN TERMS OF THE AVERAGE F-MEASURE ON CDNET 2014 DATASET

Category	HMAO	GFL	LSD	OMoGMF	PCP	DECOLOR	TVRPCA	SOIR
badWeather	0.79	0.74	0.85	0.72	0.68	0.76	**0.86**	0.80
baseline	**0.82**	0.75	0.76	0.67	0.64	0.76	0.74	0.69
cameraJitter	0.63	0.75	0.63	0.57	0.65	**0.78**	**0.78**	0.56
intermitOM	**0.72**	0.68	0.67	0.65	0.61	0.67	0.67	0.60
lowFramerate	**0.60**	0.59	0.57	0.58	0.50	0.47	0.33	0.36
nightVideos	0.36	**0.44**	0.43	0.42	0.37	0.39	**0.44**	0.37
shadow	**0.86**	0.76	0.75	0.70	0.71	**0.86**	0.75	0.73
thermal	**0.84**	0.52	0.50	0.75	0.72	0.64	0.79	0.59
turbulence	0.46	0.36	0.29	**0.52**	0.26	0.44	0.39	0.51
Average	**0.68**	0.62	0.61	0.62	0.57	0.64	0.64	0.58

TABLE VII
TIME CONSUMPTION OF THE ALGORITHMS (THE UNIT IS SECOND)

HMAO	GFL	LSD	OMoGMF	PCP	DECOLOR	TVRPCA	SOIR
6547.0	12289.7	5264.3	19.4	303.4	164.9	1898.4	60.7

In all the employed videos, the first frame usually contains no foreground object, which just meets the demand of incPCP. In OMoGMF and GRASTA, the performances of the algorithms are stable when enough number of frames for warm-start is arranged. Therefore, although online algorithms are more time-saving than HMAO, they are less effective than HMAO.

4) Comparisons for Other Categories: Furthermore, we have conducted the comparisons on all other categories in CDnet 2014 dataset (the only exclusion is PTZ category because it's not our target to model the videos with different kinds of zooms). Similarly, we still select 220 continuous frames from each video for evaluation. The average F-measures for each video and benchmark method are reported in Table VI. As can be observed from the table, HMAO still outperforms the rest in most categories - e.g., 'baseline', 'intermittentObjectMotion' and 'thermal'. Meanwhile, there are still some cases that HMAO fail to work effectively. For example, HMAO is not robust to camera jitter by natural, because no specialized component is designed for jittering camera motion. Besides, HMAO perform the worst on 'nightVideos', which is the case that our assumption (the principal component of the residuals as the detail BG) fail to work. Instead, residuals are mainly composed by the constantly changing illumination and shadows. Eventually, we can find that, in terms of the overall average F-measure, HMAO has achieved superior performance to all the other competing approaches.

5) Running Time Comparison: Finally, we report the running time comparison for each method, which is shown in Table VII. Here, the experimental results are obtained by averaging the running times of all 9 videos in I2R dataset.

We can see that OMoGMF is the fastest and GFL is the slowest; by contrast, HMAO, TVRPCA and LSD have similar complexity, which is only slightly higher than that of conventional PCA-based algorithms (PCP and DECOLOR).

V. Conclusions

In this paper, we have proposed a joint optimization model with hierarchical background and hierarchical foreground estimation. In the proposed model, hierarchical background and foreground models are developed targeting at better incorporating our a priori knowledge about those two layers. Experimental results have shown that our framework reflects the natural data organization in video containing dynamic background and achieves comparable and often superior performances to current state-of-the-art techniques.

In view of the rapid advances in the field of deep learning and deep neural networks, one cannot help wondering if data-driven (learning-based) approaches will outperform model-based approaches including this work. Deep learning for FG/BG separation is still at its infancy and there are several technical challenges (e.g., training data, computational burden and memory requirement) to overcome. Thus, a feasible approach is to employ some middle ground - i.e., hybrid approaches of combining both model-based and learning-based ones, which will be explored as our follow-up work.

Appendix

In this Appendix, we provide some background material related to the rigorous definition of 4-mode product. More general definition and details of n-mode product can be referred to Sec. 2.5 in [52].

The n-mode product of an arbitrary tensor $\mathbf{X} \in \mathbb{R}^{K_1 \times \cdots \times K_N}$ with a given matrix $\widetilde{\mathbf{T}} \in \mathbb{R}^{J \times K_n}$ is of size $K_1 \times \cdots \times K_{n-1} \times J \times K_{n+1} \times \cdots \times K_N$. Element-wise,

$$(\mathbf{X} \times_n \widetilde{\mathbf{T}})_{k_1 \cdots k_{n-1} j k_{n+1} \cdots k_N} = \sum_{k_n=1}^{K_n} \mathbf{X}_{k_1 \cdots k_{n-1} k_{n+1} \cdots k_N} \widetilde{\mathbf{T}}_{j k_n}, \quad (30)$$

where $k_n \in [1, \cdots, K_n]$, $j \in [1, \cdots, J]$.

Specifically, we have $\mathbf{B}^{l*} \in \mathbb{R}^{I \times J \times 3 \times 1}$ and $\mathbf{T} \in \mathbb{R}^{N \times 1}$. Then the 4-mode product of $\mathbf{B}^{l*}$ and $\mathbf{T}$ is of size $I \times J \times 3 \times N$. Assuming $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N]$, we have

$$(\mathbf{B}^{l*} \times_4 \mathbf{T})_{k_1 k_2 k_3 n} = \mathbf{B}^{l*}_{k_1 k_2 k_3} \mathbf{T}_n, \quad (31)$$

where $n \in [1, \cdots, N]$.

References

[1] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1305–1312.

[2] 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, 2015.

[3] L. Yang, H. Cheng, J. Su, and X. Chen, "Pixel-to-model background modeling in crowded scenes," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[4] T. Bouwmans and E. H. Zahzah, "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance," *Comput. Vis. Image Understand.*, vol. 122, pp. 22–34, May 2014.

[5] 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.

[6] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, "Saliency-aware nonparametric foreground annotation based on weakly labeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1253–1265, Jun. 2016.

[7] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 1999, pp. 246–252.

[8] H.-H. Lin, J.-H. Chuang, and T.-L. Liu, "Regularized background adaptation: A novel learning rate control scheme for Gaussian mixture modeling," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 822–836, Mar. 2011.

[9] J. K. Suhr, H. G. Jung, G. Li, and J. Kim, "Mixture of Gaussians-based background subtraction for Bayer-pattern image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 365–370, Mar. 2011.

[10] M. Chen, X. Wei, Q. Yang, Q. Li, G. Wang, and M.-H. Yang, "Spatiotemporal GMM for background subtraction with superpixel hierarchy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1518–1525, Jun. 2018.

[11] C. R. Wern, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.

[12] H. Yong, D. Meng, W. Zuo, 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.

[13] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.

[14] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Background modeling and subtraction by codebook construction," in *Proc. Int. Conf. Image Process.*, vol. 5, Oct. 2004, pp. 3061–3064.

[15] 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.

[16] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. ECCV*, 2000, pp. 751–767.

[17] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Universal background subtraction using word consensus models," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4768–4781, Oct. 2016.

[18] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, 2006.

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

[20] 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.

[21] A. Zaharescu and M. Jamieson, "Multi-scale multi-feature codebook-based background subtraction," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 1753–1760.

[22] A. Schick, M. Bäuml, and R. Stiefelhagen, "Improving foreground segmentations with probabilistic superpixel Markov random fields," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 27–31.

[23] B.-H. Do and S.-C. Huang, "Dynamic background modeling based on radial basis function neural networks for moving object detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–4.

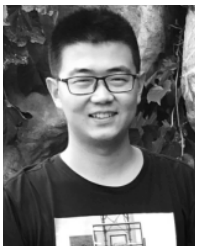
[24] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

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

[26] L. Li, P. Wang, Q. Hu, and S. Cai, "Efficient background modeling based on sparse representation and outlier iterative removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 2, pp. 278–289, Feb. 2016.

[27] 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.

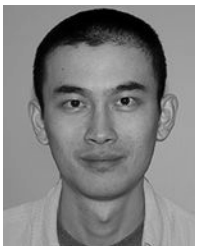
[28] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.
[29] W. Hu, Y. Yang, W. Zhang, and Y. Xie, "Moving object detection using tensor-based low-rank and saliently fused-sparse decomposition," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 724–737, Feb. 2017.
[30] C. Zhao, X. Wang, and W.-K. Cham, "Background subtraction via robust dictionary learning," *EURASIP J. Image Video Process.*, vol. 2011, no. 1, p. 972961, 2011.
[31] X. Liu *et al.*, "Background subtraction using spatio-temporal group sparsity recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1737–1751, Aug. 2018.
[32] 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.
[33] M. Piccardi, "Background subtraction techniques: A review," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 4, Oct. 2004, pp. 3099–3104.
[34] X. Cao, L. Yang, and X. Guo, "Total variation regularized RPCA for irregularly moving object detection under dynamic background," *IEEE Trans. Cybern.*, vol. 46, no. 4, pp. 1014–1027, Apr. 2016.
[35] M. Heikkila and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, Apr. 2006.
[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[37] M. Van Droogenbroeck and O. Paquot, "Background subtraction: Experiments and improvements for ViBe," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 32–37.
[38] 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*, Jun. 2012, pp. 38–43.
[39] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern Recognit.*, vol. 40, no. 3, pp. 1091–1105, Mar. 2007.
[40] J. D. Romero, M. J. Lado, and A. J. Méndez, "A background modeling and foreground detection algorithm using scaling coefficients defined with a color model called lightness-red-green-blue," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1243–1258, Mar. 2017.
[41] S. Yoshinaga, A. Shimada, H. Nagahara, and R.-I. Taniguchi, "Object detection using local difference patterns," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 216–227.
[42] G. Ramírez-Alonso and M. I. Chacón-Murguía, "Auto-adaptive parallel SOM architecture with a modular analysis for dynamic object segmentation in videos," *Neurocomputing*, vol. 175, pp. 990–1000, Jan. 2016.
[43] B. N. Subudhi, S. Ghosh, S. C. K. Shiu, and A. Ghosh, "Statistical feature bag based background subtraction for local change detection," *Inf. Sci.*, vol. 366, pp. 31–47, Oct. 2016.
[44] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, May 2011.
[45] B. Wohlberg, R. Chartrand, and J. Theiler, "Local principal component pursuit for nonlinear datasets," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2012, pp. 3925–3928.
[46] J. Huang, X. Huang, and D. Metaxas, "Learning with dynamic group sparsity," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 64–71.
[47] P. Dong, S. Wang, Y. Xia, D. Liang, and D. D. Feng, "Foreground detection with simultaneous dictionary learning and historical pixel maintenance," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5035–5049, Nov. 2016.
[48] 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. Workshops*, Dec. 2015, pp. 90–98.
[49] S. Javed, S. H. Oh, T. Bouwmans, and S.-K. Jung, "Robust background subtraction to global illumination changes via multiple features-based online robust principal components analysis with Markov random field," *J. Electron. Imag.*, vol. 24, no. 4, p. 043011, 2015.
[50] 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.
[51] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–8.
[52] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.
[53] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[54] 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.*, Jun. 2015, pp. 4676–4684.
[55] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1568–1575.
[56] P. Rodriguez and B. Wohlberg, "Incremental principal component pursuit for video background modeling," *J. Math. Imag. Vis.*, vol. 55, no. 1, pp. 1–18, 2016.



Linhao Li received the B.S. degree in applied mathematics and the M.S. degree in computational mathematics from Tianjin University in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology. His research interests focus on quantization and hashing learning, sparse signal recovery, background modeling, and foreground detection.

Qinghua Hu received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. He was a Post-Doctoral Fellow with the Department of Computing, Hong Kong Polytechnic University, from 2009 to 2011.

He is currently the Dean of the School of Artificial Intelligence, the Vice Chairman of the Tianjin Branch of China Computer Federation, the Vice Director of the SIG Granular Computing and Knowledge Discovery, and the Chinese Association of Artificial Intelligence. He is currently supported by the Key Program, National Natural Science Foundation of China. He has published over 200 peer-reviewed papers. His current research is focused on uncertainty modeling in big data, machine learning with multi-modality data, intelligent unmanned systems. He is an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Acta Automatica Sinica*, and *Energies*.



Xin Li received the B.S. degree (Hons.) in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 1996 and 2000, respectively.

He was a member of Technical Staff with Sharp Laboratories of America, Camas, WA, USA, from 2000 to 2002. Since 2003, he has been a Faculty Member with the Lane Department of Computer Science and Electrical Engineering. His current research interests include image and video coding and processing. He is currently a member of the Image, Video, and Multidimensional Signal Processing Technical Committee.