A Novel Bottom-Up Saliency Detection Method for Video With Dynamic Background

Chenglizhao Chen, Yunxiao Li, Shuai Li, Hong Qin, and Aimin Hao

***Abstract*—After years of extensive studies, the salient motion detection problem has gained plausible performance improvement that was primarily propelled by the rapid development of self-adaptive top-down modeling techniques. Nevertheless, almost all the conventional solutions are still not robust enough to handle video sequences captured by hand-hold cameras. This is mainly due to the absence of the position alignment information that is indispensable for top-down background modeling. In contrast, the bottom-up video saliency detection methods, though achieving excellent salient motion detection in either stationary or nonstationary videos, still have rather poor detection performance in scenarios with massive dynamic background. In this letter, we explore a bottom-up saliency framework by introducing a novel spatial-temporal regional filter method to handle the dynamic background problem. Our key rationale is to assign large saliency value to those regions with stable spatial-temporal coherency while eliminating irregular, repeating dynamic background. As far as we know, this is the first work to address the dynamic background problem from the perspective of the bottom-up video saliency. We conduct massive quantitative evaluations over public available benchmarks to validate the effectiveness and robustness of our method.**

***Index Terms*—Dynamic background, nonstationary video, spatial-temporal coherency, spatial-temporal regional filter, video saliency.**

I. Introduction and Motivation

THE objective of salient motion detection is to locate the most eye-attracting regions in a given scenario, which frequently contains salient objects undergoing distinct movement. And the detected salient regions can be applied in various downstream computer vision applications, including object tracking [1], video segmentation and impression [2]–[4], image fusion [5], traffic surveillance [6], etc. Although considerable efforts have been endeavored in the top-down background modeling to perform subtraction-based salient motion detection [7]–[9], it is still difficult to obtain robust detection for nonstationary videos due to the absence of pixel alignment information, which is indispensable for the robust background modeling. Different from the top-down modeling based detections [10], [11], the recently developed video saliency methods [12]–[15] commonly integrate individually estimated low-level saliency clues into the high-level spatial-temporal saliency prediction by introducing certain bottom-up strategies. To estimate low-level saliency clues, multiscale patch-like features are individually extracted in both the spatial scope and the temporal scale, which can be roughly categorized into spatial clues (e.g., color contrast [16] and compactness [17]) and temporal clues (e.g., velocity and acceleration [18]). Thus, the high-level video saliency can be computed via fusion strategies [19], [20], which formulate an complementary combination for those precomputed low-level clues. In addition, the spatial-temporal smoothness [13], [21] are also integrated as a high-level saliency constraint to filter out the false-alarm detections with weak interframe coherency. Since the entire bottom-up detection procedures involved in a video saliency detection method are totally independent of the pixel-wise position alignment information, it can produce much better detection result than the top-down modeling solutions for nonstationary videos. Nonetheless, for video scenarios with massive dynamic backgrounds, which frequently exhibit irregular movements (e.g., fountains, floating leaves, sea waves, etc.), the conventional bottom-up solutions can easily assign large saliency value to those regions containing dynamic backgrounds, due to its inability to distinguish real saliency from meaningless dynamic backgrounds.

In this letter, following the basic bottom-up framework, we devise a newly designed spatial-temporal saliency filter to solve the aforementioned challenges, which can eliminate saliency disturbances from dynamic background while pinpointing the real saliency. The key idea of our method is inspired by two observed common phenomena related to the true saliency and the nonsalient dynamic backgrounds: (1) The spatial-temporal coherency of the true saliency is much stronger than the nonsalient dynamic backgrounds; and (2) the spatial-temporal motion gradients of the salient foreground frequently exhibit better compactness than those of the nonsalient dynamic backgrounds. Therefore, based on the precomputed low-level saliency clues, such as contrast based motion saliency, we make use of an object-aware KNN-histogram (KNNH) to measure the spatial-temporal coherency among consecutive video frames, which substantially outperforms the conventional patch-wise statistics

Manuscript received August 24, 2017; revised November 3, 2017; accepted November 6, 2017. Date of publication November 20, 2017; date of current version December 13, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Edmund Y. Lam. *(Corresponding author: Dr. Shuai Li.)*

C. Chen is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the College of Computer Science and Technology, Qingdao University, Qingdao 266000, China (e-mail: cclz123@163.com).

Y. Li, S. Li, and A. Hao are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: 296367440@qq.com; lishuai@buaa.edu.cn; ham@buaa.edu.cn).

H. Qin is with the Stony Brook University, Stony Brook, NY 11794 USA (e-mail: qin@cs.stonybrook.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2775212

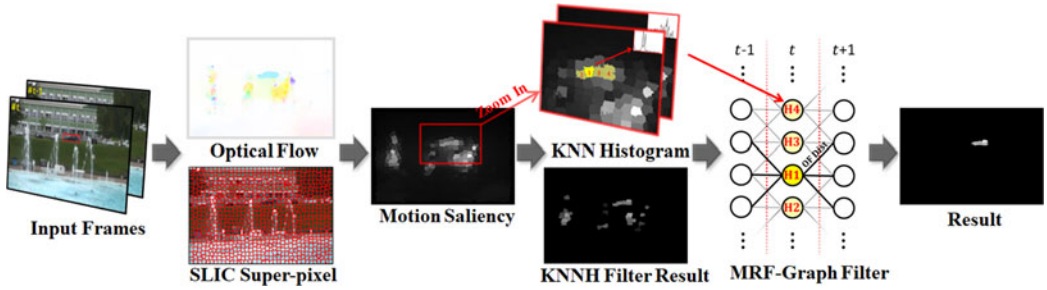

Fig. 1. Architecture overview of our bottom-up video saliency detection method. The ground truth is marked with a red rectangle in the **Input Frames**. Our method mainly consists of two components: the novel KNN-histogram (KNNH) based regional filter and the newly designed MRF-Graph (MRFG) model. As we can see from the filtering result, our method can correctly eliminate the dynamic backgrounds while retaining the true saliency.

based solutions for nonstationary videos. Next, to ensure the spatial-temporal consistency of the filtered saliency, we extend the saliency filtering range by incorporating our newly designed Markov random field graph (MRFG), wherein the graph edges (motion gradient spanned feature distance) reflect the compactness of the corresponding motion distribution. Benefiting from this, we can obtain highly accurate and robust video saliency detection results over various videos, and the quantitative evaluations confirm the effectiveness and robustness of our method.

II. Spatial-Temporal Saliency Filter

A. Method Overview

As shown in Fig. 1, we first utilize the RF [22] method to perform edge-preserved smoothing, and then adopt the SLIC [23] based superpixel decomposition method to reduce the computational burden. Meanwhile, we employ optical flow [24] to detect the motion clues in each video frame. Consequently, we can obtain the low-level saliency clue (i.e., the motion saliency **MS**) via global contrast computation, which is similar to our previous work [21]. Obviously, from the computed motion saliency in Fig. 1, massive dynamic backgrounds (i.e., fountains) may be falsely detected as the salient foregrounds due to the distinct optical flow estimation in the spatial-temporal scope. To alleviate this limitation, we propose to use the KNNH (see Section II-B) to measure the regional coherency among consecutive video frames; thus the preobtained low-level saliency clues are adjusted to bias toward those regions with strong inter-frame coherency. After that, the KNNH-filtered results are integrated into our newly designed MRFG model (see Section II-C), wherein the motion gradient spanned feature distance serves as the graph edge to represent the compactness of the gradient distribution. Therefore, those regions with either irregular saliency distribution or distinct motion flows can be easily filtered out in a spatial-temporal manner.

B. Short-Term Regional Saliency Filter

Given a smoothed input video frame $\mathbf{I}$ with n superpixel $(p_i, i \in [1, n])$, the ith superpixel motion saliency (**MS**) can be obtained via the global contrast computation (1) over the precomputed optical flow gradient $f = [vx, vy] \in \mathbb{R}^{1\times 2}$, wherein vx and vy, respectively, denotes the averaged horizontal and vertical gradients. In the rest of this letter, we use $|\cdot|$ and $||\cdot||_2$ to, respectively, denote absolute operation and l_2-norm.

$$\mathbf{MS}_i = \frac{1}{Z}\sum_{j=1}^{n} ||f_i, f_j||_2, Z = \sqrt{\sum\nolimits_{i=1}^{n}(\sum\nolimits_{j=1}^{n} ||f_i, f_j||_2)^2}. \tag{1}$$

Here, Z denotes the normalization factor. Due to the varying flows between consecutive video frames, large saliency values can be easily found in regions containing dynamic backgrounds. Thus, we propose to utilize regional metric (KNNH) to measure the variation degree for the compression of the dynamic backgrounds while highlighting the remaining true saliency. That is, for each superpixel, we formulate its corresponding KNN structure over the **MS** spanned saliency space, and then explicitly represent it based on conventional histogram statistics $(h_i \in \mathbb{R}^{1\times 1000})$ as

$$h_i = \frac{1}{Z}H(\mathbf{I} \odot \mathbf{M}), \quad Z = \sqrt{\sum\nolimits_{l}^{1000} b_l^2}. \tag{2}$$

Here, H denotes the histogram, $\odot$ is the element-wise Hadamard product, $\mathbf{M} \in \{0, 1\}$ is the binary indicator matrix, formulated in (3), wherein ϕ controls the KNN searching distance and the parameter ρ controls the KNN strength. Since both ϕ and ρ belong to pair-wise tradeoff parameters, we empirically set $\rho = 0.1$ while quantitatively selecting the optimal ϕ (see details in Section III).

$$\mathbf{M}_i = \begin{cases} 1, & \text{if } |\mathbf{MS}_i, \mathbf{MS}_{j,p_j\in\phi}| \le \rho \times \mathbf{MS}_i \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

Since the KNN structure of the salient foregrounds tend to exhibit strong spatial-temporal coherency while the dynamic backgrounds are the opposite to that, we can adopt the temporal weighting scheme to the ith superpixel h_i to measure its spatial variation degree V_i in the temporal scale. Hence, we formulate the temporal weighting scheme as

$$V_i^t = \frac{\sum_{j\in\phi} ||h_i^t, h_j^{t+/-1}||_2 \cdot exp(-w \cdot ||c_i^t, c_j^{t+/-1}||_2)}{\sum_{j\in\phi} \exp(-w \cdot ||c_i^t, c_j^{t+/-1}||_2)}. \tag{4}$$

Here, the superscript t indicates the frame index, $c \in \mathbb{R}^{1\times 3}$ denotes the RGB color information, and we set the weighting strength $w = 25$ according to our previous works [21], [25]. Hence, the computed variation degree V can be regarded as an indicator to further compress the saliency degree of the dynamic

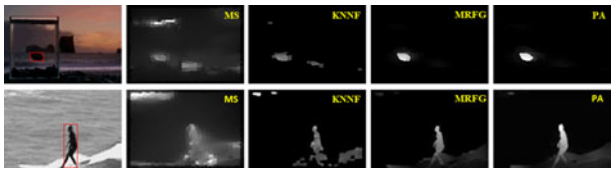

Fig. 2. Demonstrations of our intermediate results. The first column shows the source input frame, the second column demonstrates the raw motion saliency result, the third column demonstrates the saliency filtering results based on our KNNH, the fourth column demonstrates the filtering results by introducing our MRFG model, and the last column demonstrates the pixel assignment results.

TABLE I
QUANTITATIVE EVALUATION FOR DIFFERENT CHOICES OF ϕ

$\phi : \|\|p_i, p_j\|\|_2$	≤ 25	≤ 30	≤ 35	≤ 40	≤ 45
F_ Measure	.724	**.730**	.730	.714	.706

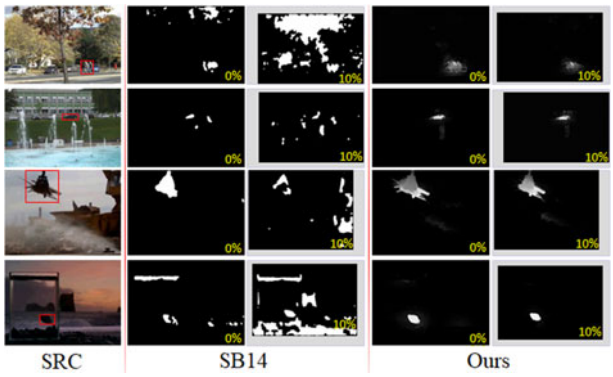

Fig. 3. Performance comparisons between the conventional top-down modeling method SB14 [8] and our bottom-up solution for the dynamic background cases with 10% camera jitter.

backgrounds via

$$\mathbf{S} \leftarrow \mathbf{MS} \odot \exp\left(-N\left(V - \frac{\alpha}{n}\sum_{i=1}^{n} V_i\right)_+\right) \tag{5}$$

where $\mathbf{S}$ denotes the filtered saliency result by KNNH, N denotes the minmax normalization function. In practice, due to the large feature margin between the true saliency and the nonsalient dynamic backgrounds, the filter performance is insensitive to the choices of threshold parameter α, and we empirically set $\alpha = 2$. So far, a substantial portion of the dynamic backgrounds, which exhibit weak spatial-temporal coherency, have been correctly compressed (see **KNNH Filter Result** in Fig. 1). The remaining dynamic backgrounds related false-alarm detections (see the third column of Fig. 2) are mainly caused by the temporarily strong coherency in a limited number of frames. Therefore, we propose to extend our local KNNH filter to the long-term spatial-temporal scope to further compress the remaining nonsalient dynamic backgrounds.

C. Long-Term Spatial-Temporal Saliency Filter

In general, since the movement direction of the salient foregrounds frequently exhibits compact spatial distribution and strong temporal coherency, we can utilize the MRFG to enhance those regions with consistent spatial-temporal movements in the long-term scope. In particular, we propose to use the feature distance over the optical flow gradients spanned feature space as the MRFG edge to measure the motion similarity, while directly employing the KNNH filter results $\mathbf{S}$ as the MRFG nodes, see the pictorial demonstration **OF Dist** in Fig. 1, wherein the graph structure follows the implementation of [26] while omitting the intraframe spatial connections to bias our filter toward the temporal direction. Thus, our long-term saliency filter revealing can be formulated as the following binary assignment ($\mathbf{B} \in \{0, 1\}$) problem:

$$\min_{\mathbf{B}} \sum_i u(\mathbf{B}_i) + \lambda \sum_{j \in \phi} |\mathbf{B}_i - \mathbf{B}_j| \cdot \exp(-w \cdot ||f_i, f_j||_2). \tag{6}$$

Here, u denotes the unary function, which can be formulated as (7). The parameters ϕ and w are identical to those in (4). The MRF performance tradeoff parameter λ controls the bias tendency toward the smooth term (i.e., the second part of (6)), and we empirically assign it to 1.5:

$$u(\mathbf{B}_i) = \begin{cases} 1, & \text{if } (\mathbf{B}_i \cdot \mathbf{S}_i) \geq 2 \times \text{std}(\mathbf{S})^2 \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

From the perspective of long-term temporal scale, those dynamic backgrounds with weak movement coherency can be effectively revealed by (6) and then be compressed by

$$\mathbf{S} \leftarrow 0.7 \times \mathbf{B} \odot \mathbf{S} + 0.3 \times \mathbf{S}. \tag{8}$$

After the above spatial-temporal saliency filtering, we utilize the pixel-wise saliency assignment strategy to further sharpen boundaries of the remaining salient foregrounds, which follows the conventional spatial weighting scheme with 20×20 rectangle mask. The corresponding intermediate results of our method can be found in Fig. 2.

Now, it sets the stage to summarize the key advantages of our MRF saliency filter in two aspects: (1) Our MRF saliency filter can simultaneously measure temporal coherency and spatial variation towards automatically pinpointing the salient foregrounds while compressing those nonsalient backgrounds; such filter can remarkably outperform the conventional fusion based methods; and (2) our graph-based saliency filter enables the filtering procedure to retain spatial-temporal smoothness, which can utilize the beyond-scope information to facilitate the current detection.

III. EXPERIMENTAL RESULTS

We quantitatively evaluate the performance of our method over CD2014 dataset (dynamic-background category [31]) and six additional video sequences (YouTube) with massive dynamic backgrounds. The ground truth of our newly adopted sequences is obtained in the same manner as [32]. Since the detection performance of our method is heavily dependent on the constructed KNN structure (see Section II-B), we quantitatively evaluate the parameter ϕ to obtain the optimal choice, and the detailed results can be found in Table I, wherein we select $\phi : ||p_i, p_j||_2 \leq 30$ as the optimal choice. To demonstrate the unique advantage of our method, we compare our method with 11 state-of-the-art methods, including FD17 [21], GF15 [13],

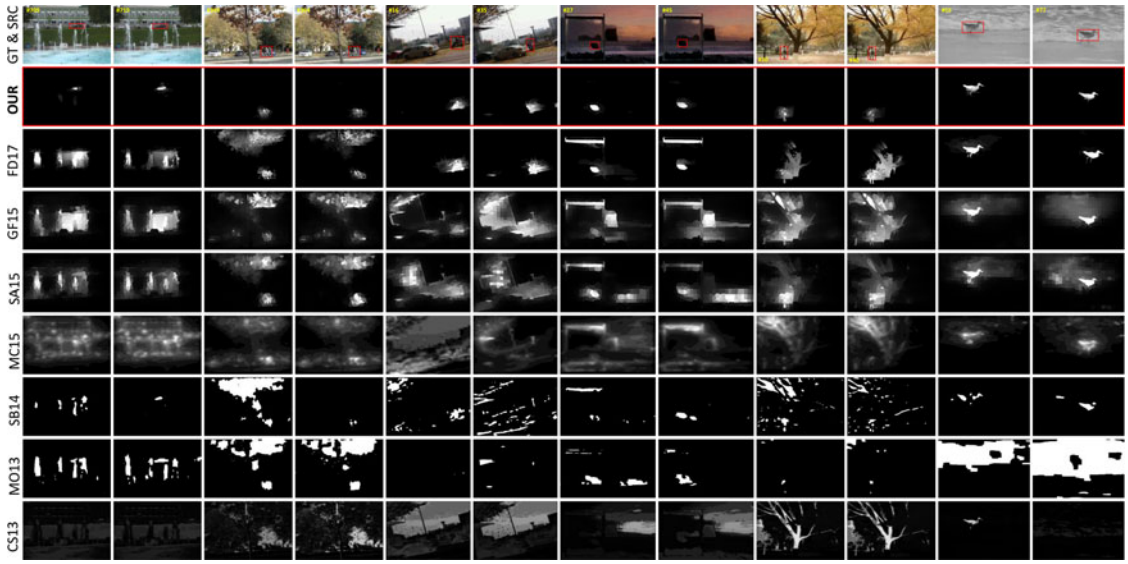

Fig. 4. Qualitative comparisons between our method and seven state-of-the-art methods, including FD17 [21], GF15 [13], SA15 [26], MC15 [27], SB14 [8], MO13 [7], and CS13 [16], wherein the ground truth is marked with rectangle in the first row.

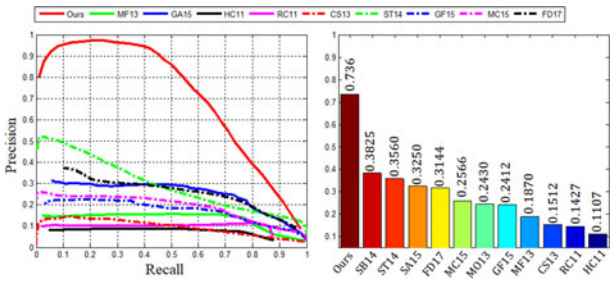

Fig. 5. Quantitative comparisons between our method and 11 state-of-the-art methods, including FD17 [21], GF15 [13], SA15 [26], MC15 [27], MF13 [28], SB14 [8], ST14 [18], MO13 [7], CS13 [16], RC11 [29], and HC11 [29], wherein the left is the precision-recall curve and the right part demonstrates the averaged F-measure (with bias term 0.3 [30]). It should be noted that there are no PR curves for SB14 and MO13, because their outputs are binary segmentations.

SA15 [26], MC15 [27], SB14 [8], ST14 [18], MF13 [28], MO13 [7], CS13 [16], RC11 [29], and HC11 [29], and the detailed quantitative results are documented in Fig. 5. We also demonstrate the qualitative comparisons in Fig. 4. It should be noted that both MO13 and SB14 are among the top-down modeling based methods while the remaining methods all belong to the bottom-up modeling based methods. Since the conventional top-down methods can well handle the dynamic background problem in stationary videos, our method only slightly outperforms SB14 and MO13 over CD2014 dataset; however, our method exhibits much better robustness over the six newly obtained nonstationary sequences. Moreover, the robustness of our method can be further confirmed in the camera jitter experiment, wherein we disturb the original sequences with 10% random vibrations, and results can be found in Fig. 3 and Table II. As for the remaining bottom-up methods, due to the lack of a proper mechanismto compress the dynamic backgrounds, all these methods exhibit much poorer performance over the tested video sequences. As for efficiency, since our framework is implemented by using MATLAB (accelerated with CUDA) in an alienware laptop with Quad Core i7-6700HQ 2.6 GHz, 16 GB

TABLE II
PERFORMANCE DEGRADATION DUE TO DISTURBANCE WITH CAMERA JITTER

Method	Camera Jitter 0%			Camera Jitter 10%		
	Pre-	Rec-	F-	Pre-	Rec-	F-
MO13 [7]	.203	.673	.243	*.160*	**.216**	.170
SB14 [8]	.384	.375	.382	**.277**	*.340*	**.290**
FD17 [21]	.291	.426	.314	.308	.317	.320
GF15 [13]	.220	.352	.241	.224	.313	.242
SA15 [26]	.292	.525	.325	.233	.565	*.269*
MC15 [27]	.231	.400	.256	.215	.381	.239
Ours	.896	.462	.736	.907	.433	.724

Bold, *italic*, and normal respectively represents the 1st, 2nd, and 3rd performance degradation.

TABLE III
TIME COST (IN SECONDS) FOR SINGLE VIDEO FRAME

Method	Ours	FD17	SA15	GF15	MC15	SB14	ST14
Cost	*3.28*	3.93	6.38	6.26	14.3	**3.09**	24.3

Bold, *italic*, and normal respectively represents the 1st, 2nd, and 3rd performance.

RAM, and GTX 970 m, it costs 3.28 s to perform saliency detection for a single video frame (with a size of 300×300), and the time cost comparison with the state-of-the-art video saliency methods are detailed in Table III.

IV. CONCLUSION

In this letter, we proposed a simple yet effective bottom-up method to conduct saliency detection over videos with massive dynamic backgrounds. Our method comprises two key novel technical elements: (1) We exploit regional KNNH to measure the spatial variations between consecutive video frames, which is independent of the position alignment information; and (2) we integrate the spatial variations into the MRFG model to simultaneously respect the long-term temporal coherency. Various experiments have confirmed the effectiveness and robustness of our method. In the near future, we will continue to improve the efficiency of our method, so that it becomes essential in many more applications with real-time requirements.

REFERENCES

[1] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognit.*, vol. 48, no. 9, pp. 2885–2905, 2015.

[2] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2017.2721544.

[3] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, and K. Rapantzikos, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[4] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 4408–4421, Nov. 2015.

[5] Q. Wang, S. Li, H. Qin, and A. Hao, "Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis," *Inform. Fusion*, vol. 26, pp. 103–121, 2015.

[6] J. Hsieh, S. Yu, Y. Chen, and W. Hu, "Automatic traffic surveillance system for vehicle tracking and classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 175–187, Jun. 2006.

[7] 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.

[8] P. StCharles, G. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[9] C. Chen, S. Li, H. Qin, and A. Hao, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognit.*, vol. 52, pp. 410–432, 2016.

[10] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split Gaussian models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 420–424.

[11] Z. Gao, L. Cheong, and Y. Wang, "Block-sparse RPCA for salient motion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1975–1987, Oct. 2014.

[12] W. Kim and J. Han, "Video saliency detection using contrast of spatiotemporal directional coherence," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1250–1254, Oct. 2014.

[13] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[14] C. Aytekin, H. Possegger, T. Mauthner, S. Kiranyaz, H. Bischof, and M. Gabbouj, "Spatiotemporal saliency estimation by spectral foreground detection," *IEEE Trans. Multimedia*, to be published, 10.1109/TMM.2017.2713982.

[15] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1063–1069.

[16] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[17] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.

[18] F. Zhou, S. Kang, and F. Michael, "Time-mapping using space-time saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3358–3365.

[19] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.

[20] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[21] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[22] E. Gastal and M. Olive, "Domain transform for edge-aware image and video processing," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–12, 2011.

[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels," École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Tech. Rep., 2010.

[24] C. Liu, "Exploring new representations and applications for motion analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., *Massachusetts Inst. Technol.*, Cambridge, MA, USA, 2009.

[25] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2303–2316, Aug. 2015.

[26] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3395–3402.

[27] H. Kim, Y. Kim, J. Sim, and C. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015.

[28] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.

[29] M. Cheng, G. Zhang, J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.

[30] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[31] Y. Wang, P. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 393–400.

[32] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.