Signal Processing: *Image Communication* 38 (2015) 45–56

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

A spatiotemporal weighted dissimilarity-based method for video saliency detection

Lijuan Duan [a,*], Tao Xi [a,b], Song Cui [a], Honggang Qi [c], Alan C. Bovik [d]

[a] *College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China*
[b] *Center for Multimdia & Network Technology, School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore*
[c] *University of Chinese Academy of Sciences, Beijing 100190, China*
[d] *Laboratory of Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712-0240, USA*

ARTICLE INFO

Available online 19 August 2015

Keywords:
Saliency detection
Video
Visual attention

ABSTRACT

Accurately modeling and predicting the visual attention behavior of human viewers can help a video analysis algorithm find interesting regions by reducing the search effort of tasks, such as object detection and recognition. In recent years, a great number and variety of visual attention models for predicting the direction of gaze on images and videos have been proposed. When a human views video, the motions of both objects in the video and of the camera greatly affect the distribution of visual fixations. Here we develop models that lead to motion features that are extracted from videos and used in a new video saliency detection method called spatial–temporal weighted dissimilarity (STWD). To achieve efficiency, frames are partitioned into blocks on which saliency calculations are made. Two spatial features are defined on each block, termed spatial dissimilarity and preference difference, which are used to characterize the spatial conspicuity of each block. The motion features extracted from each block are simple differences of motion vectors between adjacent frames. Finally, the spatial and motion features are used to generate a saliency map on each frame. Experiments on three public video datasets containing 185 video clips and corresponding eye traces revealed that the proposed saliency detection method is highly competitive with, and delivers better performance than state-of-the-art methods.

1. Introduction

Digital videos have become an increasingly important part of daily life owing to the rapid proliferation of networked video applications such as video on demand, digital television, video chatting, streaming video over the Internet, and consumer video appliances. A great deal of processing is required to support the delivery and display of this increasingly massive amount of video content. Finding ways to process these videos efficiently is crucial towards delivering them to the consumer in real time.

An important and underutilized goal of video processing research are automated anticipation of a user's gaze direction. Knowledge of the likely times and locations of visual fixations can allow the definition of a set of strategies to reduce the computational cost of search processes. Recognizing this, numerous researchers have developed saliency detection techniques for applications in multimedia processing, such as image/video compression, image segmentation, image retargeting, and advertising design.

* Corresponding author.
E-mail address: ljduan@bjut.edu.cn (L. Duan).

http://dx.doi.org/10.1016/j.image.2015.08.005
0923-5965/

To effectively locate the most visually "attractive" or "interesting" content in multimedia data, researchers have built on groundbreaking work by Treisman and Gelade [1], Koch and Ullman [2], and subsequent attention theories proposed by Itti [3] and Wolfe et al. [4] and others. Two broad classes of visual attention mechanisms predominate: top-down approaches and bottom-up approaches. The top-down approach is task-driven. This approach needs prior knowledge of the target before the detection process. It is modeled as a spontaneous and voluntary process. Traditional rule-based or training-based saliency prediction methods belong to the top-down type. Bottom-up approaches are largely driven by low-level stimuli, where prediction is based on modeling human reactions to external stimuli, such as color, shape or motion, and is a largely automatic process. Over the last decade, many bottom-up and top-down methods of approaching this problem have been proposed, such as [5–11]. Most of these methods have been developed to detect saliency on pictures rather than on video. Here, we focus on building a bottom-up saliency detection model for video. Motion is probably the key attribute for accomplishing saliency detection on video. Considerable attentional resources in the human visual system are driven by motion. As such, the success of visual attention models significantly depends on their ability to model and account for motion perception [12]. However, there are myriads of moving content and structure found in natural videos, and highly diverse shooting and editing styles. Thus, it is imperative that a success saliency model be able to adapt to highly diverse kinds of videos. In [13], we introduced a still picture saliency detection model called SWD, which computes saliency maps based on image patch differences weighted by spatial distance and center bias. Here we extend our prior model by introducing a spatio-temporal attention model for predicting video saliency. The new model deploys a simple and effective way of incorporating motion information into the saliency map. As we show in Section 4.3, this mechanism can effectively handle scene changes. When a person watches a video, the direction of gaze often tends towards regions that differ from their surrounding and hence are conspicuous. This pertains to both static (spatial) as well as temporal characteristics introduced by motion. In our proposed saliency prediction model, likely attractors of gaze direction are found by integrating three elements as follows: spatial dissimilarity of each image block which is evaluated in a reduced dimensional space and weighted by spatial relationships between image blocks, a center bias feature, and block motion features. We then discuss ways to integrate these three saliency factors and suggest strategies to adapt them to different types of video. We also carried out experiments on three saliency video datasets and compared the saliency maps generated by several state-of-the-art saliency detection approaches and the proposed method with recorded eye tracking data. We show that our method, despite its simplicity, predicts human fixations as accurately (or better) than leading methods. The remainder of this paper is organized as follows: static and dynamic attention models are presented in Section 2 and Section 3, respectively. Section 3.5 describes a fusion model to combine the two (space and time) models. Section 4 presents experimental results and a performance evaluation. We conclude in Section 5.

2. Related work

Research on visual attention has been widely conducted by experts in biological vision, perception, cognitive science and computer vision. The results of visual attention has been applied to a wide range of problems, such as tele-remote robot navigation [14], image retargeting [15,16], image compression [17,18] and video quality assessment (VQA) [19,20]. Early on, Treisman and Gelade [1] suggested that humans perceive external features, such as colors, brightness, texture, and motion in a distinct manner. Visual attention seems to be drawn towards visually different, conspicuous regions, and much of the work in bottom-up saliency analysis has focused on the detection of feature contrasts, as guided by this concept. In recent decades, a large number of visual attention models have been proposed. A classical saliency detection model was proposed by Itti et al. [21], where a feature map is calculated using the multi-scale center-surround differences of each of three image features: intensity, color, and orientation, then a linear combination of the three feature maps is utilized to obtain a final saliency map. Subsequently, Goferman et al. [7] presented a content-aware saliency detection model that considered local low-level clues, global features, visual organization rules, and high-level features. In [22] Hou and Zhang analyzed the log amplitude spectrum of an image and obtained a spectral residual which they related to saliency. Cheng et al. [23] developed a regional contrast based saliency extraction model which can be used to create high quality segmentation masks.

Visual gaze prediction on still images has been long studied, but less effort has been applied to the problem of spatiotemporal visual attention analysis. A naive way to generate the saliency map in videos is to utilize the image-based saliency model frame-by-frame. But, this neglects the influence of motions which is crucial in gaze prediction on video sequences. Based on this idea, many researchers proposed various spatial and temporal models to estimate video saliency map. Zhai and Shah [24] proposed a method of combining spatial and temporal attention models. In the temporal attention model, motion contrast was computed based on analysis of planar motions, which was estimated by applying RANSAC to establish between-frame point correspondences. A dynamic fusion technique was then used to combine the temporal and spatial saliency maps, where temporal gaze attractive was assumed to dominate spatial factors when large motion contrast exists, and vice versa. Another novel method presented by Itti and Baldi [25] elaborated the concept of "visual surprise" when there is little motion contrast, using the Kullback–Leibler divergence between the prior and posterior distributions of a feature map. Cheng et al. [26] incorporated motion information into an attention model by analyzing the magnitudes of pixel motions. Boiman and Irani [27] proposed a spatiotemporal "irregularity" detection method for video. They reinterpreted the ideas of "saliency" and "visual attention" on videos in this regard. Guo and Zhang [28] suggested that the phase spectrum of an image's Fourier transform can be effectively used to calculate the locations of salient areas, and they subsequently devised a phase-based saliency detection algorithm. Rudoy et al. [29] received a competitive result by applying the center-surround in candidate locations rather than every pixel. Rapantzikos et al. [30] treat video sequences

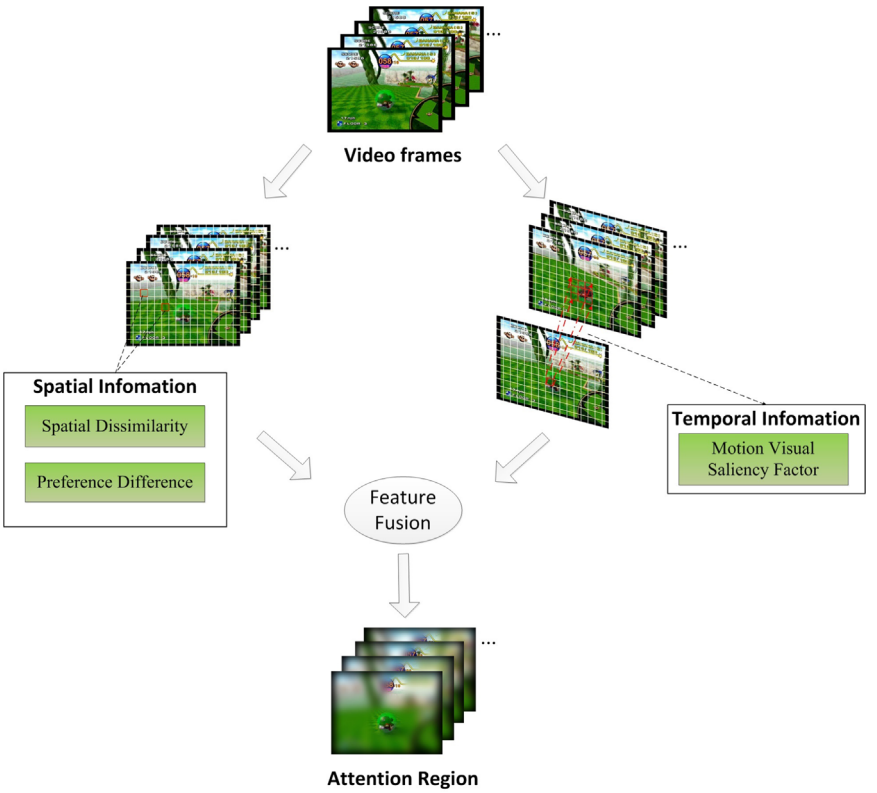

Fig. 1. Framework of proposed video saliency model.

as a volume which maintains spatial and temporal information and produce saliency measurements in different voxel levels. In [31] Kim et al. adopted the center-surround framework to obtain the spatial and temporal saliency. The spatial saliency is acquired by the association of edge and color conspicuous in local regions and temporal saliency between temporal gradients of the adjacent regions. Fang et al. [32] considered a center-surround saliency model with coded spatio and temporal features from uncompressed video. Mahapatra et al. [33] indranght the coherency information in both spatio and temporal saliency. Liu et al. [34] proposed saliency detection based on motion and color histograms at the super-pixel level. Meanwhile, several researches are inspired by biological mechanisms and under the common view of center-surround framework. Mahadevan and Vasconcelos [35] extend a discriminant formulation of center-surround saliency to obtained temporal components. In [36] Tavakoli et al. exploited a spherical representation to implement center-surround model. Zaharescu and Wildes [37] also proposed a effective method by measurements of visual spacetime orientation. Marat et al. [38] unified spatial and temporal saliency by merging both static and motion saliency map generated by cortial-like filters. Besides biological inspirations, a novel saliency detection for video and image is proposed by Mauthner et al. [39], they suggest that salient regions of videos can be modeled as a fully unsupervised encoding problem. There are also some machine learning-based methods for video saliency prediction. Liu et al. [40] utilized a machine learning method to obtain a saliency map on either a single image or on videos. They defined a new set of features and integrate them using a conditional random field (CRF) model to predict salient objects. Deploying stimulus-driven (bottom-up) and task-related (top-down) factors simultaneously, Li et al. [41] proposed a video saliency prediction model based on multi-task learning.

All of these models use temporal change or motion features such as frame differences and optical flow. However a common assumption that is made in most of these models is that the video camera is fixed; with motion arising only from moving objects. However, many real videos are captured during camera motions, such as translation, rotation, and scaling. It is important to also account for camera ego-motion in the development of video saliency prediction models. Further, real-world videos viewed by humans often contain abrupt scene changes, typically not accounted for by saliency models even though scene changes affect visual attention. The space time saliency model that we present here accounts for both of these important practical factors.

3. Video saliency detection

The saliency of an item in a video, be it an object, a person, a pixel, etc., may be viewed as a state or quality by which it stands out relative to its neighbors. Here, we propose a saliency detection model for videos, which follows this precept using both spatial and temporal information. A broad overview of our model follows. As shown in Fig. 1, we first partition a current frame into blocks of equal dimensions. Features expressing both spatial and temporal information relevant to each block's level of possible saliency are extracted. The spatial information includes a measure of spatial dissimilarity relative

to the surrounding blocks and a factor accounting for the well-known center bias. Differences between each block's motion vector relative to the motion vectors of the spatially co-located blocks in current and previous frames are used as salient temporal information. These three factors are fused to compute a saliency value for each block. Finally, all of the block saliency values are combined to define the overall saliency map. Each processing stage shown in Fig. 1 is detailed in the following.

3.1. Pre-processing

All calculations are carried out at the block level to capture local image properties and to promote computational simplicity. For example, standard block matching methods, as used for motion estimation in video compression, can be used to obtain the necessary motion information. Denote an $H \times W$ video frame as I, represented in the efficient YUV color space [42]. The frame I is partitioned into non-overlapping blocks of size of $k \times k$, hence, the number of blocks in I is approximately $L = \lfloor H/k \rfloor \cdot \lfloor W/k \rfloor$. Denote the blocks as b_i where $i = 1, 2, \ldots, L$. Each block is represented by a length $3k^2$ column vector f_i, containing the YUV block color values. Then I can be represented by the matrix $A = [f_1, f_2, \ldots, f_i, \ldots, f_L]$.

3.2. Spatial dissimilarity

For each block b_i, two aspects of spatial dissimilarity are computed: the appearance difference and the spatial location relationship between b_i and all the other blocks within the current frame. The spatial location relationship is used to weight the local dissimilarity of each block, which is motivated by the fact that the degree of saliency of a visually fixated area is related to the difference in appearance between the fixated area and also the distances to the compared (similar of dissimilar) areas [43]. The concept is related to, but used differently than, the method of selecting image patches in nonlocal mean image denoising models [44]. Thus spatial dissimilarity is measured by

$$SD(b_i) = \sum_{i \neq j} Loc(b_i, b_j) D(b_i, b_j) \tag{1}$$

where $D(b_i, b_j)$ denotes the appearance difference between b_i and b_j, and $Loc(b_i, b_j)$ denotes the distance between b_i and b_j, as defined below.

Realizing that a sufficient difference in appearance as it relates to saliency between two blocks may be more abstract and be based on just a few structural elements, the appearance difference metric $D(b_i, b_j)$ is defined by first applying a principal component (PC) decomposition to each block as a dimension reducing method. This has the virtue of eliminating factors such as unnecessary detail or noise from the saliency calculation, unlike pointwise dissimilarity metrics such as the mean-squared error (MSE) or even the perceptually relevant SSIM index [45]. The block principal components are extracted by regarding each column in the matrix $A = [f_1, f_2, \ldots, f_i, \ldots, f_L]$ as a sample. Applying PCA to matrix A yields the new $d \times L$ matrix $A^* = [f_1^*, f_2^*, \ldots, f_i^*, \ldots, f_L^*]$. Then each block b_i is represented by column f_i^* in a reduced dimensional space. Given two arbitrary blocks b_i and b_j, the appearance difference $D(b_i, b_j)$ between them is defined as

$$D(b_i, b_j) = \sum_{s=1}^{d} |f_{si}^* - f_{sj}^*|. \tag{2}$$

The spatial location relationship $Loc(b_i, b_j)$ is defined as a biased reciprocal of the Euclidean distance $D(b_i, b_j)$ between the centers of blocks b_i and b_j, which favors nearer blocks:

$$Loc(b_i, b_j) = \frac{1}{1 + Dist(b_i, b_j)}. \tag{3}$$

3.3. Preference difference

By analyzing the distribution of human's fixations on a large number of images, it has been be found that the human gaze direction when viewing displays tends forwards the center of the display, even if the central region has no particularly attractive feature [46]. In an interesting study, Judd et al. [47] analyzed fixation data on a public image library and found that 40% of the fixations fell into 11% of the image area near the image center, and that 70% of the fixations fell into the region near the image center comprising 25% of the image area. Therefore, we employ a “Center Bias” mechanism as a significant factor in the video saliency index. It increases the saliency value of patches close to the center of the image, and vice versa. In the following, $C(b_i)$ is the “center bias” weight on block b_i, and is defined:

$$C(b_i) = 1 - Dist(b_i, b_{center})/D_{max} \tag{4}$$

where b_{center} is the center block in the current frame, that is, $Dist(b_i, b_{center})$ is the spatial Euclidean distance between the center of block b_i and the center of the block that is located at the center of the current frame, and $D_{max} = \max\{Dist(b_i, b_{center})\}$ is a normalization factor such that $0 \leq C(b_i) \leq 1$. In Section 5.3, the impact of including the center bias factor will be discussed in detail.

3.4. Visual motion saliency factor

In the temporal part of our saliency model, we express the saliency degree in units of image blocks. A visual motion saliency factor (VMSF) is defined using motion vector fields computed from the video sequence. Our implementation uses the approach [48] that deploys four-step search to accomplish motion vector estimation. The VMSF is computed on the differences of motion vectors between adjacent images, estimated using Zhu et al.'s method [20]. Note that we apply VMSF from the second frame, due to there is no former frame for the first frame.

3.4.1. Motion vector computation

Motion vectors capture the displacements between patches in a current frame and corresponding best match patches from neighboring reference frames, and thus indicate the image motion of the current patches. Denote motion vectors as $(V_h(t), V_v(t))$, where $V_h(t)$ and $V_v(t)$ are horizontal and vertical displacements, respectively, of current blocks in the tth frame, relative to matched blocks in a previous frame.

We deploy a simple and fast block matching method to estimate the motion vectors. Specifically, we employ the four-step search approach [48] to obtain best matching blocks under the minimum Mean Square Error (MSE) criterion:

$$MSE(V_h(t), V_v(t)) = \sum |f_{t-1}(x+V_h(t), y+V_v(t)) - f_t(x,y)| \tag{5}$$

where f_{t-1} and f_t denote a previous frame and a current frame, and x and y indices the horizontal and vertical position of the current block, measured in block units. Here $f_t(x,y)$ represents the three RGB channels of the block at position (x, y). Supposing the position of the current block to be $(Pos_x(c), Pos_y(c))$ and the position of the best match block to be $(Pos_x(m), Pos_y(m))$, then the motion vector computed from the current block is

$$V_h(t) = Pos_x(c), Pos_x(m) V_v(t) = Pos_y(c), Pos_y(m). \tag{6}$$

Scene changes present a difficulty, since it is not possible or meaningful to compute motion features across scenes transitions. We therefore utilize a simple scene change detection mechanism to adaptively decouple the computation of motion vectors. First, we compute the magnitudes of previously obtained horizontal and vertical motion vectors:

$$V(t) = \sqrt{V_h(t)^2 + V_v(t)^2}. \tag{7}$$

Then, let

$$M = \sum_x \sum_y |V(t) - V(t-1)| \tag{8}$$

where $V(t)$ and $V(t-1)$ are computed from corresponding blocks in the current and previous frame, respectively.

Fig. 2 shows some successive frames from a video "mtvclip01" which includes two scene changes. At scene changes, M generally takes much larger values than elsewhere in a video. Thus, we use M to determine whether a scene occurs at the current frame. When the current frame is the first frame of a new scene, all motion vectors at that frame are set to 0, i.e. motion computation is decoupled. In Section 4.3, we demonstrate the effectiveness of this simple expedient for improving the prediction accuracy of the saliency model.

3.4.2. Computation of VMSF

In some temporal saliency models, a saliency map is constructed using the contrast of motion. However, we believe that moving objects should be assigned higher temporal saliency. In [20], a motion vector model was proposed to obtain frame-level saliency. Here, we modify and use this model to compute the value of VMSF for each block.

The VMSF of block b_i denoted by $SV_i(t)$ is computed as follows:

$$SV(b_i, t) = \begin{cases} V_i(t) - \left[\sum_{k=0}^{t-1} V_i(k)\right] \Big/ 3 & \text{if } 0 < t \le 3 \\ V_i(t) - \left[\sum_{k=t-3}^{t-1} V_i(k)\right] \Big/ 3 & \text{if } t > 3 \end{cases} \tag{9}$$

where $V_i(t)$ denotes $V(t)$ computed by (7) for block b_i.

Fig. 3 shows a video frame (left) and its computed temporal feature map (right). In the red rectangle in Fig. 3 (a), the hand holding the phone is a foreground moving target. In Fig. 3(b), the region corresponding to the red rectangle has a much higher average temporal saliency than the rest of the map.

3.5. Saliency fusion

Next we describe a method of integrating the various saliency factors described in the preceding to produce a final spatiotemporal saliency value for each block. Under the assumption that spatially dissimilar and moving blocks are more visually attractive, the fusion strategy used to create the overall spatiotemporal saliency value of block b_i at time t:

$$Saliency(b_i, t) = \mathbb{N}\left\{ C(b_i) \sum_{i \neq j} Loc(b_i, b_j) D(b_i, b_j) \right\} + \alpha \cdot \mathbb{N}\left\{ \left[C(b_i) \sum_{i \neq j} Loc(b_i, b_j) D(b_i, b_j) \right] \cdot SV(b_i, t)^{\beta} \right\}. \tag{10}$$

In formula (10), the first term supplies the spatial saliency value, the second term supplies the spatiotemporal saliency value arising from motion features. Also, α weights the relative contributions of the two saliency terms, β regulates the contribution of the temporal motion feature, and $\mathbb{N}(\cdot)$ is a normalizing operator which maps the value of its argument to the range [0, 1], it is defined as

$$\mathbb{N}(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}} \tag{11}$$

where $\mathbf{x}_{min}$ and $\mathbf{x}_{max}$ are the minimum and maximum values of x over all blocks in the tth frame.

In our implementation, we set $\alpha = 0.2$ and $\beta = 3$ following the discussion in Section 5.2. Finally, the saliency map is normalized using (11), resized to the scale of the original video, and smoothed with a unit-energy Gaussian function ($\sigma = 3$).

4. Experiments

We evaluated the performance of the described video saliency method on three public eye-movement datasets [9,49], ORIG-CRCNS, MTV and DIEM. The ORIG-CRCNS dataset is provided by iLab at the University of Southern California. It consists of 50 video clips which have over 46,000 video frames with a total display duration of 25 min. This dataset contains different scenes, and eight subjects viewed the video clips while being eye tracked, where at least four subjects viewed each clip. An ISCAN RK-464 eye-tracker was used to record the fixation traces. The MTV dataset was created by cutting the same set of video clips into 1–3s "clippets", randomly reassembling those clippets, and another eight subjects viewed this dataset. Other aspects of the second study were identical as the ORIG-CRCNS dataset. To analyse our method in detail, we divided the ORIG-CRCNS dataset into two subsets, ORIG-M and ORIG-N, based on whether a video clip contained obvious moving objects. All 30 video clips in ORIG-M contained obvious moving objects while all 20 video clips in ORIG-N contained none. The scenes in ORIG-M included outdoor day and night,

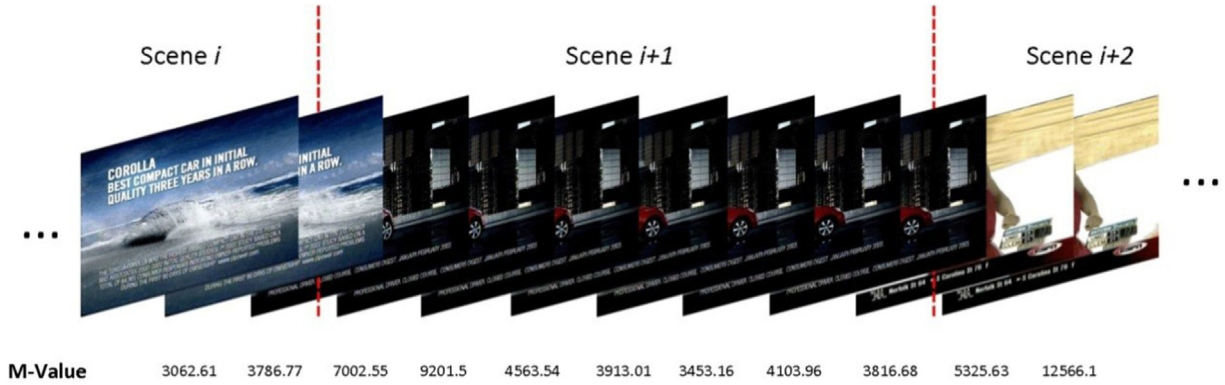

Fig. 2. Successive frames from video "mtvclip01", which includes two scene changes. At the top are the video frames, at the bottom are the values of M corresponding to each frame.

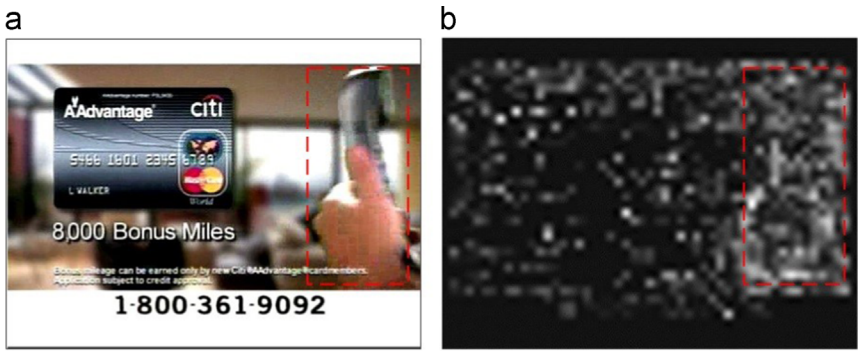

Fig. 3. (a) Original frame. (b) The temporal feature map computed from (a).

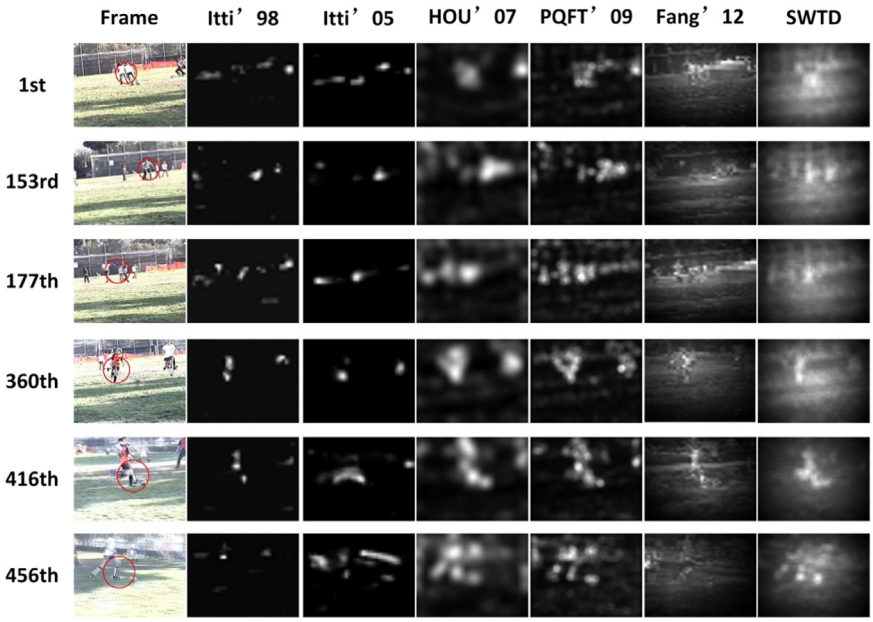

Fig. 4. Illustration of some frames in "beverly03" and corresponding saliency maps generated by the different models, where red circles are the locations of human eye fixations on each frame.

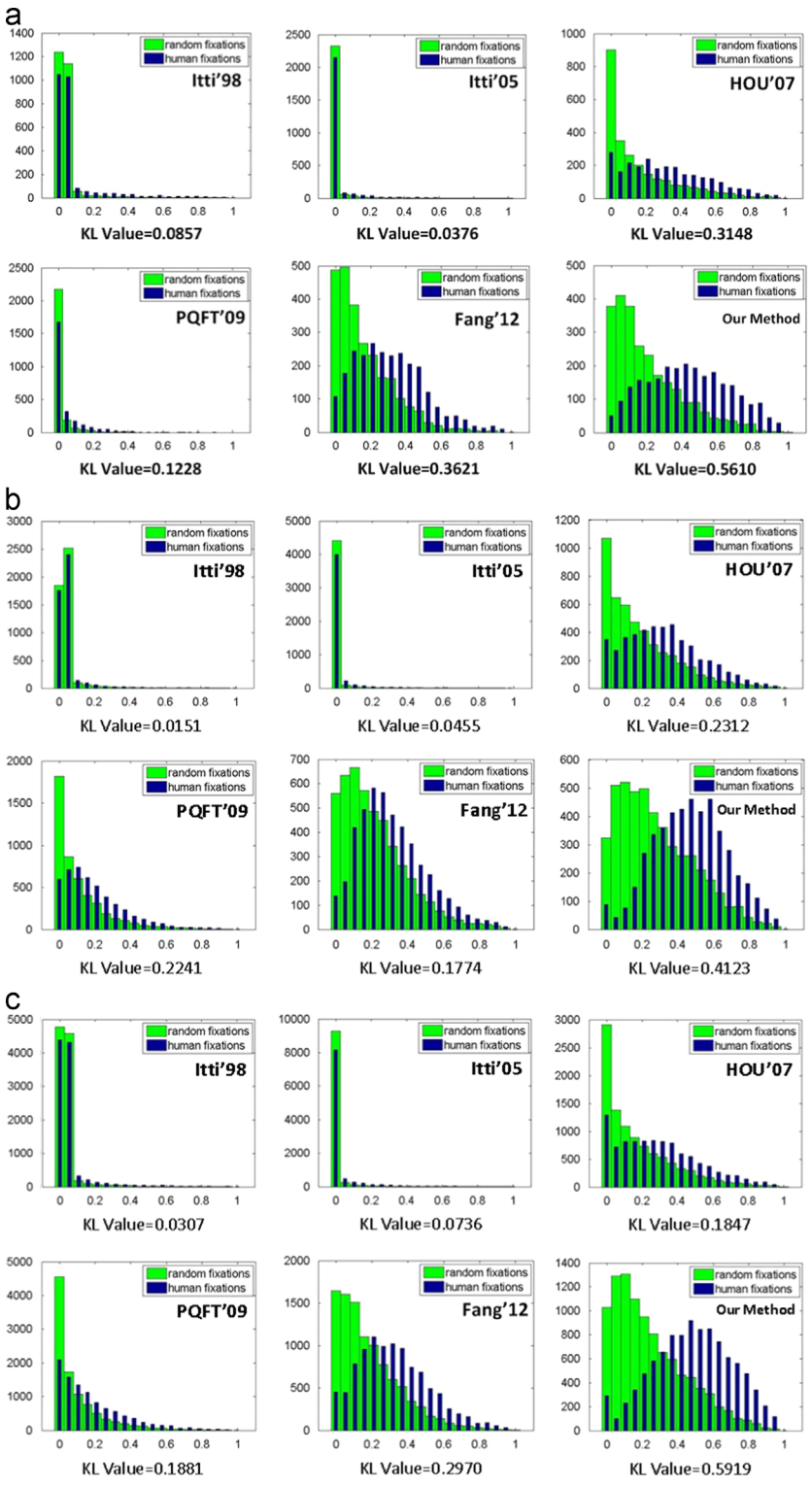

Fig. 5. Comparisons between Itti'98, Itti'05, HOU'07, PQFT'09, Fang'12 and our method. We quantify differences between histograms of saliency maps generated by these models with fixations samples collected from human and random fixation regions using the Kullback–Leibler (KL) distance. (a) Comparison on ORIG-M dataset. (b) Comparison on ORIG-N dataset. (c) Comparison on MTV dataset.

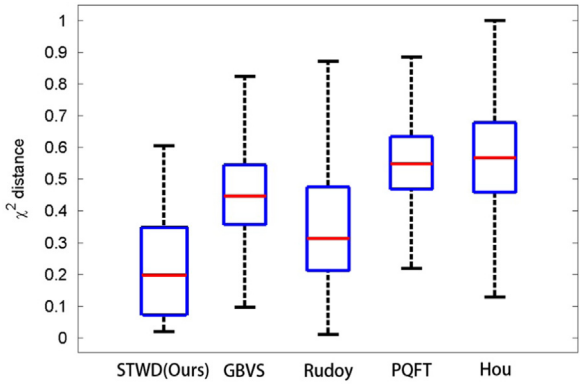

Fig. 6. Comparison with other four algorithms on DIEM dataset. We display the χ^2 distance between saliency maps and fixations which is better when the result is lower. The red lines stand for the median and blue represent the 90-th percentile.

Table 1
Comparison with [29] on DIEM.

Model	Rudoy'13 (all cues)	STWD (ours)
χ^2	0.313	0.199

Table 2
Comparison of the three methods: STWD including scene-change prediction (SCP), STWD without scene-change prediction and SWD.

Model	KL distance
STWD (with SCP)	0.5919
STWD (without SCP)	0.5699
SWD'11 [13]	0.5670

Note: Two parameters need to be discussed: (1) the dimension d to which each vector representing a block is reduced and (2) the size p of each block. The best results obtained by our method (Fig. 5) were obtained when $d=30$ and $p=14$. More details on finding good parameter values are given in Section 5.2.

Table 3
Comparison with SWD on the ORIG-M, ORIG-N and MTV.

KL distance	ORIG-M	ORIG-N	MTV
STWD	0.5610	0.4123	0.5919
SWD	0.5425	0.4165	0.5670

parks, crowds, sports, commercials, and video games. The scenes in ORIG-N include a rooftop bar scene, TV news show, and talk shows. We choose these datasets to understand the influence of moving objects on saliency and whether our method can adapt to both smoothly continuous scenes as well as scene changes. The DIEM dataset is a more challenging dataset in video saliency detection which is provided by the Dynamic Images and Eye Movements (DIEM) project. It is collected from over 250 volunteers gazing performance on 85 videos of several different types like sports, news report, and documentary film. There is quite a high resolution for most of the videos.

We compared our proposed method with five established methods of saliency detection on ORIG-CRCNS and MTV datasets: Itti'98 [21], Itti'05 [25], HOU'07 [22], PQFT'09 [28], Fang'12 [50]. Note that Itti'98, HOU'07 and Fang'12 estimate saliency on pictures (no temporal information) hence are applied on a frame-by-fame basis, whereas Itti'05 [25] and PQFT'09 [28] both use temporal information as an important factor. Each of the compared methods is representative of saliency detection in a different domains: spatial [21], spatial–temporal [25], spatial transform [22], temporal transform [28], and compressed [50] respectively. The method SWD'11 [13] is also used as a representative method that uses motion features. We will discuss the detailed result in Section 5.1. The KL distance is a popular evaluation of video saliency which measures the similarity between the predicted saliency distribution around fixations and random distribution saliency. In existing studies, KL distance can be computed in different ways. For example, in Itti et al. [25] KL distance is computed as the dissimilarity between the histogram of saliency sampled at eye fixations and that sampled at random locations, and [51] measured the KL distance between the saliency distribution of fixated points of a test image and the saliency distribution at the same pixel locations but of a randomly chosen image from the test set. In this paper, we reference the classical KL distance valuation of video saliency in [21,25,52]. Higher KL distances from random means that the corresponding method can better predict human fixations. Moreover, we evaluate the performance of our method on the DIEM. The parameters we set in this experiment are the same with the optimal parameter values selected from ORIG-CRCNS experiment. The whole DIEM dataset is seen as the test set. Our method is compared with four saliency detection algorithms, including one image saliency algorithm of GBVS'06 [5], and three video saliency methods from Rudoy'13 [29], PQFT'09 [28], and Hou'07 [22]. In this experiment we followed the evaluating indicator in Rudoy'13 [29] where χ^2 distance between saliency maps and fixations is employed to exhibit the result.

Next, we discuss the performance of our method on the datasets and the effectiveness of the motion feature for saliency detection on video.

4.1. Performance on ORIG-M dataset

The videos in this dataset contain obvious moving objects, such as joggers, cars, and cartoon characters. We randomly partitioned the dataset into 2 equal site subset. For each experiment, we selected one subset as the training set for parameters selection, and the remaining subset as the testing set. Fig. 5 provides quantitative comparison results for the compared methods. From Fig. 5(a) we can see that, on the ORIG-M dataset, our method delivered the best performance among all the compared saliency models. The results show that the performance of the proposed model is significantly different from the others. Some visual comparison samples are provided in Fig. 4. In the figure, the red circles in the first column are locations of human eye fixations. It is apparent

Table 4
(a) Relationship between β and performance, when $\alpha=0.2$. (b) Relationship between α and performance, when $\beta=3$.

(a)

β	1	2	3	4	5
KL distance	0.7310	0.7458	0.7481	0.7394	0.7406

(b)

α	0	0.2	0.4	0.6	0.8
KL distance	0.7290	0.7481	0.7218	0.7113	0.6985

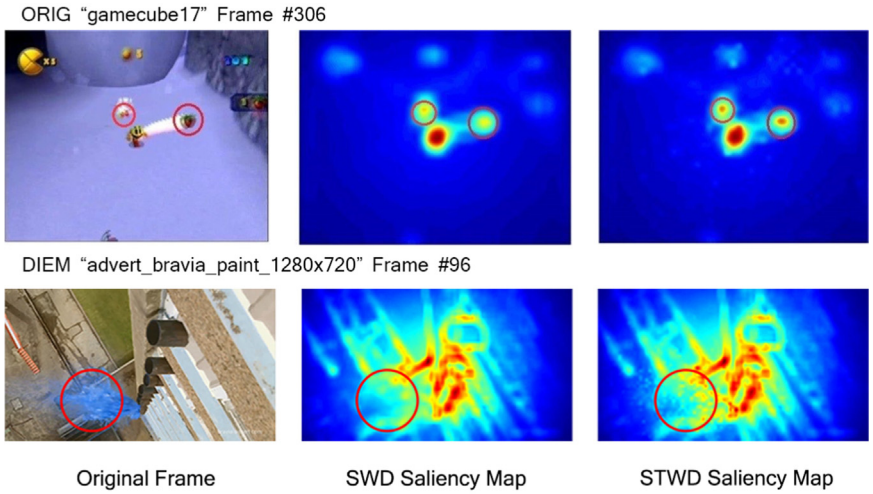

Fig. 7. Saliency maps on accelerating objects. Left: original frame, where the objects in the red circles have higher motion saliency. Middle: saliency map generated by the SWD'11 model. Right: saliency map generated by the STWD model.

that the saliency maps generated by our model on these examples were more consistent with the recorded human visual fixations.

4.2. Performance on ORIG-N dataset

This video dataset does not contain any videos where the camera tracks moving objects. Rather it contains a random selection of scenes from daily life. In this experiment, we use the same model parameters as in the previous experiment. As shown in Fig. 6(b), our method yields higher KL values than do Itti'98, Itti'05, HOU'07, PQFT'09 and Fang'12. However, the performance of our method is essentially equivalent to that of SWD'11 on this dataset.

4.3. Performance on MTV dataset

The MTV dataset also contains a wide variety of scenes. We also study the efficiency of our method when applied to scene changes, i.e. when the concatenated MTV clips are analyzed. Fig. 6(c) compared the proposed method with the other five methods. We again used the same parameters as were used in the previous experiment. The KL values from the proposed method are noticeably higher than those of the other models.

4.4. Performance on DIEM dataset

The DIEM dataset contains more complex scenes and have a higher resolution. We ran our algorithm on DIEM with the same parameter shown in Section 5.2. Our method performed well and extract the conspicuous motion successfully. We employed the results for other methods from [29] and compared with our method. It can be seen in Fig. 6 that our method presents the lowest χ^2 distance, it means the performance of our method on DIEM outperforms other algorithms. Table 1 presents the χ^2 distance compared with algorithms in [29].

5. Discussion

5.1. Comparison with SWD

The proposed method is an effective extension of SWD'11 [13]. We compared our method with SWD from two aspects, the effectiveness of the motion feature and our scene change handling mechanism.

To intuitively demonstrate the effectiveness of the motion feature, Fig. 7 depicts the result of saliency processing at a single frame using SWD'11 model and STWD respectively. In the original frame (Fig. 7, row 1 left, ORIG dataset "gamecube17" Frame ♯306), the cartoon character has a more

Table 5
(a) Relationship between retained dimension and performance, when block size is 14×14. (b) Relationship between block size and performance, when retained dimension is 30.

(a)

Retained dimension	6	10	14	18	22	26	30	34
KL distance	0.6237	0.6910	0.7167	0.7229	0.72637	0.7333	0.7477	0.7151

(b)

Block size	8×8	10×10	12×12	14×14	16×16	18×18
KL distance	0.6692	0.6784	0.6985	0.7477	0.7427	0.7144

Table 6
Performance on the three datasets (including center bias vs. not including center bias).

KL distance	Including center bias	Not including center bias
ORIG-M	0.5610	0.3997
ORIG-N	0.4123	0.3490
MTV	0.5699	0.2935

distinctive conspicuous appearance than the fruits (red circles), but the fruits have an accelerating motion at that moment which is more likely to attract visual fixations. The saliency map output by the STWD model (Fig. 7, row 1 right) takes much larger values on the fruits than does the saliency map generated by the SWD'11 model (Fig. 7, row 1 middle) does. In Fig. 7 row 2, the spouting paint received higher saliency from STWD model.

To further verify the effectiveness of our scene change handling mechanism, we studied the performance of our method both with and without it against the SWD'11 model (Table 2) on MTV dataset. When the scene-change mechanism is not used, the performance of our model falls below that of the SWD'11 model. However, when it is used, the performance of our model is much higher than that of the SWD'11 model. Indeed, handling scene changes on such videos appears to be an important ingredient of saliency prediction. The results of our method and SWD on ORIG-M, ORIG-N and MTV are present in Table 3. The KL distance of STWD is higher on both ORIG-M and MTV datasets which contain obvious moving objects. On ORIG-N dataset STWD performs a flat level with SWD. And this is sensible because the expand strategies focus on moving objects.

5.2. Parameter selection

In the proposed method, four parameters need to be discussed, which are α and β in formula 10, the block sizes and the dimensions to be reduced. Since the number of possible combinations of these parameters exceeds 10,000, computing the KL distance for every combination would be excessively time-consuming. As a more parsimonious solution, we use locally optimized parameters to approximate the global ones. The training set selected from ORIG-M dataset was used to accomplish parameters selection.

We first analyze α and β. In the testing phase, the block size and the retained dimensions were initialized at 16×16 and 32, respectively. When α was fixed to 0.2 ($\alpha=0.2$), the β was varied from 1 to 5 in increments of 1. Table 4(a) shows the performance of the proposed method at each increment. It may be seen that when $\beta=3$, the performance obtained was most accurate. To obtain the best value of α, we then fix $\beta=3$. Six different values of α (ranging from 0 to 0.8, in increments of 0.2) were tested. The results are shown in Table 4(b), from which it may be observed that the KL distance was the highest when $\alpha=0.2$.

Also, the block sizes and the dimensions to be reduced need to be discussed. We tested multiple combination of them to find the best block sizes and retained dimensions as we did when selecting α and β, as shown in Table 5. Here, we set $\alpha=0.2$ and $\beta=3$ as above. In the training phase, the block size was fixed to 14×14 ($p=14$), while the number of retained dimensions was varied from 6 to 34 in increments of 4. Table 5(a) shows the performance of the proposed method at each increment. It may be seen that when the number of retained dimensions is 30, the performance obtained was most accurate. Then, we fixed the number of retained dimensions at $d=30$. Six different patch sizes (ranging from 8 to 18, in increments of 2) were tested. The results are shown in Table 5(b), from which it may be observed that the KL distance was the highest when the block size is 14.

Although $\alpha=0.2$, $\beta=3$, $d=30$ and $p=14$ are not necessarily the optimal parameter values, our method achieves the best performance on the ORIG-M, ORIG-N and MTV datasets, as compared with the other five models using this assignment of parameters.

5.3. Effectiveness of center bias

The center bias factor is an important contribution to saliency prediction performance. We tested the performance of the proposed method both with center bias and without center bias, respectively. As shown in Table 6, on the ORIG-M dataset, eliminating the center bias causes the performance to decrease by 0.2754 ($\approx 28.8\%$). On the ORIG-N dataset, the performance decreases by 0.0633 ($\approx 15.4\%$). On the MTV datasets, the performance decreases by 0.2764 ($\approx 48.5\%$). This strongly indicates that center bias is a significant contributor to the efficiency of our method.

6. Conclusion

We proposed a new video saliency detection model for detecting salient regions on video, which combines two spatial features with one motion feature. We demonstrated the effectiveness of our model on four kinds of video datasets (ORIG-M, ORIG-N, MTV and DIEM) and found that it delivers highly competitive performance. We showed that accounting for sudden scene changes can boost the performance of visual saliency prediction, and hence we developed a model that is sensitive to scene changes. By employing this strategy, we found that our model can effectively locate salient regions even in the presence of scene changes. A novel feature fusion technique was applied to combine the proposed temporal and spatial features. Experimental results on four video datasets, ORIG-M, ORIG-N, MTV and DIEM, show that our model outperforms state-of-the-art video saliency detection approaches when predicting human fixations.

Acknowledgments

This research is partially sponsored by Natural Science Foundation of China (Nos. 61175115, 61370113, 61272320, 61472387 and 61472388), Beijing Municipal Natural Science Foundation (4152005 and 4152006), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (CIT&TCD201304035), Jing-Hua Talents Project of Beijing University of Technology (2014-JH-L06) and the International Communication Ability Development Plan for Young Teachers of Beijing University of Technology (No. 2014-16).

References

[1] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[2] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, in: Matters of Intelligence, Springer, Netherlands, 1987, pp. 115–141.
[3] L. Itti, Models of bottom-up and top-down visual attention, Ph.D. Thesis, California Institute of Technology, 2000.
[4] J.M. Wolfe, S.J. Butcher, C. Lee, M. Hyle, Changing your mind: on the contributions of top-down and bottom-up guidance in visual search for feature singletons, J. Exp. Psychol.: Hum. Percept. Perform. 29 (2) (2003) 483.
[5] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems, 2006, pp. 545–552.
[6] Y.-F. Ma, H.-J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: Proceedings of the Eleventh ACM International Conference on Multimedia, November 2–8, ACM, Berkeley, CA, USA, 2003, pp. 374–381.
[7] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1915–1926.
[8] J. Han, K.N. Ngan, M. Li, H. Zhang, Towards unsupervised attention object extraction by integrating visual attention and object growing, in: Proceedings of the IEEE International Conference on Image Processing, ICIP'04, vol. 2, Singapore, October 24–27, 2004, pp. 941–944.
[9] Z. Lu, W. Lin, X. Yang, E. Ong, S. Yao, Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation, IEEE Trans. Image Process. 14 (11) (2005) 1928–1942.
[10] R. Milanese, H. Wechsler, S. Gill, J.-M. Bost, T. Pun, Integration of bottom-up and top-down cues for visual attention using non-linear relaxation, in: Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'94, 1994, Seattle, WA, USA, June 21–23, 1994, pp. 781–785.
[11] A. Oliva, A. Torralba, M.S. Castelhano, J.M. Henderson, Top-down control of visual attention in object detection, in: Proceedings of the IEEE International Conference on Image Processing. vol. I, Barcelona, Spain, 2003, September 14–17.
[12] K. Seshadrinathan, A.C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE Trans. Image Process. 19 (2) (2010) 335–350.
[13] L. Duan, C. Wu, J. Miao, L. Qing, Y. Fu, Visual saliency detection by spatially weighted dissimilarity, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, June 20–25, 2011, pp. 473–480.
[14] J.-C. Baccon, L. Hafemeister, P. Gaussier, A context and task dependent visual attention system to control a mobile robot, in: Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 1, Lausanne, September 30–October 4, Switzerland, 2002, pp. 238–243.
[15] L.-Q. Chen, X. Xie, W.-Y. Ma, H. Zhang, H.-Q. Zhou, Image adaptation based on attention model for small-form-factor device, in: Proceedings of the 9th International Conference on Multi-Media Modeling, MMM, 2003, Taiwan, January 7–10, 2003, pp. 421–439.
[16] T. Lu, Z. Yuan, Y. Huang, D. Wu, H. Yu, Video retargeting with nonlinear spatial–temporal saliency fusion, in: Proceeding of the 2010 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, China, September 26–29, 2010, pp. 1801–1804.
[17] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, F. Pellandini, Adaptive color image compression based on visual attention, in: Proceedings of the 11th International Conference on Image Analysis and Processing, 2001, Palermo, Italy, 2001, September 26–28, pp. 416–421.
[18] F. Stentiford, A visual attention estimator applied to image subject enhancement and colour and grey level compression, in: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, Cambridge, UK, August 23–26, 2004, pp. 638–641.
[19] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, D. Kukolj, Salient motion features for video quality assessment, IEEE Trans. Image Process. 20 (4) (2011) 948–958.
[20] L. Zhu, L. Su, Q. Huang, H. Qi, Visual saliency and distortion weighting based video quality assessment, in: Proceedings of the Advances in Multimedia Information Processing—PCM 2012, Springer, 2012, Singapore, December 4–6, 2012, pp. 546–555.
[21] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[22] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07, Minneapolis, Minnesota, USA, June 18–23, 2007, pp. 1–8.
[23] M.-M. Cheng, G.-X. Zhang, N.J. Mitra, X. Huang, S.-M. Hu, Global contrast based salient region detection, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, Colorado Springs, CO, USA, June 20–25, 2011, pp. 409–416.
[24] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: Proceedings of the 14th Annual ACM International Conference on Multimedia, ACM, 2006, Santa Barbara, CA, USA, October 23–27, 2006, pp. 815–824.
[25] L. Itti, P.F. Baldi, Bayesian surprise attracts human attention, in: Advances in Neural Information Processing Systems, 2005, pp. 547–554.
[26] W.-H. Cheng, W.-T. Chu, J.-H. Kuo, J.-L. Wu, Automatic video region-of-interest determination based on user attention model, in: Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS 2005, Kobe, Japan, May 23–26, 2005, pp. 3219–3222.
[27] O. Boiman, M. Irani, Detecting irregularities in images and in video, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005, vol. 1, Beijing, China, October 17–21, 2005, pp. 462-469.
[28] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (1) (2010) 185–198.
[29] D. Rudoy, D. Goldman, E. Shechtman, L. Zelnik-Manor, Learning video saliency from human gaze using candidate selection, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, June 23-28, 2013, pp. 1147–1154, 〈http://dx.doi.org/10.1109/CVPR.2013.152〉.
[30] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, S. Kollias, Spatiotemporal saliency for video classification, Signal Process.: Image Commun. 24 (7) (2009) 557–571.

[31] W. Kim, C. Jung, C. Kim, Spatiotemporal saliency detection and its applications in static and dynamic scenes, IEEE Trans. Circuits Syst. Video Technol. 21 (4) (2011) 446–456.
[32] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, C.-W. Lin, A video saliency detection model in compressed domain, IEEE Trans. Circuits Syst. Video Technol. 24 (1) (2014) 27–38.
[33] D. Mahapatra, S. Gilani, M. Saini, Coherency based spatio-temporal saliency detection for video object segmentation, Sel. Top. Signal Process. 8 (3) (2014) 454–462.
[34] Z. Liu, X. Zhang, S. Luo, L.M. Olivier, Superpixel-based spatiotemporal saliency detection, IEEE Trans. Circuits Syst. Video Technol. 24 (9) (2014) 1522–1540.
[35] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.
[36] H.R. Tavakoli, E. Rahtu, J. Heikkilä, Spherical center-surround for video saliency detection using sparse sampling, in: Proceedings of the Advanced Concepts for Intelligent Vision Systems, Springer, Poznań, Poland, October 28–31, 2013, pp. 695–704.
[37] A. Zaharescu, R. Wildes, Spatiotemporal salience via centre-surround comparison of visual spacetime orientations, in: Lecture Notes in Computer Science, 2013, pp. 533–546.
[38] S. Marat, T.H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos, Int. J. Comput. Vis. 82 (3) (2009) 231–243.
[39] T. Mauthner, H. Possegger, G. Waltner, H. Bischof, Encoding based saliency detection for videos and images, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2494–2502.
[40] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[41] J. Li, Y. Tian, T. Huang, W. Gao, Probabilistic multi-task learning for visual saliency estimation in video, Int. J. Comput. Vis. 90 (2) (2010) 150–165.
[42] V. Gopalakrishnan, Y. Hu, D. Rajan, Random walks on graphs to model saliency in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, IEEE, 2009, Miami, Florida, USA, June 20–25, 2009, pp. 1698–1705.
[43] U. Rajashekar, I. van der Linde, A.C. Bovik, L.K. Cormack, Foveated analysis of image features at fixations, Vis. Res. 47 (25) (2007) 3160–3172.
[44] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, Beijing, China, October 17–21, 2005, pp. 60–65.
[45] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[46] B.W. Tatler, R.J. Baddeley, I.D. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vis. Res. 45 (5) (2005) 643–659.
[47] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, September 27–October 4, 2009, pp. 2106–2113.
[48] L.-M. Po, W.-C. Ma, A novel four-step search algorithm for fast block motion estimation, IEEE Trans. Circuits Syst. Video Technol. 6 (3) (1996) 313–317.
[49] P.K. Mital, T.J. Smith, R.L. Hill, J.M. Henderson, Clustering of gaze during dynamic scene viewing is predicted by motion, Cogn. Comput. 3 (1) (2011) 5–24, http://dx.doi.org/10.1007/s12559-010-9074-z.
[50] Y. Fang, Z. Chen, W. Lin, C.-W. Lin, Saliency detection in the compressed domain for adaptive image retargeting, IEEE Trans. Image Process. 21 (9) (2012) 3888–3901.
[51] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a Bayesian framework for saliency using natural statistics, J. Vis. 8 (7) (2008) 1–20.
[52] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, in: Advances in Neural Information Processing Systems, 2008, pp. 681–688.