Video Saliency Detection via Spatial-Temporal Fusion and Low-rank Coherency Diffusion

Chenglizhao Chen[1] Shuai Li[1*] Yongguang Wang[1] Hong Qin[2] Aimin Hao[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2]Stony Brook University

***Abstract*—This paper advocates a novel video saliency detection method based on the spatial-temporal saliency fusion and low-rank coherency guided saliency diffusion. In sharp contrast to the conventional methods, which conduct saliency detection locally in a frame-by-frame way and could easily give rise to incorrect low-level saliency map, in order to overcome the existing difficulties, this paper proposes to fuse the color saliency based on global motion clues in a batch-wise fashion. And we also propose low-rank coherency guided spatial-temporal saliency diffusion to guarantee the temporal smoothness of saliency maps. Meanwhile, a series of saliency boosting strategies are designed to further improve the saliency accuracy. First, the original long-term video sequence is equally segmented into many short-term frame batches, and the motion clues of the individual video batch are integrated and diffused temporally to facilitate the computation of color saliency. Then, based on the obtained saliency clues, inter-batch saliency priors are modeled to guide the low-level saliency fusion. After that, both the raw color information and the fused low-level saliency are regarded as the low-rank coherency clues, which are employed to guide the spatial-temporal saliency diffusion with the help of an additional permutation matrix serving as the alternative rank selection strategy. Thus, it could guarantee the robustness of the saliency map's temporal consistence, and further boost the accuracy of the computed saliency map. Moreover, we conduct extensive experiments on 5 public available benchmarks, and make comprehensive, quantitative evaluations between our method and 16 state-of-the-art techniques. All the results demonstrate the superiority of our method in accuracy, reliability, robustness, and versatility.**

***Index Terms*—Spatial-temporal Saliency Fusion, Low-rank Coherency Guided Saliency Diffusion, Video Saliency, Visual Saliency.**

I. Introduction and Motivation

THE detection of video saliency aims to locate the most eye attractor in a given video sequence, which is extremely valuable in many downstream applications, such as video segmentation [1], video object tracking [2], and video expression [3]. Different from image saliency detection, which has already achieved great success in recent years, video saliency is a relatively new topic. Compared to image saliency detection over spatial domain only, the incursion of video motion information is the critical factor to make this task challenging. Currently, how to properly exploit and use the spatial-temporal information has become a recognized research trend in video saliency field. Here, we will provide a brief introduction to the state-of-the-art methods related to video saliency detection.

Corresponding author: Shuai Li, lishuai@buaa.edu.cn

Given a single static image, its saliency is the most conspicuous content that tend to draw human attention. After years of extensive research works, people have reached a consensus that, the rational core of saliency computation is the "contrast" [4]. That is, the more an object is different from its surroundings, the higher saliency degree it should have. Although various salient object detection methods have been proposed in recent years, the main differences among these state-of-the-art methods commonly lie in two aspects: the definition of the feature space [5], [6] and the formulation of the contrast computation [4], [7]. In fact, although the reported detection accuracy has been gradually increased by introducing more complicated and specific saliency mechanism (e.g., priors [8], constraints [9], bionics clues [10], etc.), the severely bad cases (which are completely in contrary to the ground truth, and the proofs could be found in Fig. 14(b)) occur more frequently than ever before (see details in our experimental section). Therefore, instead of naively employing the saliency results of the state-of-the-art methods as low-level saliency clues, a motion clue guided low-level saliency fusion is much more desirable for robust video saliency detection.

The purpose of salient motion detection is to locate the moving object in the given video sequence, which is seemingly similar to video saliency detection. The key rationality of salient motion detection is "modeling", which intends to extract the background appearance and regards the residuals (between the established background model and the current video frame) as the salient motion detection results. In fact, the modeling-driven methods have two-fold effects. First, it requires a long learning/updating period to establish a stable background model, which easily gives rise to poor performance for short-term video sequences. Second, although various regional modeling solutions [11], [12] can be integrated to handle camera movements (e.g., camera jitter), the modeling-based methods seem to be feasible only for stationary videos. Specifically, several low-rank analysis based salient motion detection methods have been proposed in recent years, which can achieve state-of-the-art performance [13], [14], [15]. However, these methods are mainly based on the assumption that the input video sequences will be relatively stationary after various frame-level pre-processing (e.g., affine transformation, background tracking, etc.), and thus it could easily introduce additional errors. Therefore, there has been a strong expectation for a newly-designed low-rank analysis method, so that it can simultaneously accommodate both stationary and non-stationary videos regardless of the video

length.

In fact, different from the top-down salient motion detection methods, the video saliency methods commonly employ the bottom-up image saliency as the basic saliency clues, which can well handle the saliency detection in non-stationary videos. Yet, the detection performances over stationary videos are inferior to those of low-rank analysis based salient motion detection methods [14] [12]. Meanwhile, because the motion information can be regarded as an additional trustful saliency clue to facilitate the video saliency detection, many of the state-of-the-art video saliency methods tend to fuse the color saliency with the motion saliency. However, the fusion procedures adopted by these methods [3], [16], [17] are temporally too local (i.e., in a frame-by-frame manner) to obtain robust low-level saliency, and lack subtle way to solve it when the motion saliency is in contrary to the color saliency. Furthermore, almost all existing video saliency methods neglect the fact that, the obtained saliency map should keep temporal smoothness, which can be leveraged as an important constraint to further boost the detection accuracy. Most recently, some methods [18], [19] take into account even global temporal clues to compute robust low-level saliency, however, the subsequent energy minimization framework, which is designed to exploit the saliency consistency over temporal scale, can easily cause the accumulation error of the incorrect low-level saliency, and thus lead to massive false-alarm detections. Therefore, it is critical to design a proper solution to guarantee the temporal saliency consistence while being robust enough to limit accuracy deterioration.

To tackle the aforementioned limitations, our research endeavors focus on designing a video saliency detection method with excellent performance for both stationary (it should be better than state-of-the-art low-rank analysis based salient motion detection methods) and non-stationary videos. In sharp contrast to the traditional video saliency methods, which employ the state-of-the-art image saliency detection results as basic saliency clues, we reconsider the most straightforward local contrast as the low-level saliency while involving no high-level priors or constraints. And the spatial-temporal gradient map is integrated into color contrast computation to avoid the hollow effect. Meanwhile, the appearance/background modeling, which can be regarded as the temporal-level global clue, is also considered to guide the fusion of color saliency and motion saliency. Specifically, the salient contributions of this paper can be summarized as follows:

- We propose a novel spatial-temporal gradient definition to guide contrast computation, which can assign high saliency value around foreground object but simultaneously avoid the obstinate hollow effects.
- We formulate a series of saliency adjustment strategies to guide the fusion of color saliency and motion saliency, which outperforms the traditional fusion solutions adopted by previous works in terms of both accuracy and robustness.
- We propose to explore the spatial-temporal low-rank coherency to construct the temporal saliency correspondences among cross-frame super-pixels, which can guarantee the temporal smoothness of the resulted video saliency map.
- We leverage the temporal smoothness to further boost the saliency accuracy via "one-to-one" spatial-temporal saliency diffusion based on the constructed temporal saliency correspondences, which works much better than the traditional, unconstrained "many-to-many" cases.

II. Background and Related Works

A. Image Saliency Detection Methods

The central idea of image saliency detection is to extract the most eye-attracting object that is significantly distinctive (i.e., uniqueness) from its non-salient surroundings. To represent the uniqueness, most of the earlier saliency methods directly employ the global contrast as the saliency criterion, either in the raw color feature space [20] or in the frequency domain [21]. Following the same rationality, the improved multi-scale solutions dominate the saliency detection field for a long period of time, which explore the global saliency over the color information spanned feature spaces, including sparse dictionary based method [22], multi-level super-pixel feature based method [23], image boundary based method [5], etc. Although the global contrast based saliency methods have achieved remarkable accuracy, they may easily miss some important sub-parts in the salient object because of the feature (color) overlapping between foreground and background. Then, the local contrast methods are resorted to conquer this limitation, however, it tends to bring in new hollow effect problems, which leave the inside regions of the salient object being undetected but assign high saliency value around the salient object boundaries [24]. The hybrid solutions (considering both local and global contrast) are also proposed to alleviate these limitations [4] by adopting more meaningful feature space [25], [26] and even high discriminative descriptor [27], [6] to perform multi-scale contrast computation [7], [28]. Also, some high-level shape/structure based constraints [9] and priors [8] are introduced to sharpen the boundary of the salient object. Although the state-of-the-art single image saliency detection methods have already achieved great success, they are still struggling to make trade-off between the local contrast and the global contrast. In particular, their detection performance over videos is extremely poor.

B. Salient Motion Detection Methods

Since the salient motion detection method is originally designed for the stationary video surveillance, almost all salient motion detection methods leverage the modeling based framework. From the earliest Gaussian model (e.g., single Gaussian model [29], Gaussian mixture model (GMM) [30], extended GMM model [31]) based methods to current structure topology based modeling solution [11], most of the methods regard the residuals between the established background model and the current video frame as the salient motion clues. Although the Gaussian-like modeling methods can well handle the background variations, their isolated pixel-level modeling method tends to frequently encounter massive false-alarm detections due to their slow-adaptation ability in handling sudden camera movements (e.g., camera jitter,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing

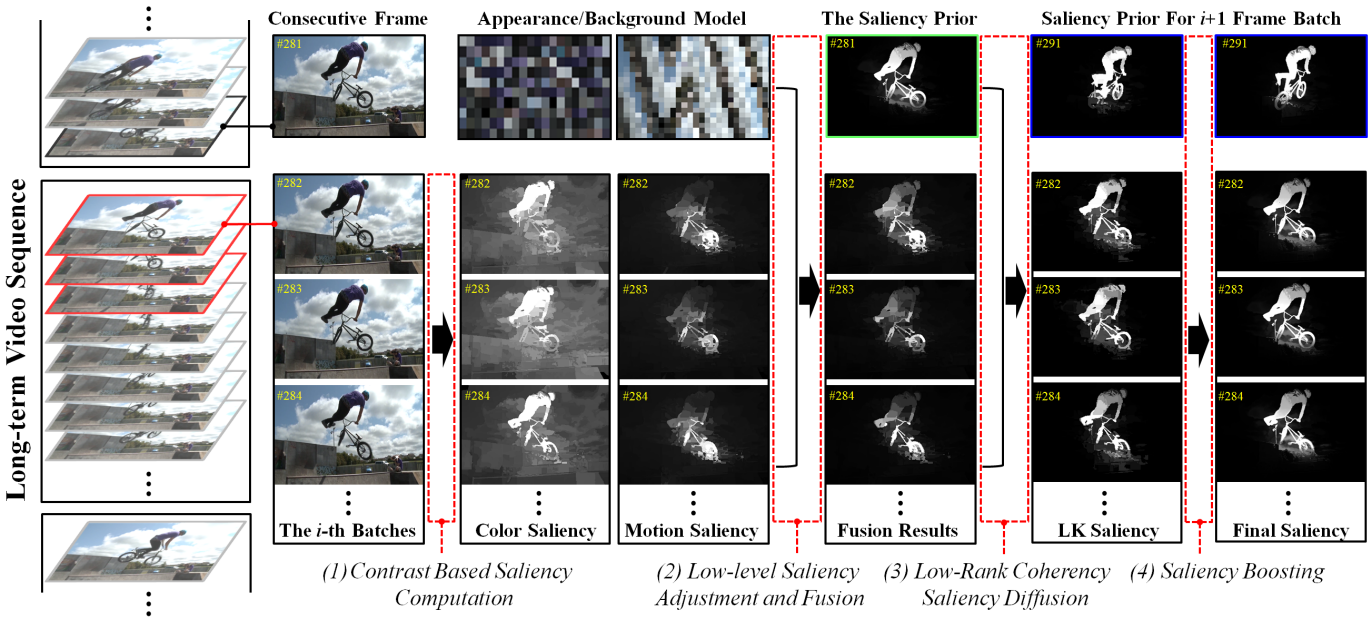


Fig. 1. The architectural overview of our video saliency detection method. The yellow number located in the top-left corner of each sub-image indicates the frame index. "The Current Saliency Prior" (marked with green boundary) is the final saliency detection result ($\mathbf{s}_l$, see Eq. 36) of the last frame in the previous frame batch. And the saliency value of $\mathbf{s}_l$ will be temporally diffused over the entire frame batch according to the established cross-frame super-pixels' low-rank coherency correspondences. Similarly, the saliency detection result of the last frame in the current frame batch will also be diffused over the next frame batch (marked with blue boundary).

fast dynamic backgrounds, etc.). Therefore, texture-sensitive or structure-sensitive [11], [32], [12] feature representations are proposed to enlarge the inter-class feature distance (i.e., the feature distance between the backgrounds and the foreground moving object) while shortening the intra-class (either backgrounds or foregrounds) distance. Meanwhile, the optical flow like temporal motion detector [33] is also proposed to suppress the ghost effect, which can be frequently observed in the scenarios with intermittent object movements. Also, to suppress the false-alarm detections induced by dynamic backgrounds, [14] proposes a multi-level low-rank solution for the detection of salient object in a coarse-to-fine manner. To adapt the modeling based salient motion detection methods to the non-stationary videos, [15] resorts to frame-level affine registration, and [13] employs high-level background tracking as the pre-processing procedure to obtain relatively-stationary short-term video sequences. However, because these modeling based methods usually require long-period video frames to gradually learn the background model, the obstinate challenges still exist when the input video sequences only have limited frames.

C. Video Saliency Detection Methods

Video saliency detection is to extract the most distinctive motion-related salient object from videos. The state-of-the-art video saliency detection methods can be roughly divided into two categories: fusion based methods and spatial-temporal contrast based methods. Since the motion clues can be easily obtained from the optical flow methods, the fusion based methods mainly focus on the combination of the color saliency and the motion saliency. Rahtu et al. [34] propose to use the conditional random field (CRF) to integrate the motion clues and the color saliency. Similarly, Fang et al. [17] proposes to use entropy-based uncertainty weights to merge the spatial saliency and the temporal saliency. And Liu et al. [16] resort to the mutual consistence between the spatial saliency and temporal saliency to guide the fusion process. Although fusion based methods can identify the most trustful saliency clue alternatively from the spatial or temporal saliency clues, failure cases still frequently occur when either the spatial saliency or temporal saliency is incorrect. Different from fusion based methods, spatial-temporal contrast based methods usually compute the low-level saliency clues in a spatial-temporal manner. For example, Seo et al. [35] propose to compute the contrast based saliency in a pre-defined spatial-temporal surroundings. Fu et al. [28] propose to estimate the temporal correspondence to guide the computation of the spatial saliency clue in a cluster-wise manner. Similarly, Zhou et al. [3] propose to compute multi-scale saliency in a region based spatial-temporal manner. Although such methods can achieve much better saliency detection performance than most of the image saliency methods, the obtained saliency maps usually have bad temporal consistence due to the frame-by-frame saliency computation. Also, Zhong et al. [36] propose to utilize the spatial-temporal info between consecutive video frames to construct their newly designed attention model based on optical flow, which fully take the advantages of the motion continuity nature to eliminate false-alarm detections. Similarly, with the low-level saliency clues based graph model, Kim et al. [37] propose to restart the random walk's stationary status among consecutive video frames to capture the real video saliency, which can fully respect the continuity of the spatial-temporal info. Most recently, Wang et al. [18], [19] propose to use the motion clue based geodesic distance (or

gradient flow) as the low-level saliency, and they adopt a global saliency energy function to guarantee the temporal smoothness of final saliency map. However, their global saliency energy function is too global to accurately diffuse saliency along the temporal axis, thus, incorrect low-level saliency can be easily accumulated, which finally gives rise to false-alarm video saliency detection.

D. Brief Summary

Although the performances of the salient motion detection methods are competitive over stationary long-term videos, their performances over short-term non-stationary videos are extremely poor. Meanwhile, considering the motion's influences, the image saliency detection methods are also ineffectual for the video saliency detection. Specifically, as for the existing video saliency detection methods, the unconstrained pursuit for the temporal saliency consistence tends to easily cause massive false-alarm detections (see the performance of SA15 in Fig. 17 and Fig. 18). Therefore, inspired by the aforementioned methods, we propose to integrate the motion clues and the foreground/background models based global spatial-temporal information to guide the low-level saliency fusion. And the well-designed low-rank analysis (i.e., seeking correspondences among super-pixels along the temporal axis) is proposed to perform the spatial-temporal saliency diffusion, which can make the saliency maps keep temporal smoothness at the cost of slight detection-accuracy degradations (proofs can be found in Fig. 14(a)). Furthermore, based on the established low-rank coherency correspondences, we will use the saliency boosting strategies to further improve the saliency accuracy in a strictly-constrained manner. To the best of our knowledge, our paper makes the first attempt to leverage the temporal coherency to improve the accuracy of video saliency detection. The overview of our entire method will be described in the following section.

TABLE I
LIST OF THE KEY MATHEMATICAL SYMBOLS INVOLVED IN OUR MATHEMATICAL FORMULATIONS.

Symbols	Symbol Interpretations
m	Total frame number of the short-term frame batch
c	Super-pixel's average RBG color
α, β	Learning factor of θ and Saliency adjustment factor
ϵ	The estimated saliency degree of the salient object
γ	Filter strength of the coarse salient object region
σ, θ	Spatial-temporal diffusion strength and range
u	Super-pixel number of the coarse salient object region
n	Super-pixel number of the original input video frame
ϑ	The selection or permutation matrix
μ_1, μ_2	The strength parameter of the low-rank component
λ_1, λ_2	The strength parameter of the sparse component
ρ	The iteration step size of the low-rank optimization
$\mathbf{P}$	Super-pixel's location matrix, $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$
$\mathbf{L}_c, \mathbf{L}_s$	The low-rank component of the color/saliency space
$\mathbf{E}_c, \mathbf{E}_s$	The sparse component of the color/saliency space
MS, **CS**	The raw motion saliency and color saliency
FC, **FS**	Color/motion feature subspace
FM, **BM**	Foreground model and background model
ST	Spatial-temporal gradient map
LS	The fused low-level saliency
F	Coarse foreground region

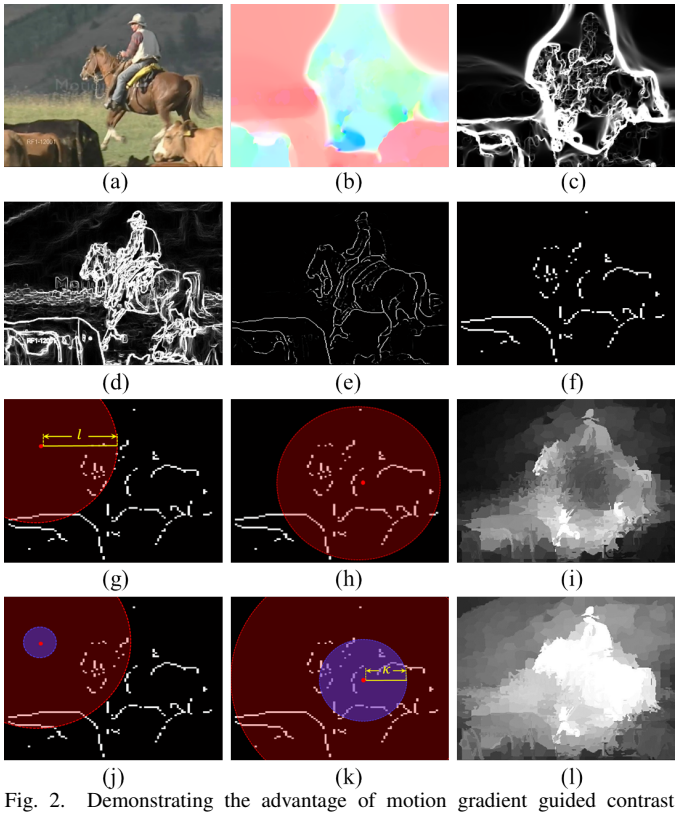


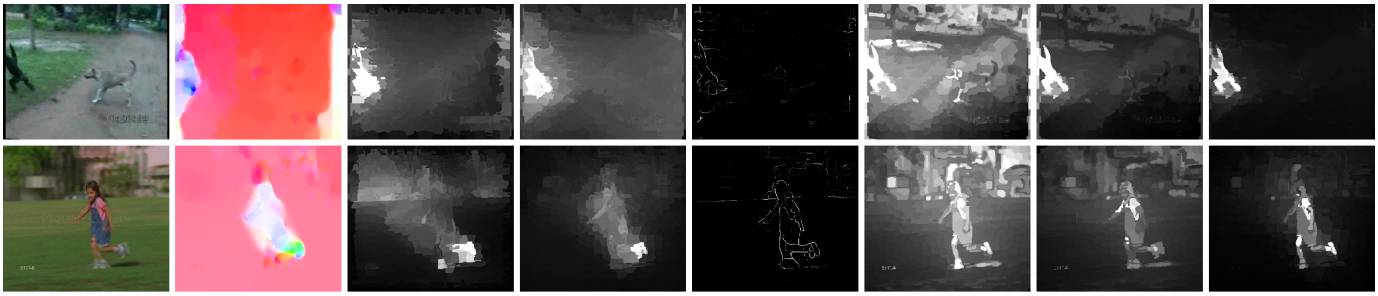
Fig. 2. Demonstrating the advantage of motion gradient guided contrast computation. (a) shows the source video frames, (b) shows the optical flow detection result, (c) and (d) respectively demonstrate the motion gradient and the color gradient, (e) shows the spatial-temporal gradient map obtained via Eq. 1, (f) shows the down-sampled **ST** gradient result to alleviate the computation burden, (g) and (h) demonstrate the contrast computation region of the traditional local method with corresponding motion saliency result demonstrated in (i), (i) and (h) demonstrate the **ST** gradient map guided contrast computation with corresponding result demonstrated in (l). The red dot denotes the position of the given superpixel, the red circle denotes the valid contrast computation region, and blue circle denotes the excluded region.

III. Method Overview

As shown in Fig. 1, our method mainly consists of four steps: (1) Contrast based saliency computation (Section IV-A); (2) Low-level saliency adjustment and fusion (Section IV-B and IV-C); (3) Low-rank coherency correspondence construction (Section V-B and V-E); and (4) Low-rank guided saliency diffusion and boosting (Section V-F).

Our method equally segments the entire long-term video sequence into several short-term frame batches to avoid the error accumulation of the false-alarm detections, wherein multiple diffusion constraints (i.e., Step 3 and Step 4 in Fig. 1) are employed to guide the saliency diffusion. We propose spatial-temporal gradient map **ST** (Eq. 1 in Section IV-A) to control the range of contrast computation, which is expected to roughly assign high saliency value around the foreground object. It is demonstrated in the columns of "Color Saliency" and "Motion Saliency" in Fig. 1. Also, a series of saliency adjustment (Section IV-B) and smoothing (Section IV-C) strategies are designed to guide the color and motion saliency fusion. From the column of "Fusion Results", it can be clearly observed that, the fused saliency map is much better than that resulted from only using "Color Saliency" or "Motion Saliency". Specifically, we propose to construct

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Illustration of the low-level saliency computation. (a) shows the input source images; (b) demonstrates the optical flow result; (c) demonstrates the obtained contrast based motion saliency and its corresponding adjusted saliency; (d) and (e) show the spatial-temporal gradient map (**ST**, see Eq. 1); (f) shows the raw color saliency map obtained by replacing optical flow gradient with *RGB* color in Eq. 2; (g) demonstrates the color saliency map adjusted by the guidance of the foreground/background model (Eq. 5); the fused saliency maps via Eq. 12 are demonstrated in (h).

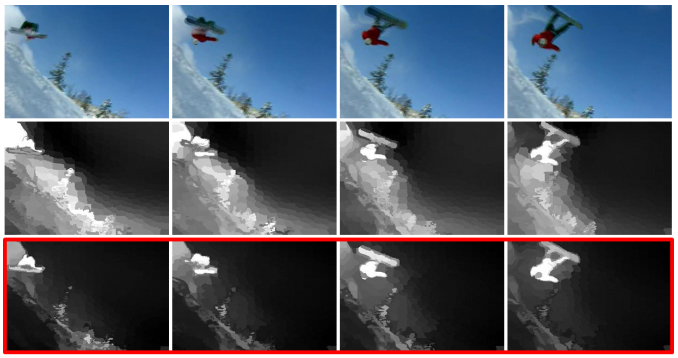


Fig. 4. Illustration of the impacts when adopting different color saliency computation methods. The first row demonstrates the input source images, the second row demonstrates the color saliency maps computed by the method proposed in [19], and the bottom row demonstrates the color saliency maps of our method, which are computed via Eq. 2 and Eq. 5.

low-rank spatial-temporal correspondences, which can boost the intra-batch saliency detection accuracy while retaining the temporal smoothness of saliency map. It will be detailed in Section V. Compared to the traditional methods, which simply diffuse low-level saliency clues over one or several consecutive frames to pursue the saliency map's temporal consistence, our method can not only achieve the temporal consistence but also further boost the video saliency detection accuracy (see the "Final Saliency" column in Fig. 1).

Table I summarizes the key symbols involved in the following mathematical derivations, wherein the normal-case letters denote scalars, bold lower-case letters denote finite dimensional vectors, and bold upper-case letters denote matrices.

IV. Spatial-Temporal Saliency Fusion

A. Contrast-based Saliency Clues

Give a long-term video sequence, video saliency detection is to find the salient object in each frame. Since the coarse/initial clues of the salient object can be well revealed by the contrast based saliency computation, in this section we will conduct detailed discussion on the low-level saliency clues adopted by this paper. Different from the conventional video saliency detection methods [17], [28], [18], [19], which reveal the saliency in a frame-wise or sequence-wise manner, we equally decompose the original long-term video sequence into many short-term frame batches $\mathbf{B}_i = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_m\}$. Here $\mathbf{I}_k$ denotes the k-th video frame. For each video frame $\mathbf{I}_t$ in the current frame batch $\mathbf{B}_i$, we employ the boundary-aware smoothing method [38] to eliminate unnecessary details, and simple linear iterative clustering (SLIC) based super-pixel decomposition [39] is also adopted to reduce the computational burden. Meanwhile, according to the rationality mentioned in [1], [40], the motion gradient map is much more robust and trustful than the motion saliency clues, we combine the motion gradient with the color gradient to obtain the spatial-temporal gradient map to guide the low-level contrast computation, which can be computed by Eq. 1, including the color contrast and the motion contrast.

$$\mathbf{ST} = ||\nabla(\mathbf{I})||_2 \odot ||vx, vy||_2, \tag{1}$$

where $\odot$ denotes the element-wise Hadamard product, and $\nabla(\mathbf{I})$ denotes the color gradient map, vx and vy respectively denote the horizontal and vertical gradient of the optical flow results. Then, the motion contrast of the i-th super-pixel can be computed via Eq. 2.

$$\mathbf{MS}_i = \sum_{\mathbf{p}_j \in \psi_i} \frac{||\mathbf{V}_i, \mathbf{V}_j||_2}{||\mathbf{p}_i, \mathbf{p}_j||_2}, \psi_i = \{\kappa \leq ||\mathbf{p}_i, \mathbf{p}_j||_2 \leq \kappa + l\}, \tag{2}$$

Here $||\cdot||_2$ denotes the l_2-norm, $\mathbf{p}_i \in \mathbb{R}^{2\times 1}$ denotes the position center of the i-th super-pixel, $\mathbf{V} \in \mathbb{R}^{2n\times 1}$ denotes the two-direction optical flow gradients, n denotes the super-pixel number in the current video frame, and ψ_i denotes the contrast range used in computation, which is determined by the shortest Euclidean distance between the i-th super-pixel and the spatial-temporal gradient map **ST** (Eq. 3).

$$\kappa = \frac{l}{||\Lambda(\mathbf{ST})||_0} \sum_{k \in ||k,i||_2 \leq l} ||\Lambda(\mathbf{ST}_k)||_0. \tag{3}$$

Here we empirically set l as the initial local contrast computation range $l = \frac{1}{2} min\{W, H\}$, W and H separately denote the image width and height, and $\Lambda(\cdot)$ denotes the down sampling function (30%) to alleviate the computation burden. Also, the pictorial demonstration can be found in Fig. 2.

Although our method also adopts the motion information to control the contrast computation range as that in [19], the underlying rationality of our method is totally different from [19] in two aspects. First, the inner salient object regions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing

are only compared to the outer non-salient background to avoid hollow effects (see demonstration in Fig. 2(i)), but the contrast computation range of [19] heavily depends on the assumption of the background regions, wherein incorrect approximation (i.e., error accumulations) of the background regions may easily make color contrast computation fail. Second, since the optical flow based motion clue is trustful in most scenarios, our method automatically assigns larger saliency value to the foreground object by discarding the distance penalty $exp(-||\mathbf{p}_i, \mathbf{p}_j||_2/\psi_i)$ adopted in [19]. Benefitting from this, the performance improvements can be found in Fig. 4 and the quantitative results are shown in Fig. 14(a). Similarly, the color saliency **CS** can be computed by simply replacing the optical flow gradient with the *RGB* color value in Eq. 2, and the detailed formulation can be found in Eq. 4.

$$\mathbf{CS}_i = \sum_{\mathbf{p}_j \in \psi_i} \frac{||(\mathbf{R}_i, \mathbf{G}_i, \mathbf{B}_i), (\mathbf{R}_j, \mathbf{G}_j, \mathbf{B}_j)||_2}{||\mathbf{p}_i, \mathbf{p}_j||_2}, \tag{4}$$

where the definition of the ψ_i is identical to Eq. 2, and $(\mathbf{R}_i, \mathbf{G}_i, \mathbf{B}_i)$ denote the corresponding averaged *RGB* color of the i-th superpixel.

B. Modeling-based Saliency Adjustment

When the current motion clue is incorrect, purely considering the short-term contrast information (intra-batch's contrasts) is insufficient to produce robust saliency map, we integrate the long-term inter batch information into the computation of color contrast to suppress the saliency degree of non-salient backgrounds. That is, we keep updating the salient foreground model and the non-salient background model (see the Appearance/Background Model in Fig. 1) when accomplishing the saliency detection of each frame batch.

Suppose $\mathbf{FM} \in \mathbb{R}^{3\times fn}, \mathbf{BM} \in \mathbb{R}^{3\times bn}$ respectively denote the foreground appearance model and the background model, which record the super-pixel's mean *RGB* color history of all the foreground/background regions over the entire frame batch, we employ both the average and the minimum super-pixel's feature distances (i.e., *RGB* color) as the inter-batch indicators $Cinter$ to adjust the color saliency value. According to our experimental observations, the motion saliency **MS** is much more meaningful and trustful than the color saliency **CS**. Thus, we propose to adopt the relative discrepancy degree $Cintra$ to refine the color saliency **CS** (see details in Eq. 5).

$$\mathbf{CS}_i \leftarrow \mathbf{CS}_i \cdot Cinter_i \cdot Cintra_i, \tag{5}$$

$$Cinter_i = \phi\left(\frac{min||\mathbf{c}_i, \mathbf{BM}||_2 \cdot \frac{1}{bn}\sum||\mathbf{c}_i, \mathbf{BM}||_2}{min||\mathbf{c}_i, \mathbf{FM}||_2 \cdot \frac{1}{fn}\sum||\mathbf{c}_i, \mathbf{FM}||_2}\right), \tag{6}$$

$$Cintra_i = exp(\delta - |\phi(\mathbf{MS}_i) - \phi(\mathbf{CS}_i)|). \tag{7}$$

Here $\mathbf{c}_i = (R_i, G_i, B_i) \in \mathbb{R}^{3\times 1}$ denotes the average *RGB* value of the i-th super-pixel, fn and bn respectively denotes the size of the foreground and background model, $\phi(\cdot)$ is the $minmax$ normalization function, which strictly normalizes the color saliency adjustment degree into $[0.5, 1]$. And δ is the upper bound of the discrepancy degree between **CS** and **MS**, which is empirically set to be 0.5.

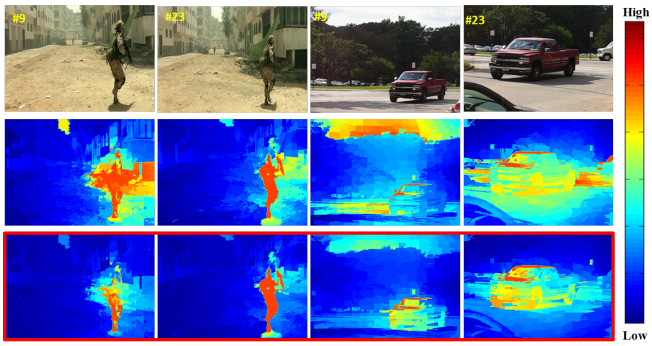


Fig. 5. Demonstrations of the performance improvements benefitted from our color saliency adjustment strategy. The top row shows the source images with the frame number marked in the left-top corner; the middle row shows the color saliency without our adjustment strategy, wherein massive false alarm detections can be easily observed; and the bottom row shows the color saliency after applying Eq. 5.

Obviously, the underlying rationality of color saliency adjustment is echoed in two aspects. First, both the salient object's appearance and the non-salient backgrounds (i.e., its corresponding color distributions) principally tend to stay unchanged in limited consecutive frames, which facilitates to adjust the saliency degree according to the previously-established appearance models (i.e., the foreground model FM and the background model BM). Second, the color saliency **CS** can be regarded as the complementary part of the motion saliency, whose main effect is to boost or sharpen the tiny details of the salient object when the motion saliency indicating high saliency degree. Meanwhile, we have demonstrated the performance improvements benefitted from Eq. 5 in Fig. 5.

Following the first rationale, because our video saliency extraction method is conducted in batch-wise manner, the previously-obtained saliency detection results (i.e., the last frame batch), which fully respect the spatial-temporal info via low-rank coherency saliency diffusion and boosting, already have both high recall and precision rate. Hence, the appearance models (i.e., the foreground model **FM** and the background model **BM**) can be gradually perfected via considering the previous detection results, and it is reasonable to utilize the corresponding *RGB* color histories to adjust the color saliency value via Eq. 6. As shown in Eq. 6, we mainly consider the minimum ($min||\mathbf{c}_i, \mathbf{FM}||_2$, $min||\mathbf{c}_i, \mathbf{BM}||_2$) and the average l_2 *RGB* color distance ($\frac{1}{bn}\sum||\mathbf{c}_i, \mathbf{BM}||_2$, $\frac{1}{fn}\sum||\mathbf{c}_i, \mathbf{FM}||_2$) from the i-th superpixel to the entire appearance model as the main criterion to guide the adjustment of color saliency. In fact, without considering the minimum model distance, the value of $Cinter$ is constantly larger than 1 when the i-th superpixel's *RGB* distance is closer to the foreground model **FM** than the background model **BM**. It means that, the probability of the i-th superpixel belonging to the salient foreground object is larger than the non-salient backgrounds, and the saliency degree of the color saliency **CS** should be increased accordingly, and vice versa. However, it is apparently not discriminative enough to obtain correct color saliency adjustment if we only consider the average model distance, because the existing false-alarm non-salient backgrounds may have closer l_2 *RGB* distance to the appearance model than the background model.

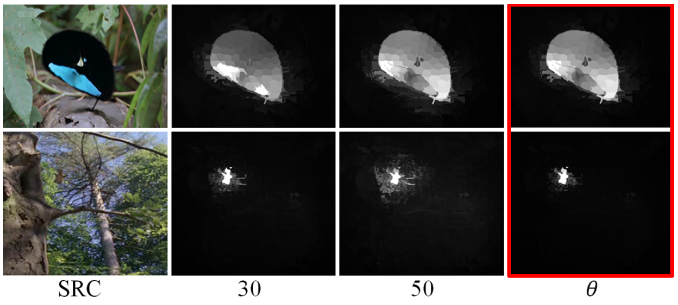


Fig. 6. Illustration of impacts when adopting different spatial-temporal smoothing ranges. Different smooth ranges are marked at the bottom row of each column (i.e., 30, 50, and θ), which state that assigning θ as the smooth range could produce the best saliency results (the last column).

Furthermore, introducing the minimum model distance into Eq. 6 can definitely alleviate these phenomenon via increasing the discriminative power from the perspective of l_2 *RGB* color distance, which nearly has no side effects.

Following the second rational, Eq. 7 is a hinge function constrained by the predefined hard threshold δ, which serves as inverse penalty to the discrepancy degree between the motion saliency and the color saliency. In fact, as the complementary part used to sharpen the tiny details of the salient object, the contribution of the color saliency becomes valueless when the discrepancy degree is large. According to Eq. 7, the precondition of the large discrepancy degree obviously relies on two aspects: either the motion saliency **MS** is extremely larger than the color saliency **CS**, or the color saliency **CS** is extremely larger than the motion saliency **MS**. Since it is a common sense that trustful optical flow detection always results in large absolute motion saliency degree, it is reasonable to penalize the color saliency degree with respect to the discrepancy degree between **MS** and **CS** when **MS** $\gg$ **CS**. Meanwhile, the overall motion saliency frequently stays at a relatively low level when only existing small motions. When **MS** $\ll$ **CS**, the "multiplicative" based low-level saliency fusion strategy (Eq. 12) can also guarantee those regions with large color saliency degree to remain at a relatively high level after the adjustment toward the flat distributed motion saliency.

C. Low-level Saliency Fusion

Since both the spatial (i.e., super-pixel's topology information in color feature space) and temporal information (i.e., color l_2 distance based inter-frame saliency smoothing) can be integrated to further refine the saliency value, we propose to refine both the low-level saliency **CS** and the **MS** (the smoothing of **MS** is identical to **CS**, as shown in Eq. 8) via spatial-temporal smoothing first as:

$$\mathbf{CS}_{t,i} \leftarrow \frac{\sum_{k=t-1}^{t+1} \sum_{\mathbf{p}_{k,j} \in \varphi} exp(-||\mathbf{c}_{t,i}, \mathbf{c}_{k,j}||_1/\sigma) \cdot \mathbf{CS}_{k,j}}{\sum_{k=t-1}^{t+1} \sum_{\mathbf{p}_{k,j} \in \varphi} exp(-||\mathbf{c}_{t,i}, \mathbf{c}_{k,j}||_1/\sigma)}. \quad (8)$$

Here $c_{t,i}$ denotes the averaged *RGB* color value of the i-th superpixel in the t-th video frame, φ denotes the spatially local neighbor region that satisfies $||\mathbf{p}_{t,i}, \mathbf{p}_{k,j}||_2 \leq \theta$, θ is dynamically controlled by Eq. 9, and σ controls the smoothing strength, which will be further discussed in Section VI-A.

$$\theta = \frac{1}{m \times n} \sum_{t=1}^{m} \sum_{i=1}^{n} ||\frac{1}{n} \sum_{i=1}^{n} E(\mathbf{ST}_{t,i}), E(\mathbf{ST}_{t,i})||_1, \quad (9)$$

$$E(\mathbf{ST}_i) = \begin{cases} \mathbf{p}_i, & \mathbf{ST}_i \leq \epsilon \times \frac{1}{n} \sum_{i=1}^{n} \mathbf{ST}_i \\ 0, & otherwise \end{cases} \quad (10)$$

m and n separately denote the frame number in the current batch and the super-pixel number in the current frame, $E(\cdot)$ is an indicator function (see Eq. 10) used to select super-pixels with large **ST** values (Eq. 1), ϵ is a parameter to control the filter strength, which is empirically set to be 10. And $\mathbf{p}_i$ denotes the mean center coordinates of the i-th super-pixel. As shown in Fig. 6, after spatial-temporal smoothing, the quality of the saliency map heavily depends on the selection of the smoothing range. That is, a small smooth range is better than a larger one for a tiny salient object, but in contrast it has better effects when assigning large smooth range to huge salient object. And the advantages of introducing θ can be obviously found in the last column of Fig. 6. Meanwhile, to guarantee the saliency consistence at a frame batch level, we dynamically update the q-th frame batch's smoothing range θ_q via

$$\theta_q \leftarrow \alpha\theta_q + (1-\alpha)\theta_{q-1}, \quad (11)$$

where α is the learning weight, and we empirically set it to be 0.2. Finally, we combine the color saliency **CS** with the motion saliency **MS** to obtain the fused low-level saliency **LS** via

$$\mathbf{LS} = \mathbf{CS} \odot \mathbf{MS}. \quad (12)$$

Here $\odot$ denotes the element-wise Hadamard product. As shown in Fig. 3(h), the fused saliency map is much better than those that solely utilize color saliency (Fig. 3(g)) or motion saliency (Fig. 3(d)). And the quantitative proofs of the performance improvement are documented in Fig. 14(a). It should be noted that, the fused saliency map significantly increases the accuracy, but the recall rate also decreases much compared to the motion saliency. Now, in order to alleviate this problem, it sets the stage for us to introduce our newly-designed low-rank coherency based spatial-temporal saliency diffusion and boosting in the next section.

V. Low-rank Coherency Guided Spatial-Temporal Video Saliency Detection

Although the fused low-level saliency **LS** (Eq. 12) is much better than pure color or motion saliency map, there still exist many false-alarm detections, and the saliency distributions are not temporally consistent (see the "Fused Result" in Fig. 1). Thus, in this section, we propose to boost the saliency map accuracy while keeping its temporal smoothness based on our proposed low-rank coherency analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing

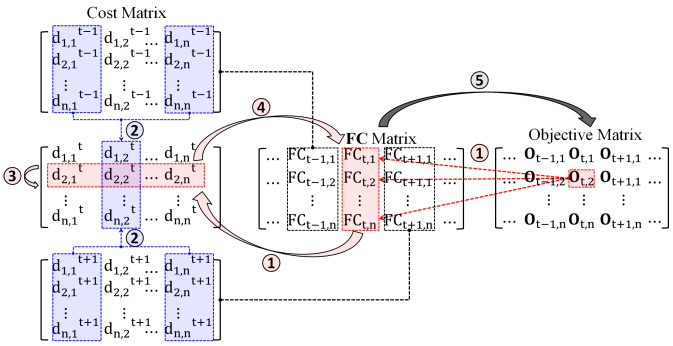


Fig. 7. Illustration of our low-rank correspondence revealing procedure (Eq. 18). "Mark 1" shows the computation of the cost matrix (Eq. 24 and Eq. 25); "Mark 2" denotes the weak structure constraints (Eq. 27); "Mark 3" indicates the global minimum of the cost matrix **M** (Hungarian algorithm); "Mark 4" represents the feature space updating (using Eq. 28 to update **FC** and **FS**); "Mark 5" represents the updating of the objective matrix with newly-updated **FC** and **FS** (Eq. 20 and Eq. 26).

A. Brief Review of the Low-rank Revealing Methods

The low-rank revealing problem aims to decompose the original input matrix $\mathbf{D}$ into the low-rank part $\mathbf{L}$ and the sparse part $\mathbf{E}$ as $\mathbf{D} = \mathbf{L} + \mathbf{E}$. Thus, the problem formulation can be regarded as:

$$\min_{\mathbf{L},\mathbf{E}} \ rank(\mathbf{L}) + \lambda||\mathbf{E}||_0 \quad subj \quad \mathbf{D} = \mathbf{L} + \mathbf{E}. \tag{13}$$

Eq. 13 is a non-convex optimization problem (NP-hard), however, it can be approximately solved via its relaxing convex envelope as:

$$\min_{\mathbf{L},\mathbf{E}} \ ||\mathbf{L}||_* + \lambda||\mathbf{E}||_1 \quad subj \quad \mathbf{D} = \mathbf{L} + \mathbf{E}. \tag{14}$$

$||\cdot||_*$ indicates the nuclear norm of $\mathbf{L}$. In fact, the above convex problem can be effectively solved by the Robust Principal Component Analysis (RPCA) [41], and the key solution of the RPCA low-rank revealing consists of two steps: the singular value thresholding based low-rank part estimation (Eq. 15), the soft thresholding based sparse part computation (Eq. 16).

$$\mathbf{L} \leftarrow \mathbf{U}[\Sigma - \mu\mathbf{I}]_+\mathbf{V}, \quad (\mathbf{U}, \Sigma, \mathbf{V}) \leftarrow svd(\mathbf{Y}). \tag{15}$$

$$\mathbf{E} \leftarrow sign(\mathbf{D} - \mathbf{E} - \mathbf{L})[|\mathbf{D} - \mathbf{E} - \mathbf{L}| - \lambda\mu]_+. \tag{16}$$

Here $\mathbf{I}$ denotes the identity matrix, $\mathbf{D}$ denotes the original input matrix, and the svd denotes the SVD decomposition, $\mathbf{Y}$ denotes the Lagrange multiplier, μ and λ respectively denote the low-rank threshold parameter and the sparse threshold parameter, which will be further discussed later. The RPCA low-rank revealing iterates these two steps to gradually obtain the low-rank part $\mathbf{L}$ and the sparse part $\mathbf{E}$. It should also be noted that, although both our method and the low-rank related methods [15], [14], [13] are based on the low-rank revealing, here we further clarify the difference as follows:

1. Our newly proposed method only utilize the sparse component to eliminate the side-effects induced by incorrect alignments, while the previous low-rank methods all take this sparse component as the only saliency indicator, which easily results in massive false-alarm detections.

2. Although all these low-rank related methods adopt alignment steps to handle the non-stationary video problems, the graph model based method [15] is too local to detect tiny movement. The affine transformation [42], [14] and the background tracking [13] are too global to handle the camera movement induced non-rigid variations (i.e., view angle change), which may easily cause massive false-alarm detections. In sharp contrast, our method proposes to conduct mid-level alignment (i.e., the superpixel level) and employs the low-rank coherency constraint to suppress the non-salient backgrounds while enhancing the salient foregrounds, which is more suitable to handle the aforementioned limitations in the non-stationary videos.

Fig. 8. The obtained foreground masks (the middle row) under different choices of γ (Eq. 17) over *cheetah* sequence. The top row and bottom row demonstrate the overlapped areas between the foreground masks separately with $\gamma = 6$ (an aggressive choice) and $\gamma = 2.2$ (our best choice), and the yellow **B** and **E** in the top-left corner respectively indicate the beginning frame and the ending frame of the identical frame batch.

3. Different from previous RPCA based low-rank revealing methods that only consider the color info, our method consider both the color info and the low-level saliency info to represent the low-rank coherency among consecutive video frames, which gives rise to more robust video saliency results for the salient objects with extremely slow movements. However, the low-rank revealing based background modeling method can not well approximate the background info, because the overlapped foreground regions brought by the extremely slow movements can easily be incorrectly taken as the low-rank backgrounds.

B. The Low-Rank Analysis Framework for Saliency Coherency

From the perspective of single video frame, the motion clues captured by optical flow usually contain many false-alarm detections. In contrast, the regions constantly with low **LS** throughout the entire frame batch should be excluded

from the scope of our low-rank based saliency boosting. That is, to eliminate most of the false-alarm detections induced by incorrect optical flow, we initially locate the coarse foreground regions that contain all the super-pixels of the potential salient object. Given the k-th frame batch $\mathbf{B}_k$ with m video frames, the t-th video frame's feature subspace spanned by low-level saliency $\mathbf{LS}_t$ can be represented by $\mathbf{fI}_t = \{\mathbf{LS}_{t,1}, \mathbf{LS}_{t,2}, ..., \mathbf{LS}_{t,n}\}$, here n denotes the super-pixel number. Thus, the saliency feature space of the entire frame batch $\mathbf{fB}_k$ can be written as the matrix form $\mathbf{fB}_k = \{\mathbf{fI}_1, \mathbf{fI}_2, ..., \mathbf{fI}_m\}$, and the coarse foreground regions $\mathbf{F} \in \mathbb{R}^{n\times 1}$ can be determined by Eq. 17.

$$\mathbf{F}_i = [\sum_{t=1}^{m} \mathbf{LS}_{t,i} - \frac{\gamma}{m \times n} \sum_{t=1}^{m}\sum_{i=1}^{n} \mathbf{LS}_{t,i}]_+. \quad (17)$$

Here γ is a parameter to control the reliability of the obtained coarse foreground region, which will be further discussed in Section VI-A. By adopting $\mathbf{F}$ to indicate the coarse foreground regions, we get two feature subspaces separately spanned by the *RGB* color and low-level saliency **LS**, which can be respectively represented as $\mathbf{FC} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_m\} \in \mathbb{R}^{3u\times m}$. Of which, $\mathbf{C}_i = \{vec(R_{i,1}, G_{i,1}, B_{i,1}, R_{i,2}, G_{i,2}, B_{i,2}, ..., R_{i,n}, G_{i,n}, B_{i,n})^{\mathbf{T}} \in \mathbb{R}^{3u\times 1}$ and $R_{i,j}, G_{i,j}, B_{i,j}$ respectively represents the *RGB* color channel of the j-th superpixel in i-th video frame, $\mathbf{FS} = vec(\mathbf{LS}_1), vec(\mathbf{LS}_2), ..., vec(\mathbf{LS}_m)\} \in \mathbb{R}^{u\times m}$. Here u denotes the total positive elements in $\mathbf{F}$, and $vec(\cdot)$ denotes the vectorizing function. In fact, since the coarse foreground region (Fig. 8) is fixed for the entire frame batch, for the foreground object, there will exist strong low-rank coherency in the feature subspaces **FC** and **FS**. Therefore, we can leverage low-rank analysis to construct "one-to-one" correspondence among cross-frame super-pixels, and then the low-level saliency can be diffused (i.e., to achieve temporal consistence) and boosted (i.e., to obtain better accuracy) globally over the entire frame batch.

However, due to the movements of foreground salient object and the variations of non-salient surroundings, it is infeasible to directly apply the traditional low-rank analysis over the **FC** of the **FS**. Inspired by the recently-proposed inner alignment involved low-rank solutions [43], [15], [44], we propose an alternative low-rank selecting strategies, which alternatively re-order the row-wise structures of **FC** and **FS** during RPCA [45] based low-rank revealing. Hence, the low-rank coherency problem can be formulated as:

$$\begin{aligned} \min_{\mathbf{D}_{c,s},\mathbf{E}_{c,s},\vartheta,\mathbf{P}\odot\vartheta} & ||\mathbf{D}_c||_* + ||\mathbf{D}_s||_* + ||\mathbf{P}\odot\vartheta||_2 \\ & + \lambda_1||\mathbf{E}_c||_1 + \lambda_2||\mathbf{E}_s||, \\ s.t. \mathbf{D}_c = \mathbf{L}_c + \mathbf{E}_c, & \mathbf{D}_s = \mathbf{L}_s + \mathbf{E}_s, \mathbf{D}_c = \mathbf{FC}\odot\vartheta, \mathbf{D}_s = \mathbf{FS}\odot\vartheta, \\ \vartheta = \{\mathbf{Q}_1, \mathbf{Q}_2, ... & \mathbf{Q}_m\}, \mathbf{Q}_i \in \{0,1\}^{n\times n}, 1^{\mathrm{T}}\mathbf{Q}_i = 1. \end{aligned} \quad (18)$$

Here $||\cdot||_*$ denotes the nuclear norm, $\mathbf{P} \in \mathbb{R}^{2n\times m}$ is the position matrix, $\mathbf{L}_c$ and $\mathbf{L}_s$ respectively represent the estimated low-rank component over the color feature space and the saliency feature space, $\mathbf{E}_c$, $\mathbf{E}_s$ respectively represent the sparse component over the color feature space and the saliency feature space, $\odot$ denotes element-wise Hadamard product. ϑ is the selection matrix (i.e., permutation matrix [46], [43]), which encodes the constructed correspondences among super-pixels. In Eq. 18, $||\mathbf{D}_c||_*$ is the primary clue to construct the correspondence among super-pixels in $\mathbf{F}$ (Eq. 17). However, due to the weak discriminative ability of *RGB* color information ($\mathbf{D}_c$), to bridge the gap between salient foreground object and non-salient surroundings, we define the low-rank constraint of low-level saliency clue **LS** (Eq. 12) as a weak classifier, see $||\mathbf{D}_s||_*$ in Eq. 18, and we will give a brief discussion in the next section to explain the behind rational of our saliency coherency.

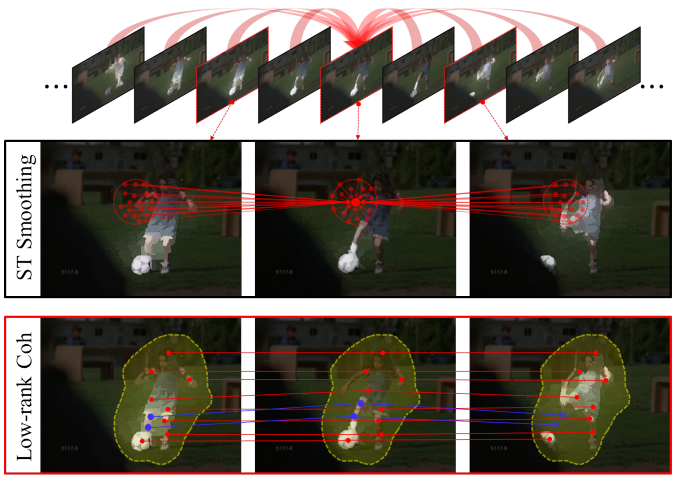


Fig. 9. Demonstration of the differences between the traditional spatial-temporal smoothing method (ST Smoothing) and our low-rank coherency guided method. The yellow region in the bottom row indicates the estimated foreground motion mask.

Meanwhile, since the priority of *RGB* color information is higher than the saliency clues, we constrain the sparse parameter to satisfy $\lambda_1 > \lambda_2$, which will be further discussed in Section VI-A. The $||\mathbf{P}\odot\vartheta||_2$ in Eq. 18 is a weak structure/location constraint, which could ensure the selection/alignment result (i.e., the row-wise elements) consecutively located in the overlapped neighboring regions. Here the motivation of adopting a "weak" structure/location constraint instead of a "strong" one is crucial, because the variations induced by SLIC super-pixel decomposition, camera/foreground object movements and view angle/appearance will definitely make the optimal "one-to-one" correspondences infeasible. In fact, the core rationality of the second part and the third items of Eq. 18 is to prevent incorrect super-pixel correspondences, which tend to align salient foreground super-pixels with non-salient background ones, yet, the structure sensitive one-to-one corresponding is somewhat unnecessary.

C. Advantage of Our Low-rank Coherency

From the perspective of video frame batch, although the appearance of the salient moving object may vary among consecutive video frames, the global low-rank estimation of its overall appearance is relatively stable (we call it "low-rank temporal coherence"), and it heavily relates to the global consistency of the salient foreground object in the *RGB* color spanned feature space. Thus, to improve the overall performance of video saliency detection, we propose to use this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing

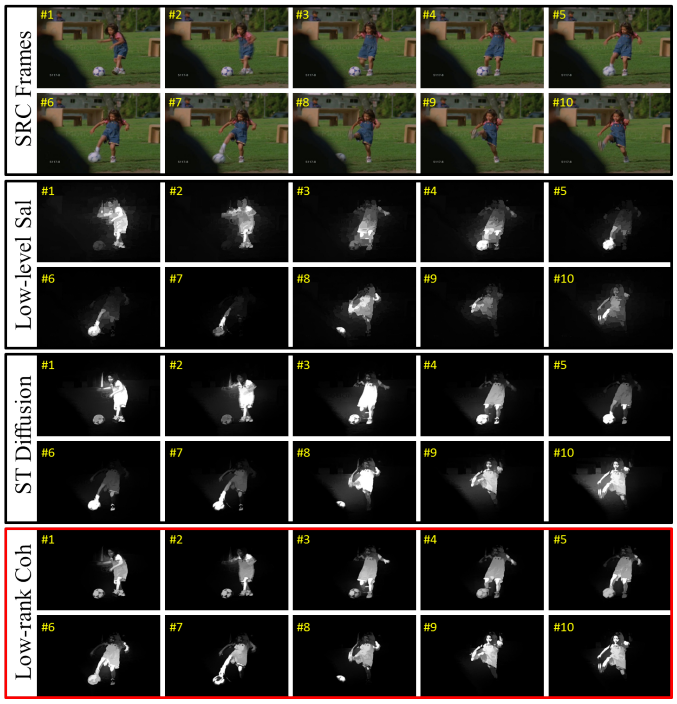


Fig. 10. Performance comparison between the traditional spatial-temporal weighting method (ST Smoothing) and our low-rank coherency guided saliency diffusion and boosting method (Low-rank Coh). The initial inputs are the low-level saliency fusion results (Low-level Sal) over the *kicking* sequence, and all other parameters are identical.

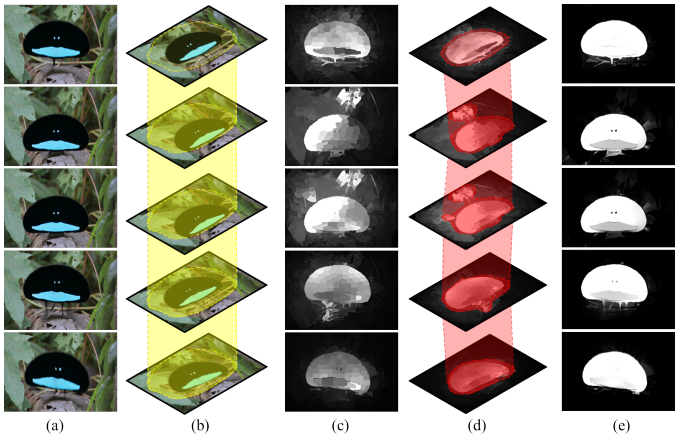


Fig. 11. Demonstrating the benefits by introducing the saliency coherency. (a) demonstrates the source video frames, (b) demonstrates the initial correspondences between consecutive video frames, where the yellow ring denotes the previously determined coarse foreground region, (c) shows the low-level saliency detection results, (d) demonstrates the coarse foreground region constrained by the saliency coherency, (e) demonstrates the final video saliency detection results.

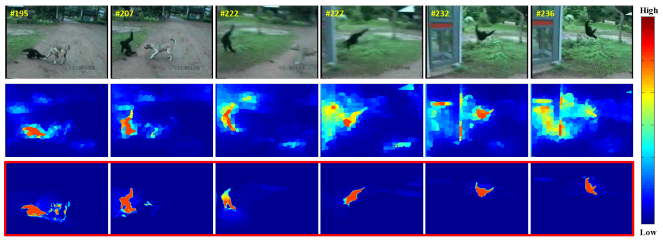


Fig. 12. Demonstrating the advantages of our method over the complex case with suddenly appeared non-salient backgrounds. The top row shows the source images, the middle row shows the detection results of the graph based method [18], and the bottom row shows the detection results of our method. Obviously, the newly appeared glass door (e.g., red bar on the door) from the frame #227 causes massive false-alarm detections in the graph based method.

low-rank coherency to guide the low-level saliency diffusion and boosting. As shown in Fig. 9, the top row demonstrates the basic rational of the spatial-temporal smoothing method (e.g., the graph model used in [40] and [18]), which attempts to reveal the undetected regions of the current video frame via spatial-temporal saliency transferring (L_2 color distance based majority weighting scheme) among all the video frames belonging to an identical video frame batch. And the detailed saliency transferring is also demonstrated in the "**ST Smoothing**" in Fig. 9. As for the **ST Smoothing** scheme, the final saliency value of each superpixel is determined by the weighted average of all surrounding superpixels's saliency degrees in both the spatial and temporal scales, as a "many-to-many" scheme, which can easily cause the accumulation of false-alarm detections. Because the behind rational of the **ST Smoothing** scheme is to make average, it can not suppress those false-alarm detections of non-salient surroundings, especially for those surroundings sharing similar *RGB* color distributions (e.g., the green grass in Fig. 9), and it will finally cause the accumulation of false-alarm detections, see #8#9#10 frame of **ST Smoothing** in Fig. 10. In sharp contrast to the conventional **ST Smoothing** method, we propose to utilize the low-rank revealing solution to estimate the global appearance model (the low-rank component) while suppressing those false-alarm detections via our unique "one-to-one" saliency diffusion and boosting scheme. The detailed saliency diffusion scheme is demonstrated in the "**Low-rank Coh**" in Fig. 9. Different from the "many-to-many" diffusion scheme, our low-rank guided saliency diffusion is conducted in a "one-to-one" manner, which means all superpixels belonging to different frames are lined up, and the saliency degree of superpixels are assigned to be identical to the other superpixels belonging to the same temporal coherency "line". Therefore, benefitting from this strategy, the advantages of our low-rank coherency guided saliency diffusion over the traditional **ST Smoothing** scheme can be summarized as: 1. The false-alarm detections in non-salient background regions won't accumulate, see demonstrations in Fig. 12. 2. Because the superpixels belonging to identical temporal "line" share the same saliency degree, the undetected salient object can be better revealed by our method than **ST Smoothing** scheme, and the saliency degree between consecutive video frames can keep temporal smoothness. 3. The "one-to-one" strategy provides the foundation to perform low-rank coherency guided saliency boosting, which can further enhance the saliency of foreground salient object while suppressing the saliency degree of the non-salient backgrounds.

D. Advantage of Our Saliency Coherency

Since the spatial displacement of the moving salient foreground object may lead to weak superpixel correspondence (see the yellow region in Fig. 11(b)) at the very beginning of

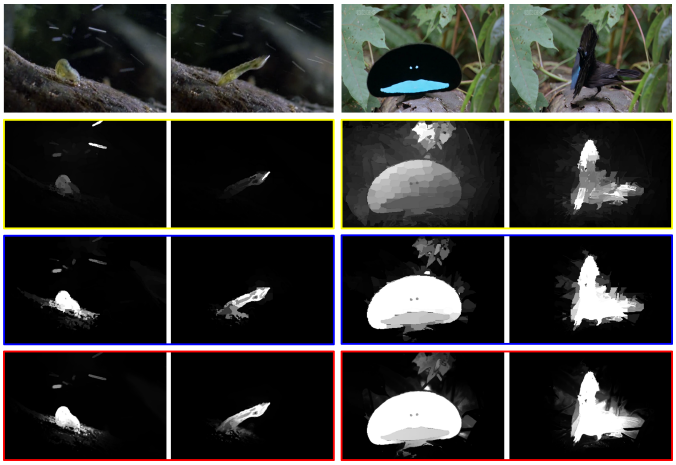


Fig. 13. Demonstration of the performance improvement. The first row shows the source video frames; the second row highlighted with yellow line lists some bad cases of fused low-level saliency **LS** (Eq. 12); the third row highlighted with blue line shows the saliency map obtained after conducting low-rank coherency based spatial-temporal saliency diffusion and boosting; the last row highlighted with red line demonstrates the final saliency map after pixel-wise saliency refinement.

the optimization procedure in Eq. 18, the obtained color low-rank component is extremely untrustful to be regarded as the alignment indicator, which may easily make the optimization be trapped into the local minimum and produce incorrect alignment results. However, since the saliency degree is a single value, the corresponding feature space of the saliency coherency can be regarded as the constrained version of the color feature space (see demonstration in Fig. 11(d)), and the approximated saliency low-rank component is relatively trustful to guide the alignments at early iterations.

Although those incorrect low-level saliency (e.g., the incorrectly detected leafs in Fig. 11(c)) can definitely affect the low-rank accuracy of the Hungarian algorithm based alignments at early iterations, the color low-rank component will ultimately dominate the alignment procedure, because we assign $\lambda_1 > \lambda_2$ in Eq. 18. That is, those data conflicts related to the saliency coherency may easily be regarded as the sparse noises, which guarantees the entire optimization to bias toward the color info at the later iterations, and the quantitative proofs can be found in Fig. 14(g). Also, the feasible solver of Eq. 18 will be detailed in next section.

E. Mathematical Solver

Since the low-rank revealing problem of Eq. 18 is non-convex, we resort to the ADMM [47] framework to convert it into several convex sub-problems. Of which, the augmented Lagrangian can be represented as:

$$\begin{aligned}\mathbf{L}(\mathbf{D}_{c,s}, \mathbf{E}_{c,s}, \vartheta, \mathbf{P}\odot\vartheta) &= \mu_1||\mathbf{D}_c||_* + \mu_2||\mathbf{D}_s||_* \\ &+ \lambda_1||\mathbf{E}_c||_1 + \lambda_2||\mathbf{E}_s||_2 + ||\mathbf{P}\odot\vartheta||_2 \\ &+ tr(\mathbf{Y}_1^{\mathrm{T}}(\mathbf{D}_c - \mathbf{L}_c - \mathbf{E}_c)) + tr(\mathbf{Y}_2^{\mathrm{T}}(\mathbf{D}_s - \mathbf{L}_s - \mathbf{E}_s)) \\ &+ \frac{\rho}{2}(||\mathbf{D}_c - \mathbf{L}_c - \mathbf{E}_c||_2 + ||\mathbf{D}_s - \mathbf{L}_s - \mathbf{E}_s||_2).\end{aligned} \tag{19}$$

Here $\mathbf{D}_{c,s}$ ($\mathbf{E}_{c,s}$) is the abbreviation to represent $\mathbf{D}_c$ or $\mathbf{D}_s$ ($\mathbf{E}_c$ or $\mathbf{E}_s$), $tr(\cdot)$ represents the matrix trace, $\mathbf{Y}_{1,2}$ is the Lagrangian multiplier, and ρ denotes the iteration step size. To this end, Eq. 19 can be iteratively solved via the strategy of solving one by fixing the others. The optimization solution via partial derivative on $\mathbf{L}_{c,s}$ and $\mathbf{E}_{c,s}$ can be separately written as:

$$\begin{aligned}\mathbf{L}_{c,s}^{t+1} &= \underset{\mathbf{L}_{c,s}^t}{\operatorname{argmin}}\, \mu_{1,2}||\mathbf{L}_{c,s}^t||_*/\rho_t \\ &+ 1/2||\mathbf{L}_{c,s}^t - (\mathbf{D}_{c,s}^t - \mathbf{E}_{c,s}^t + \mathbf{Y}_{1,2}^t/\rho_t)||_2^2,\end{aligned} \tag{20}$$

$$\begin{aligned}\mathbf{E}_{c,s}^{t+1} &= \underset{\mathbf{E}_{c,s}^t}{\operatorname{argmin}}\, \lambda_{1,2}||\mathbf{E}_{c,s}^t||_1/\rho_t \\ &+ 1/2||\mathbf{E}_{c,s}^t - (\mathbf{D}_{c,s}^t - \mathbf{L}_{c,s}^t + \mathbf{Y}_{1,2}^t/\rho_t)||_2^2.\end{aligned} \tag{21}$$

And $\mathbf{Y}_{1,2}$ is the abbreviation to represent $\mathbf{Y}_1$ or $\mathbf{Y}_2$ accordingly. In fact, since both Eq. 20 and Eq. 21 are the convex surrogates, these optimization sub-problems can be separately solved by RPCA singular value thresholding and soft thresholding. Thus, the low-rank components $\mathbf{L}_c$ and $\mathbf{L}_s$ can be iteratively updated via Eq. 22.

$$\begin{aligned}\mathbf{L}_{c,s}^{t+1} &\leftarrow \mathbf{U}[\Sigma - \mu_{1,2}/\rho_t]_+\mathbf{V}^T, \\ (\mathbf{U}, \Sigma, \mathbf{V}) &\leftarrow svd(\mathbf{D}_{c,s}^t - \mathbf{E}_{c,s}^t + \mathbf{Y}_{1,2}^t/\rho_t).\end{aligned} \tag{22}$$

Here the superscript t and $t+1$ denote the iteration times, and svd denotes the SVD decomposition. Similarly, the soft thresholds to iteratively reveal the sparse components $\mathbf{E}_c$ and $\mathbf{E}_s$ are formulated as Eq. 23.

$$\begin{aligned}\mathbf{E}_{c,s}^{t+1} &\leftarrow sign(|\mathbf{H}|/\rho_t)[\mathbf{H} - \lambda_{1,2}/\rho_t]_+, \\ \mathbf{H} &= \mathbf{D}_{c,s}^t - \mathbf{L}_{c,s}^t + \mathbf{Y}_{1,2}^t/\rho_t.\end{aligned} \tag{23}$$

Since λ_1 is larger than λ_2, the outliers induced by movements and variations in **FS** are much easier to be identified as the sparse component, which automatically enables the low-rank revealing procedure to bias toward the color clues (**FC**).

Then, the low-rank assumption (see the objective matrix in Fig. 7), which determines the selection result $\mathbf{Q}$, is adopted to compute the l_2-norm cost matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ (see "Mark 1" in Fig. 7).

$$d_{i,j}^t = ||\mathbf{O}_{t,i} - G(\mathbf{U}_1, j)||_2, \mathbf{U}_1 = G(\mathbf{FC}, t) \odot \mathbf{Q}_t, \tag{24}$$

$$r_{i,j}^t = ||\mathbf{O}_{t,i} - G(\mathbf{U}_2, j)||_2, \mathbf{U}_2 = G(\mathbf{FS}, t) \odot \mathbf{Q}_t, \tag{25}$$

where $\mathbf{O}$ is the column-wise low-rank objective function (i.e., low-rank residual/margin, see Eq. 26), $G(\mathbf{F}_{c,s}, t)$, which returns the t-th column of the $\mathbf{F}_{c,s}$ matrix as:

$$\mathbf{O}_{t,i} = \mathbf{L}_{c,s}(t,i) + \mathbf{E}_{c,s}(t,i) - \mathbf{Y}_{1,2}(t,i)/\rho_t. \tag{26}$$

Specifically, since the item $min||P\odot\vartheta||_2$ is NP-hard, and it is also very hard to approximate, we relax it by converting the consecutive cost matrices $\mathbf{M}_k = \{d_{1,1}^k + r_{1,1}^k, d_{1,2}^k + r_{1,2}^k, ..., d_{n,n}^k + r_{n,n}^k\} \in \mathbb{R}^{n\times n}, k \in [t-1, t+1]$ to the current cost matrix $\mathbf{M}_t$. That is, the j-th column of the t-th cost matrix $G(\mathbf{M}_t, j)$, which represents the l_2 distance between the j-th super-pixel's feature distance to the t-th column of the objective function $G(\mathbf{O}, t)$ (totally inverse to the computation of the cost element d, and see Eq. 24), can be jointly determined by Eq. 27.

$$G(\mathbf{M}_t, j) \leftarrow \sum_{k=t-1}^{t+1} \sum_{\mathbf{p}_{k,u}\in\xi} G(\mathbf{M}_k, u) \cdot exp(-||\mathbf{c}_{k,u}, \mathbf{c}_{t,j}||_1/\sigma), \tag{27}$$

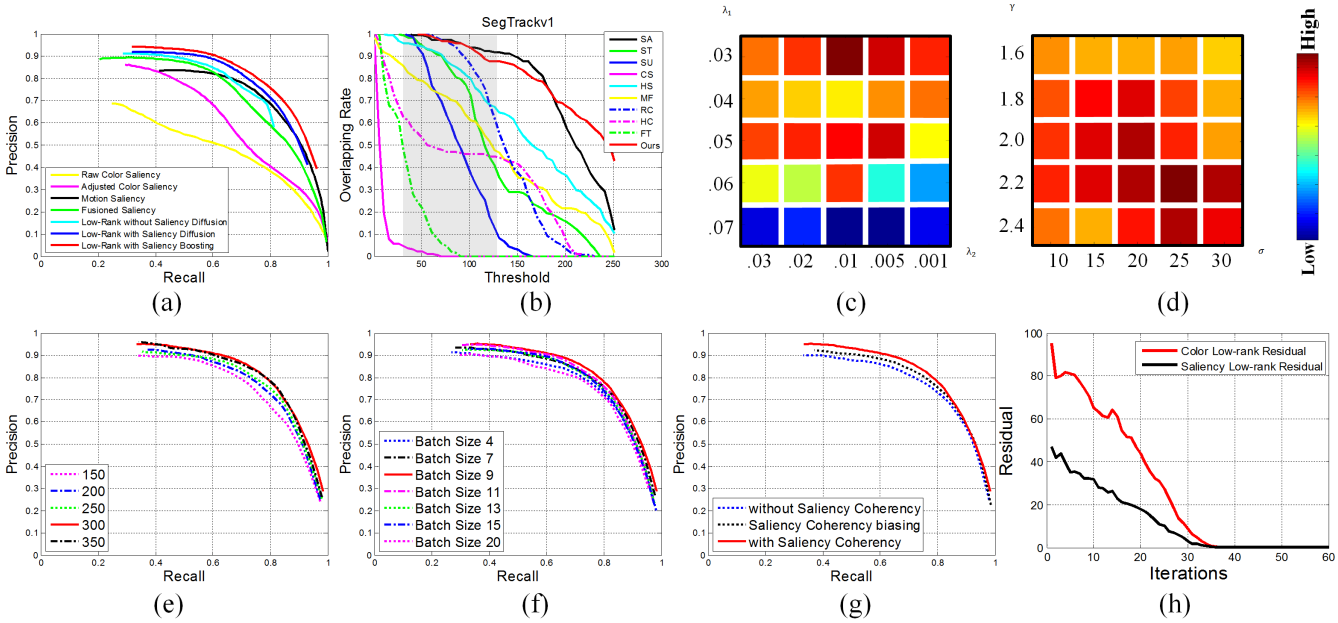


Fig. 14. (a) Precision-recall curves of our method combining with different components over SegTrackv1 [48] and SegTrackv2 [49] datasets, wherein the raw color saliency is computed with the method proposed in [19] based on the features proposed in [28]; (b) demonstrates the overlapping rate comparison results, which indicate the "bad cases" frequency (see detailed discussions in Section VI-A); (c) and (d) are the average F-measure results by adopting different parameter combinations, wherein the color from blue to yellow indicates the performance from worse to better; (e) is the quantitative evaluation of our method over SegTrack v1 [48], SegTrack v2 [49], and BMS [50] dataset when adopting different superpixel numbers; (f) shows the quantitative evaluation of our method under different batch sizes over SegTrack v1 [48], SegTrack v2 [49], BMS [50] dataset; (g) shows the quantitative proofs (over SegTrack v1 [48], SegTrack v2 [49], and BMS [50] dataset) of the performance improvement brought by introducing the saliency coherency, wherein the "Saliency Coherency biasing" means to set $\lambda_1 < \lambda_2$, i.e., $\lambda_1 = 0.003$ and $\lambda_1 = 0.01$, and (h) shows the convergency analysis of our low-rank coherency revealing method.

where θ is identical to Eq. 9, σ is identical to Eq. 8, and ξ controls the neighborhood distance, which satisfies $||\mathbf{p}_{k,u}, \mathbf{p}_{t,j}||_2 \leq \theta/5$. By incorporating Eq. 27 as the weak structure-aware constraint, the super-pixels belonging to the identical spatial-temporal neighborhood will be highly likely selected in the same row of selection matrix ϑ. Then, the global low-rank selection optimization problem can be separately regarded as several independent binary assignment problems as $\min_{\mathbf{Q}_{t+1}} ||\mathbf{M}_t||_2$, which can be solved by Hungarian algorithm [51] in polynomial time (see "Mark 3" in Fig. 7). After obtaining the optimal selection matrices ϑ, the feature submatrix **FC** and **FS** can be updated accordingly via

$$\mathbf{FC}^{t+1} \leftarrow \mathbf{FC}^t \odot \vartheta, \mathbf{FS}^{t+1} \leftarrow \mathbf{FS}^t \odot \vartheta. \tag{28}$$

Finally, the Lagrangian multipliers $\mathbf{Y}_{1,2}$ ($\mathbf{Y}_1$ and $\mathbf{Y}_2$) can be updated by Eq. 29, and ρ can be updated by Eq. 30, wherein no upper bound is needed because of the low discriminative ability of **FC** and **FS**, see the convergence demonstration in Fig. 14(h). And the intermediate low-rank revealing results can be found in Fig. 15.

$$\mathbf{Y}_{1,2}^{t+1} \leftarrow \mathbf{Y}_{1,2}^t + \rho_t(\mathbf{D}_{c,s}^t - \mathbf{L}_{c,s}^t - \mathbf{E}_{c,s}^t), \tag{29}$$

$$\rho_{t+1} \leftarrow \rho_t \times 1.05. \tag{30}$$

F. Low-rank Saliency Diffusion and Boosting

Now the super-pixel's alignment results can be obtained from the selection matrices ϑ, and the i-th video frame's saliency map with temporal smoothness (the final saliency value of the i-th video frame $\mathbf{fS}_i$ in Eq. 31) can be easily revealed by uniformly assigning the averaged low-level saliency value over the temporal scale using Eq. 31.

$$\mathbf{fS}_i = \frac{1}{m-1} \sum_{k=1, i \neq k}^{m} G(\mathbf{FS} \odot \vartheta, k), \tag{31}$$

where G is a column-wise function being identical to Eq. 24, $\mathbf{FS} = \{vec(\mathbf{LS}_1), vec(\mathbf{LS}_2), ..., vec(\mathbf{LS}_m)\}$, and vec denotes the vectorizing function. Although Eq. 31 can guarantee the temporal smoothness of the obtained saliency maps, the incorrect alignments induced by both SLIC super-pixel decomposition and the foreground/background variations, may still easily lead to false-alarm detections. In fact, the sparse component $\mathbf{E}_c$ is a good indicator to distinguish the correct super-pixel's alignments from the incorrect ones. Thus, the non-zero elements in $\mathbf{E}_c$ can be used as the true indicator for the potential incorrect correspondence assignments. Therefore, we can adjust the fused low-level saliency feature matrix **FS** via

$$\widetilde{\mathbf{FS}} \leftarrow \mathbf{FS} \odot \vartheta \tag{32}$$

$$\mathbf{FS} \leftarrow \widetilde{\mathbf{FS}} \cdot (1^{n \times m} - N(\mathbf{E}_c)) + \beta \cdot \bar{\mathbf{FS}} \cdot N(\mathbf{E}_c), \tag{33}$$

$$\beta_{i,j} = \begin{cases} 0.5, & \widetilde{\mathbf{FS}}_{i,j} > \frac{1}{m} \sum_{j=1}^{m} \widetilde{\mathbf{FS}}_{i,j} \\ 2, & otherwise \end{cases}. \tag{34}$$

Here $\bar{\mathbf{FS}} = \frac{1}{m} \sum_{j=1}^{m} \widetilde{\mathbf{FS}}_{i,j}$, $N(\cdot)$ is a normalization function, and $\beta \in \mathbb{R}^{n \times m}$ is a balance-factor matrix, which can be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing

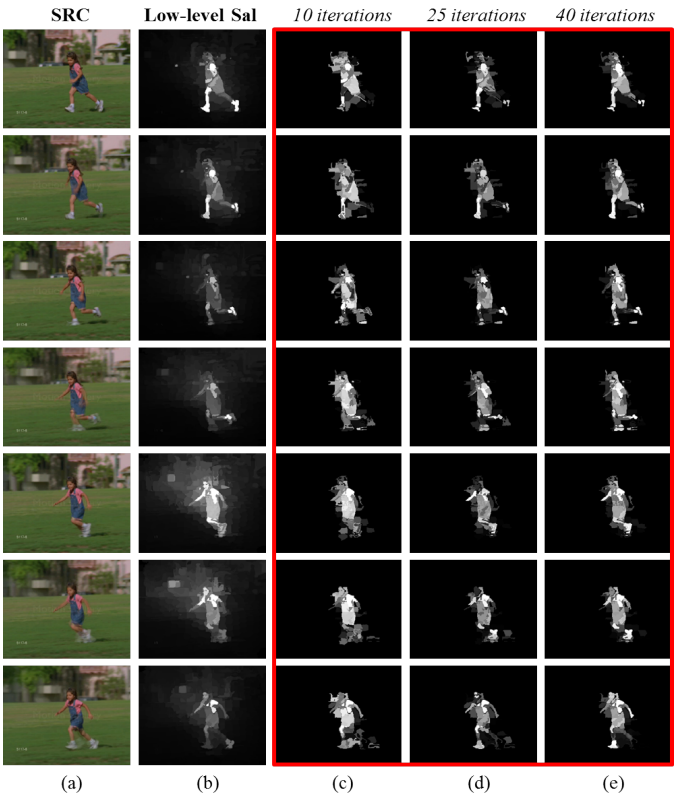


Fig. 15. Demonstration of our low-rank alignment processing. (a) shows the input video frames; (b) shows the obtained low-level saliency clue based on our proposed color and motion fusion strategy; (c),(d) and (e) respectively demonstrate the intermediate low-rank revealing results under different iteration times.

formulated as Eq. 34. In fact, the underlying rationality of this saliency boosting strategy is that, the low-level saliency **LS** should be compensated (if smaller than the average saliency) or penalized (if larger than the average saliency) according to the average saliency degree (i.e., the average saliency value of the aligned super-pixels), if its corresponding color clue is regarded as the sparse part during our low-rank revealing. Then, the final saliency value of the i-th video frame convert into the Eq. 35, where $\bar{\beta} = G(\beta, i)$, and $\bar{\mathbf{E}}_c = G(\mathbf{E}_c, i)$.

$$\mathbf{fS}_i = \frac{\bar{\beta} - (\bar{\beta} - 1) \cdot N(\bar{\mathbf{E}}_c)}{\bar{\beta}(m-1)} \sum_{k=1, i \neq k}^{m} G(\mathbf{FS} \odot \vartheta, k), \qquad (35)$$

Meanwhile, the inter-batch temporal prior $\mathbf{s}_l \in \mathbb{R}^{n \times 1}$, which can be obtained by setting the last video frame of the previous frame batch as the first frame of the current frame batch, is also valuable to ensure the temporal consistence of the saliency over different frame batches. Therefore, we diffuse the batch-level temporal prior $\mathbf{s}_l$ to all the frames of the current frame batch according to the color similarity degree, and the final saliency value of the j-th super-pixel in i-th video frame $\mathbf{fS}_{i,j}$ can be obtained via

$$\mathbf{fS}_{i,j} = \frac{\mathbf{s}_l \cdot w_l + \sum_{i=1}^{m} w_i \cdot \mathbf{fS}_{i,j}}{w_l + \sum_{i=1}^{m} w_i}. \qquad (36)$$

Here both w_l and w_i are the l_2 color distance based weights, $w_l = exp(-||c_{l,j}, c_{i,j}||_2/\sigma)$. Benefitting from our low-rank coherency saliency boosting and diffusion, the performance improvements are demonstrated in Fig. 13, and the quantitative proofs can be found in Fig. 14(a). To sharpen the salient object's boundary and slightly suppress the false-alarm detections, we also conduct pixel-level spatial-temporal smoothing (see the last row in Fig. 13), which is identical to Eq. 8.

VI. Experiments and Evaluations

A. Implementation Details and Parameter Selection

In principle, there are two sets of parameters influencing the performance of our method: the filter strength of the coarse salient object region γ (Eq. 17) and the spatial-temporal diffusion strength σ (Eq. 8), the strength parameters of the sparse component λ_1, λ_2 (Eq. 18). As the first parameter set (γ and σ) can directly affect the quality of the subsequent low-rank coherency revealing procedure's inputs, in order to obtain the optimal solution at the beginning, we comprehensively test their whole effects on the overall performance, and then deal with the optimal selection of the strength parameters of the sparse component λ_1, λ_2.

Parameter γ. In fact, γ in Eq. 17 is an important hard threshold to determine the potential foreground regions. A large γ tends to obtain high-accuracy saliency maps at the expense of recall rate, which can easily filter out some parts around the salient object's boundary. However, a small γ will easily result in the instant frame-level false-alert detections, which would definitely affect the performance of the subsequent spatial-temporal saliency revealing, not to mention the newly-introduced additional computations. Meanwhile, since the movement of the foreground salient object commonly has regular trajectory, it is desirable to select a mild γ, as shown in Fig. 14(d), we set $\gamma = 2.2$ as the optimal choice.

Parameter σ. Actually, both the spatial-temporal diffusion strength σ and its range θ complementarily control the saliency consistence. Since the selection of θ is determined by Eq. 9, the value of σ has direct influence on the consistency of the spatial-temporal saliency. Obviously, a large σ can sharpen the boundary of the foreground object (high precision rate), but it can also easily make the saliency detection performances sensitive to the texture information (poor recall rate), and vice versa. Therefore, as shown in Fig. 14(d), we select 25 as the optimal choice of σ.

Parameters λ_1, λ_2. As shown in Eq. 19, the alternative low-rank revealing process is complementarily determined by both the strength parameter of the low-rank component (μ_1, μ_2) and the strength parameter of the sparse component (λ_1, λ_2). Since the color low-rank constraint ($\mathbf{L}_c$) has much higher priority than the low-level saliency information ($\mathbf{L}_s$), we empirically set $\mu_1 = 0.1$ and $\mu_2 = 0.05$, so that it can lead the low-rank revealing procedure to bias toward the color component via adopting an aggressive singular value thresholding step size (see details in Eq. 22). Therefore, considering the selection of λ_1, λ_2 is done, the optimal choice of λ_1, λ_2 can be obtained based on extensive quantitative experiments, see details in Fig. 14(c). Besides, according to our observations,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing

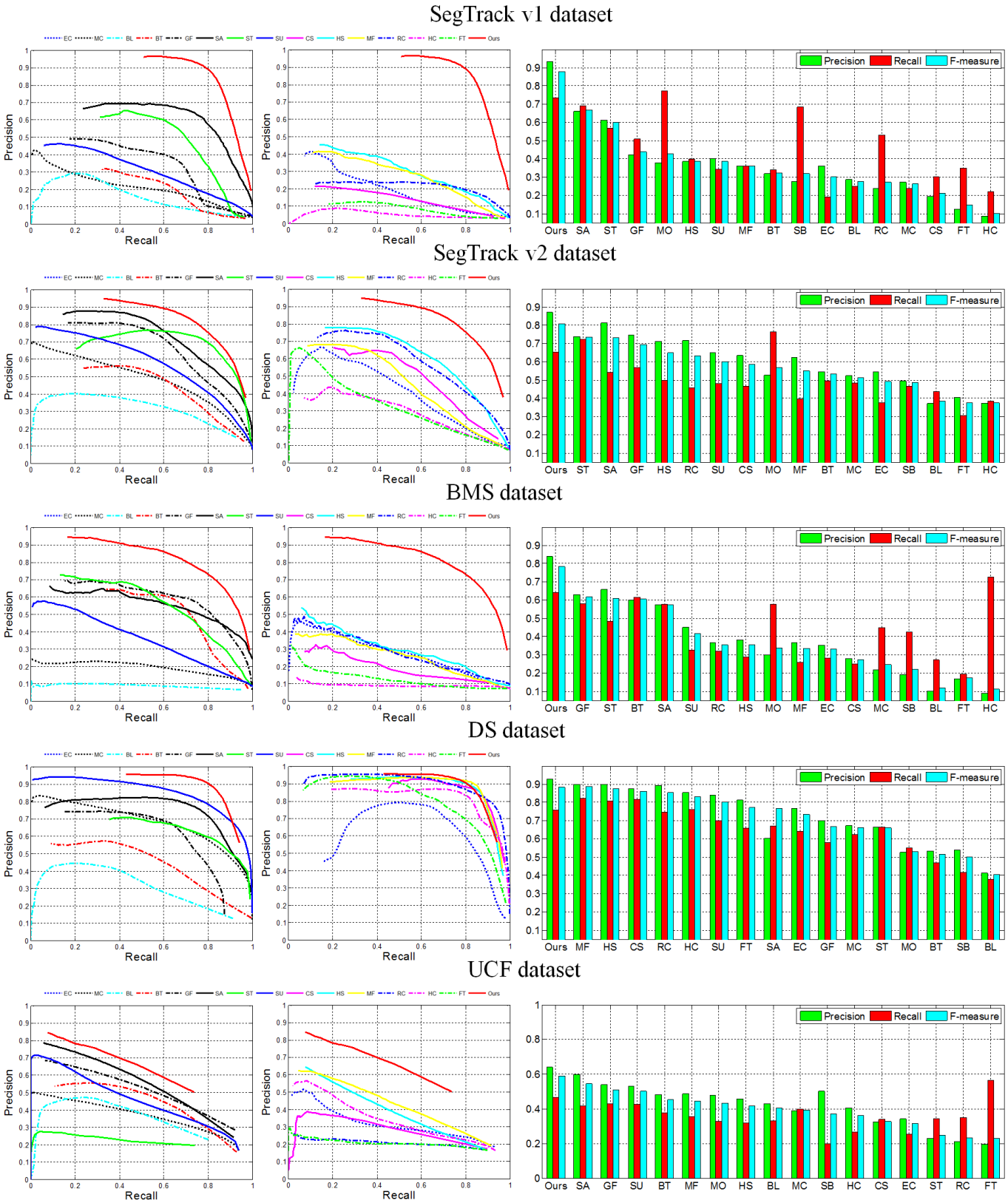


Fig. 16. Quantitative comparisons between our methods and 16 state-of-the-art methods over SegTrack v1 [48], SegTrack v2 [49], BMS [50], DS [52] and UCF [53] dataset (almost 200 video sequences). Those state-of-the-art methods include: SA15 [18], GF15 [40], BT16 [13], ST14 [3], BL14 [14], MC15 [37], SU14 [17], CS13 [28], HS13 [23], MF13 [5], SB14 [12], MO13 [15], EC10 [54], RC11 [4], HC11 [4], and FT09 [21]. The left parts show the Precision-Recall curves, and the right parts show the averaged Precision, Recall and F-measure with fixed thresholds according to the largest F-measure.

the parameters of the SLIC super-pixel decomposition can also affect the performance of our method, thus, we empirically set the super-pixel number to be around 300 (set 15 as the minimum pixel number of each super-pixel) with a mild super-

pixel regularity (0.01).

Also, both the choices of superpixel number and the frame batch size are slight effecting the overall performance of our method, we separately test the performance of our method over different choices toward these two parameters, and the detailed quantitative results can be found in Fig. 14(e) and (f). Obviously from Fig. 14(e), it gives rise to remarkable performance improvement via increasing the superpixel number at the expense of the computation cost. However, the quantitative evaluation results indicate that the optimal superpixel number is 300. Here it should be noted that, the performance with 300 superpixels slightly outperforming 350 superpixels is mainly caused by other parameters (e.g., λ_1, λ_2, γ, etc.), which are optimally selected based on the assumption that the total superpixel number is 300. As for the frame batch size, since the benchmarks adopted in our paper have several short video sequences with total frame number around 20 (e.g., *girl*, *cars1*, etc.), we empirically select a mild batch size (i.e., we set the minimum batch size to 9) in our implementation, and the detailed batch size computation is dynamically determined by the pseudo-code in Algorithm 1.

Algorithm 1. Detailed Batch Size Computation

Input: Total frame number FN;
Minimum batch size $BS = 9$.
Output: Batch size assignment b_t.
Initialization: Batch Number $BN = \lfloor FN/BS \rfloor$;
Batch Residual $BR = BN - BN \times BS$;
$b_t = BS, t \in [1, 2, ..., BN]$; $t = 0$.

While (1)
1. $if(BR == 0)$ $break$; end
2. $t = t + 1$;
3. $b_t = b_t + 1$;
4. $BR = BR - 1$;
5. $if(t > BN)$ $t = 1$; end
End While

Here, we quantitatively evaluate the performance of our method when adopting different batch size ranging from 4 to 20, and the results can be found in Fig. 14(f). Obviously, the overall performance of our method is insensitive to the minimum batch size, because the batch size ranging from 7 to 15 has few influence on the overall performance. Thus, these tiny differences are mainly caused by other parameters (e.g., the hard threshold γ of the coarse foreground mask $\mathbf{F}$, the sparse parameters λ_1 and λ_2 in our low-rank revealing procedure, etc.), which have been optimally selected based on our initial assignment of batch size 9. Specially, we notice that there will be a performance degradation when the minimum batch size is too small (i.e., 4) or too large (i.e., 20). For a small minimum batch size, the low-rank temporal coherency in each batch will become too local to suppress those false-alarm detections. As for a large batch size, since both the salient foreground object and the non-salient backgrounds may vary too much, it will definitely affect the convergency of our low-rank estimation, and thus results in poor video saliency detections.

After determining the aforementioned parameters, we quantitatively evaluate the overall performance of our method by testing different combinations of the components involved in our method, and the results can be found in Fig. 14(a). Obviously, **Raw Color Saliency** exhibits the worst precision-recall (PR) curve, but the performance can be remarkably improved as it could benefit from our **Adjusted Color Saliency** (Section IV-B). Meanwhile, **Fused Saliency** is much better than pure **Adjusted Color Saliency** or pure **Motion Saliency**. Actually, due to the stubborn deficiency of the spatial-temporal consistence constraint, naive low-rank coherency based video saliency (**Low-rank without Saliency Diffusion**) is just a little better than **Fused Saliency**. However, after introducing our low-rank saliency diffusion (**Low-rank with Saliency Diffusion**, see details in Section V) and boosting (**Low-rank with Saliency Boosting**), the overall performance (especially the accuracy rate) is greatly improved without any recall rate degradation.

Particularly, the overlapping rate$\geq 50\%$ (OR, Eq. 37) can be used to measure the success rate of the detection, which can truly indicate the bad-case frequencies (i.e., with a small segmentation threshold (≤ 100), the lower the overlapping rate, the higher the bad-case frequencies).

$$\mathrm{OR} = \frac{area\{\mathrm{ROI_T} \bigcap \mathrm{ROI_G}\}}{area\{\mathrm{ROI_T} \bigcup \mathrm{ROI_G}\}}, \tag{37}$$

where $\mathrm{ROI_T}$ denotes the saliency segmentation results under dynamic threshold settings, and $\mathrm{ROI_G}$ denotes the corresponding ground truth mask. Just like the gray region shown in Fig. 14(b), the RC11 [4] method, which is the most simple regional contrast based image saliency detection method, outperforms all the other image saliency methods.

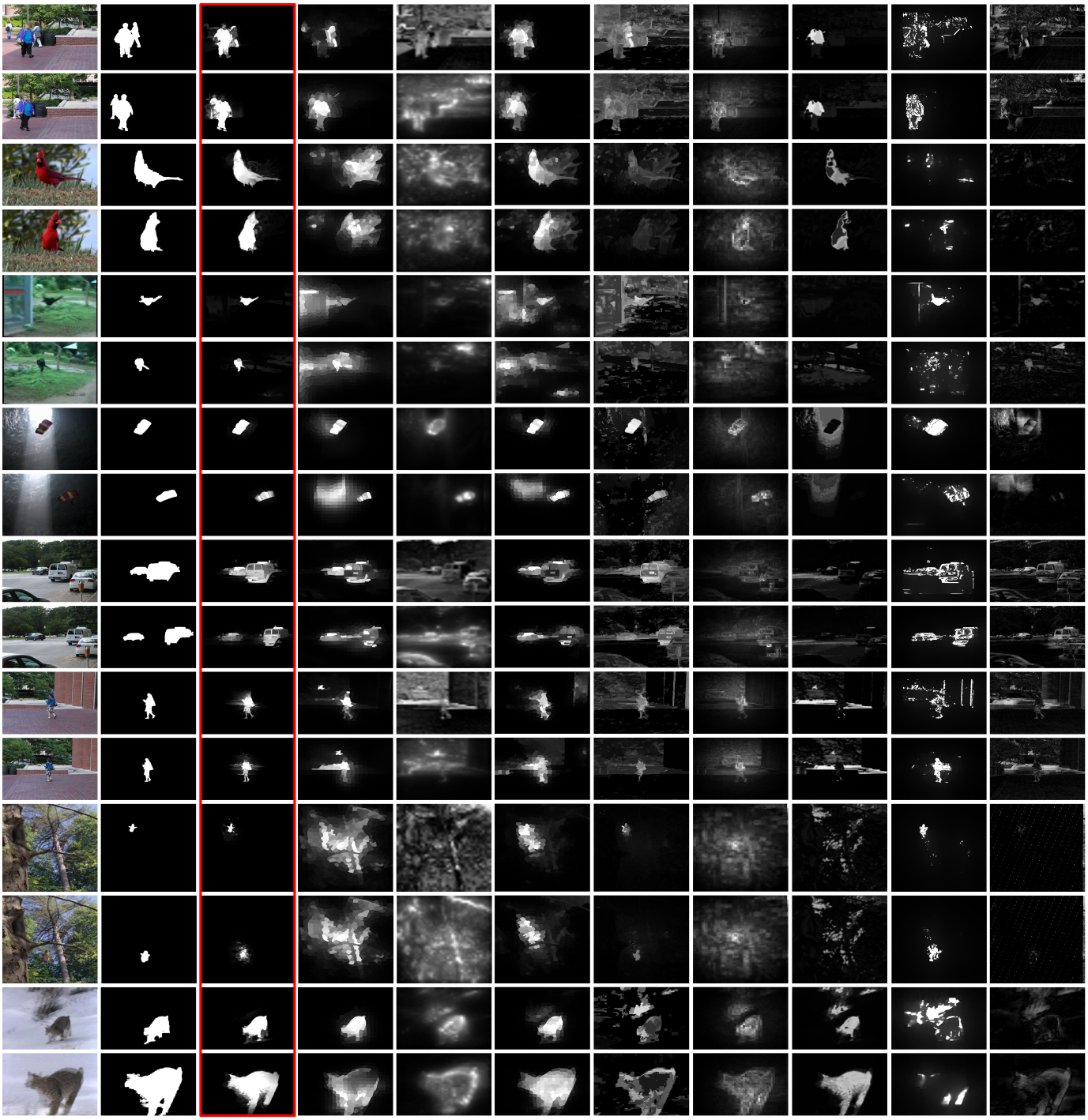
B. Quantitative Evaluations

In this paper, we evaluate the performance of our method over 5 public benchmarks, including SegTrack v1 [48], SegTrack v2 [49], BMS [50], DS [52] and UCF [53] dataset. The SegTrackv1 dataset contains 6 short video sequences with fast object movements compounded by complex surroundings. The SegTrackv2 dataset contains 10 video sequences with mild-level object movements in either stationary or non-stationary scene. The DS dataset contains 10 video sequences with slow object movements and dynamic backgrounds. The BMS dataset contains 26 diverse-length video sequences with various movements. The UCF dataset [53], which is guided by the human eye fixations containing almost 150 sport related video sequences.

It should be noted that, we exclude the *penguin* sequence from the SegTrackv1 dataset and the *marple* sequence from the BMS dataset, because both of these sequences are designed to evaluate video segmentation.

We compare our method with 16 state-of-the-art methods, including SA15 [18], GF15 [40], BT16 [13], ST14 [3], BL14 [14], MC15 [37], SU14 [17], CS13 [28], HS13 [23], MF13 [5], SB14 [12], MO13 [15], EC10 [54], RC11 [4], HC11 [4], and FT09 [21]. To better verify and validate the performance of our method, we leverage the well-recognized precision-recall (PR) as evaluation indicator. To this end, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2670143, IEEE Transactions on Image Processing



(a) SRC (b) GT (c) Ours (d) GF15 (e) MC15 (f) SA15 (g) ST14 (h) SU14 (i) CS13 (j) BT16 (k) BL14

Fig. 17. Qualitative comparisons over SegTrack v1 [48], SegTrack v2 [49], BMS [50] and DS [52] datasets, where (a) denotes the source input video frame, (b) is the ground truth (GT), (c) demonstrates the results obtained by our method our method (highlighted with red rectangle), and some state-of-the-art methods, including GF15 [40],MC15 [37], SA15 [18], ST14 [3],SU14 [17], CS13 [28], BT16 [13], BL14 [14].

alternatively segment the video saliency detection results of different methods with the same threshold ($\mathbf{T} \in [0, 255]$), and the regions whose saliency values are larger than $\mathbf{T}$ are labeled as foreground. If the obtained foreground is consistent with the ground truth mask, it is deemed as successful detection, and the final precision-recall curves are obtained by varying $\mathbf{T}$ from 0 to 255. As the recall rate is inversely proportional to the precision, the tendency of the trade-off between precision and recall can truly indicate the overall video saliency detection performance. As we can see from the PR curves over SegTrackv1, SegTrackv1, and BMS dataset in Fig. 16, our method outperforms all other methods by a large margin. Specifically, the temporal minimization framework of SA15 and GF15, which were originally designed to maintain the saliency consistence, can easily cause error accumulation of false-alarm detections (see Fig. 17). This is why even the sole

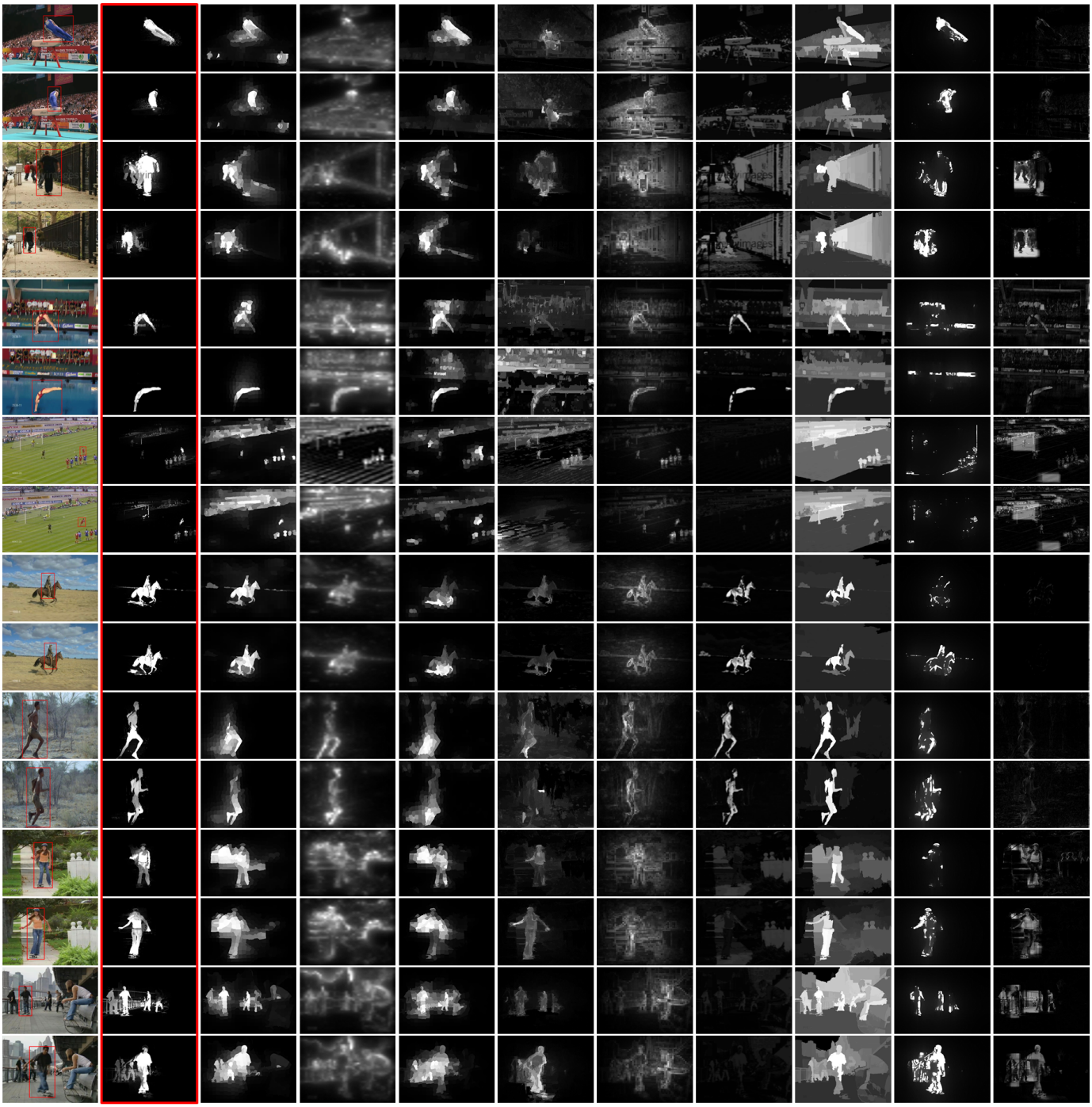


Fig. 18. Qualitative comparisons over UCF dataset [53], where (a) denotes the source input video frame, and the human eye fixation based ground truths (GT) are demonstrated as the red rectangle in (a), (b) demonstrates the results obtained by our method (highlighted with red rectangle), and some state-of-the-art methods, including GF15 [40], MC15 [37], SA15 [18], ST14 [3], SU14 [17], CS13 [28], HS13 [23], BT16 [13], BL14 [14].

Fused Saliency of our method (Fig. 14(a)) still outperforms the SA15 method (Fig. 16). As for ST14 method, because it focuses on contrast based saliency in spatial-temporal scope, it achieves good detection performance over *birdfall* sequence (Fig. 17). However, it also leads to massive false-alarm detections mainly caused by the neglect of the saliency coherency (saliency consistency). Similarly for SU14 and CS13 methods, which only consider the isolated temporal information between two consecutive video frames, both of them perform much worse over the *monkeydog*, *birdfall*, and *parachute* sequences. Meanwhile, since the SB14 and BL14 belong to the modeling based methods, which require a sequence of long period to construct the robust background model, these methods exhibit good performance over stationary video but poor performance over non-stationary videos (e.g., the *birdfall* sequence). Although both MO13 and BT16 adopted the alignment steps to handle the salient motion detections in non-stationary videos, the alignment steps adopted by these methods are either too

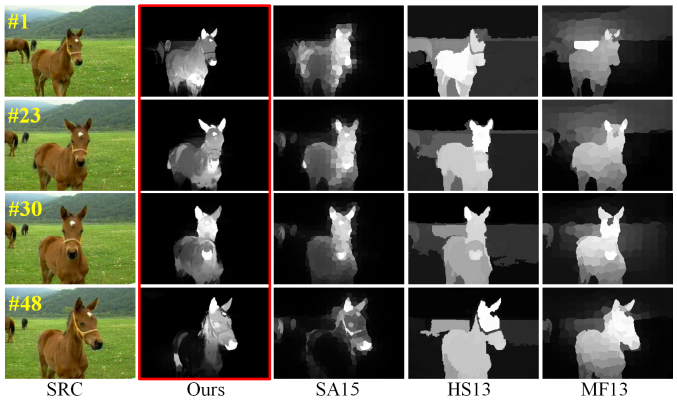


Fig. 19. Illustration of the limitation of our method. The hollow effect easily occurs when the foreground object remains static for a longer period of time, see the bottom row.

local or too global to obtain robust detections, which has been detailed in Section V-A. Furthermore, because of the motion clue deficiency, the performance of the image saliency detection methods (HS13, MF13, RC11, HC11, FT09) are even worse. However, there exists a turning point over the DS (Fig. 16) dateset, wherein these image saliency detection methods achieve better saliency detection performance than SA15 and ST14. This is mainly caused by the improper color saliency and motion saliency fusion. As for the comparison results over the UCF dataset [53], all these compared methods exhibit low recall rate because the human eye fixation guided ground truths are marked by a rectangle box. However, because our method utilize both the color coherency and the saliency coherency to represent the most eye attractor, which is determined by the Optokinetic Reflex system in human brain [55], our method outperforms the other methods by a large margin.

Moreover, we leverage the average precision, recall, and F-measure indicators to demonstrate the superiority of our method. And the F-measure can be computed via

$$\text{F-measure} = \frac{(\beta^2 + 1) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}}, \qquad (38)$$

where Precision denotes the average precision rate, Recall denotes the average recall rate, and the $\beta^2 = 0.3$. It can be easily found in Fig. 16 that our method apparently outperforms other state-of-the-art methods.

C. Limitation and Discussion

Because our method incorporates motion clues into color saliency computation, incorrect low-level saliency clues (the fused saliency) would cause hollow effect when the intermittent foreground object remains static for a longer period of time (i.e., more than 30 frames). As shown in the last row of Fig. 19, the main body of the horse is undetected by our method. In fact, although our method already incorporates previous saliency prior into the current saliency boosting (Eq. 36), it can only alleviate this situation to certain extent, because the current foreground mask will finally filter out those "standstill" regions. Meanwhile, integrating previous foreground motion mask region into the current computation may easily affect the convergency speed of the low-rank revealing procedure and cause false-alarm detections. Also, similar situation can be found for SA15 method, in contrast, both HS13 and MF13 methods can well handle such case, because these image saliency detection methods never consider the motion clues at all.

TABLE II
AVERAGE TIME COST (IN SECONDS) FOR A SINGLE VIDEO FRAME.

Method	Ours	SA15	ST14	SU14	CS13	HS13
Time cost	3.61	2.43	22.1	82.4	3.27	0.432
Method	MF13	SB14	MO13	RC11	HC11	FT09
Time cost	0.213	0.038	0.291	0.213	0.031	0.017
Method	BL14	BT16	MC15	GF15	EC10	NULL
Time cost	48.5	3.16	50.3	12.1	5.36	NULL

Another limitation of our method is that, our method tends to be time-consuming in some sense. Table II documents the average time expense of each method (note that, the runtime of the optical flow computation is excluded). All the methods are run on a computer with Quad Core i7-4790k 4.0 GHz, 16GB RAM and NVIDIA GeForce GTX 970. For a single 300*300 video frame, the low-level saliency computation costs about 0.12s, the low-rank revealing costs about 2.49s (the major bottle neck), the saliency diffusion costs about 0.54s (CUDA accelerated), and the pixel-wise refinement costs about 0.45s (CUDA accelerated). For some cases, high accuracy is actually somehow less desirable in the interest of efficiency, so we suggest reducing the SLIC super-pixel number (e.g., reducing from 300 in the original setting to 200), so that the total time costs can be decreased to about 1.6s.

VII. Conclusion and Future Work

In this paper, we have advocated a novel video saliency detection method, which could produce high-accuracy saliency maps while retaining the temporal saliency consistent. Our method involves several novel technical elements, including: (1) The motion clue guided color contrast computation, which can automatically assign high saliency value to the foreground salient object; (2) The modeling based low-level saliency fusion and diffusion, which guarantees to complementarily leverage both color and motion saliency clues towards producing high-accuracy low-level saliency; and (3) the low-rank coherency based spatial-temporal saliency diffusion and boosting, which gives rise to intrinsic video saliency exploration from the perspective of temporal scope. Moreover, comprehensive experiments and extensive comparisons with the state-of-the-art methods have demonstrated our method's distinct advantages in terms of accuracy and reliability.

As for our near future works, we are particularly interested in reconsidering the low-rank coherency guided motion clue to improve the background extraction techniques, which is expected to conquer several obstinate difficulties in the video surveillance applications (either in the stationary videos or the non-stationary Pan-Tilt-Zoom cameras), including long-period intermittent motions, slow movements in surroundings with dramatic variations, the salient motion detection in low frame rate videos, etc. At the same time, generalizing our key

ideas to facilitate the modeling-based change detection in non-stationary scenarios with complex surroundings also deserves our immediate research endeavor.

REFERENCES

[1] Z. Dong, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 628–635.

[2] 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 Recognition*, vol. 48, no. 9, pp. 2885–2905, 2015.

[3] F. Zhou, S. Kang, and F. Michael, "Time-mapping using space-time saliency," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3358–3365.

[4] M. Cheng, G. Zhang, J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 409–416.

[5] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

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

[7] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 733–740.

[8] Y. Xie, H. Lu, and M. Yang, "Bayesian saliency via low and mid level cues," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 314–325, 2013.

[9] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *British Machine Vision Conference*, 2011, pp. 1–12.

[10] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 853–860.

[11] L. Dong, S. Kaneko, M. Hashimoto, K. Iwata, X. Zhao, and Y. Satoh, "Robust object detection in severe imaging conditions using co-occurrence background model," *International Journal of Optomechatronics*, vol. 8, no. 1, pp. 14–29, 2014.

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

[13] 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 Recognition*, vol. 52, pp. 410–432, 2016.

[14] Z. Gao, L. Cheong, and Y. Wang, "Block-sparse rpca for salient motion detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1975–1987, 2014.

[15] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.

[16] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, 2014.

[17] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[18] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

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

[20] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[21] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[22] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang, "Saliency detection via dense and sparse reconstruction," in *IEEE International Conference on Computer Vision*, 2013, pp. 2976–2983.

[23] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[24] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2376–2383.

[25] J. Li, L. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the secret of image saliency in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2428–2440, 2015.

[26] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[27] J. Yan, M. Zhu, H. Liu, and Y. Liu, "Visual saliency detection via sparsity pursuit," *IEEE Signal Processing Letters*, vol. 17, no. 8, pp. 739–742, 2010.

[28] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[29] S. Huwer and H. Niemann, "Adaptive change detection for real-time surveillance applications," in *IEEE International Workshop on Visual Surveillance*, 2000, pp. 37–46.

[30] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[31] S. Varadarajan, P. Miller, and H. Zhou, "Spatial mixture of gaussians for dynamic background modelling," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2013, pp. 63–68.

[32] G. Bilodeau, J. Jodoin, and N. Saunier, "Change detection in feature space using local binary similarity patterns," in *International Conference on Computer and Robot Vision*, 2013, pp. 106–112.

[33] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split gaussian models," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 420–424.

[34] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*, 2010, pp. 366–379.

[35] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, vol. 9, no. 12, pp. 1–27, 2009.

[36] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *AAAI Conference on Artificial Intelligence*, 2013, pp. 1063–1069.

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

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

[39] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels," *EPFL Technical Report*, 2010.

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

[41] J. Wright, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *Advances in Neural Information Processing Systems*, 2009, pp. 2080–2088.

[42] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2233–2246, 2012.

[43] Z. Zeng, T. Chan, K. Jia, and D. Xu, "Finding correspondence from multiple images via sparse and low-rank decomposition," in *European Conference on Computer Vision*, 2012, pp. 1016–1021.

[44] P. Ji, H. Li, and M. S. Y. Dai, "Robust motion segmentation with unknown correspondences," in *European Conference on Computer Vision*, 2014, pp. 204–219.

[45] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Robust, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Neural Information Processing Systems*, 2009, pp. 2080–2088.

[46] R. Oliveira, J. Costeira, and J. Xavier, "Optimal point correspondence through the use of rank constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 1016–1021.

[47] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, pp. 1–122, 2011.

[48] D. Tsai, M. Flagg, A. Nakazawa, and M. James, "Motion coherent tracking using multi-label mrf optimization," *International journal of computer vision*, vol. 100, no. 2, pp. 190–202, 2012.

[49] F. Li, T. Kim, A. Humayun, D. Tsai, and R. James, "Video segmentation by tracking many figure-ground segments," in *IEEE International Conference on Computer Vision*, 2013, pp. 2192–2199.

[50] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *European Conference on Computer Vision*, 2010, pp. 282–295.

[51] M. James, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

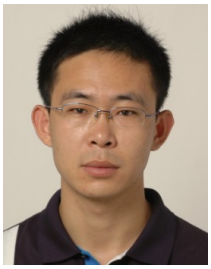
[52] F. Ken, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 638–641.

[53] M. Stefan and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *European Conference on Computer Vision*, 2012, pp. 842–856.

[54] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*, 2010, pp. 366–379.

[55] D. Robinson, "The mechanics of human saccadic eye movement," *The Journal of physiology*, vol. 174, no. 2, pp. 245–264, 1964.

Chenglizhao Chen received the M.S. degree in computer science from Beijing University of Chemical Technology, in 2012. He is currently pursuing the Ph.D. degree in Technology of Computer Application from Beihang University, Beijing, China. His research interests include pattern recognition, computer vision, and machine learning.



Shuai Li received the Ph.D. degree in computer science from Beihang University. He is currently an associate professor at the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include computer graphics, pattern recognition, computer vision, physics-based modeling and simulation, and medical image processing.

Yongguang Wang received the B.S. degree in computer science from Wuhan University of Technology, in 2014. He is currently pursuing the M.S. degree in Technology of Computer Application from Beihang University, Beijing, China. His research interests include pattern recognition, computer vision, and machine learning.

Hong Qin received the B.S. and M.S. degrees in computer science from Peking University. He received the Ph.D. degree in computer science from the University of Toronto. He is a professor of computer science in the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing. He is a senior member of the IEEE.

Aimin Hao is a professor in Computer Science School and the Associate Director of State Key Laboratory of Virtual Reality Technology and Systems at Beihang University. He received his B.S., M.S., and Ph.D. in Computer Science at Beihang University. His research interests are on virtual reality, computer simulation, computer graphics, geometric modeling , image processing, and computer vision.