EXTRACT FOREGROUND OBJECTS BASED ON SPARSE MODEL OF SPATIOTEMPORAL SPECTRUM

Zhangjian Ji *Weiqiang Wang* *Ke Lu*

University of Chinese Academy of Sciences, School of Computer and Control Engineering, Beijing, China
Email: *jizhangjian08@mails.ucas.ac.cn* *wqwang@ict.ac.cn* *luk@gucas.ac.cn*

ABSTRACT

In this paper, we present a novel foreground object detection method based on the sparse model of the spectrum of spatiotemporal DCT domain, which is robust for high dynamic scenes. First, we adopt the three-dimensional Discrete Cosine Transform (DCT) to calculate the spatiotemporal spectrum representation of the current frame. Then, identification of foreground pixels is formulated as the analysis of the sparse solution of an optimization problem, where foreground pixels correspond to an outlier of the sparse model. Finally, the background updating method is presented to adaptively update the dictionary of sparse model corresponding to background representation. The experimental results on four challenging video sequences show that the proposed method is more robust to high dynamic changes of scenes compared with four representative methods.

Index Terms— Spatiotemporal spectrum, Sparse model, foreground object detection

1. INTRODUCTION

Moving object extraction in complex scenes is widely used in the automated surveillance, which is the primary technique of many high-level tasks, such as object tracking, action recognition and understanding. In fixed camera surveillance systems, background subtraction techniques are popular. In the real world, the background of most scenes contains complex dynamic changes, i.e., gradual or sudden illumination variation or repetitive motion, such as rippling water, moving vegetation in the wild, spouting fountain. The traditional pixel-based method [1, 2] cannot well model these changes. Recently, there is a tendency of using the neighborhood information to enhance the power of background model, i.e., spatial neighborhood[1, 3, 4] or temporal neighborhood[5, 6]. However, these methods only considering the information of spatial neighborhood or temporal neighborhood may also produce a lot of false detections due to complex dynamic scenes.

Intuitively, spatial neighborhood information and temporal neighborhood information are complementary to each other. Thus, the integration of spatiotemporal information is a better way to model dynamic scenes. Doretto *et al.* [7] exploit the dynamic texture to model a spatiotemporal volume from a linear dynamic system. Moshe *et al.* [8] directly adopt the spatiotemporal volume. Generally, this category of methods extract the features from the pixels in a volume as its representation [9, 10]. Its limitation lies in the existence of obvious block effect and representation complexity.

In the proposed approach, we first compute the local spatiotemporal spectrum of the current frame, and then a low-dimension compact representation is constructed through selecting the spectral coefficients with intensive energy statistically. Next, foreground pixels are identified through analyzing the sparse solution of an optimization problem about the sparse model. Finally, the background model is updated by updating the atoms of dictionary in the sparse model if necessary. Compared with the existing approaches of foreground detection, our approach utilizes a more compact and computationally efficient spatiotemporal spectrum representation based on DCT instead of local Fourier transform(LFT) [11], and we presents a novel foreground object detection scheme based on the sparse model, as well as background update strategy.

The rest of the paper is organized as follows. Section 2 presents the background model based on the DCT spectrum of spatiotemporal volumes. The foreground object extraction approach is detailed in Section 3. The experimental results are reported in Section 4. Section 5 concludes the paper.

2. BACKGROUND MODEL BASED ON LOCAL SPATIOTEMPORAL SPECTRUM

For a pixel in a video frame, its spatiotemporal neighboring pixels determine that it is foreground or background. Let $\mathbf{p} = (x, y, t)$ denote a location in the spatiotemporal volume, and then its neighboring spatiotemporal cuboid is defined as $\Omega(\mathbf{p}) = \{p' | p' = (x', y', t'), x' \in [x - N_x, x + N_x], y \in [y - N_y, y + N_y], t' \in [t - N_t, t + N_t]\}$. For a gray image sequence $I(\mathbf{p})$, the local three-dimensional Discrete Cosine Transform (DCT) centering on pixel $\mathbf{p}$ is defined as

$$S(u, v, \tau) = c(u; N_x)c(v; N_y)c(\tau; N_t)S'(u, v, \tau), \quad (1)$$

This work was supported by the National Natural Science Foundation of China under Grant No. 61232013, No. 61271434, No. 61175115.

$$S'(u,v,\tau) = \sum_{x=0}^{2N_x}\sum_{y=0}^{2N_y}\sum_{t=0}^{2N_t} I(x,y,t)\phi(u,v,\tau,x,y,t), \quad (2)$$

$$\phi(u,v,\tau,x,y,t) = \psi(u,x;N_x)\psi(v,y;N_y)\psi(\tau,t;N_t), \quad (3)$$

$$\psi(u',x';L) = \cos[\frac{\pi(2x'+1)u'}{2(2L+1)}], \quad (4)$$

$$c(u';L) = \begin{cases} \sqrt{\frac{1}{L}}, & u' = 0 \\ \sqrt{\frac{2}{L}}, & u' = 1,2,\cdots,L-1 \end{cases}, \quad (5)$$

where u, v, τ denote the frequency variables corresponding to x, y, t. Thus, for each pixel (x, y) on frame t, there is a corresponding spatiotemporal spectrum $S(u, v, \tau)$ to represent it. In our system , $N_x = N_y = N_t = 2$, i.e, the size of the spectrum cuboid is $5 \times 5 \times 5$. To explore the energy distribution on different frequency axes, we sample about 3×10^5 spatiotemporal volumes and calculate the mean energy on each 3D DCT coefficients. Just as shown in Fig.1(b)(c), our experimental results show that most of the energy concentrates on the spectrum slice indexed by $\tau = 0$, and further almost 99.8% energy is captured by the first 20 DCT coefficients according to the snake order [12]. Thus, we only use the first 20 spectral coefficients $S_{\mathbf{p}}(k)$,$k = 0, 1, \cdots, 19$, to represent the spatiotemporal feature of point $\mathbf{p}$ instead of 125 coefficients, i.e., $\mathbf{v}(\mathbf{p})$=$[S_{\mathbf{p}}(0), S_{\mathbf{p}}(1), \cdots, S_{\mathbf{p}}(19)]$, which can effectively reduce the complexity of background model.

Given a video sequence, the initial background model can be established. Concretely, w spatiotemporal points can be uniformly sampled along the temporal axis at pixel location (x, y), and the corresponding spectral representations $\mathbf{v}(x, y, t_i)$, $i = 1, \cdots, w$, are computed. Let $\mathbf{v}_i(x, y)$ denote $\mathbf{v}(x, y, t_i)$, and the initial background model of point $\mathbf{p}$ in spatiotemporal DCT domain can be represented as

$$B_0(\mathbf{p}) = \{\mathbf{v}_i(x,y) | i = 1, \cdots, w\} \quad (6)$$

In order to adapt the dynamic background, it is crucial to select the appropriate parameter N_t, which is related to the period of background repetitive motion. If the period of repetition motion (*i.e.* fast background motion) is short in an application, a small value for N_t is suitable; otherwise, a large value for N_t is a good choice (*i.e.* slow background motion).

3. FOREGROUND OBJECT DETECTION AND BACKGROUND UPDATE

In our approach, the background is represented as the set of atoms of dictionary in the sparse model, and the foreground is taken as an outlier with respect to the dictionary. Then, foreground pixels are identified through analyzing the sparse solution of an optimization problem about the sparse model. Further, the atoms of the dictionary as background model are updated based on their importance. The concrete technical details are elaborated in the following subsections.

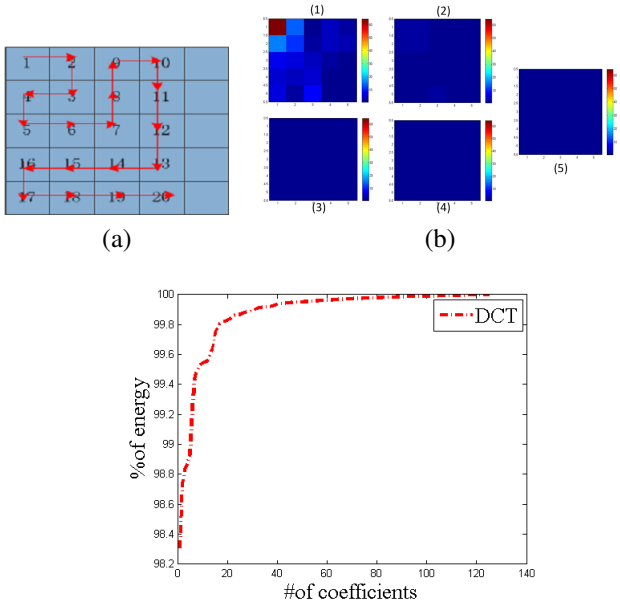

Fig. 1. Graphical representations of the spatiotemporal spectrum: (a) snake order; (b) the slicing spectral energy of DCT coefficients; (c)The percentage of average accumulated spectral energy of 3D DCT coefficients computed on about 3×10^5 spatiotemporal volumes

3.1. Sparse Dictionary Learning

In the sparse dictionary learning problem, given a group of instances $\mathbf{y}_i, i = 1, 2, \cdots, n$, a compact dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m]$ is sought to sparsely represent each instance, i.e., $\mathbf{y}_i = \sum_{j=1}^{m} x_{j,i}\mathbf{d}_j = \mathbf{D}\mathbf{x}_i$, and most entries $x_{j,i}$ of column vector $\mathbf{x_i}$ are zeros. Formally, it can be expressed as an optimization problem with the objective function,

$$\min_{\mathbf{D},\mathbf{X}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \quad s.t. \quad ||\mathbf{x}_i||_0 \le k, \forall i \quad (7)$$

where $||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 = \sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{D}\mathbf{x}_i||_2^2$, matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ and $\mathbf{x}_i$ is the coefficient vector of $\mathbf{y}_i$ with respect to dictionary $\mathbf{D}$. $||\mathbf{x}_i||_0$ denotes the number of nonzero elements of $\mathbf{x}_i$. Since the ℓ_0-norm is a non-convex NP-hard problem, the optimal solution are generally obtained through solving the following relaxed problem,

$$\min_{\mathbf{D},\mathbf{X}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 + \lambda||\mathbf{X}||_1 \quad (8)$$

where $||\mathbf{X}||_1 = \sum_{i,j} |x_{j,i}|$. Apparently, the ℓ_1-norm replaces ℓ_0-norm.

3.2. Identification of Foreground Pixels

For a given pixel location $\mathbf{q} = (x, y)$, let $B_t(\mathbf{q}) = \{\mathbf{b}_i(x, y)|i = 1, 2, \cdots, w\}$ denote the corresponding background model at time t, and each $\mathbf{b}_i(x, y)$ is a spectral feature vector computed according to Eq.(1). Then, for a coming frame $\mathbf{I}_t$, the problem is how to reliably identify foreground pixels in it based on the background model.

In the proposed approach, the sparse dictionary model is exploited to model the current frame $\mathbf{I}_t$ as the linear combination of the atoms of dictionary $\mathbf{D}_t$ and foreground image $\mathbf{f}_t$, i.e.,

$$\mathbf{I}_t = \mathbf{D}_t\mathbf{x}_t + \mathbf{f}_t \tag{9}$$

where $\mathbf{x}_t \in \mathbb{R}^m$ ($||\mathbf{x}_t||_0 \ll m$) is a sparse vector, and $\mathbf{I}_t, \mathbf{f}_t$ are column vectors. Specifically, in our system, for pixel location $\mathbf{q} = (x, y)$ at time t, the corresponding spatiotemporal spectrum $\mathbf{S}_t(\mathbf{q})$ can be written as

$$\mathbf{S}_t(\mathbf{q}) = \mathbf{D}_t(\mathbf{q})\mathbf{x}_t(\mathbf{q}) + \mathbf{S}_t^f(\mathbf{q}), \tag{10}$$

where $\mathbf{D}_t(\mathbf{q}) = [\mathbf{b}_1(\mathbf{q}), \cdots, \mathbf{b}_w(\mathbf{q})]$, $\mathbf{S}_t^f(\mathbf{q})$ denotes the spatiotemporal spectrum for foreground, and $\mathbf{x}_t(\mathbf{q}) \in \mathbb{R}^w$ ($||\mathbf{x}_t(\mathbf{q})||_0 \ll w$). Ideally, if the pixel $\mathbf{q}$ corresponds to foreground, it cannot sparsely be represented by the atoms of background dictionary, so $\mathbf{x}_t(\mathbf{q}) = \mathbf{0}$ according to Eq.(10); otherwise, we have $\mathbf{S}_t^f(\mathbf{q}) = \mathbf{0}$ instead. In practice, although we cannot obtain the ideal solution generally, it is true that the spectrum feature of foreground pixels is incoherent with that of background pixels, and can be regarded as an outlier of the collection of background pixels. Now, the problem of identifying foreground pixels is converted into that of judging whether the current pixel is the outlier of background dictionary.

Inspired by EIhamifar *et al.*'s work [13], we combine the current background model $\mathbf{D}_t(\mathbf{q})$ and the spectrum of current pixel $\mathbf{S}_t(\mathbf{q})$ together to construct a new matrix $\mathbf{Y} = [\mathbf{b}_1(\mathbf{q}), \cdots, \mathbf{b}_w(\mathbf{q}), \mathbf{S}_t(\mathbf{q})]$. Further, we formulate an optimal problem through modifying Eq.(8) to judge whether the spectrum $\mathbf{S}_t(\mathbf{q})$ of current pixel is an outlier with respect to the background model $\mathbf{D}_t(\mathbf{q})$, i.e.,

$$\min_{\mathbf{X}} ||\mathbf{Y} - \mathbf{Y}\mathbf{X}||_F^2 + \lambda||\mathbf{X}||_{1,2} \tag{11}$$

where $||\mathbf{X}||_{1,2} = \sum_{j=1}^{w+1} ||x^j||_2$, and here x^j denotes the j-th row vector of coefficient matrix $\mathbf{X}$, which can minimize the number of non-zeros rows of $\mathbf{X}$. To make the data invariant with respect to the global translation, an affine constraint $\mathbf{1}^T\mathbf{X} = \mathbf{1}^T$ is append on the Eq. (11) as [13]. Compared with Eq.(8), we use $\mathbf{Y}$ to replace dictionary $\mathbf{D}$. If pixel $\mathbf{q}$ corresponds to a foreground pixel, it prefers to write itself as itself; otherwise, it is represented by an affine combination of some atoms of background dictionary $\mathbf{D}$. In other words, if the pixel $\mathbf{q}$ belongs to the foreground, the corresponding row of $\mathbf{X}$ should have very few nonzero entries. Thus, we define a row-sparsity metric for the atom corresponding to current pixel, i.e.,

$$\eta = \frac{(w+1)||x^{w+1}||_\infty - ||x^{w+1}||_1}{w||x^{w+1}||_1} \in [0, 1], \tag{12}$$

to compute the probability of pixel $\mathbf{q}$ as a foreground pixel. For a foreground pixel, the η value is close to 1, and for a background pixel, the η value is close to 0. Hence, if the η value of pixel $\mathbf{q}$ is larger than a predefining threshold ε, it is classified as foreground.

Our experiments show that a fixed threshold cannot obtain the good adaptation to the variance of scenes. Thus, an adaptive threshold is used in our system and computed by $\varepsilon = \mu + \kappa\sigma$, where μ, σ are the average and the standard variance of η values of all the pixels in the current frame, and parameter κ in our system is 1.8.

3.3. Update of Background model

Since the scenes often change slowly over time, it is necessary to adaptively update background models. For pixel $\mathbf{q}$ in the current frame, if it is marked as foreground, its background model $\mathbf{D}_t(\mathbf{q})$ keeps unchanged; Otherwise, if it is consecutively marked as background during N_t frames, the background model needs to be updated through replacing the atom with the lowest efficacy by the spatiotemporal spectrum $\mathbf{S}_t(\mathbf{q})$ of current frame. Concretely, the efficacy of each atom in the current model is evaluated according to

$$\rho(j) = \sum_{i=1}^{w+1} |x_{j,i}| - |x_{j,j}| \quad j = 1, 2, \cdots, w \tag{13}$$

where $x_{j,i}$ denotes the entry of matrix $\mathbf{X}$ at the j-th row and the i-th column. Then, the index k of the atom with the lowest efficacy is determined by $k = \arg\min_j \rho(j)$, and the corresponding atom $\mathbf{b}_k(\mathbf{q})$ is replaced.

4. EXPERIMENTAL RESULTS

To evaluate the performance of our method, the experiments are performed on four challenging sequences from three datasets publicly available, i.e., I2R, Wallflower, and Monnet's dataset. All of the video sequences contain complex scene changes. The frame size of the video sequences are all normalized into 120×180. All the experiments run on the PC with a 3.1GHz CPU in the Matlab environment. Since some of the datasets have no ground truth of foreground objects in the frame, we label some ground- truths manually using the PhotoShop in order to compare the performance of the proposed method with other representative algorithms, including GMM[2, 14], Ali'method[11], Cui'method[15], kernel density estimation(KDE)[16]. For the GMM and KDE methods, we adopt the default parameters in OpenCV2.2 and the parameter setting in the paper[17] respectively. Ali's method

Methods	Campus			Waving Trees			Curtain			Water Surface			Average		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
MOG	.324	.930	.481	**.925**	.488	.639	.770	.414	.538	.070	.441	.121	.522	.568	.544
KDE	.693	**.961**	.805	.845	.895	.869	**.899**	.885	.892	.920	**.986**	**.952**	.839	**.932**	.883
Ali's method	.235	.558	.331	.662	.793	.722	.818	.663	.732	.716	.720	.718	.608	.684	.644
FMSD	.623	.625	.624	.158	.630	.253	.198	.682	.307	**.958**	.346	.508	.484	.571	.524
Our method	**.973**	.753	**.849**	.866	**.983**	**.921**	.830	**.961**	**.891**	.885	.972	.926	**.889**	.917	**.903**

Table 1. Quantitative comparison of several algorithms on the four datasets

and ours both use the $5 \times 5 \times 5$ spatiotemporal volume to extract spatiotemporal spectrum features.

4.1. Qualitative Comparison

First, we qualitatively compare the performance of the four methods mentioned above with ours on the four challenging sequences containing dynamic scenes, as shown in Fig.2. We can see that the GMM method can adapt to illumination

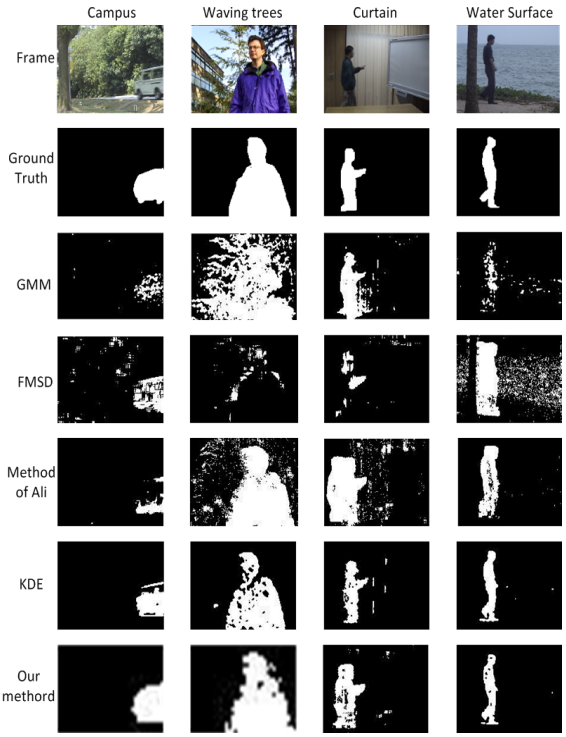

Fig. 2. Qualitative comparison of several algorithms on the four datasets

changes to some extent, but the heavy missing detections exist for high dynamic scenes (e.g.,Campus, Water surface). We implemented Cui's method(FMSD), but the experimental results are not as good as those given by the authors, since lots of severe missing detections and false detections exist. Ali's method can handle the dynamic background at a certain degree, but it needs to adjust the scale size of spatiotemporal volumes according to the motion period of background. Otherwise, it can induce some false detections (e.g., Waving tree, Curtain). Thus, it affects the practical application of the method. Overall, the proposed method achieves the best performance, and it can adapt to high dynamic scenes and illumination variance. Although no postprocessing like morphological operations is used, the least noisy pixels exist with respect to the other four method.

4.2. Quantitative Comparison

We perform the quantitative comparison between the mentioned four methods and ours based on precision (P), recall(R), and F-score (F), $F = \frac{2 \cdot R \cdot P}{R+P}$. The recall is the ratio of the number of foreground pixels correctly segmented to the number of the ground truth, and the precision is the ratio of the number of foreground pixels correctly segmented to the number of foreground pixels outputted by the system. The final experimental results are summarized in Table 1. As shown in the Table 1, the best F-scores of MoG and FMSD have only about 63% on the four video sequences. Additionally, Ali's method has the poor performance for Campus sequence where the illumination varied much, since their method does not update the background model over time. But, for other sequences with no large illumination changes, the performance improves significantly, which is attributed to the adoption of spatiotemporal information. Both our method and KDE have very good performance, and the average F-Score of ours is 0.903 which is a little better than that of KDE (0.883) on the four sequences.

5. CONCLUSIONS

In this paper, we present a novel approach to extract foreground objects in surveillance videos. The proposed method exploits the spatiotemporal spectrum in 3D DCT domain to represent background, and then identify foreground pixels through solving an optimal problem. The experimental results show that our approach is robust to dramatic change of scenes, and it can obtain more complete and accurate foreground objects. But, our method has the high computation complexity compared with FMSD and GMM, because it involves lots of computation on solving the optimization problem.

6. REFERENCES

[1] W.Z.Hu, H.F.Gong, S.C.Zhu, and Y.T.Wang, "An integrated background model for video surveillance based on primal sketch and 3d scene geometry," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[2] C. Stauffer and WEL. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, vol. 2, pp. 246–252.

[3] Anurag Mittal, Antoine Monnet, and Nikos Paragios, "Scene modeling and change detection in dynamic scenes: A subspace approach," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 63–79, 2009.

[4] Weiqiang Wang, Jie Yang, and Wen Gao, "Modeling background and segmenting moving objects from compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 670–680, 2008.

[5] L. Wixson, "Detecting salient motion by accumulating derectionally sonsistent flow," *Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 774–780, 2000.

[6] Gerald Dalley, Joshua Migdal, and W. Eric L. Grimson, "Background subtraction for temporally irregular dynamic textures," *Workshop on Applications of Computer Vision*, 2008.

[7] Gianfranco Doretto, Alessandro Chiuso, Yingnian Wu, and Stefano Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 52, no. 1, pp. 91–109, 2003.

[8] Yair Moshe, Hagit Her-or, and Yacov Hel-Or, "Foreground detection using spatiotemporal projection kernels," *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[9] Youdong Zhao, Haifeng Gong, Liang Lin, and Yunde Jia, "Spatio-temporal pathes for night background modeling by subspace learning," *In International Conference on Pattern Recognition*, 2008.

[10] Atsushi Shimada and Rin ichiro Taniguchi, "Hybrid background model using spatial-temporal lbp," *Advanced Video and Signal Based Surveillance*, 2009.

[11] Imitiaz Ali, Julien Mille, and Laure Tougne, "Space-time spectral model for object detection in dynamic textured background," *Pattern Recognition Letters*, vol. 33, pp. 1710–1716, 2012.

[12] Yair Moshe and Hagit Hel-Or, "Video block motion estimation based on gray-code-kernels," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2243–2254, 2009.

[13] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal, "See all by looking at a few:sparse modeling for finding representative objects," *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[14] Jodoin P, Mignotte M, and Konrad J, "Statistical background subtraction methods using spatial cues," *IEEE Trans on Circuits and Systems for Video Technology*, vol. 17, no. 12, pp. 1758–1764, 2007.

[15] Xinyi Cui, Qingshan Liu, and Dimitris Metaxas, "Temporal spectral residual: Fast motion saliency detection," in *ACM,Multimedia*, 2009.

[16] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction.," *In European Conference on Computer Vision*, pp. 751–767, 2000.

[17] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1778–1792, 2005.