Consistent Video Saliency Using Local Gradient Flow Optimization and Global Refinement

Wenguan Wang, Jianbing Shen, *Senior Member, IEEE*, and Ling Shao, *Senior Member, IEEE*

***Abstract*— We present a novel spatiotemporal saliency detection method to estimate salient regions in videos based on the gradient flow field and energy optimization. The proposed gradient flow field incorporates two distinctive features: 1) intra-frame boundary information and 2) inter-frame motion information together for indicating the salient regions. Based on the effective utilization of both intra-frame and inter-frame information in the gradient flow field, our algorithm is robust enough to estimate the object and background in complex scenes with various motion patterns and appearances. Then, we introduce local as well as global contrast saliency measures using the foreground and background information estimated from the gradient flow field. These enhanced contrast saliency cues uniformly highlight an entire object. We further propose a new energy function to encourage the spatiotemporal consistency of the output saliency maps, which is seldom explored in previous video saliency methods. The experimental results show that the proposed algorithm outperforms state-of-the-art video saliency detection methods.**

***Index Terms*— Video saliency, energy optimization, gradient flow field, spatiotemporal saliency energy.**

I. Introduction

THE human vision system is remarkably effective in selecting visually important regions in its visual field. This cognitive process enables humans to easily interpret complex scenes in real time without training. Saliency detection is originally a task of predicting scene locations where a human observer may fixate [3], [9]. Recently, it has been extended to detect the salient object, which is the focus of our paper. The output of salient object detection is usually a saliency map where the intensity of each pixel represents the probability of that pixel belonging to the salient object. Over the last few decades, salient object detection has gained much attention for its wide applications, such as unsupervised image segmentation [35], [36], image retargeting [13], [23], [24], object recognition and detection [31], [34], [42], video compression [41] and summarization [30], [32].

The task of salient object detection for still images is to identify the most salient and attention-grabbing object in a static scene. To accomplish this, many image saliency algorithms have been proposed. These methods in general can be categorized as either bottom-up or top-down approaches. Top-down approaches [12], [19], [26], [27], [47] are goal-directed and require an explicit understanding of the context of the image. Supervised learning with a specific class is therefore a frequently adopted principle. Most of the saliency estimation methods are based on bottom-up visual attention mechanisms, which are independent of the knowledge of the content in the image and utilize various low level features, such as intensity, colour and orientation.

While image saliency detection has been extensively studied, computing spatiotemporal saliency for videos is a relatively new problem. For saliency detection in videos, the motion cues of objects provide indication for the visual foregrounds; however, various background motions also bring difficulties for locating the motion objects. There are a few methods [7], [8], [11], [14], [20], [37], [50] designed for video saliency till now and most of them simply add the motion feature to image saliency models (e.g., [7], [8], and [37]) to cope with the extra temporal dimension. Additionally, these methods generally neglect the fact that video saliency should be spatiotemporally consistent, i.e., the saliency values of foreground/background regions should not change dramatically along the time axis. Most of these methods process the input video in a frame-by-frame basis without a measure for global saliency computation across the whole video sequence. Real spatiotemporal processing should design a global saliency algorithm with the consideration of space-time consistency.

Perceptual research [2], [4], [5] indicates that the most important factor in low-level visual saliency is *contrast*; an assumption, called *contrast prior*, is used in almost all bottom-up saliency detection methods, no matter for image saliency [6], [9], [17], [28], [29], [38], [51] or video saliency [7], [8], [15], [16], [33], [49]. This fundamental assumption states that a pixel/patch is salient if its appearance is high contrast within a certain context. Although these contrast prior based methods have achieved success in their own aspects, a few commonly noticeable and critically influencing issues still exist. Typically, these methods detect only high-contrast edges and have their difficulty in highlighting the entire object uniformly, attenuating the smooth object interior.

Manuscript received September 28, 2014; revised March 22, 2015 and May 26, 2015; accepted July 20, 2015. Date of publication July 22, 2015; date of current version August 10, 2015. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB328805, in part by the National Natural Science Foundation of China under Grant 61272359, and in part by the Fok Ying-Tong Education Foundation for Young Teachers through the Joint Building Program of the Beijing Municipal Education Commission. *(Corresponding author: Jianbing Shen.)*

W. Wang and J. Shen are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wangwenguan@bit.edu.cn; shenjianbing@bit.edu.cn).

L. Shao is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. (e-mail: ling.shao@ieee.org).

Another related issue is to choose an appropriate surrounding context range. Too large a surrounding range causes difficulty in distinguishing among similar colors in both foreground and background, while the opposite approach leads to object attenuation.

Aiming to solve these open problems, we propose a novel saliency optimization framework for detecting salient objects in videos. Firstly, we address these problems with the proposed gradient flow field utilizing different distinctive features. According to the famous feature integration theory (FIT) proposed by Tresiman and Gelade [1], basic visual features such as motion and edges are processed at the human pre-attentive stage for visual attention. Following this theory, our gradient flow field deeply incorporates edge and motion features, which considers different foreground motion patterns, resists wrong motion estimation and indicates the locations of salient areas. Benefiting from the estimation for foreground and background, we further introduce local as well as global contrast saliency measures, which are able to remedy the shortcoming of contrast prior based methods and uniformly highlight an entire object. Finally, we present a spatiotemporal energy function to encourage the spatiotemporal consistency of the saliency maps, which is critical for video saliency detection. The source code of this work will be available at.[1]

Compared to the existing approaches, the proposed method offers the following contributions:

- A novel video saliency detection method is proposed for automatically locating the visual foregrounds with a low constraint on their appearances and motion patterns.
- We propose a gradient flow field, which deeply incorporates intra-frame and inter-frame information to efficiently detect the salient regions.
- A spatiotemporal saliency energy function is presented to encourage the spatiotemporal consistency of the output video salience maps.

II. Related Work

We now briefly review previous work along two lines: bottom-up image saliency and video saliency, as our work is on bottom-up salient object detection method for video sequences.

Bottom-Up Image Saliency: Image saliency has been extensively studied for decades. Instead of surveying the large volume of literature, we focus on recent works of bottom-up image saliency that are most related to our method, and analyze their properties and limitations. Bottom-up approaches resemble mechanisms in pre-attentive vision and are largely independent of the knowledge of content in the image. These saliency methods can be broadly classified into two categories: frequency-domain methods and spatial-domain methods.

For frequency-domain methods, it is usually assumed that globally less frequent features are more salient, and frequency analysis is carried out in the spectral domain. For example, Hou and Zhang [28] proposed a saliency detection algorithm using spectral residual on the log spectra representation for images. Guo *et al.* [16] claimed that the phase spectrum of the Fourier transform is the key in obtaining the location of salient regions. Later, Achanta *et al.* [22] introduced a frequency-tuned approach to estimating center-surround contrast using color and luminance features. Fang *et al.* [10], [13] introduced saliency models in compressed domain for adaptive retargeting.

Our method detects saliency in the spatial-domain. Spatial-domain usually adopt several visual cues (color, contrast, etc.) for salient object detection. As one of the earliest methods, Itti *et al.* proposed a saliency model based on the neuronal architecture of the primates' early visual system [29]. Specifically, they proposed a set of center-surround operations as local feature contrast in the color, intensity, and orientation. These visual features are then linearly combined to generate the final saliency map. Based on Itti's model, Walther and Koch [31] created SaliencyToolBox (STB) and Gao *et al.* [6], [7] further utilized the center-surround mechanism for both image and video saliency detection. Harel *et al.* [27] developed a saliency detection model by using a graph-based dissimilarity measure. In [38], Goferman *et al.* built a content-aware saliency detection with the consideration of the contrast from both local and global perspectives. Klein and Frintrop [40] presented a framework for saliency detection based on the efficient fusion of different feature channels and the local center-surround hypothesis. In [48], Cheng *et al.* aimed at two saliency indicators: global appearance contrast and spatially compact distribution. Recently, several methods [46], [51] exploited some information of the background, called *boundary prior*. These methods treat image boundaries as background, further enhancing saliency computation.

Video Saliency: Video saliency detection aims to identify the most salient object from video sequences. To the best of our knowledge, there are only a few methods specifically designed to address this problem till now, and most of them are based on bottom-up mechanisms. Different from image saliency detection, moving objects catch more attention of human beings than static ones, even if the objects have large contrast to their neighbors in static images. In other words, motion is the most important cue for video saliency detection, which makes deeper exploration of the inter-frame information crucial. The existing methods, however, usually build their system with a simple combination of image saliency models with motion cues. For example, Gao *et al.* [7] extended their image saliency model [6] by adding the motion channel for prediction of human eye fixations in dynamic scenes based on the center-surround hypothesis. Similarly, Mahadevan and Vasconcelos [8] combined center-surround saliency with the dynamic textures for spatiotemporal saliency using the saliency model in [6]. In [37], Seo *et al.* computed the so-called local regression kernels from the given video, measuring the likeness of a pixel (or voxel) to its surrounding. They extended their model for video saliency detection straightforwardly by extracting a feature vector from each spatiotemporal 3D cube. Recently, Rahtu *et al.* [20] used a statistical framework and local feature contrast in illumination, color, and motion for formulating final saliency maps. Fu *et al.* [49] proposed a cluster-based saliency method, where three visual attention cues: contrast, spatial, and global

[1] http://github.com/shenjianbing/videosal

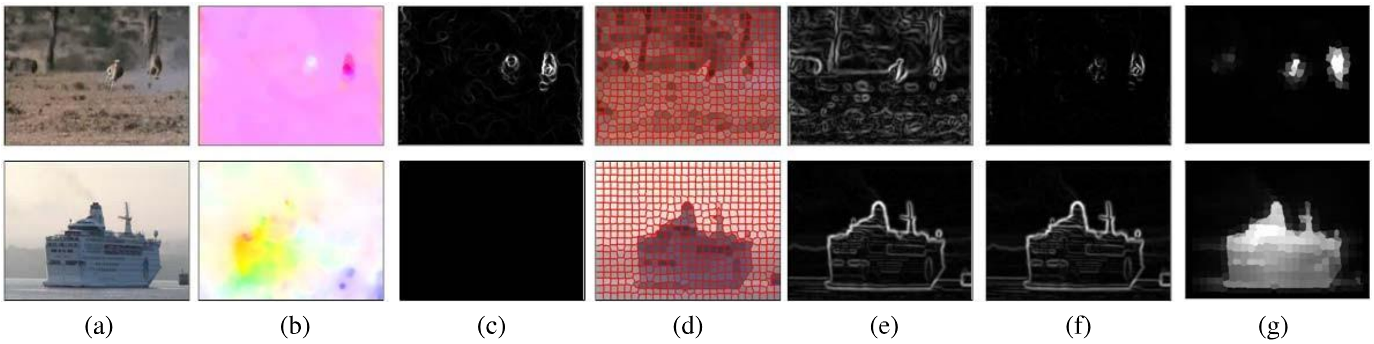


Fig. 1. Illustration of our saliency estimation steps. (a) Two frames from different input videos; (b) optical flow fields v of frames in (a); (c) the optical flow gradient magnitude M^o of v by (2); (d) abstraction of frames in (a) through SLIC; (e) the color gradient magnitude M^c of abstraction in (d) by (1); (f) spatiotemporal gradient field M by combining M^c and M^o; (g) our saliency detection results computed by the gradient flow field.

correspondence, are devised to measure the cluster saliency. Zhou *et al.* [33] adopted space-time saliency to generate a low-frame-rate video from a high-frame-rate input using various low-level features and region-based contrast analysis. Nevertheless, these approaches process the input video sequence in a frame-by-frame basis, ignoring the fact that video saliency maps should be spatiotemporally consistent. It can be seen that video saliency detection is still an emerging and challenging research problem to be further investigated.

III. Our Approach

The goal of our work is to produce the accurate spatiotemporal saliency maps, where the objects of interest are discovered and the foreground and background are separated over the whole video. Our method has three main steps: saliency estimation, saliency cues refinement and spatiotemporal saliency optimization.

A. Saliency Estimation

Given an input video sequence, we first obtain the superpixels for each frame to preserve the initial structure elements of video contents, while the undesirable details are efficiently simplified and ignored. Strong edges or contours in the frame are preserved as boundaries between superpixels. These boundaries and discontinuities reveal the important content of the video frame (see bottom image of Fig. 1(e)). However, the color discontinuities are not discriminative enough in a complex scenario with highly textured background areas (see top image of Fig. 1(e)). Motion information can be reasonably assumed to contribute to salient region detection, since the pixels which change abruptly in the optical flow field often attract more attention by people (see top image of Fig. 1(b)). Nevertheless, motion information alone is insufficient for identifying the salient regions since the moving objects may have very small optical flow, or the background is dynamic (see bottom image of Fig. 1(b)).

According to the aforementioned analysis, we integrate both discontinuity and motion information into our saliency optimization framework, which is more reliable than either alone. Let $\mathbf{I} = \{I_1, I_2, \cdots\}$ be a set of frames of the input video $\mathbf{I}$. We first apply the SLIC [25] to abstract each frame I_k into superpixels $\mathbf{R}_k = \{R_{k,1}, R_{k,2}, \cdots\}$, and the corresponding abstraction of frame I_k is denoted by I^x_k (see Fig. 1(d)). Then we compute the color gradient magnitude M^c_k of abstraction frame I^x_k at position $\mathbf{x}(x, y)$:

$$M^c_k(\mathbf{x}) = \star\nabla I^x_k(\mathbf{x})\star. \qquad (1)$$

We adopt the large displacement motion estimation algorithm [39] to compute the optical flow. Let v_k be the optical flow field of I_k, we then compute the magnitude of the gradient of v_k:

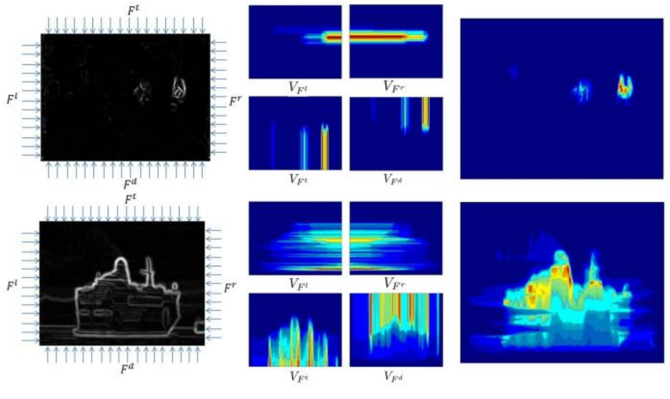
$$M^o_k(\mathbf{x}) = \star\nabla v_k(\mathbf{x})\star. \qquad (2)$$

The color gradient magnitude M^c_k and optical flow gradient magnitude M^o_k are integrated into a spatiotemporal gradient field M_k of frame f_k as follows:

$$M_k(\mathbf{x}) = \begin{cases} M^c_k(\mathbf{x}) \cdot (1 - exp(-\lambda \cdot M^o_k(\mathbf{x}))) & \text{if } \max(M^o_k) > 1; \\ M^c_k(\mathbf{x}) & \text{if } \max(M^o_k) \leq 1. \end{cases} \qquad (3)$$

In practice, we have found that M^o has more discriminative ability when $\max(M^o) > 1$. Therefore, an exponential function is employed to emphasize M^o_k. λ is a scaling factor for the exponential function, and we set $\lambda = 1$ in all our experiments. When $\max(M^o) \leq 1$, the scene is nearly static and the very small optical flow is not discriminative. Fig. 1(c) shows the optical flow gradient magnitude M^o between two frames from different videos. From the top one we can find that it is very helpful to reveal the moving object when the max value of M^o is larger than 1. When the object is nearly static, *e.g.,* the boat in Fig. 1 is moving slowly, M^o is very small and becomes less effective (see the bottom image of Fig. 1(c)). Clearly, the spatiotemporal gradient field (see Fig. 1(f)) is able to reveal the locations of visually important regions, which are better compared to considering either color discontinuities or motion information only. Based on this effort, we present an efficient and robust saliency estimation algorithm in the following.

We can imagine that many flows start from the four sides of the frame and end at the opposite sides along the vertical/horizontal directions (see Fig. 2(a)). When the flow passes through the current frame, the value of the flow will increase with the value of the corresponding spatiotemporal gradient field. Let the size of the frame I_k be $n \times m$, we define



(a) (b) (c)

Fig. 2. Illustration of our gradient flow field using (6) and (7). (a) Gradient flows start from the four sides of the frame and end at the opposite sides along the vertical/horizontal directions. When the flow passes through the spatiotemporal gradient field, the value of the flow will increase. (b) Four value maps $V_F(\mathbf{x})$ correspond to four kind of flows along different directions through (6). (c) Gradient flow field computed by (7).

a gradient flow F^l from the left side of I_k to the right side as follows:

$$F^l = \{\mathbf{f}^l_j\}^m_{j=1} = \{(i, j)\}^m_{j=1}, \quad 1 \leq i \leq n. \tag{4}$$

A left-to-right gradient flow F^l contains all the pixels in the same row of the frame. Similar to the definition of F^l, we define right-to-left gradient flow F^r, top-down gradient flow F^t and down-top gradient flow F^d:

$$\begin{aligned} F^r &= \{\mathbf{f}^r_j\}^m_{j=1} = \{(i, m - j + 1)\}^m_{j=1}, \quad 1 \leq i \leq n, \\ F^t &= \{\mathbf{f}^t_i\}^n_{i=1} = \{(i, j)\}^n_{i=1}, \quad 1 \leq j \leq m, \\ F^d &= \{\mathbf{f}^d_i\}^n_{i=1} = \{(n - i + 1, j)\}^1_{i=n}, \quad 1 \leq j \leq m. \end{aligned} \tag{5}$$

For frame I_k, the pixels in the path of flow F (e.g. left-to-right gradient flow f^l) will be $I^F_k = \{I_k(\mathbf{f}^l_j)\}^m_{j=1} = \{I_k(i, j)\}^m_{j=1}$. We then formulate the value of the flow at position $\mathbf{x}(x, y)$ as follows:

$$\begin{aligned} V_{F^l}(\mathbf{x}) = V_{F^l}((x, y)) &= \overset{y}{\underset{j=1}{}} M(\mathbf{f}^l_j) = \overset{y}{\underset{j=1}{}} M((x, j)) \\ &= M(\mathbf{x}) + V_{F^l}((x, y - 1)). \end{aligned} \tag{6}$$

The value of a left-to-right gradient flow F^l will increase when it passes through the spatiotemporal gradient field (see Fig. 2(b)). Based on the gradient flow, we build a gradient flow field T as follows:

$$\begin{aligned} T(\mathbf{x}) &= \min(V_{F^l}(\mathbf{x}), V_{F^r}(\mathbf{x}), V_{F^t}(\mathbf{x}), V_{F^d}(\mathbf{x})) \\ &= M(\mathbf{x}) + \min \; V_{F^l}((x, y - 1)), V_{F^r}((x, y + 1)), \\ &\qquad V_{F^t}((x + 1, y)), V_{F^d}((x - 1, y)) \; . \end{aligned}$$

some noisy points. That is because the gradient flow only considers all the points in the same straight line, and the gradient flow cannot pass these noisy points. Therefore, we redefine our definition of the gradient flow in (6) as follows:

$$\begin{aligned} V_{F^l}(\mathbf{x}) = M(\mathbf{x}) + \min(&V_{F^l}((x - t, y - 1)), \\ &\ldots, V_{F^l}((x, y - 1)), \\ &\ldots, V_{F^l}((x + t, y - 1))). \end{aligned} \tag{8}$$

In this way, a gradient flow considers all the surrounding $(2t + 1)$ gradient flows. (6) can be viewed as a special case of (8) when $t = 0$. Fig. 3 gives an illustration of gradient flow fields computed by (7) and (8) with different values of t. When $t \in \{1, 2, 3\}$, the estimation for the visual importance map by the gradient flow field is more correct than the estimated one when $t = 0$. However, the performance of the gradient flow field will decrease when the value of t increases too much. That is because more object boundaries are ignored when more gradient flows are taken into account (see Fig. 3(e) and (f)). Therefore, we set the value of t as 2 in all our experiments.

We then average the gradient flow field in a region level. Given the gradient flow field T_k of I_k, we can get the following region-averaged gradient flow field T^R_k:

$$T^R_k(\mathbf{x}) = \frac{_{\mathbf{x_r} \in R^{\mathbf{x}}_k} T_k(\mathbf{x_r})}{|R^{\mathbf{x}}_k|}, \tag{9}$$

where $R^{\mathbf{x}}_k$ indicates the region that pixel $\mathbf{x}$ belongs to, $\mathbf{x_r}$ indicates the pixels in region $R^{\mathbf{x}}_k$, and $|R^{\mathbf{x}}_k|$ denotes the number of pixels in region $R_{\mathbf{x}}$. Then we normalize T^R_k with values in [0, 1].

Our gradient flow field can obtain satisfying results in most situations, while it will introduce inaccurate saliency results when the optical flow is not correct. As shown in Fig. 4(b), the top image shows an example of the optical flow gradient magnitude M^o with acceptable motion estimation. While the motion estimation in the bottom one is incorrect, this situation makes the corresponding spatiotemporal gradient field M (see the bottom image of Fig. 4(c)) unreliable. Fig. 4(d) shows our gradient flow field based saliency results corresponding to two adjacent frames. Two saliency maps are quite different - the top one is more accurate than the bottom one. That is because of the unsatisfactory optical flow estimation for the bottom frame.

For an input video sequence, we can observe that the visual backgrounds are consistent between adjacent frames. Therefore, we can safely assume that the visual background regions in one frame are also not salient in the next frame. Based on this assumption, we first rewrite (1) as follows:

$$M^c_k(\mathbf{x}) = \begin{matrix} \varepsilon \cdot \ast\nabla I^c_k(\mathbf{x})\ast & \text{if } T^R_{k-1}(\mathbf{x}) < \eta; \\ \ast\nabla I & \quad R \end{matrix} \tag{10}$$

(7)

From (7), we can find that the value of T at $\mathbf{x}$ is the value of M at $\mathbf{x}$ by adding the minimum value of its 4-neighbors. Fig. 2(c) shows that our gradient flow field T can correctly estimate the visually important regions. However, the gradient flow field T will be influenced by the outlier of its neighbors

because we only consider 4-neighbors. T is also sensitive to

$$_k(\mathbf{x})* \qquad \text{if } T_{k-1}(\mathbf{x}) \geq \eta.$$

That means we will decrease the color gradient magnitude at position $\mathbf{x}$ in current frame I_k by multiplying ε^c ($\varepsilon^c \in [0, 1]$), when the visual importance at position $\mathbf{x}$ is less than η ($\eta \in [0, 1]$) in previous frame I_{k-1}. Similarly, (2) is rewritten as:

$$M^o_k(\mathbf{x}) = \begin{matrix} \varepsilon^Q * \nabla v_k(\mathbf{x})* & \text{if } T^R_{k-1}(\mathbf{x}) < \eta; \\ * \nabla v_k(\mathbf{x})* & \text{if } T^R_{k-1}(\mathbf{x}) \geq \eta. \end{matrix} \qquad (11)$$

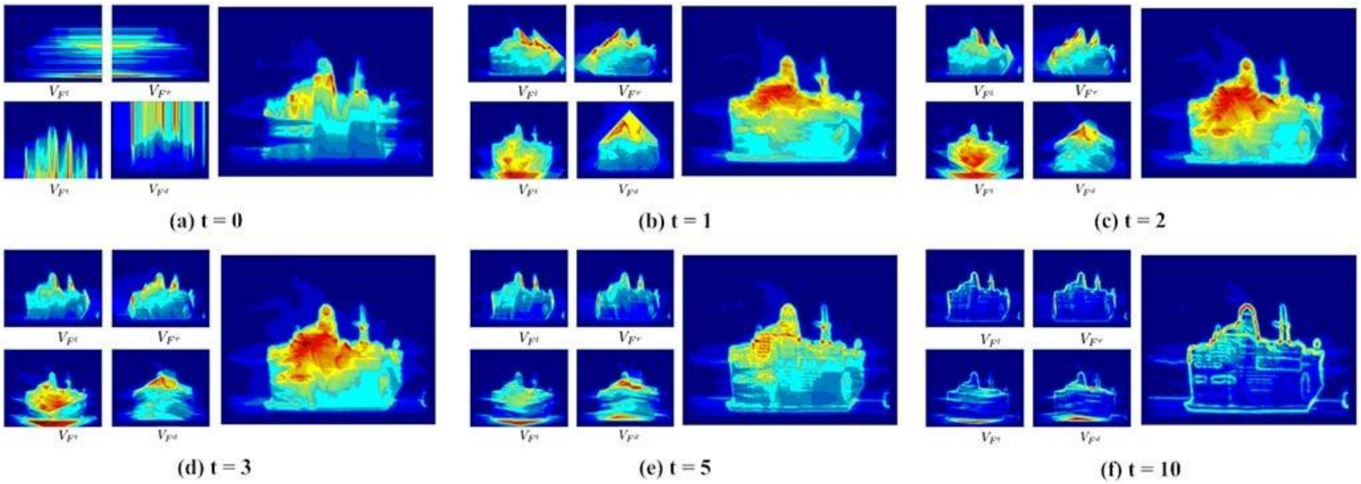


Fig. 3. Our gradient flow field and saliency maps generated by previous representative contrast prior based methods. (a)-(f) Our gradient flow fields computed by (7) and (8) with different values of t.

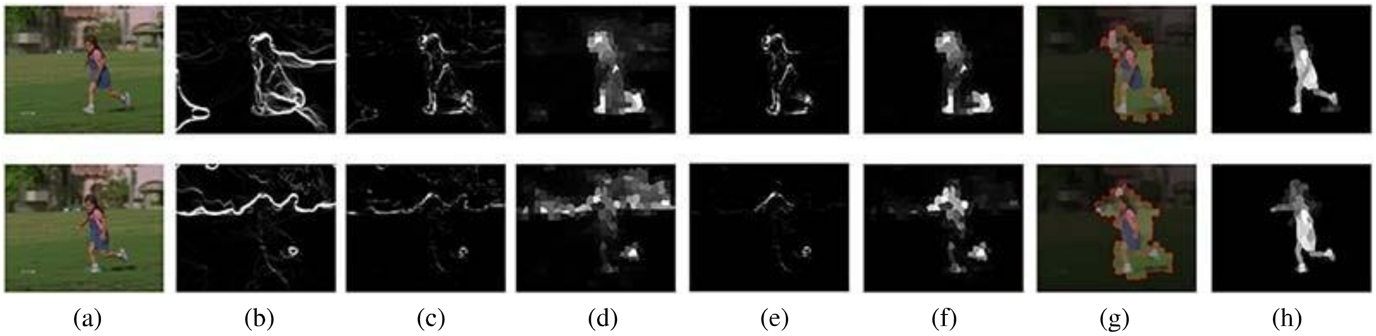


Fig. 4. Illustration of our performance when the optical flow is inaccurate. (a) Two adjacent frames I_{k-1} and I_k from the input video. (b) Optical flow gradient magnitude M^o via (2). From the bottom image we can find that the motion estimation is incorrect. (c) Corresponding spatiotemporal gradient field M to (b). (d) Our gradient flow field based saliency map for frames I_{k-1} and I_k by (9). (e) Spatiotemporal gradient field M with refined M^c by (10) and M^o by (11), which is better than (c). (f) More correct saliency results from (e). (g) The dark regions indicate the virtual background regions. (h) Our saliency map S^C combining global and local saliency cues by (17).

We set the value of parameter η as 0.2. It means that the pixel/region is seen as visual background when its corresponding value in the gradient flow field is less than 0.2. We use the set $\mathbf{B}_k$ to indicate the background regions in the k-th frame. Fig. 4(g) shows examples for the background region $\mathbf{B}_k$. Fig. 4(f) shows our gradient flow field based saliency map using (10) and (11). Comparing with the results in Fig. 4(d), the results in Fig. 4(f) give a significant improvement. Although a part of foreground is not salient, it is observed that the objects and backgrounds are roughly separated. In the next step, we will further introduce global and local saliency measures to generate more uniform saliency maps by utilizing this property.

B. Saliency Cues Refinement

Local Saliency Cue: Many works [6], [9], [17], [28], [29], [38], [51] use the region contrast against its surrounding scales as a saliency cue, which is computed as the summation of its color differences from other regions and weighted by their spatial distances. In this way, the contrast saliency cue for superpixel $R_{k,p}$ in frame I_k can be written as

$$C(R_{k,p}) = \sum_{p^{\prime}=1}^{|\mathbf{R}_k|} \varphi(R_{k,p}, R_{k,p^{\prime}}) * c_{k,p} - c_{k,p^{\prime}} *_2, \qquad (12)$$

where $c_{k,p}$ and $c_{k,p^{\prime}}$ are colors of regions $R_{k,p}$ and $R_{k,p^{\prime}}$ respectively. $\varphi(R_{k,p}, R_{k,p^{\prime}}) = exp\{-D(R_{k,p}, R_{k,p^{\prime}})/\sigma^2\}$ controls the spatial influence between two regions $R_{k,p}$ and $R_{k,p^{\prime}}$. $D(R_{k,p}, R_{k,p^{\prime}})$ is a square of Euclidean distance between region centers of $R_{k,p}$ and $R_{k,p^{\prime}}$.

With the $\varphi(R_{k,p}, R_{k,p^{\prime}})$ term, close regions have a larger impact than distant ones. Parameter σ controls the range of neighborhood. Clearly, (12) measures color contrast of all the surrounding regions. However, some limitations of this region contrast based saliency measure are obvious. The first one is that the regions distinct from the surrounding should be highlighted not matter whether this region belongs to the foreground or the background. Secondly, this strategy causes the object attenuation problem. Thirdly, the parameter σ, which is the particular scale threshold and quite important for the contrast based saliency cue, is difficult to set a suitable value. If σ is large, all regions will be compared in a near-global manner. When σ is set a small value, small and unsuitable neighborhoods will be considered. In previous approaches, the value of σ is manually set. This illustrates that the definition of the range of the neighborhood is not clear, which needs to be further exploited. These limitations are mainly because that this saliency measure lacks a method to confirm where the surrounding background is.

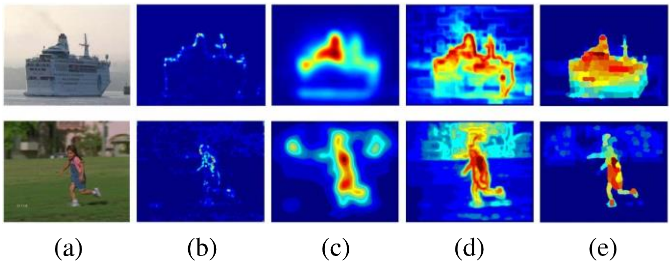


Fig. 5. Comparison between our proposed surrounding background contrast based saliency measure with traditional contrast prior based saliency measures. (a) Input frame. (b)-(d) Saliency maps computed by different state-of-the-art contrast prior based methods: Hou and Zhang [28], Itti *et al.* [29] and Goferman *et al.* [38]. (e) Saliency maps computed by our local saliency cue with (12).

A few examples for these aforementioned limitations are illustrated in Fig. 5(b)–(d), which shows the saliency results generated by three representative contrast prior based methods [28], [29], [38]. In order to overcome these disadvantages, we reintroduce the foreground-background contrast assumption into the saliency detection problem by using our gradient flow field to effectively detect background regions. We assume that a region highly different from its surrounding background is salient. The enhanced saliency measurement, called *surrounding background contrast*, is defined as

$$lC(R_{k,p}) = \frac{\sum_{p'=1, R_{k,p'} \in \mathbf{B}_k}^{|\mathbf{R}_k|} \varphi'(R_{k,p}, R_{k,p'}) \| c_{k,p} - c_{k,p'} \|_2}{\sum_{p'=1, R_{k,p'} \in \mathbf{B}_k}^{|\mathbf{R}_k|} \varphi'(R_{k,p}, R_{k,p'})}, \tag{13}$$

(13) measures color contrast to surrounding background regions, which can be viewed as the average appearance distance from a region to its surrounding background. The function $\varphi'(R_{k,p}, R_{k,p'})$ is defined as

$$\varphi'(R_{k,p}, R_{k,p'}) = exp\{-D(R_{k,p}, R_{k,p'})/D_s(R_{k,p})\},$$
$$\text{where } D_s(R_{k,p}) = \arg\min_{p'} D(R_{k,p}, R_{k,p'}),\ R_{k,p'} \in \mathbf{B}_k. \tag{14}$$

The function $D_s(R_{k,p})$ indicates the shortest Euclidean distance between region centers of $R_{k,p}$ and other background regions of $\mathbf{B}_k$. According to (13), the object regions of $\mathbf{R}_k - \mathbf{B}_k$ receive higher saliency values compared with the background regions with their high color contrast. And the saliency values of background regions of $\mathbf{R}_k - \mathbf{B}_k$ should be effectively suppressed by only considering the surrounding background regions of $\mathbf{B}_k$. According to (14), if a region is spatially close to the background, the value of $D_s(\cdot)$ is small and the contrast measurement is mainly influenced in a relatively smaller scope. If the value of $D_s(\cdot)$ is larger, which means a region is far from the background, the region will be compared with background regions in a larger extent. (14) enforces the regions of $\mathbf{R}_k - \mathbf{B}_k$ to be compared with surrounding backgrounds and function $D_s(\cdot)$ is able to automatically adjust the neighborhood scope of every region, which remedies the limitations of previous contrast prior based methods.

Global Saliency Cue: The surrounding background contrast measure is a kind of local saliency cue, which considers the averaged distance between a region and surrounding backgrounds. We further define a global saliency measure of a superpixel as the length of its shortest distance to the virtual backgrounds. The distance between any two superpixels $D_g(R_{k,p}, R_{k,p'})$ considers the color distance and the gradient flow field distance, which is defined as

$$D_g(R_{k,p}, R_{k,p'}) = \min_{r_1 = R_{k,p}, r_2, \ldots, r_w = R_{k,p'}} \sum_{i=1}^{w-1} \| c_i - c_{i+1} \|_2 \cdot \exp(|T_k^R(r_i) - T_k^R(r_{i+1})|),$$
$$\text{s.t. superpixel } r_i \text{ and } r_{i+1} \text{ are adjacent,} \tag{15}$$

where c_i ($T_k^R(r_i)$) and $c_{i'}$ ($T_k^R(r_{i+1})$) are the gradient flow field values of regions r_i and r_{i+1} respectively. The distance D_g can be efficiently solved by Johnson's algorithm [18]. The global saliency cue $gC(R_{k,p})$ of a superpixel $R_{k,p}$ is the shortest distance from $R_{k,p}$ to the background superpixels $\mathbf{B}_k$, which is offered by our gradient flow field.

$$gC(R_{k,p}) = \min_{p'} D_g(R_{k,p}, R_{k,p'}), \quad \text{s.t. } R_{k,p'} \in \mathbf{B}_k. \tag{16}$$

If superpixel $R_{k,p}$ is outside the desired object, its foreground probability is small because there possibly exists a relatively shorter pathway to backgrounds. The superpixels which are connected by this pathway have less difference with background superpixels in color space and gradient flow field. Whereas, if superpixel $R_{k,p}$ is inside the object, this superpixel is distinct from background superpixels, which increases the distance $gC(R_{k,p})$.

For frame I_k, we normalize both global saliency cue gC and local saliency cue lC to the range of [0, 1]. Hence we combine these terms to compute a saliency value S^C for each superpixel in $\mathbf{B}_k$ as follows:

$$S^C(R_{k,p}) = \min\{lC(R_{k,p}), gC(R_{k,p})\}. \tag{17}$$

In practice, we find that setting the saliency as the minimal value of surrounding background contrast bC and global saliency cue gC will produce outstanding and uniform saliency results (see Fig. 4(h)).

C. Spatiotemporal Saliency Optimization

In the previous step, we have detected the salient regions and obtained the satisfying results by considering the local spatiotemporal consistency for each frame. Therefore, we further propose a spatiotemporal saliency energy function to keep the whole video saliency temporally consistent.

Let Q be a set of all the superpixels of a video ($Q = \mathbf{R}_1 \cup \mathbf{R}_2 \cup \cdots$). For convenience, we take $Q = \{1, 2, \cdots, N\}$, $|Q| = N$. For superpixel $q \in Q$, we define

$$q = (\mathbf{x}_q, k_q), \quad 1 \leq x_q \leq n, 1 \leq y_q \leq m$$

where $\mathbf{x}_q = (x_q, y_q)$ indicates the coordinate of the center point of superpixel q and k_q is the index of the frame superpixel q that belongs to.

k

k

The saliency of superpixel q therefore is $S^{C}_{k_q}(\mathbf{x}_q)$ computed by the last step. We propose an energy function to encourage the spatiotemporal consistency of the whole video saliency map. The final saliency of each superpixel is defined as s_q, which is further optimized through the proposed spatiotemporal saliency energy function as follows:

$$\begin{aligned} F &= F_{unary} + F_{smooth} \\ &= \alpha \quad (s_q - S^{C}_{k_q}(\mathbf{x}_q))^2 + \qquad w_{q,q'}(s_q - s_{q'})^2. \qquad (18) \\ &\qquad\quad q \qquad\qquad\qquad\qquad q,q'\in\aleph \end{aligned}$$

where the set $\aleph$ contains all the spatially adjacent superpixels within one frame and the temporally adjacent superpixels in a neighborhood: if $* \mathbf{x}_q - \mathbf{x}_{q'} * \leq 800$ and $| k_q - k_{q'} | = 1$, superpixels q and q' are temporally adjacent. The parameter α is the positive coefficient for balancing the relative influence between F_{unary} and F_{smooth}.

The first term F_{unary} defines an unary constraint that each superpixel tends to have the initial estimation for its saliency $S^{C}_{k_q}(\mathbf{x}_q)$. The smooth term F_{smooth} gives spatiotemporal consistency constraint that all the spatiotemporally adjacent superpixels of the whole video sequence should have the same saliency when they are similar. $w_{q,q'} = \exp(-*c_q - c_{q'}*)$ is a weighting function that gives a similarity measure for spatiotemporally adjacent superpixels q and q', and c_q indicates the mean CIELab color value of superpixel q.

Based on $s = [s_q]_{N\times1}$ and $s^* = [S^{C}_{k_q}(\mathbf{x}_q)]_{N\times1}$, the quadratic energy function F can be formulated as the following matrix forms:

$$\mathbf{F} = \alpha(s - s^*)^T(s - s^*) + s^T(\mathbf{B} - \mathbf{W})s, \qquad (19)$$

where $\mathbf{W} = [w_{q,q'}]_{N\times N}$ and $\mathbf{B} = diag([b_1, \cdots, b_N])$.

The diagonal elements of the metric $\mathbf{B}$ are the degree values of the weight matrix $\mathbf{W}$: $b_q = \ ^{N}_{q'=1} w_{q,q'}$. (19) can be solved by convex optimization and we finally obtain the saliency of superpixels s as follows:

$$s = \alpha(\mathbf{B} - \mathbf{W} + \alpha\mathbf{E})^{-1} s^*, \qquad (20)$$

where $\mathbf{E}$ is an identity matrix and we set parameter $\alpha = 0.5$ for all the test videos in our experiments.

IV. Experimental Results

Our approach automatically detects salient regions in video clips using spatial gradients and temporal motion features between frames. In this section, we provide the experimental comparison results to demonstrate the benefits of our approach. We first evaluate our approach on the well-known SegTrack dataset [21] and Freiburg-Berkeley Motion Segmentation Dataset (FBMS) [43], [44]. To deeper explore the issue of video saliency detection and establish a benchmark for future work, we further introduce a video saliency dataset, called ViSal,[2] which is collected from existing video databases and YouTube.

For all these databases: SegTrack, FBMS and ViSal, three measures are employed for the quantitative evaluation. We first use precision versus recall curves (PR curves) for performance evaluation. Given a saliency map with saliency values in the range of [0, 255], we vary a threshold from 0 to 255 to obtain different binary masks from the saliency map. Then we compute the precision and recall at each value of the threshold for comparing the quality of different saliency maps. The precision value corresponds to the ratio of salient pixels correctly assigned to all the pixels of extracted regions, while the recall corresponds to the fraction of detected salient pixels in relation to the ground truth of salient pixels. The curves are then averaged on each video sequence. We also estimate F-Measure [22] for considering both precision and recall:

$$\text{F-score} = \frac{(1 + \gamma^2) \cdot \text{precision} \cdot \text{recall}}{\gamma^2 \cdot \text{precision} + \text{recall}}. \qquad (21)$$

Thresholding is applied and γ^2 is set to 0.3 as suggested in [22]. We further introduce the mean absolute error (MAE) into the evaluation. The MAE estimates the approximation degree between the saliency map and the ground truth, which is normalized to [0, 1]. MAE provides a new means of evaluation, which directly measures how close a saliency map is to the ground truth. We measure the performance of the proposed algorithm, and compare with competitive image and video saliency methods, such as frequency-tuned saliency [22] (IG), saliency filter [45] (SF), sliding window based saliency [20] (SS), cluster-based co-saliency [49] (CS), self-resemblance based saliency detection [37] (SD), Quaternion Fourier Transform based saliency [16] (QS) and space-time saliency [33] (ST). Finally, we report the run time of our method and the state-of-the-art video saliency methods.

A. Comparisons on SegTrack and FBMS Datasets

The SegTrack database [21] was originally introduced to

And the corresponding qualitative and quantitative experimental results are also reported.

[2] http://github.com/shenjianbing/videosal/ViSal.zip

evaluate tracking algorithms and then widely used for video segmentation, and it is also suitable for evaluating video saliency detection. There are six videos (*birdfall, cheetah, girl, monkeydog, parachute, and penguin*) that range in length from 21 to 70 frames in this dataset. The FBMS dataset [43], [44] contains 59 video sequences. For both datasets, a pixel-level segmentation ground-truth for each video is available. The videos in these two datasets present various challenges such as large foreground and background appearance variation, significant shape deformation, and large camera motion.

As mentioned before, we first give the qualitative comparisons with two image saliency methods: IG [22] and SF [45], and five video saliency works: SS [20], CS [49], SD [37], QS [16] and ST [33]. We then provide quantitative performance comparisons with these methods, which demonstrate that our method has the ability to generate more accurate saliency results. Fig. 6 shows a visual comparison between our method and [16], [20], [22], [33], [37], [45], [49] for selected frames of different test sequences. IG [22] is a frequency- tuned approach that computes saliency in images using low level features of color and luminance. SF [45] proposes a contrast-based saliency estimation by computing two measures

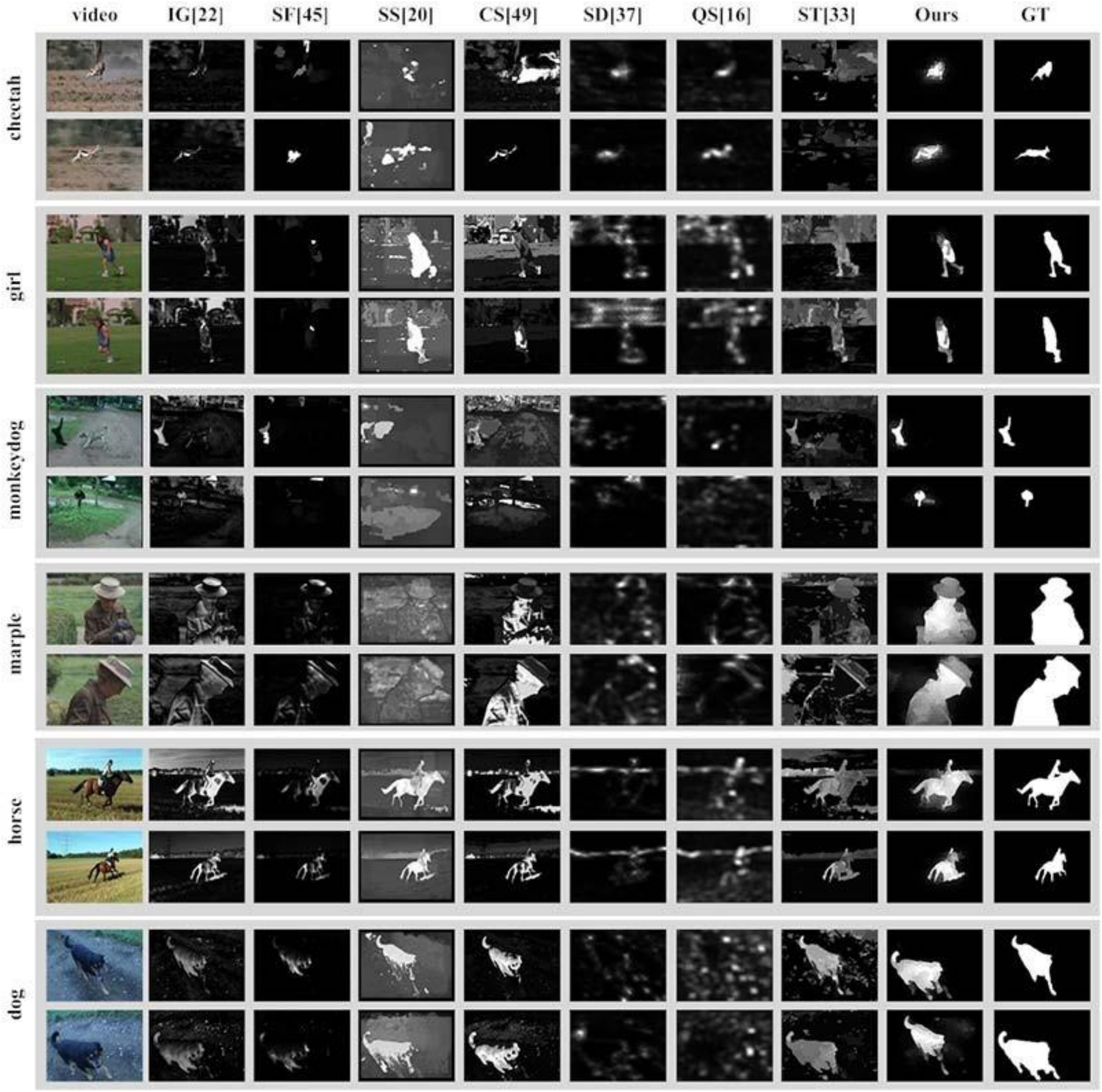


Fig. 6. Video saliency results on the SegTrack dataset (top three videos) and the FBMS dataset (bottom three videos). From left to right: indicative frames of input videos, IG [22], SF [45], SS [20], CS [49], SD [37], QS [16], ST [33], our method and ground truth (GT).

of contrast that rate the uniqueness and the spatial distribution of superpixels. However, both IG and SF do not perform well, as they both lack inter-frame information. For example, the backgrounds of the *cheetah* video is pretty complex, which places large difficulties for these methods to correctly detect the motion objects. Our algorithm integrates intra-frame information and motion features between frames into our gradient flow field, which makes our method discriminative enough in these scenes.

SS [20] locates the salient objects more precisely than IG [22] and SF [45], because it utilizes different intra-frame information and motion cues. However, it tends to highlight background pixels due to the use of sliding windows. Furthermore, as a video saliency detection method, it does not consider temporal consistency for video saliency. As a result, the saliency value of the same region may be dramatically varied in different frames (e.g., on *monkeydog*). Similar conclusions are also observed on results of other video saliency methods: CS [49], SD [37], and QS [16]. Based on the proposed spatiotemporal saliency energy function, our method fully explores the temporal consistency property of video saliency in our saliency optimization step.

The performance of CS [49] is also not satisfactory, even though it considers the correspondence of objects across

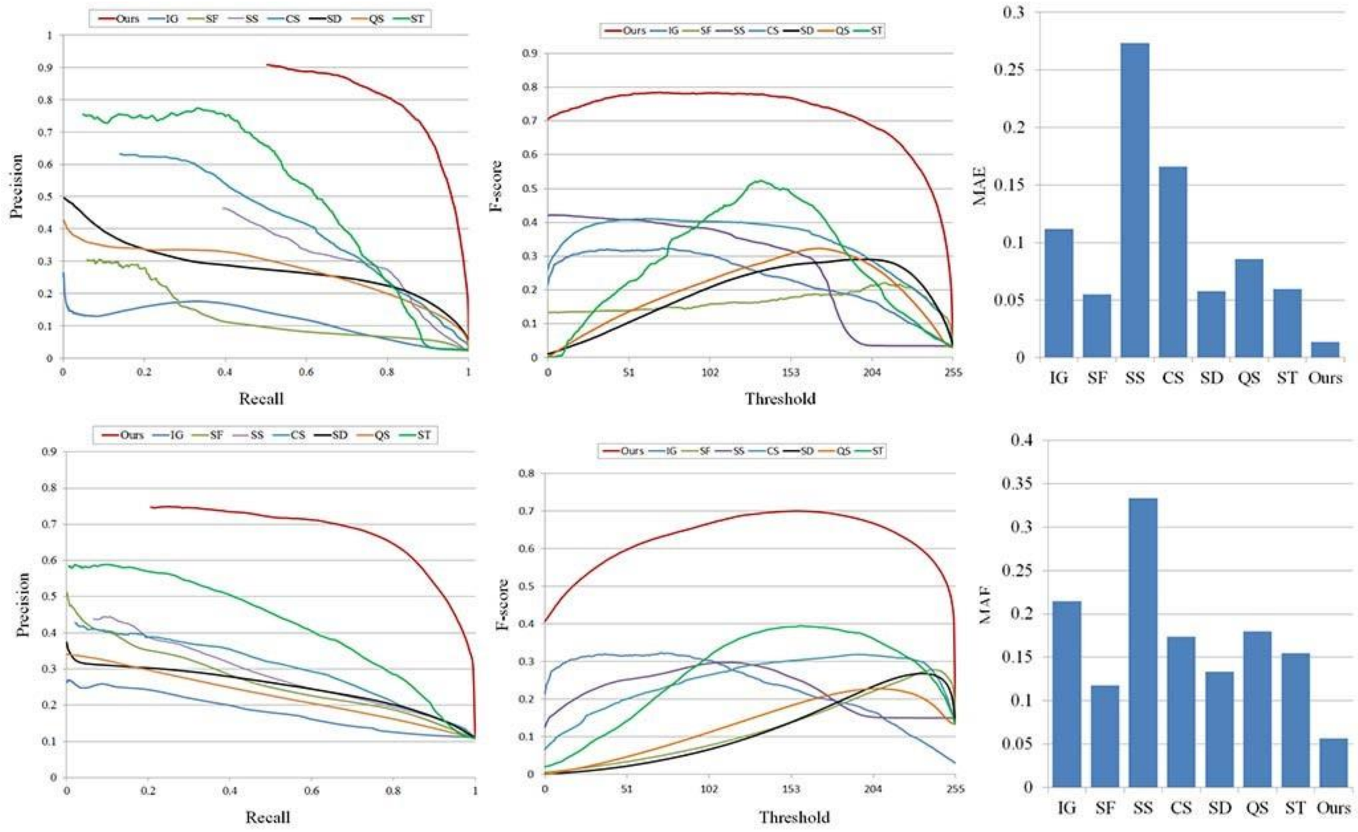


Fig. 7. Comparison of PR curves (left), F-measure (middle) and MAE (right) on the SegTrack dataset (top row) and the FBMS dataset (bottom row).

the video sequence. The global correspondence between the multiple frames is learned during a clustering process in CS. However, this clustering process can become less effective in complex videos, especially when the non-salient background involves the similar appearance (e.g., color) as the salient areas, as also mentioned in [49]. Our gradient flow field efficiently utilizes the intra-frame and inter-frame information, which provides powerful constraints that help to avoid this issue and leads to significant improvements over other methods.

SD [37] proposes a bottom-up saliency detection algorithm by employing local steering kernels and using a nonparametric kernel density estimation based on the matrix cosine similarity. In most cases, SD is able to locate the salient objects with the consideration of temporal information, but the saliency maps are generated in low resolution and some moving objects are assigned low saliency values (e.g., on *monkeydog* and *horse*). Therefore, deeper exploration of motion features will be needed to improve the performance. Similarly, in QS [16], motion features are also considered into their Quaternion Fourier Transform process for obtaining spatiotemporal saliency maps. Some results of QS are impressive (e.g., on *cheetah*), while this method overemphasizes small and local features rather than highlighting the whole object (e.g., on *monkeydog*). Additionally, this method may fail when the motion information is not correct (e.g., on *girl*).

As a video saliency method, ST [33] combines various low-level features including motion features for predicting the spatiotemporally salient object. However, this method still does not consider enforcing temporal coherence of the saliency map across the video (e.g., on *marple*). Another limitation is that the simple mechanism for utilizing motion features can not correctly locate salient objects with complex motion patterns (e.g., on *girl*). The difference between our method and others is significant. Our method exhibits substantial robustness and produces correct saliency maps, even for complex scenes.

We present quantitative comparisons with previous well-known methods: IG [22], SF [45], SS [20], CS [49], SD [37], QS [16] and ST [33] on the SegTrack and FBMS datasets. Precision versus recall curves (PR curves), F-score and the mean absolute error (MAE), are employed for the evaluation. The precision-recall and F-measure curves on these two databases are plotted in Fig. 7 (left) and Fig. 7 (middle), respectively. These curves have demonstrated that our method significantly outperforms the other seven methods. The SegTrack and FBMS video databases present large difficulties for previous saliency methods, which reflects the importance of utilizing motion information for video saliency detection. The comparison results in Fig. 7 (right) show that our method achieves the lowest MAE values, which reflects that our algorithm produces results closer to ground truth.

B. Comparisons on ViSal Database

Although videos from SegTrack [21] and FBMS [43], [44] databases span a large range of difficulties, the amount

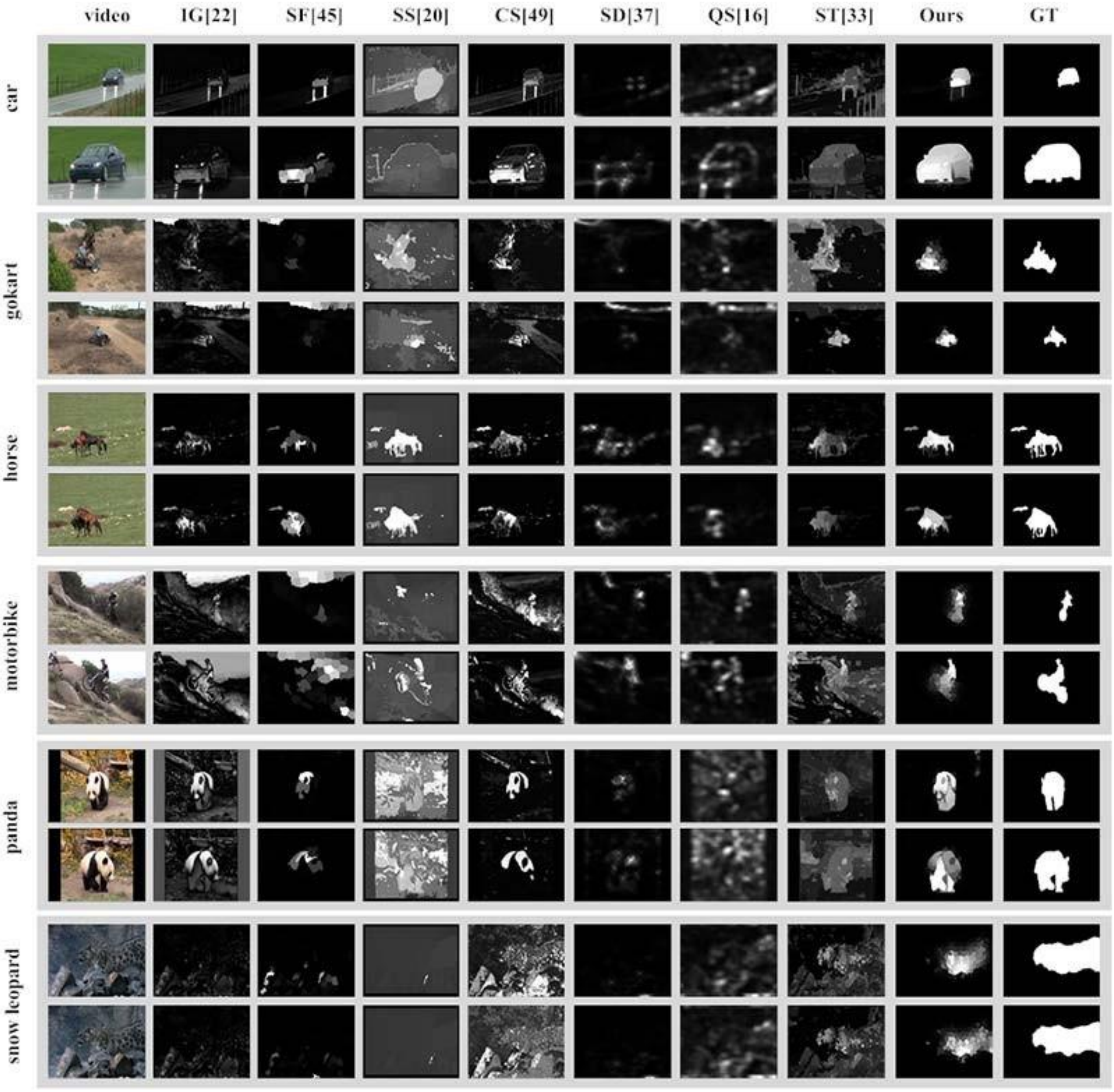


Fig. 8. Visual comparison of previous approaches to our method and ground truth (GT) on the ViSal dataset. From left to right: indicative frames of input videos, IG [22], SF [45], SS [20], CS [49], SD [37], QS [16], ST [33], our method and ground truth (GT). As can be seen, our method produces saliency maps closest to ground truth.

of video clips in SegTrack is small and motion patterns of objects in FBMS are simple. To deeper explore general situations with various foreground/background motion patterns and complex color distributions and to establish a benchmark for future work, we construct a new dataset (ViSal) that is specially designed for video saliency detection. This dataset consists of 17 challenging video sequences containing complex color distributions (*motorbike*, *cow*, etc.), highly cluttered background (*man*, *panda*, etc.), various object motion patterns (static: *boat*, fast: *car*), rapid topology changes (*cat*, *motorbike*, etc.) and camera motion (*gokart*, *motorbike*, etc.). The length of these videos ranges from 30 to 100 frames and all clips are manually annotated as the given classes.

In order to demonstrate the effectiveness of our method, we test the proposed method on our ViSal dataset. A visual comparison of different video saliency methods with six typical videos is shown in Fig. 8. 1) *car*. The foreground is moving fast and variation in foreground scale is substantial.
2) *gocart*. The foreground moves with dynamic (nonuniform) motion and this example exhibits camera motion. 3) *horse*. The foreground is more stable and the scene is nearly static.

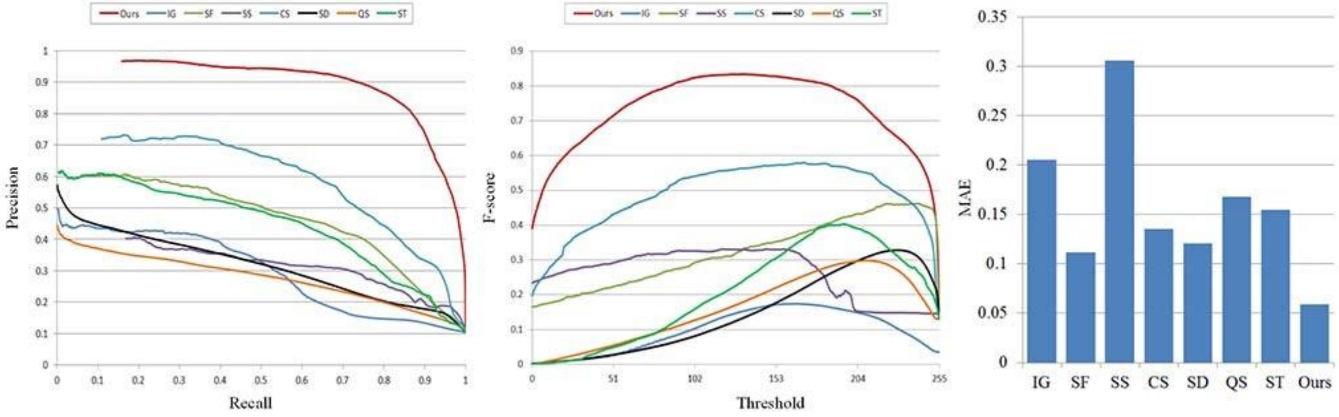


Fig. 9. Comparison of PR curves (left), F-measure (middle) and MAE (right) on our ViSal dataset.

TABLE I
COMPARISON OF AVERAGE RUN TIME (SECONDS PER FRAME) ON THE SEGTRACK DATASET

Method	Ours	SS[20]	CS [49]	SD [37]	QS [16]	ST [33]
Time(s)	0.891	6.186	3.695	5.376	0.064	25.292

Our method automatically adapts to these three examples with very different motion patterns and produces reliable saliency results. Additionally, our method performs well in highlighting large salient objects, such as on *car*. That is because our method uses surrounding background contrast instead of traditional contrast prior. For this scene, SD [37] and QS [16] tend to highlight the boundaries and assign relatively low probabilities to pixels inside the objects. 4) *motorbike*. The fourth example shows rapid topology changes and erratic motion. 5) *panda*. Highly cluttered background presents difficulties for foreground detection. 6) *snow leopard*. Complex color distributions and high similarity between foreground and background make it hard to locate the salient object. On these difficult examples, the saliency maps calculated by the proposed method are more visually consistent with the shape and location of the ground truth than the saliency maps generated by other methods. To compare the aforementioned methods quantitatively, PR curves, F-measure and the MAE are used again for the evaluation. As shown in Fig. 9, the performance of our method is superior to those of previous well-known methods.

C. Run Time Statistics

The average run time of currently top-performing video saliency methods: SS [20], CS [49], SD [37], QS [16] and ST [33] on the Segtrack database are presented in Table 1. All the saliency maps are produced by directly running their implementation codes by the authors. All the tests were performed on a Windows platform and under the same computer configuration Intel Xeon E5-2609 @2.40 GHz with 32.0 GB RAM. The run time excludes optical flow computation, which all methods require as input. As shown in Table 1, our method is much faster than the others, but is only slower than frequency domain based QS [16].

V. DISCUSSIONS AND CONCLUSION

In this paper, we proposed a novel video saliency detection method to produce high-quality and spatiotemporally consistent saliency maps. A novel gradient flow field method was introduced into our framework, which fully incorporates inter-frame and intra-frame information such as edges and motion features between neighboring frames for detecting the locations of visual foregrounds. Additionally, two enhanced contrast saliency cues: local and global contrast, were introduced to bias foreground objects with higher saliency, which is built upon the visual importance detection results from the gradient flow field. These two discriminative saliency cues overcome the shortage of traditional contrast prior based saliency methods and uniformly highlight the entire object. Furthermore, a spatiotemporal saliency energy function was proposed to refine the spatiotemporal consistency of the output salience maps, which can be efficiently solved by convex optimization. Based on these efforts, our algorithm is applicable for complex scenes even with dramatic foreground and background appearances or motion pattern variations. Experimental results show the superiority of the proposed video saliency approach to predict the salient objects over three different datasets with a large amount of data. This approach is applicable to many tasks when objects are processed sequentially in a spatiotemporal manner.

REFERENCES

[1] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[2] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Netw., Comput. Neural Syst.*, vol. 10, no. 4, pp. 341–350, 1999.

[3] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proc. IEEE*, vol. 101, no. 9, pp. 2058–2067, Sep. 2013.

[4] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vis. Res.*, vol. 42, no. 1, pp. 107–123, 2002.

[5] W. Einhäuser and P. König, "Does luminance-contrast contribute to a saliency map for overt visual attention?" *Eur. J. Neurosci.*, vol. 17, no. 5, pp. 1089–1097, 2003.

[6] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–6.

[7] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Proc. Adv. NIPS*, 2007, pp. 497–504.

[8] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[9] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[10] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

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

[12] J. Li, Y. Tian, T. Huang, and W. Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 623–636, May 2011.

[13] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

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

[15] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3137–3148, Oct. 2015.

[16] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[17] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.

[18] D. B. Johnson, "Efficient algorithms for shortest paths in sparse networks," *J. ACM*, vol. 24, no. 1, pp. 1–13, 1977.

[19] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[20] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. 11th ECCV*, 2010, pp. 366–379.

[21] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. BMVC*, 2010, pp. 1–11.

[22] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1597–1604.

[23] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. ID 16.

[24] J.-S. Kim, J.-H. Kim, and C.-S. Kim, "Adaptive image and video retargeting technique based on Fourier analysis," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1730–1737.

[25] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels," EPFL, Lausanne, Switzerland, Tech. Rep. EPFL-REPORT-149300, 2010.

[26] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. IEEE ICIP*, Sep. 2003, pp. I-253–I-256.

[27] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. NIPS*, 2006, pp. 545–552.

[28] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[29] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[30] B. Mertsching, M. Bollmann, R. Hoischen, and S. Schmalz, "The neural active vision system NAVIS," in *Handbook of Computer Vision and Applications*. New York, NY, USA: Academic, 1999, pp. 543–568.

[31] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.

[32] W.-H. Cheng, C.-W. Wang, and J.-L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.

[33] F. Zhou, S. B. Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 3358–3365.

[34] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. Berlin, Germany: Springer-Verlag, 2006.

[35] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[36] B. C. Ko and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *J. Opt. Soc. Amer. A*, vol. 23, no. 10, pp. 2462–2470, 2006.

[37] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 15.1–15.27, 2009.

[38] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2376–2383.

[39] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2010.

[40] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. IEEE ICCV*, Nov. 2011, pp. 2214–2219.

[41] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[42] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vis. Res.*, vol. 50, no. 22, pp. 2233–2247, 2010.

[43] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. IEEE 11th ECCV*, Sep. 2010, pp. 282–295.

[44] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.

[45] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 733–740.

[46] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. IEEE 12th ECCV*, Oct. 2012, pp. 29–42.

[47] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 470–477.

[48] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1529–1536.

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

[50] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE. CVPR*, Jun. 2015, pp. 3395–3402.

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

Wenguan Wang is currently pursuing the Ph.D. degree with the School of Computer Science, Beijing Institute of Technology, Beijing, China. His current research interests include video saliency and segmentation.

Jianbing Shen (M'11–SM'12) is currently a Professor with the School of Computer Science, Beijing Institute of Technology, Beijing, China. His research interests include computer vision and multimedia processing. He has authored about 50 journal and conference papers, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE Conference on Computer Vision and Pattern Recognition, and the IEEE International Conference on Computer Vision. He has also obtained many flagship honors, including the Fok Ying Tung Education Foundation from the Ministry of Education, the Program for Beijing Excellent Youth Talents from the Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from the Ministry of Education. His research interests include computer vision and multimedia processing. He is on the Editorial Board of *Neurocomputing*.

Ling Shao (M'09–SM'10) is currently a Full Professor and the Head of the Computer Vision and Artificial Intelligence Group with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, and an Advanced Visiting Fellow with the Department of Electronic and Electrical Engineering, The University of Sheffield. His research interests include computer vision, image processing, pattern recognition, and machine learning. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, and other journals. He is also a fellow of the British Computer Society, and the Institution of Engineering and Technology, and a Life Member of the Association for Computing Machinery.