

# Spatial-Temporal Gaussian Scale Mixture Modeling for Foreground Estimation

Qian Ning,<sup>1,2</sup> Weisheng Dong,<sup>1,2\*</sup> Fangfang Wu,<sup>2</sup> Jinjian Wu,<sup>2</sup> Jie Lin,<sup>2</sup> Guangming Shi<sup>2</sup>

<sup>1</sup>State Key Laboratory on Integrated Services Networks (ISN)

<sup>2</sup>School of Artificial Intelligence, Xidian University, Xi'an 710071, China

## Abstract

Subtracting the backgrounds from the video frames is an important step for many video analysis applications. Assuming that the backgrounds are low-rank and the foregrounds are sparse, the robust principle component analysis (RPCA)-based methods have shown promising results. However, the RPCA-based methods suffered from the scale issue, i.e., the  $\ell_1$ -sparsity regularizer fails to model the varying sparsity of the moving objects. While several efforts have been made to address this issue with advanced sparse models, previous methods cannot fully exploit the spatial-temporal correlations among the foregrounds. In this paper, we proposed a novel spatial-temporal Gaussian scale mixture (STGSM) model for foreground estimation. In the proposed STGSM model, a temporal consistent constraint is imposed over the estimated foregrounds through nonzero-means Gaussian models. Specifically, the estimates of the foregrounds obtained in the previous frame are used as the prior for these of the current frame, and nonzero means Gaussian scale mixture models (GSM) are developed. To better characterize the temporal correlations, the optical flow has been used to model the correspondences between foreground pixels in adjacent frames. The spatial correlations have also been exploited by considering that local correlated pixels should be characterized by the same STGSM model, leading to further performance improvements. Experimental results on real video datasets show that the proposed method performs comparably or even better than current state-of-the-art background subtraction methods.

## 1 Introduction

Estimating the foregrounds in a video sequence is an important step for many video analysis applications, e.g., video surveillance (Liu, Wu, and Lin 2015), vehicle navigation (Unzueta et al. 2012), and object detection (Zhou, Yang, and Yu 2013). As a fundamental research problem, foreground estimation (also called background subtraction) has attracted increasing attentions in the past decade. Due to the dynamic backgrounds, camera motion, illumination change, and irregular objects motion issues, separating video frames into the foreground and the background components is rather challenging.

Numerous background subtraction methods have been proposed. Early strategies for this task tried to classify the background and foreground pixels by learning pixel classification models, such as the Gaussian Mixture Models (GMM) ((Stauffer and Grimson 1999);(Chen et al. 2018)), local binary pattern (LBP) models (Heikkilä and Pietikäinen 2006), and the neural network models (Gemignani and Rozza 2016). Despite the efficiency of these pixel-based methods, they have limitations in exploiting global video structures, e.g., the spatial-temporal correlations among the backgrounds, leading to degraded performance for complex scenarios. Based on the assumption that the backgrounds are low-rank and the foregrounds are sparse, the robust principle component analysis (RPCA) model (Candès et al. 2011) has been used for background subtraction ((Xin et al. 2015);(Gao, Cheong, and Wang 2014);(Cao, Yang, and Guo 2016)). By exploiting the global video structures and the sparse prior, the RPCA-based methods have achieved promising results.

Despite their effectiveness, the RPCA-based methods suffered from two major issues, i.e., the scale issue and the ignorance of the dependencies among the foreground pixels. The scale issue is due to the fact that the fixed  $\ell_1$ -norm sparsity regularizer fails to characterize the varying sparsity of the moving objects of different sizes. To address this issue, the RPCA-based methods consisting of two-pass processes have been proposed ((Gao, Cheong, and Wang 2014);(Liu et al. 2015);(Liu et al. 2018)). The supports of the underlying foregrounds were initially estimated in the first pass. Based on the detected supports, the structured sparsity was then exploited with tuned regularization parameters in the second pass for better performance. In (Yong et al. 2018), Hong *et al.* proposed a mixture of Gaussian (MoG) model to model the noise and the foregrounds. By updating the model parameters frame by frame, the MoG-based method performs better than classic RPCA-based methods. However, it cannot fully exploit the spatial correlations between the foreground pixels. To exploit the spatial correlations of the foreground pixels, in (Ebadi and Lzquierdo 2018) tree-structured sparsity regularizer was proposed. However, the temporal correlations has not been exploited in (Ebadi and Lzquierdo 2018). More recently, Gaussian scale mixture (GSM) mod-

\*Corresponding author (wsdong@mail.xidian.edu.cn).

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

els have also been proposed to model the foregrounds (Shi et al. 2018), where each foreground pixel was modeled as a point-wise product of a hidden positive multiplier and a Gaussian variable. The hidden positive multipliers were jointly estimated along with the foregrounds with a structured sparse prior. Despite its effectiveness, it relied on accurate segmentation of the video frames and also ignored the temporal dependence between adjacent frames.

In this paper, we propose a spatial-temporal Gaussian scale mixture (STGSM) model for foreground estimation. Different from (Shi et al. 2018), nonzero-mean GSM models are proposed, where the foregrounds estimated in previous frames are used as a prior for the estimates of the current frame. To better model the temporal correlations, optical flow is used to model the correspondences between the foreground pixels in adjacent frames. Moreover, instead of using the sparse hyper prior as in (Shi et al. 2018), we propose a nonzero-mean Gaussian prior for the hidden positive multipliers, using the estimates of the hidden positive multipliers estimated in the previous frame. The maximum a posterior (MAP) estimation problem with the proposed STGSM model can be efficiently solved via alternating minimizations. Experimental results on the real video datasets show that the proposed method performs comparable or even better than current state-of-the-art background subtraction methods.

## 2 Related Works

In this section, we briefly review the works related to the proposed STGSM model for background subtraction, i.e., the RPCA-based methods and the Gaussian scale mixture model based methods for image restoration.

### 2.1 RPCA-based background subtraction methods

Let  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$  denote the input video frames, where  $\mathbf{x}_i \in \mathbb{R}^N$  is the vectorized  $t$ -th frame, and  $\mathbf{L} = [\mathbf{l}_1, \dots, \mathbf{l}_T] \in \mathbb{R}^{N \times T}$  and  $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_T] \in \mathbb{R}^{N \times T}$  are the background and foreground components, respectively. Assuming that the backgrounds of the video frames are low-rank and the foregrounds are sparse, the RPCA model attacked the background subtraction problem by solving the following objective function

$$(\mathbf{L}, \mathbf{S}) = \underset{\mathbf{L}, \mathbf{S}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{L} - \mathbf{S}\|_2^2 + \eta \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1, \quad (1)$$

where  $\|\cdot\|_*$  denotes the nuclear norm. The above optimization problem can be solved by alternatively optimizing  $\mathbf{L}$  and  $\mathbf{S}$ . Though the RPCA method is effective for background subtraction, it suffers from the scale issue, i.e., the  $\ell_1$ -sparsity regularizer with fixed parameter  $\lambda$  cannot well characterize the varying sparsity of the moving objects of different sizes. To address this issue, two-passes RPCA methods has been proposed (Gao, Cheong, and Wang 2014);(Liu et al. 2015), where the supports of the moving objects were first detected and the regularization parameters were set based on the detected moving objects. However, these methods were heuristic and their performances depended on the accuracy of the initial estimates of the foregrounds. In ((Yong et al.

2018)) the mixture of Gaussian model has been proposed to model the foregrounds. By jointly learning the model parameters and the foregrounds, it performed better than the RPCA method. However, the spatial correlations among the foreground pixels have not been considered. In (Ebadi and Lzquierdo 2018) the tree-structured sparsity model has been used to exploit the spatial correlations among foreground pixels. To group the correlated pixels, a super-pixel based segmentation method was used. A structured sparsity norm was then imposed over the group pixels. However, this method ignored the strong temporal correlations among the foreground pixels of adjacent frames.

### 2.2 GSM models for sparse signal modeling

As an effective statistical modeling tool, GSM models have shown promising results for image restoration (Portilla et al. 2003);(Dong et al. 2015). In (Portilla et al. 2003), the GSM models were used to model the wavelet coefficients, and the wavelet coefficients were recovered by a Bayesian least-square estimator. In (Dong et al. 2015), with the hyper prior (i.e., the Jeffrey’s prior) Dong *et al.* proposed a GSM model-based image sparse coding method, where both the multipliers and the sparse coefficients were jointly estimated from the degraded image.

Recently, the zero-mean GSM model has also been employed to model the foregrounds (Shi et al. 2018). The spatial correlations of the foreground pixels have also been considered by imposing the same sparse prior over the pixels belong to the same homogeneous regions. Though promising results have been obtained in (Shi et al. 2018), the rich temporal correlations have not been exploited. In this paper, we propose a spatial-temporal GSM (STGSM) model to characterize both the spatial and temporal correlations among the foregrounds of adjacent frames. Specifically, the foregrounds pixels estimated in previous frames are used as a prior in the proposed STGSM model, leading to significant improvements of the background subtraction performance.

## 3 Spatial-temporal GSM model for foreground estimation

In this section, we first briefly review the GSM model for foreground modeling (Shi et al. 2018) and then present the proposed spatial-temporal GSM model for foreground estimation.

### 3.1 GSM modeling of foregrounds

In GSM modeling, each foreground pixel  $s_{t,i}$  is expressed as a point-wise product of a positive scalar multiplier and a Gaussian variable, i.e.,  $s_{t,i} = \theta_{t,i} \alpha_{t,i}$ , where  $s_{t,i}$  is the  $i$ -th entry of  $\mathbf{s}_t$ , and  $s_{t,i}$  is modeled with a zero-mean Gaussian distribution with standard deviation  $\theta_{t,i}$ . A hyper prior  $p(\theta_{t,i})$  is further used to model  $\theta_{t,i}$ . Then, the GSM model of  $s_{t,i}$  can be expressed as,  $p(s_{t,i}) = \int_0^\infty p(s_{t,i}|\theta_{t,i})p(\theta_{t,i})d\theta_{t,i}$ , which in general cannot be expressed in an analytical form for most choices of  $p(\theta)$ . Thus, the joint estimation of  $\mathbf{s}_t$  and  $\boldsymbol{\theta}_t$  are considered in the maxi-

mum a posterior (MAP) estimator by maximizing

$$\begin{aligned} \log p(\mathbf{S}, \Theta, \mathbf{L} | \mathbf{X}) &\propto \log p(\mathbf{X} | \mathbf{S}, \mathbf{L}) + \log p(\mathbf{S}, \Theta, \mathbf{L}) \\ &= \log p(\mathbf{X} | \mathbf{S}, \mathbf{L}) + \log p(\mathbf{L}) + \log p(\mathbf{S} | \Theta) + \log p(\Theta), \end{aligned} \quad (2)$$

where  $p(\mathbf{X} | \mathbf{S}, \mathbf{L})$  is the Gaussian likelihood term and  $\Theta = [\theta_1, \dots, \theta_T] \in \mathbb{R}^{N \times T}$  denote the matrix of the positive multipliers. In (Shi et al. 2018) the Jeffrey's prior  $p(\theta_{t,i}) = \frac{1}{\theta_{t,i} + \epsilon}$  was used to model  $\theta_{t,i}$ , where  $\epsilon$  is a small constant for numerical stability, and the prior of  $\mathbf{L}$  was modeled as  $p(\mathbf{L}) \propto \exp(-\eta \|\mathbf{L}\|_*)$ . By assuming  $\mathbf{S}$  are i.i.d., substituting  $p(\theta_{t,i})$  and  $p(s_{t,i} | \theta_{t,i})$  into Eq. (2) leads to the following objective function

$$\begin{aligned} (\mathbf{A}, \Theta, \mathbf{L}) = \operatorname{argmin}_{\mathbf{A}, \Theta, \mathbf{L}} &\|\mathbf{X} - \mathbf{L} - \Theta \odot \mathbf{A}\|_F^2 + 2\eta\sigma_n^2 \|\mathbf{L}\|_* \\ &+ \sigma_n^2 \|\mathbf{A}\|_F^2 + 4\sigma_n^2 \sum_{t=1}^T \sum_{i=1}^N \log(\theta_{t,i} + \epsilon), \end{aligned} \quad (3)$$

where  $\Theta \odot \mathbf{A} = \mathbf{S}$ ,  $\mathbf{A} = [\alpha_1, \dots, \alpha_T] \in \mathbb{R}^{N \times T}$  denote the matrix of the Gaussian variables  $\alpha_{t,i}$ ,  $\odot$  is the point-wise product, and  $\sigma_n^2$  is the variance of the Gaussian noise. By jointly estimating  $\Theta$  and  $\mathbf{A}$ , the scale issue of the RPCA-based methods can be well addressed, leading to significantly improvements.

### 3.2 Temporal GSM model for foreground estimation

In (Shi et al. 2018) the foreground images  $s_t$  were estimated independently, ignoring the temporal correlations. To address this issue, we propose temporal GSM models for foreground estimation. Instead of modeling  $s_t$  with zero-mean Gaussian functions, we model  $s_t$  with nonzero-mean Gaussian functions, as

$$p(s_t | \theta_t) = \prod_{i=1}^N p(s_{t,i} | \theta_{t,i}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\theta_{t,i}} \exp\left(-\frac{(s_{t,i} - \mu_{t,i})^2}{2\theta_{t,i}^2}\right), \quad (4)$$

where  $\mu_{t,i}$  is the mean of  $s_{t,i}$ . Considering the strong correlations between  $s_{t-1}$  and  $s_t$ ,  $t = 2, \dots, T$ , we propose to compute  $\mu_t$  from the estimate of  $s_{t-1}$ , as  $\mu_t = f(\hat{s}_{t-1})$ , where  $\hat{s}_{t-1}$  denotes an estimate of  $s_{t-1}$  and  $f(\cdot)$  is a predefined function that will be discussed later. For the hyper prior  $p(\theta_t)$ , instead of using the Jeffrey's prior we also propose to model  $\theta_t$  with nonzero-mean Gaussian functions as

$$p(\theta_t) = \prod_{i=1}^N p(\theta_{t,i}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(\theta_{t,i} - \nu_{t,i})^2}{2\delta^2}\right), \quad (5)$$

where  $\nu_t$  denotes the expectations of  $\theta_t$  and  $\delta$  denotes the variance of  $\theta_{t,i}$ , which is fixed for all  $\theta_{t,i}$ . Similar to  $s_t$ , we also compute  $\nu_t$  from the estimate of  $\theta_{t-1}$  of the previous frame, as  $\nu_t = f(\hat{\theta}_{t-1})$ . By substituting  $p(s_t)$  and  $p(\theta_t)$  defined in Eqs. (4) and (5) into the MAP estimator of Eq. (2), we can obtain a novel objective function, as follows

$$\begin{aligned} (\mathbf{A}, \Theta, \mathbf{L}) = \operatorname{argmin}_{\mathbf{A}, \Theta, \mathbf{L}} &\|\mathbf{X} - \mathbf{L} - \Theta \odot \mathbf{A}\|_F^2 + 2\eta\sigma_n^2 \|\mathbf{L}\|_* \\ &+ \sigma_n^2 \sum_{t=1}^T \|\alpha_t - \beta_t\|^2 + 2\sigma_n^2 \sum_{t=1}^T \log(\theta_t + \epsilon) \\ &+ \frac{\sigma_n^2}{\delta} \sum_{t=1}^T \|\theta_t - \nu_t\|_2^2, \end{aligned} \quad (6)$$

where we have used  $\mathbf{S} = \Theta \odot \mathbf{A}$ , and  $\beta_t$  denotes the mean of  $\alpha_t$  that is computed as  $\beta_{t,i} = \mu_{t,i} / (\hat{\theta}_{t,i} + \epsilon)$ , wherein  $\hat{\theta}_{t,i}$  denotes an estimate of  $\theta_{t,i}$ . In our implementation, we recursively calculate  $\beta_{t,i}$  using the estimate of  $\theta_{t,i}$  in each iteration.

### 3.3 Estimation of the expectations

As described in the above subsection, we use the estimates of  $s_{t-1}$  and  $\theta_{t-1}$  of the previous frame to compute the means of  $s_t$  and  $\theta_t$ , as  $\mu_t = f(\hat{s}_{t-1})$  and  $\nu_t = f(\hat{\theta}_{t-1})$ . The simplest function of  $f(\cdot)$  is the identity function, i.e.,  $\mu_t = \hat{s}_{t-1}$  and  $\nu_t = \hat{\theta}_{t-1}$ . However, due to the motions between the objects in the adjacent frames, the identity function is inaccurate. To compensate the motions, we use the optical flow to warp the previously estimated foregrounds for alignment and set  $\mu_t$  and  $\nu_t$  as

$$\mu_t = w_{t,t-1}(\hat{s}_{t-1}), \quad \nu_t = w_{t,t-1}(\hat{\theta}_{t-1}), \quad (7)$$

where  $w_{t,t-1}$  denotes the operator that warps an image using the optical flow estimated between  $x_{t-1}$  and  $x_t$ . For the first frame, we set  $\mu_1 = \mathbf{0}$  and  $\nu_1 = \mathbf{0}$  and use the Jeffrey's prior for  $\theta_1$  for the first frame. For efficiency, in this work we use the total variation based optical flow estimation method (Liu et al. 2008) to compute the optical flow.

### 3.4 Spatial-temporal GSM model for foreground estimation

To exploit the spatial correlations among the neighboring pixels, we try to group the foreground pixels  $s_t$  into  $K$  groups using a pixel grouping method and characterize the pixels of the same group by the same GSM model. The Gaussian prior exploiting the spatial-temporal correlations can then be expressed as

$$p(s_t | \theta_t) = \prod_{k=1}^K \prod_{j \in \mathcal{G}_k} \frac{1}{\sqrt{2\pi}\theta_{t,k}} \exp\left(-\frac{(s_{t,j} - \tilde{\mu}_{t,k})^2}{2\theta_{t,k}^2}\right), \quad (8)$$

where  $\mathcal{G}_{t,k}$  denotes the  $k$ -th pixel group of the  $t$ -th frame,  $(\tilde{\mu}_{t,k}, \theta_{t,k})$  are the mean and variance of the pixels of  $\mathcal{G}_{t,k}$ . Accordingly, the structured hyper prior of  $\theta_t$  can be formulated as

$$p(\theta_t) = \prod_{k=1}^K \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(\theta_{t,k} - \tilde{\nu}_{t,k})^2}{2\delta^2}\right). \quad (9)$$

The expectations  $\tilde{\mu}_{t,k}$  and  $\tilde{\nu}_{t,k}$  for  $\mathcal{G}_{t,k}$  are computed as

$$\tilde{\mu}_{t,k} = \frac{1}{|\mathcal{G}_{t,k}|} \sum_{j \in \mathcal{G}_{t,k}} \mu_{t,j}, \quad \tilde{\nu}_{t,k} = \frac{1}{|\mathcal{G}_{t,k}|} \sum_{j \in \mathcal{G}_{t,k}} \nu_{t,j}, \quad (10)$$

where  $|\mathcal{G}_{t,k}|$  denotes the number of pixels in  $\mathcal{G}_{t,k}$ .

A simple pixel grouping method is to divide the image into non-overlapped blocks. A better pixel grouping method can be obtained by a segmentation method, e.g., the superpixel-based method. For simplicity, here each foreground image is divided into many non-overlapped blocks (in our implementation  $5 \times 5$  blocks were adopted). Then, the proposed spatial-temporal GSM (STGSM) model based foreground estimation method can be obtained by substituting Eqs. (8) and (9) into the MAP estimator of Eq. (2), as

$$\begin{aligned}
(\mathbf{A}, \Theta, \mathbf{L}) = & \operatorname{argmin}_{\mathbf{A}, \Theta, \mathbf{L}} \sum_{t=1}^T \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (x_{t,j} - l_{t,j} - \theta_{t,k} \alpha_{t,j})^2 \\
& + 2\eta \sigma_n^2 \|\mathbf{L}\|_* + \sigma_n^2 \sum_{t=1}^T \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (\alpha_{t,j} - \tilde{\beta}_{t,k})^2 \\
& + 2\sigma_n^2 \sum_{t=1}^T \sum_{k=1}^K |\mathcal{G}_{t,k}| \log(\theta_{t,k} + \epsilon) \\
& + \frac{\sigma_n^2}{\delta} \sum_{t=1}^T \sum_{k=1}^K |\mathcal{G}_{t,k}| (\theta_{t,k} - \tilde{\nu}_{t,k})^2,
\end{aligned} \tag{11}$$

where  $\tilde{\beta}_{t,k} = \tilde{\mu}_{t,k} / (\hat{\theta}_{t,k} + \epsilon)$ , wherein  $\hat{\theta}_{t,k}$  denotes an estimate of  $\theta_{t,k}$  that is recursively updated in each iteration.

## 4 Optimization Algorithm

The proposed objective function of Eq. (11) can be solved by alternatively solving two subproblems, i.e., the  $\mathbf{S}$ -subproblem and  $\mathbf{L}$ -subproblem, as will be described as follows.

### 4.1 The $\mathbf{S}$ -subproblem

For fixed  $\mathbf{L}$ ,  $\mathbf{S}$  can be updated by solving a sequence of the following optimization problems

$$\begin{aligned}
(\alpha_t, \theta_t) = & \operatorname{argmin}_{\alpha_t, \theta_t} \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (r_{t,j} - \theta_{t,k} \alpha_{t,j})^2 \\
& + \sigma_n^2 \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (\alpha_{t,j} - \tilde{\beta}_{t,k})^2 + \frac{\sigma_n^2}{\delta} \sum_{k=1}^K |\mathcal{G}_{t,k}| (\theta_{t,k} - \tilde{\nu}_{t,k})^2 \\
& + 2\sigma_n^2 \sum_{k=1}^K |\mathcal{G}_{t,k}| \log(\theta_{t,k} + \epsilon), \quad t = 1, \dots, T,
\end{aligned} \tag{12}$$

where  $r_{t,j} = x_{t,j} - l_{t,j}$ . The above optimization problem can be solved by alternatively updating  $\theta_t$  and  $\alpha_t$ .

**The update of  $\theta_t$ .** For fixed  $\alpha_t$ ,  $\theta_t$  can be updated by minimizing

$$\begin{aligned}
\theta_t = & \operatorname{argmin}_{\theta_t} \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (r_{t,j} - \theta_{t,k} \alpha_{t,j})^2 \\
& + 2\sigma_n^2 \sum_{k=1}^K |\mathcal{G}_{t,k}| \log(\theta_{t,k} + \epsilon) + \frac{\sigma_n^2}{\delta} \sum_{k=1}^K |\mathcal{G}_{t,k}| (\theta_{t,k} - \tilde{\nu}_{t,k})^2
\end{aligned} \tag{13}$$

Since each  $\theta_{t,k}$  is independent, solving the above equation equals to solving a sequence of scalar optimization problems, as

$$\begin{aligned}
\theta_{t,k} = & \operatorname{argmin}_{\theta_{t,k}} \frac{1}{|\mathcal{G}_{t,k}|} \sum_{j \in \mathcal{G}_{t,k}} (r_{t,j} - \theta_{t,k} \alpha_{t,j})^2 \\
& + 2\sigma_n^2 \log(\theta_{t,k} + \epsilon) + \frac{\sigma_n^2}{\delta} (\theta_{t,k} - \tilde{\nu}_{t,k})^2,
\end{aligned} \tag{14}$$

which can be further re-written as

$$\begin{aligned}
\theta_{t,k} = & \operatorname{argmin}_{\theta_{t,k}} a\theta_{t,k}^2 + 2b\theta_{t,k} + 2c \log(\theta_{t,k} + \epsilon) + d, \\
& \text{s. t. } \theta_{t,k} \geq 0,
\end{aligned} \tag{15}$$

wherein  $d$  denotes a constant independent on  $\theta_{t,k}$ ,  $a$ ,  $b$ , and  $c$  are given by

$$\begin{aligned}
a = & \frac{1}{|\mathcal{G}_{t,k}|} \sum_{j \in \mathcal{G}_{t,k}} \alpha_{t,j}^2 + \frac{\sigma_n^2}{\delta}, \\
b = & -\left( \frac{1}{|\mathcal{G}_{t,k}|} \sum_{j \in \mathcal{G}_{t,k}} r_{t,j} \alpha_{t,j} + \frac{\sigma_n^2}{\delta} \tilde{\nu}_{t,k} \right), \\
c = & \sigma_n^2.
\end{aligned} \tag{16}$$

To solve Eq.(15), we take  $\frac{f(\theta_{t,k})}{d\theta_{t,k}} = 0$  and a closed-form solution can be derived, where  $f(\theta_{t,k})$  denotes the right hand side of Eq. (15). The solution to Eq. (15) can then be obtained as

$$\theta_{t,k} = \begin{cases} 0, & \text{if } (b + a\epsilon)^2 - 4a(c + b\epsilon) < 0 \\ \min_{\theta_{t,k}} \{f(0), f(\theta^*)\}, & \text{otherwise} \end{cases} \tag{17}$$

where  $\theta^*$  is given by

$$\theta^* = \frac{-(b + a\epsilon) \pm \sqrt{(b + a\epsilon)^2 - 4a(c + b\epsilon)}}{2a}. \tag{18}$$

**The update of  $\alpha_t$ .** For fixed  $\theta_t$ ,  $\alpha_t$  can be updated by minimizing

$$\min_{\alpha_t} \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (r_{t,j} - \theta_{t,k} \alpha_{t,j})^2 + \sigma_n^2 \sum_{k=1}^K \sum_{j \in \mathcal{G}_{t,k}} (\alpha_{t,j} - \tilde{\beta}_{t,k})^2, \tag{19}$$

which can be easily solved in a closed-form, i.e.,

$$\alpha_t = (\mathbf{\Lambda}_t^\top \mathbf{\Lambda} + \sigma_n^2 \mathbf{I})^{-1} (\mathbf{\Lambda}_t^\top \mathbf{r}_t + \sigma_n^2 \tilde{\beta}_t), \tag{20}$$

where  $\mathbf{\Lambda}_t = \operatorname{diag}(\theta_{k_i})$  is a diagonal matrix,  $k_i$  denotes the  $k_i$ -th group that the  $i$ -th foreground pixel belongs to. Since the matrix to be inverted is diagonal, Eq. (20) can be computed efficiently.

### 4.2 The $\mathbf{L}$ -subproblem

With fixed  $\mathbf{S}$ , the background component  $\mathbf{L}$  can be updated by solving

$$\mathbf{L} = \operatorname{argmin}_{\mathbf{L}} \|\mathbf{X} - \mathbf{S} - \mathbf{L}\|_F^2 + 2\eta \sigma_n^2 \|\mathbf{L}\|_*, \tag{21}$$

which can be solved by a singular value thresholding (SVT) operator (Cai, Candès, and Shen 2010). However, the computational complexity of SVT is very high, as the singular value decomposition (SVD) has to be performed on the

entire matrix  $L$ . Instead, we employ the stochastic optimization method (Feng, Xu, and Yan 2013) to solve the  $L$ -subproblem, and reexpress the nuclear-norm as

$$\|L\|_* = \inf_{D \in \mathbb{R}^{N \times r}, Z \in \mathbb{R}^{T \times r}} \left\{ \frac{1}{2} (\|D\|_F^2 + \|Z\|_F^2) \right. \\ \left. s. t. L = DZ^\top \right\}, \quad (22)$$

where  $L$  is factorized into a product of two matrixes, i.e.,  $L = DZ$ , and  $r$  denotes the up bound of the rank of  $L$ .  $D$  is treated as the bases of a low-dimensional space and  $Z$  the representation coefficients. Then, the  $L$ -subproblem can be reformulated as

$$(D, z_t) = \operatorname{argmin}_{D, z_t} \sum_{t=1}^T \left\{ \|\mathbf{x}_t - Dz_t - s_t\|_2^2 + \eta \sigma_n^2 \|z_t\|_2^2 \right\} \\ + \eta \sigma_n^2 \|D\|_F^2, \quad (23)$$

where  $z_t$  is the  $t$ -th row of  $Z$ .

**The update of  $z_t$ .** With a fixed  $D$ ,  $z_t$  can be updated by minimizing

$$z_t = \operatorname{argmin}_{z_t} \|\mathbf{x}_t - Dz_t - s_t\|_2^2 + \eta \sigma_n^2 \|z_t\|_2^2, \quad (24)$$

which admits a closed-form solution

$$z_t = (D^\top D + \eta \sigma_n^2 I)^{-1} D^\top (\mathbf{x}_t - s_t). \quad (25)$$

**The update of  $D$ .** Similar to (Shi et al. 2018), we use the online learning method of (Feng, Xu, and Yan 2013) to update  $D$ . With an estimate of  $z_t$ , the objective function for updating  $D$  in the  $t$ -th iteration can be defined as

$$D^{(t)} = \operatorname{argmin}_D \frac{1}{t} \sum_{i=1}^t \frac{1}{2} \|\mathbf{x}_i - Dz_i - s_i\|_F^2 + \frac{\eta \sigma_n^2}{2t} \|D\|_F^2, \quad (26)$$

which can be re-written as

$$D^{(t)} = \operatorname{argmin}_D \frac{1}{2} \operatorname{Tr}[D^\top (A_t + \eta \sigma_n^2 I) D] - \operatorname{Tr}(D^\top B_t), \quad (27)$$

where  $A_t$  and  $B_t$  are computed via

$$A_t \leftarrow A_{t-1} + z_t z_t^\top, \\ B_t \leftarrow B_{t-1} + (\mathbf{x}_t - s_t) z_t^\top, \quad (28)$$

where  $A_0 = \mathbf{0}$  and  $B_0 = \mathbf{0}$ . Note that Eq. (27) can be efficiently solved by using the block-coordinate descent with warm restarts (Bertsekas 1999). Specifically, each column of  $D$  can be updated individually while keeping other columns fixed, as

$$d_i^{(t+1)} \leftarrow d_i^{(t)} + \frac{1}{\tilde{A}_{i,i}} (\mathbf{b}_i - D^{(t)} \tilde{\mathbf{a}}_i), \quad (29)$$

where  $d_i$  denotes the  $i$ -th column of  $D$ ,  $\tilde{A} = A_t + \eta \sigma_n^2 I$ ,  $\mathbf{b}_i$  and  $\tilde{\mathbf{a}}_i$  are the  $i$ -th column vectors of  $B_t$  and  $\tilde{A}$ , respectively. The bilateral random projections method of (Zhou and Tao 2012) is used to initialize  $D^{(0)}$ , as

$$D^{(0)} = Y_1 (P_2^\top Y_1)^{-1} Y_2^\top, \quad (30)$$

where  $P_1 \in \mathbb{R}^{W \times r}$ ,  $P_2 \in \mathbb{R}^{H \times r}$ ,  $Y_1 = Y P_1$ ,  $Y_2 = Y^\top P_2$ ,  $Y \in \mathbb{R}^{H \times W}$  denotes the first frame in matrix form, and  $H$  and  $W$  are the height and width of the frames. The overall algorithm of the proposed foreground estimation method is summarized in **Algorithm 1**. In **Algorithm 1**, we set  $K = 2$  for the inner loop and found that it was sufficient for convergence and good results.

---

**Algorithm 1** Proposed STGSM model based Foreground Estimation Method

---

• **Initialization:**

(1) Set parameter  $\eta$ ,  $\sigma_n^2$ ,  $\delta$ ,  $A_0 \in \mathbb{R}^{r \times r} = \mathbf{0}$  and  $B_0 \in \mathbb{R}^{N \times r} = \mathbf{0}$ ;

(2) Initialize  $D^{(0)} \in \mathbb{R}^{N \times r}$  via Eq.(30).

• **Outer loop:** for  $t = 1, 2, \dots, T$ , **do**

(1) **Inner loop** (solving Eqs. (12) and (24) ) for  $k = 1, 2, \dots, K$ , **do**

(a) Update  $\theta_t$  via Eq.(17);

(b) Update  $\alpha_t$  via Eq.(20);

(c) Compute  $s_t = \theta_t \odot \alpha_t$ ;

(d) Update  $z_t$  via Eq.(25).

**End for**

(2) Update the background component  $l_t$ .

(a) Update  $D_t$  via Eq. (29);

(b) Compute  $l_t = D_t z_t$ .

• **End for**

---

## 5 Experiment results

The proposed methods were implemented under the Matlab platform. The major parameters of the proposed methods were empirically set as,  $r = 15$ ,  $\eta = \frac{400}{N}$ ,  $\sigma_n^2 = 1.05 \times 10^{-3}$ , and  $\delta = 0.01$ , which are the same for all test sequences. The performances of the proposed methods are evaluated on two benchmark datasets, i.e., the perception test image sequences (PTIS) (Li et al. 2004) and the change detection (CD) 2012 (Goyette et al. 2012) datasets, containing 40 videos. The pixel values of the videos are normalized to the range of  $[0, 1]$ . The F-measure is used to evaluate the performance of the proposed methods, defined as

$$\text{F-measure} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (31)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP (true positives) is the number of pixels correctly classified as foreground, FP (false positives) the number of pixels incorrectly classified as foreground, and FN (false negatives) the numbers of foreground pixels incorrectly classified as background.

Category	Baseline	Dynamic Bac.Gro.	Camera Jitter	Shadows	Thermal	Inter. Obj. Mot.	average
GSM	93.00	83.00	81.00	87.00	82.00	69.00	82.50
<b>Proposed TGSM</b>	93.4	83.91	81.24	89.41	82.30	86.47	86.12
<b>Proposed STGSM</b>	95.92	86.22	85.63	92.98	85.07	89.67	89.25

Table 1: *F-measure* (%) performance of the variants of the proposed methods on CD 2012 dataset.

Video Method	Shopping Mall	Lobby	Curtain	Fountain	Campus	Hall	Bootstrap	Escalator	Water Surface	Average
PCP (Candès et al. 2011)	66.78	62.18	81.28	63.74	40.69	50.25	56.78	63.54	39.76	58.33
PBAS(Hofmann et al.2012)	71.77	56.74	72.48	62.04	69.50	61.92	55.38	31.13	89.02	63.33
DECOLOR(zhou et al.2013)	68.06	56.52	81.58	82.76	73.29	53.20	56.86	<b>75.46</b>	79.65	69.71
LRSSD (Liu et al. 2015)	73.62	73.13	83.57	83.71	76.13	72.22	58.42	<u>72.14</u>	90.50	75.93
TVRPCA(gao et al.2016)	—	75.00	—	80.00	77.00	69.00	69.00	66.00	88.00	—
OMoGMF+TV (Yong et al. 2018)	72.43	83.47	92.54	82.53	65.88	77.20	61.17	66.37	<u>93.14</u>	77.19
SGSM-BS-block (Shi et al. 2018)	77.70	83.78	79.12	86.44	77.93	74.96	<b>79.05</b>	69.11	71.21	77.70
SGSM-BS (Shi et al. 2018)	79.89	<u>86.02</u>	<b>93.54</b>	<b>87.16</b>	<b>83.15</b>	<u>77.59</u>	<u>78.75</u>	71.70	92.88	<b>83.41</b>
<b>Proposed STGSM</b>	<b>80.29</b>	<b>86.53</b>	<u>92.56</u>	<u>86.44</u>	<u>80.35</u>	<b>80.49</b>	77.07	71.55	<b>93.23</b>	<u>83.17</u>

Table 2: Performance of  $F$ -measure (%) on Perception Test Images Sequences dataset.

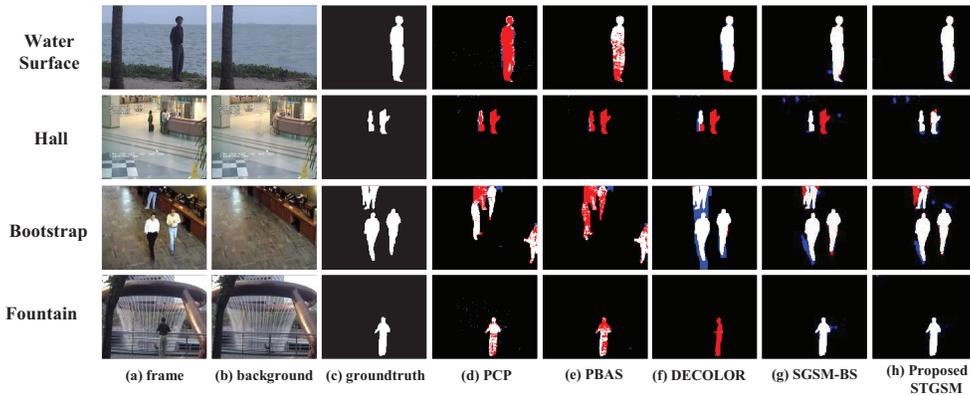


Figure 1: Background subtraction results on 4 videos from PTIS dataset by the test methods. (a) Original frame; (b) background; (c) Groundtruth foreground mask ;(d)PCP (Candès et al. 2011);(e)PBAS(Hofmann, Tiefenbacher, and Rigoll 2012) (f) DECOLOR (Zhou, Yang, and Yu 2013); (g) SGSM-BS (Shi et al. 2018); (h) **Proposed STGSM**. (White represents correctly detected foreground, red represents missing pixels, and blue represents false alarm pixels.)

Category Method	Baseline	Dynamic Bac.Gro.	Camera Jitter	Shadows	Thermal	Inter. Obj. Mot.	average
SOBS(Maddalena 2012)	93.33	66.86	70.51	77.86	69.23	59.18	72.83
TVRPCA(gao et al.2016)	91.33	65.92	74.82	79.50	70.19	58.78	73.42
DECOLOR(zhou et al.2013)	92.15	70.84	74.77	54.18	81.27	81.34	77.63
WeSamBE(Jiang et al.2018)	94.13	74.40	79.76	86.86	79.62	73.92	81.48
SuBSENSE(St-Charles2015)	95.03	81.77	81.52	89.86	81.71	65.69	82.60
CNN(Babae et al.2018)	95.80	87.61	<b>89.90</b>	<b>93.04</b>	75.83	60.98	83.86
SGSM-BS (Shi et al. 2018)	95.00	85.00	82.00	89.00	85.00	82.00	86.00
DSPSS(Ebadi et al.2018)	<b>96.64</b>	<u>90.57</u>	<u>86.62</u>	91.77	73.28	78.70	86.26
STSHBM (Chen et al. 2018)	95.34	<b>91.20</b>	85.03	83.49	<b>89.30</b>	<u>85.79</u>	<u>88.36</u>
<b>Proposed STGSM</b>	<u>95.92</u>	86.22	85.63	<u>92.98</u>	<u>85.07</u>	<b>89.67</b>	<b>89.25</b>

Table 3: Performance of  $F$ -measure (%) of the test methods on CD 2012 dataset.

## 5.1 Ablation study

To verify the effects of the proposed temporal and spatial regularization terms, we implemented three variants of the proposed method, i.e., the GSM model without considering the spatial-temporal correlations, the temporal GSM model based foreground estimation method (denoted as TGSM), and the proposed spatial-temporal GSM model based method (denoted as STGSM). The experimental results on CD 2012 dataset are shown in Table 1. From Table 1, it can be seen that the TGSM method outperforms the GSM method by a large margin. By exploiting the local spatial correlations, the STGSM performs better than TGSM.

## 5.2 CD 2012 dataset

We compared the proposed STGSM method with several state-of-the-art background subtraction methods, including two RPCA-based methods, i.e., the SGSM-BS (Shi et al. 2018) and the DSPSS (Ebadi and Lzquierdo 2018) methods, and the pixel classification based method, i.e., STSHBM method (Chen et al. 2018). For fair comparisons, the results of the other methods were obtained from their papers or the website of the CD 2012 dataset, or generated by running their source codes. From Table 3, we can see that the proposed method outperformed all the other competing methods. The proposed STGSM method outperformed the SGSM-BS method of (Shi et al. 2018) by up to 3.25% in average, verifying the effectiveness of the proposed STGSM models. Fig. 2 shows some visual results generated by the test methods on several videos of CD 2012 dataset<sup>1</sup>. It can be seen that the proposed method performed better than the other methods.

## 5.3 PTIS dataset

We also compared the proposed method with other methods on the PTIS dataset (Li et al. 2004), which contains 9 real videos covering a variety of scenarios including dynamic background and static background with short-time

<sup>1</sup>We only show the visual results of the test methods that we can find from their websites.

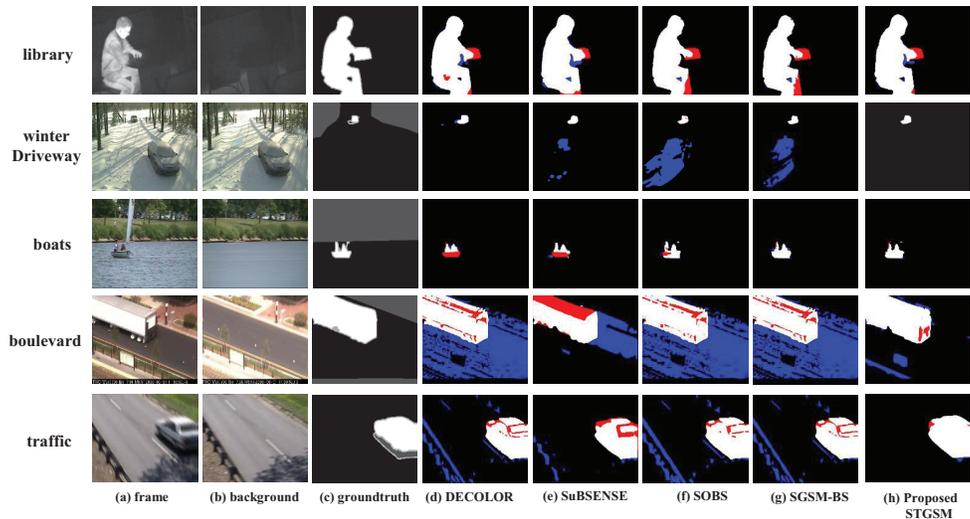


Figure 2: Background subtraction results on 5 videos from CD 2012 dataset by the test methods. (a) Original frame; (b) background; (c) Groundtruth foreground mask ; (d) DECOLOR (Zhou, Yang, and Yu 2013); (e) SuBSENSE (St-Charles, Bilodeau, and Bergevin 2015);(f) SOBS (Maddalena and Petrosino 2012); (g) SGSM-BS (Shi et al. 2018); (h) **Proposed STGSM**. (White represents correctly detected foreground, red represents missing pixels, and blue represents false alarm pixels.)

lingering objects. Table 2 shows the F-measure results of the test methods. From Table 2, we can see that the proposed STGSM method outperformed the other methods, except the SGSM-BS method (Shi et al. 2018) that used a superpixel-based segmentation method for grouping correlated pixels. The proposed STGSM method performed much better than the SGMS-BS-block method (Shi et al. 2018) that also used the block-based pixel grouping method. More experimental results are referred to the supplementary material.

#### 5.4 Computational complexity

The complexity of the proposed method consists of three parts: 1) the computing  $s_t$  that is  $O(N)$ ; 2) the updating  $z_t$  that is  $O(Nr^2)$ ; 3) the updating  $D_t$  that is  $O(Nr^2)$ . The overall complexity of the proposed method for  $T$  frames is  $O(TNr^2)$ . Here, we have not included the complexity of estimating optical flow, which is not the focus of this work.

Table 4 shows the running time of the test methods (only the methods whose codes are available are shown) on 100 rgb frames of size  $128 \times 160$ , where STGSM-wo denotes the proposed method where optical flows have been already estimated. The main computational cost of the proposed method is the estimation of the optical flow, which takes up about 75% time cost. The proposed methods were implemented with Matlab platform on an Intel Core i7-8700K 3.4GHz CPU, and can be accelerated with parallel computation technique for real-time applications.

## 6 Conclusions

To address the scale issue of the RPCA model for foreground estimation, we proposed a novel spatial-temporal Gaussian scale mixture models (GSM) for foreground estimation.

Method	GMM (Zivkovic 2004)	DECOLOR (zhou et al.2013)	LRSSD (Liu et al. 2015)	OMoGMF+TV (Yong et al. 2018)
Time(s)	0.18	3455.62	255.31	14.13
Method	TVRPCA (gao et al.2016)	SGSM-BS (Shi et al. 2018)	<b>Proposed STGSM</b>	<b>Proposed STGSM-wo</b>
Time(s)	84.40	27.39	41.49	10.32

Table 4: Running time on 100 rgb frames of size  $128 \times 160$ .

Specifically, nonzero-means Gaussian functions were used to model both the foreground pixels and the hidden positive multipliers. By estimating the means of the foreground pixels and multipliers from the estimates of the previous frame, the temporal correlations of the foregrounds can be effectively exploited. Moreover, the GSM models were further extended to exploit the local spatial correlations by considering that correlated pixels should be characterized by the same distribution. Experimental results on real challenging video datasets show that the proposed method outperforms most of current background subtraction methods.

## Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1402603 and the Natural Science Foundation of China under Grant 61622210, Grant 61632019, Grant 61621005, and Grant 61836008.

## References

- Bertsekas, D. 1999. *Nonlinear programming*. Athena scientific Belmont.
- Cai, J.-F.; Candès, E.; and Shen, Z. 2010. A singular value

- thresholding algorithm for matrix completion. *SIAM J. Optimization* 20(4):1956–1982.
- Candès, E.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis. *Journal of ACM* 58(3):11.
- Cao, X.; Yang, L.; and Guo, X. 2016. Total variation regularized RPCA for irregularly moving object detection under dynamic background. *IEEE Trans. Cybern.* 46(4):1014–1027.
- Chen, M.; Wei, X.; Yang, Q.; Li, Q.; Wang, G.; and Yang, M.-H. 2018. Spatiotemporal gmm for background subtraction with superpixel hierarchy. *IEEE trans. on pattern analysis and machine intelligence* 40(6):1518–1525.
- Dong, W.; Shi, G.; Ma, Y.; and Li, X. 2015. Image restoration via simultaneous sparse coding: Where structured sparsity meets gaussian scale mixture. *Int. J. Comput. Vis.* 114(2-3):217–232.
- Ebadi, S. E., and Lzquierdo, E. 2018. Foreground segmentation with tree-structured sparse rpca. *IEEE trans. on pattern analysis and machine intelligence* 40(9):2273–2280.
- Feng, J.; Xu, H.; and Yan, S. 2013. Online robust PCA via stochastic optimization. In *Adv. Neural Inf. Process. Syst.*, 404–412.
- Gao, Z.; Cheong, L.-F.; and Wang, Y.-X. 2014. Block-sparse rpca for salient motion detection. *IEEE trans. on pattern analysis and machine intelligence* 36(10):1975–1987.
- Gemignani, G., and Rozza, A. 2016. A robust approach for the background subtraction based on multi-layered self-organizing maps. *IEEE Trans. Image Process.* 25(11):5239–5251.
- Goyette, N.; Jodoin, P.; Porikli, F.; Konrad, J.; and Ishwar, P. 2012. Changedetection.net: A new change detection benchmark dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 1–8.
- Heikkilä, M., and Pietikäinen, M. 2006. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(4):657–662.
- Hofmann, M.; Tiefenbacher, P.; and Rigoll, G. 2012. Background segmentation with feedback: The pixel-based adaptive segmenter. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 38–43.
- Li, L.; Huang, W.; Gu, I.; and Tian, Q. 2004. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.* 13(11):1459–1472.
- Liu, C.; Freeman, W.; Adelson, E. H.; and Weiss, Y. 2008. Human-assisted motion annotation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*
- Liu, X.; Zhao, G.; Yao, J.; and Qi, C. 2015. Background subtraction based on low-rank and structured sparse decomposition. *IEEE Trans. Image Process.* 24(8):2502–2514.
- Liu, X.; Yao, J.; Hong, X.; Huang, X.; Zhou, Z.; Qi, C.; and Zhao, G. 2018. Background subtraction using spatio-temporal group sparsity recovery. *IEEE Transactions on Circuits and Systems for Video Technology* 28(8):1737–1751.
- Liu, N.; Wu, H.; and Lin, L. 2015. Hierarchical ensemble of background models for ptz-based video surveillance. *IEEE Trans Cybern* 45(1):89–102.
- Maddalena, L., and Petrosino, A. 2012. The SOBS algorithm: What are the limits? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 21–26.
- Portilla, J.; Strela, V.; Wainwright, M. J.; and Simoncelli, E. P. 2003. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process.* 12(11):1338–1351.
- Shi, G Huang, T.; Dong, W.; Wu, J.; and Xie, X. 2018. Robust foreground estimation via structured gaussian scale mixture modeling. *IEEE Trans. Image Process.* 25(11):5239–5251.
- St-Charles, P.-L.; Bilodeau, G.-A.; and Bergevin, R. 2015. Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing* 24(1):359–373.
- Stauffer, C., and Grimson, W. 1999. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 246–252.
- Unzueta, L.; Nieto, M.; Cortes, A.; Barandiaran, J.; Otaegui, O.; and Sanchez, P. J. 2012. Adaptive multicue background subtraction for robust vehicle counting and classification. *IEEE Transactions on Intelligent Transportation Systems* 13(1):527–540.
- Xin, B.; Tian, Y.; Wang, Y.; and Gao, W. 2015. Background subtraction via generalized fused lasso foreground modeling. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4676–4684.
- Yong, H.; Meng, D.; Zuo, W.; and Zhang, L. 2018. Robust online matrix factorization for dynamic background subtraction. *IEEE trans. on pattern analysis and machine intelligence* 40(7):1726–1740.
- Zhou, T., and Tao, D. 2012. Bilateral random projections. In *Proc. IEEE ISIT*, 1286–1290.
- Zhou, X.; Yang, C.; and Yu, W. 2013. Moving object detection by detecting contiguous outliers in the low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(1):597–610.
- Zivkovic, Z. 2004. Improved adaptive Gaussian mixture model for background subtraction. In *Proc. Int. Conf. Pattern Recognit.*, volume 2, 28–31.