Pattern Recognition 96 (2019) 106987

Contents lists available at ScienceDirect

Pattern Recognition

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

Video saliency detection by gestalt theory

Yuming Fang[a], Xiaoqiang Zhang[a], Feiniu Yuan[b,*], Nevrez Imamoglu[c], Haiwen Liu[d]

[a] School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, Jiangxi, China
[b] College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China
[c] Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan
[d] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China

ARTICLE INFO

Article history:
Received 4 September 2018
Revised 21 July 2019
Accepted 31 July 2019
Available online 2 August 2019

Keywords:
Visual attention
Video saliency detection
Gestalt theory
Uncertainty weighting
Spatiotemporal saliency

ABSTRACT

Image saliency detection has been widely explored in recent decades, but computational modeling of visual attention for video sequences is limited due to complicated temporal saliency extraction and fusion of spatial and temporal saliency. Inspired by Gestalt theory, we introduce a novel spatiotemporal saliency detection model in this study. First, we compute spatial and temporal saliency maps by low-level visual features. And then we merge these two saliency maps for spatiotemporal saliency prediction of video sequences. The spatial saliency map is calculated by extracting three kinds of features including color, luminance, and texture, while the temporal saliency map is computed by extracting motion features estimated from video sequences. A novel adaptive entropy-based uncertainty weighting method is designed to fuse spatial and temporal saliency maps to predict the final spatiotemporal saliency map by Gestalt theory. The Gestalt principle of similarity is used to estimate spatial uncertainty from spatial saliency, while temporal uncertainty is computed from temporal saliency by the Gestalt principle of common fate. Experimental results on three large-scale databases show that our method can predict visual saliency more accurately than the state-of-art spatiotemporal saliency detection algorithms.

1. Introduction

Visual attention is a significant mechanism in the Human Visual System (HVS) and it has been widely investigated in research areas of neuroscience and visual perception [1]. When we view a visual scene, the visual attention mechanism filters out most irrelevant information and focuses on salient regions. Basically, visual attention can fall into two kinds of approaches: bottom-up and top-down. The bottom-up approach, known as a stimulus-driven mechanism, usually predicts salient regions automatically based on feature contrast from low-level features such as luminance, color, and texture in visual scenes, while top-down attention, also known as a task-driven mechanism, is determined by specific prior knowledge, such as tasks, expectations and current goals.

By visual attention modeling, saliency regions can be extracted from visual scenes. Generally, the saliency of an image pixel is defined as the probability for this pixel being looked at. Existing saliency detection models can be divided into two types: eye fixation prediction and salient object detection models. Human fixation prediction models aim to locate fixation regions where human eyes look at during scene viewing, while salient object detection models aim to locate the whole salient objects in visual scenes. In the past decade, there have been various kinds of saliency detection models explored due to their wide applications [2–9]. Itti et al. proposed a computational model of visual saliency by multi-scale low-level features including color, luminance, and orientation [3]. Following Itti's work, a simple and biologically inspired model was proposed using graph theory in the study [4]. Some psychological studies [10,11] show that the HVS deals with targets from multiple scales. Inspired by these studies [10,11], Yan et al. [12] designed a hierarchical saliency prediction model to solve the problem of salient object detection at a small scale.

Besides these commonly used low-level features in saliency detection due to the sensitivity of the HVS to them [13], some new features such as image boundaries have also been investigated in saliency detection [14,15]. In [14], the authors considered image boundaries as parts of background regions, which are segmented out for saliency prediction by manifold ranking based on graph theory. Based on the study [14], Zhu et al. [15] designed a method to compute a background prior called boundary connectivity by estimating the degree of an image block connecting an image boundary. The authors in [16] explored color histogram features for saliency detection. Goferman et al. [17] introduced a context-aware saliency prediction model, which claims that salient objects cannot exist without context information. Gopalakrishnan

* Corresponding author.
E-mail addresses: fa0001ng@e.ntu.edu.sg (Y. Fang), yfn@ustc.edu.cn (F. Yuan).

https://doi.org/10.1016/j.patcog.2019.106987
0031-3203/

et al. [18] defined visual attention modeling as a labelling problem based on graph theory by exploiting color and orientation entropy features. Some other studies try to extract saliency information in the frequency domain [19–22]. Hou et al. explored spectral residual features of an image for saliency prediction by converting the input image into the frequency domain [19]. Guo et al. developed a saliency prediction model by phase spectrum [23]. Schauerte et al. used the quaternion Fourier spectrum for saliency prediction [24]. Some superpixel-based saliency detection models have been developed for performance improvement of visual saliency detection [25–30]. Recently, there have been some machine learning based models built for visual saliency detection [31,32].

The image saliency detection models mentioned above only need to extract spatial features, while video saliency detection models have to exploit the complicated motion features in video sequences. Traditional studies of video saliency detection try to detect moving objects as salient regions in video sequences. In [33], Sun et al. summarized unconstrained video sequences using salient montages by finding "montageable moments", which is adopted to identify salient people and actions in video sequences. Some studies use background priors to locate salient regions in video sequences. Xi et al. [34] proposed a video saliency prediction model based on spatial and temporal background priors. In the study [34], superpixel-level boundary connectivity is computed as the spatial background prior, while background homography is used to estimate the temporal background prior. Le Meur et al. designed a spatiotemporal saliency detection method based on perceptual characteristics, including visual masking, perceptual decomposition, contrast sensitivity functions, and center-surround interactions [35]. Some studies explore visual saliency detection based on superpixel segmentation for video sequences [36–38]. Liu et al. computed a superpixel-based spatiotemporal saliency method for video sequences based on global contrast, spatial sparsity, and object prior [36]. In that study [36], superpixel-level spatial saliency and temporal saliency are predicted based on spatial and temporal features, respectively. For generating pixel-level saliency results, they adapted a saliency derivation approach to compute the final spatiotemporal saliency for video sequences [36]. In [37], the authors predicted video saliency by two steps: first, it estimates locations of saliency regions based on graph theory; second, it refines saliency results on these locations to generate the final saliency map [37]. In [39], Leboran et al. developed a spatiotemporal saliency detection model by the assumption that perceptual information is related to high-order statistical structures.

Some studies try to explore effective fusion of spatial and temporal saliency for saliency detection of video sequences. Lee et al. extracted various kinds of features and fused those features for video saliency prediction using a SVM model [40]. Kim et al. adopted a random walk with restarting to detect spatial and temporal saliency maps, and fuse these two saliency maps by a constant distribution of the walker [41]. Based on feature integration theory [42], Le Meur et al. [43] combined achromatic, chromatic and temporal saliency maps to predict video saliency. Mahadevan et al. [44] calculated feature maps based on mechanisms of feature perception, and combined these feature maps by modeling dynamical textures. Chen et al. [45] computed video saliency in a batchwise way. The authors in [46] presented a video saliency detection approach based on an energy function and optical flow. The proposed model includes three steps: first, locate salient regions with optical flow gradient; second, improve saliency results with local and global contrast information; finally, refine spatiotemporal saliency with an energy function.

Recently, some deep learning based models have been proposed for saliency detection [47,48]. In [47], the authors proposed two deep learning based models for saliency detection of images including DeepGaze II and ICF. DeepGaze II uses high-level features trained on databases of object recognition, while ICF is designed based on low-level features (one luminance and two color). These two models predict saliency results by passing features to the same readout network. In [49], the authors built a saliency detection model by two modules: a network for temporal saliency extraction and a network for static saliency extraction. In [50], Yuan et al. constructed a deep neural network based denser and sparse labeling framework for saliency detection. Kruthiventi et al. explored fully convolutional networks for saliency detection in an end-to-end manner [48]. Wang et al. [51] proposed a video saliency detection model by using a convolutional LSTM (Long Short-Term Memory) architecture. In [52], the authors designed a saliency detection model based on a skip-layer network structure. Recently, Wang *et al.* designed an salient object detection model based on deep neural network [53] by two subnetworks.

As introduced above, traditional methods of video saliency detection are designed by linearly combining spatial and temporal saliency for spatiotemporal saliency prediction. However, the simple linear combination method for spatial and temporal saliency is not reasonable due to the differences between perception of spatial and temporal saliency in the HVS. To address this problem, we propose an adaptive fusion method of spatial and temporal saliency based on Gestalt theory [54,55]. The theory is introduced in Section 2.1. To summarize, the Gestalt theory of similarity is adopted to estimate the spatial uncertainty, while the temporal uncertainty is computed by the Gestalt theory of common fate. Based on uncertainty weighting, we calculate the final spatiotemporal saliency by combining spatial and temporal saliency. Experimental results show that the proposed model performs better than the state-of-art video saliency detection algorithms on three public available datasets.

Please note, the proposed model is quite different from our previous work [56]. In the proposed method, we calculate temporal saliency uncertainty using the Gestalt theory of common fate, while Fang [56] computed the temporal uncertainty based on psychophysical experiments on motion perception [57]. The uncertainty of temporal saliency by Fang [56] mainly detects local texture information, and it fails to detect complicated motion features in video frames. The reason for this failure might be that the consideration of local contrast in Fang [56] would influence the uncertainty weighting estimation of temporal saliency. Thus, temporal saliency uncertainty predicted by the proposed method is more robust than that in [56].

In the study [56], spatial saliency uncertainty was computed based on the Gestalt theory of continuity and proximity. In the proposed method, we calculate the spatial saliency uncertainty using the Gestalt theory of similarity. Furthermore, we propose a unified framework for saliency prediction in video sequences based on two laws of similarity and common fate in Gestalt theory. The uncertainty of temporal saliency is estimated by the law of common fate in Gestalt theory, which is the first attempt in the research community. Experimental results in Section 3 also show that the proposed method using Gestalt theory can predict salient regions more accurately than the study [56].

The detailed steps of the proposed model are shown as follows: (1) we compute spatial and temporal saliency as a function of spatial distance and feature difference between image regions as introduced in Section 2.2; (2) we use this saliency measure to directly compute two probabilities for the weighting of spatial and temporal saliency inspired by Gestalt grouping as introduced in Section 2.3; (3) lastly, we make an approximation of these two probability distributions and then convert them to an uncertainty and entropy measure in Section 2.4; (4) the final spatiotemporal saliency is computed by fusing spatial saliency and temporal saliency with uncertainty based weighting in Section 2.5.

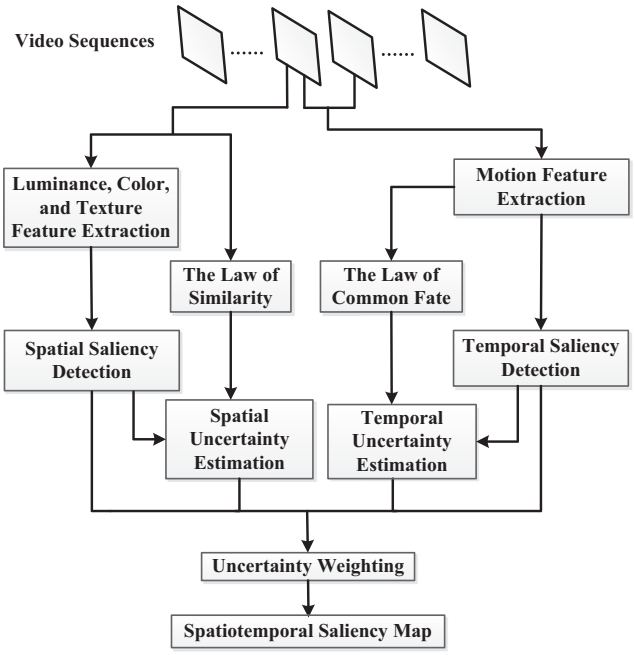

Fig. 1. The framework of the proposed model.

We have released the source code of this work on the following link http://sim.jxufe.cn/JDMKL/ymfang_EN/GelstaltSaliency.zip

2. Proposed method

We show the proposed framework in Fig. 1, which includes two main steps: first, spatial and temporal saliency maps are predicted by low-level features; second, the final spatiotemporal saliency is predicted by combining the spatial and temporal saliency maps using uncertainty weighting. The Gestalt law of similarity is used to estimate spatial uncertainty, while temporal uncertainty is computed by Gestalt law of common fate. We detect salient objects in each frame individually without considering the results in prior frames for computation efficiency.

2.1. Gestalt theory

Gestalt psychology aims to understand theories behind the ability to obtain and maintain meaningful perceptions in an apparently chaotic world. The main Gestalt principle is that the perception forms a global whole with self-organizing tendencies. The fundamental principle of Gestalt perception is the law of grouping aka *Prägnanz*, including eight rules of grouping: proximity, similarity, closure, symmetry, common fate, continuity, good gestalt, and past experience. In this work, we use the laws of similarity and common fate to compute the uncertainty of spatial saliency and temporal saliency, respectively. The law of similarity claims that elements would tend to be perceived within one group if they are similar to each other. As shown in Fig. 2, the HVS would instinctively divide the image content into three groups owing to similar color properties. Besides color attributes, other visual properties can also be used in the Gestalt theory of similarity. The common fate principle states that elements tend to be perceived as a group if they move in the same way. As shown in Fig. 3, if elements move in the same direction and speed, they would be perceived as a group, even across large distances. These two laws are used to estimate the uncertainty of spatial saliency and temporal saliency as follows: first, an image pixel that is more similar to the saliency center in an image is more likely to be a salient pixel; second, an image pixel in the same way of moving with the saliency center is more likely to be a salient pixel.

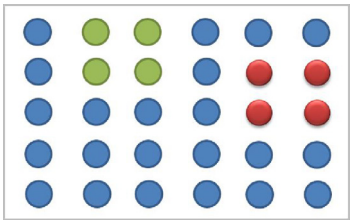

Fig. 2. Similarity principle.

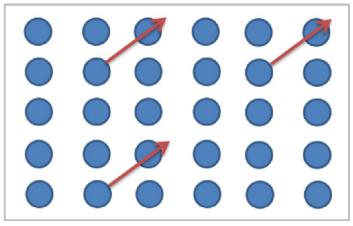

Fig. 3. Common fate principle.

2.2. Spatial and temporal saliency evaluation

In the proposed method, three kinds of static features are extracted for spatial saliency detection: luminance, color, and texture. First, we transform the color space from RGB to YCbCr. The Y channel contains luminance information, while the Cb and Cr channels represent color information for video sequences. Discrete cosine transform (DCT) of an image block is adopted to compute feature contrast for video frames [20], which would get DCT coefficients including DC coefficient and AC coefficients. These low-level features are computed as follows: one luminance feature L is extracted from the DC coefficient of the Y channel; two color features (C_b and C_r) are computed from the DC coefficients of the Cb and Cr channels; one texture feature T is computed from the AC coefficients of the Y channel. In the study [58,59], AC coefficients contain high frequency information, which can be used to represent texture features for video sequences. Based on these four features, we calculate the feature map S_i^f for image block i as follows:

$$S_i^f = \sum_{j \neq i} \frac{1}{\sigma_s \sqrt{2\pi}} e^{-d_{ij}^2/2\sigma_s^2} D_{ij}^f, \tag{1}$$

where $f \in \{L, C_b, C_r, T\}$; σ_s is the parameter of the Gaussian kernel function and used to balance local and global feature contrast; d_{ij} denotes the Euclidean distance between image blocks i and j; D_{ij}^f denotes the feature difference between image blocks i and j. For luminance and color features, feature differences can be predicted by the difference between DC coefficients of the corresponding blocks as follows.

$$D_{ij}^f = \frac{|DC_i^f - DC_j^f|}{|DC_i^f| + |DC_j^f|} \tag{2}$$

where $f \in \{L, C_b, C_r\}$; DC denotes DC coefficient.

The texture feature difference can be computed using the difference between the AC coefficients of the corresponding blocks as follows.

$$D_{ij}^T = \frac{\sqrt{\sum_t (T_i^t - T_j^t)^2}}{\sum_t (T_i^t + T_j^t)} \tag{3}$$

where t denotes the index for AC coefficients in an image block.

The final spatial saliency map S_s can be obtained as:

$$S_s = \frac{1}{n}\sum_{f\in\{L,C_b,C_r,T\}} N(S^f). \tag{4}$$

where n denotes the number of features ($n = 4$); $N(\cdot)$ represents the operation of normalization. We use Min-Max normalization as shown below.

$$N(S^f) = \frac{S^f - min(S^f)}{max(S^f) - min(S^f)}, \tag{5}$$

where the functions $max(\cdot)$ and $min(\cdot)$ is used to find maximum and minimum values in the saliency map S^f, respectively.

Here, we use optical flow to estimate a motion feature for video sequences [60]. The motion saliency can be also treated as the prior probability distribution about perceptual motion speed [57]. An object with strong motion compared to the background would be perceived as salient object for HVS. According to the study [57], the prior probability distribution can be fitted by the function:

$$p(v) = a/v^{b_1}, \tag{6}$$

where a and b_1 are two positive constants and set as $e^{-0.09}$ and 0.2, respectively; v denotes motion speed. The temporal saliency S_t can be calculated by using its self-information as follows:

$$S_t = -\log p(v) = b_1 \log v + b_2. \tag{7}$$

where b_1 and b_2 are set as 0.2 and 0.09, respectively; $b_2 = -\log a$ is a constant. These parameters are set based on the study [61], where these parameters are fitted by large-scale experimental data. v denotes the relative motion between the object and background [62], which can be computed as:

$$v_i = \sum_{j\neq i} \frac{1}{\sigma_s\sqrt{2\pi}} e^{-d_{ij}^2/2\sigma_s^2} D_{ij}^v, \tag{8}$$

where D_{ij}^v denotes the motion difference between blocks i and j. In this work, we compute D_{ij}^v as follows.

$$D_{ij}^v = \sqrt{(G_i^x - G_j^x)^2 + (G_i^y - G_j^y)^2}, \tag{9}$$

where G_i^x and G_j^x represent horizontal motion vectors of image patches i and j, respectively; G_i^y and G_j^y denote vertical motion vectors of image patches i and j, respectively. Here, i and j index represent image patch. For the spatial saliency map and temporal saliency map, we resize these saliency maps into the same size of original video frames for the final saliency prediction.

2.3. Saliency probability estimation

Compared with saliency detection of images, saliency detection of video sequences is more complicated due to the additional dimension of motion existing in video sequences. An object might be regarded as salient with high certainty if the object has high color contrast relative to the background, while the certainty would descend greatly if the object has smaller motion compared to other objects in the same video sequence. In other words, an object might be salient from the perspective of spatial features, while the object might be non-salient from the perspective of temporal features. Therefore, how to balance and integrate spatial and temporal saliency to compute the final spatiotemporal saliency of video sequences is very challenging. We introduce an uncertainty measure as the weighting to combine these two types of saliency maps based on Gestalt theory [54,55].

The visual acuity in the HVS decreases with increasing eccentricity from the fovea [63]. The HVS is more sensitive to center-surround difference from the patch with nearer distance compared with those from farther patches. Here, we use a Gaussian model to simulate this mechanism to weight center-surround differences for saliency detection, as shown in Eq. (1). Therefore, feature differences from a nearer patch would get a larger weight during saliency computation. The saliency value is proportional to the feature difference. For uncertainty estimation of spatial saliency, we compute the similarity between the saliency values of a pixel and the saliency center based on the law of similarity in Gestalt theory. With larger similarity of an image pixel to the saliency center, the uncertainty of this image pixel is smaller.

The uncertainty of spatial saliency is estimated based on the Gestalt theory of similarity, which states that elements with similar properties should be perceived as a group. The uncertainty of temporal saliency is predicted based on the Gestalt theory of common fate, which states that elements with similar motion should be perceived as a group. We apply these two kinds of principles to visual attention modeling as follows: an image pixel with similar properties to that of the saliency center tends to be a salient pixel; an image pixel that has the same way of moving with that of the saliency center tends to be a salient pixel. These properties are consistent with the statistics of an eye tracking database of video sequences built by Lee et al. [40], which are used to obtain fitting curves for the proposed model. These fitting curves are shown in Fig. 4(a) and (b). We can see that the probability of an image pixel being looked at is proportional to the feature similarity and inversely proportional to the motion difference.

Here, the feature similarity of the current superpixel region denotes similarity degree of feature vectors of this current superpixel region and other superpixel regions. It is calculated by the difference between feature vectors of the current superpixel region and other superpixel regions. With larger feature differences between the current superpixel region and other superpixel regions in the input image, the saliency value of this current superpixel region is larger.

In the proposed method, we first use simple linear iterative clustering algorithm (SLIC) [64] to segment video frames into superpixels. With spatial saliency results computed in Eq. (4), the saliency center represented by a superpixel region can be obtained by:

$$c = \underset{p}{argmax}\left(\frac{1}{L(Q_p)}\sum_{(r,l)\in Q_p} S_{r,l}\right),\ s.t.\ 1 < p < M, \tag{10}$$

where S represents spatial saliency map computed in Section 2.2; Q_p denotes the pth superpixel in the segmented video frame; $S_{r,l}$ is the saliency value at the location (r, l) in saliency map S corresponding to location (r, l) in the superpixel Q_p; M denotes the superpixel number of the segmented video frame; $L(\cdot)$ is used to count the pixel number within the superpixel Q_p. We calculate the average saliency value of the region in S corresponding to superpixel Q_p. The resultant saliency center represents a superpixel region. We compute the similarity between different superpixel regions by color statistics. Here, the degree of similarity R_p of each superpixel Q_p compared to the saliency center Q_c is calculated by:

$$R_p = e^{(-H(Q_p,Q_c))}, \tag{11}$$

where $H(\cdot)$ is the function that computes the $L2$ distance of the histogram-based features between superpixels Q_p and Q_c. For $H(\cdot)$, we use the histograms of color and luminance features for its simplicity and effectiveness in saliency prediction [16]. $H(\cdot)$ is defined as:

$$H(Q_p, Q_c) = \sqrt{\sum_{k=1}^{K}((F_{Q_p})_k - (F_{Q_q})_k)^2}, \tag{12}$$

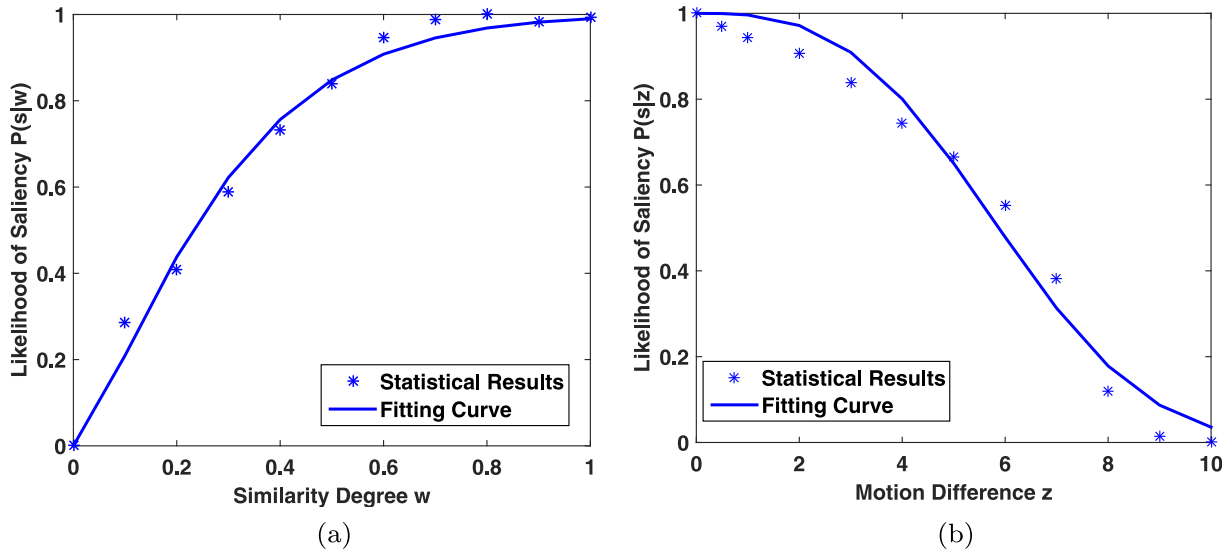

Fig. 4. (a) Likelihood of saliency as a function of similarity. The fitting function for this figure is given in Eq. (14). Here, the saliency of an image pixel is defined as a probability of being looked at with the condition of feature similarity. (b) Likelihood of saliency as a function of common fate. The fitting function for this figure is given in Eq. (16). Here, the saliency of an image pixel is defined as a probability of being looked at with the condition of motion difference.

where F_{Q_p} and F_{Q_c} are feature vectors of superpixels Q_p and Q_c based on color/luminance histograms, respectively; k denotes the kth dimension of feature vector; c represents the saliency center calculated by Eq. (10); K denotes the dimension size of histogram feature vector. We set each feature (R, G, B, and Luminance) with 10 bins, and thus, K is equal to 40. The contribution of the degree of similarity to saliency prediction decreases when the image pixel m (the distance is computed on pixel level) is far away from the image pixel n. Therefore, we use an exponential function to weight the degree of similarity as below.

$$w_m = e^{-(\frac{1}{\alpha})g_{mn}} R_m, \tag{13}$$

where R_m denotes the degree of similarity of the pixel m to the saliency center and m belongs to superpixel Q_p; the image pixel n is the center of superpixel Q_c; g_{mn} represents the Euclidean distance between image pixels m and n; α is a parameter and we set $\alpha = 60$; w_m is the final the degree of similarity. We compute statistics on the likelihood of a pixel being salient as a function of the degree of similarity w on the video dataset [40], and the results are shown in Fig. 4(a). As shown in this figure, we can see that the probability of an image pixel being salient is proportional to the similarity between the image pixel and the saliency center. This relationship can be summarised by using a fitting function as follows:

$$P(s|w) = 1 - exp\left[-\left(\frac{w}{\alpha_1}\right)^{\alpha_2}\right], \tag{14}$$

where $P(s|w)$ denotes the probability of an image pixel being salient given its degree of similarity w; α_1 and α_2 are fitting parameters fitted as $\alpha_1 = 0.3062$ and $\alpha_2 = 1.2930$, respectively, based on the video database [40]. The fitting curve is shown in Fig. 4(a).

We estimate the uncertainty for temporal saliency by the Gestalt theory of common fate, which states that an image pixel with more similar motion features to that of the saliency center would be more likely to be salient. We represent the motion feature of the saliency center as (V_{xc}, V_{yc}) and compute the motion difference z_m between motion vectors of any image pixel m and saliency center in a video frame as follows:

$$z_m = \sqrt{(V_{xc} - V_{x_m})^2 + (V_{yc} - V_{y_m})^2}, \tag{15}$$

where (V_{x_m}, V_{y_m}) denotes the motion feature of the image pixel m in the video frame.

To figure out the relationship between the probability of an image pixel being salient and the motion difference, we compute statistics on the probability as a function of motion difference z_m, and results are shown in Fig. 4(b). From this figure, we can see that the probability of an image pixel being salient is inversely proportional to the motion difference between the image pixel and the saliency center. This relationship could be summarised by an empirical function as:

$$P(s|z) = exp\left[-\left(\frac{z}{\beta_1}\right)^{\beta_2}\right], \tag{16}$$

where z denotes motion difference; $P(s|z)$ denotes the probability of an image pixel being salient with its motion difference z; β_1 and β_2 are fitting parameters and found to be $\beta_1 = 6.6528$ and $\beta_2 = 2.9547$, respectively, based on the video database [40]. The fitting curve is shown in Fig. 4(b).

2.4. Uncertainty estimation

Based on the probability of saliency calculated in Section 2.3, we estimate the uncertainty for spatial and temporal saliency. The uncertainty of spatial saliency is quantified by the entropy of the likelihood:

$$U^s = E(P(s|w)). \tag{17}$$

where $E(P)$ denotes the entropy function computed as: $-Plog_2P - (1-P)log_2(1-P)$. Similar with Eq. (17), we formulate the uncertainty of temporal saliency as:

$$U^t = E(P(s|z)). \tag{18}$$

2.5. Spatiotemporal saliency estimation

Since larger uncertainty should be given lower weight in the final saliency prediction, we let the weight be inversely proportional to uncertainty. Thus, the weight given to spatial saliency is U^t, while the weight given to temporal saliency is U^s. The spatiotemporal saliency map of each video frame can be estimated by

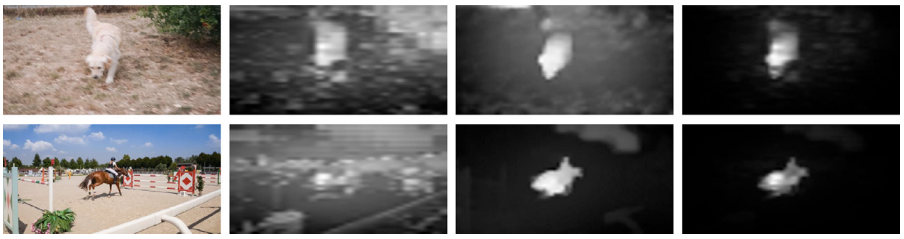

Fig. 5. Saliency map samples. Column 1–4: original video frames, spatial saliency maps, temporal saliency maps, and spatiotemporal saliency maps.

combining spatial and temporal saliency with uncertainty weighting as follows.

$$S = \frac{U^t S_s + U^s S_t}{U^s + U^t}. \tag{19}$$

From Eq. (19), we can see that the corresponding weights of spatial saliency map S_s and temporal saliency map S_t are spatiotemporally adaptive, which are different from linear or fixed weight to fuse spatial and temporal saliency maps [41,65]. As shown in Fig. 5, the fused spatiotemporal saliency map from the proposed method predicts more accurate saliency locations than spatial/temporal saliency map.

3. Experimental evaluation

3.1. Evaluation methodology

We perform extensive experiments on FBMS [66], DAVIS [67], and ViSal [46]. FBMS is a publicly available salient object database, containing 59 video sequences (30 test video sequences). DAVIS contains 3455 video frames (50 video sequences) with common video saliency detection challenges such as fast-motion, occlusions, no-linear deformation and motion blur. ViSal database includes 17 challenging video sequences with highly cluttered background, multiple objects with various motion patterns, complex color distributions and camera motion. These video sequences range from 30 to 100 frames, and all these frames are manually labelled as binary ground truth maps.

In this experiment, three kinds of commonly used evaluation criteria are used: Precision-Recall (PR) curve, *F*-measure, and mean absolute error (MAE) [49]. PR curve, determined by precision and recall, is widely used to evaluate the performance of saliency prediction models. We can sort the salient values in the saliency map as a list from the largest value to the smallest value, and use values from the first to the last in the list as thresholds to classify image pixels in the saliency map into salient pixels and non-salient pixels. Precision is the percentage of correctly detected image pixels to detected image pixels by a saliency detection model, while recall is the percentage of correctly detected image pixels to the ground-truth salient image pixels. Specifically, Precision and Recall are defined as:

$$Precision = \frac{TP}{TP + FP}, \tag{20}$$

$$Recall = \frac{TP}{TP + FN}. \tag{21}$$

where TP, FP, TN, and FN represent the number of correctly detected salient pixels, falsely detected salient pixels, correctly detected non-salient pixels, and falsely detected non-salient pixels. $TP + FP$ and $TP + TN$ represent the number of being predicted to be salient pixels and true salient pixels. F-measure is a comprehensive consideration of precision and recall:

$$F - measure = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{22}$$

where we set $\beta^2 = 0.3$ as suggested in [49]. Each PR pair corresponds to an F-measure. Therefore, using different thresholds from 0 to 255 for a saliency map will result in a series of F-measure values.

MAE is the average per-pixel difference between saliency map *S* and ground truth map *G*:

$$MAE = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} |S_{ij} - G_{ij}| \tag{23}$$

where saliency map *S* and ground truth map *G* are normalized to [0, 1]. *m* and *n* represent the number of row and column for the saliency map *S*, respectively.

3.2. Comparison experiments

3.2.1. Visual comparison

We compare the proposed model against the state-of-the-art studies including RWRV [41], CE [65], Seo [68], Fang [56], SGSP [38], LGGR [46], DeepGaze II [47], and ICF [47]. The first six are traditional models for saliency detection, while the last two are deep learning based saliency detection models. In Figs. 6–8, we provide some visual samples of saliency results on FBMS, DAVIS, and ViSal to evaluate the performance of the proposed model. As can be seen from these figures, it is obvious that the saliency maps from existing models have many false saliency detection results, and some background pixels are falsely detected as salient by these models.

In Fig. 6, we can see that CE would detect some background regions as salient due to the failure of motion feature extraction. As shown by the fourth and seventh rows of Fig. 6, RWRV, Seo, and Fang would also wrongly detect some background regions with rich texture information as salient due to their unreasonable combination methods of spatial and temporal saliency. SGSP extracts color and motion information as features for saliency prediction and ignore other important information, such as texture and luminance. Therefore, SGSP cannot suppress residual saliency in background region, as shown in the last two rows of Fig. 6. LGGR detects salient regions by considering local and global saliency cues. LGGR combines these two terms by simple linear operation, which is not adaptive. DeepGaze II and ICF only extract static features and ignore motion feature. In this work, we compute spatiotemporal saliency by fusing spatial and temporal saliency with uncertainty weighting estimated by the two laws of similarity and common fate in Gestalt theory. The proposed fusion method by uncertainty is adaptive, and thus, it can obtain better performance of saliency detection over other existing ones on three public databases.

Fig. 7 shows some visual samples from different video saliency detection models on DAVIS. From the first, second and fifth rows in this figure, we can observe that RWRV, Seo, and Fang wrongly detect some background regions (wall and building) as salient with their unreasonable combination of spatial and temporal saliency. Without considering motion features, CE cannot retain the boundaries of salient objects in video sequences. As indicated previously, SGSP only uses color and motion information for saliency measurement and would ignore other important information. Therefore,

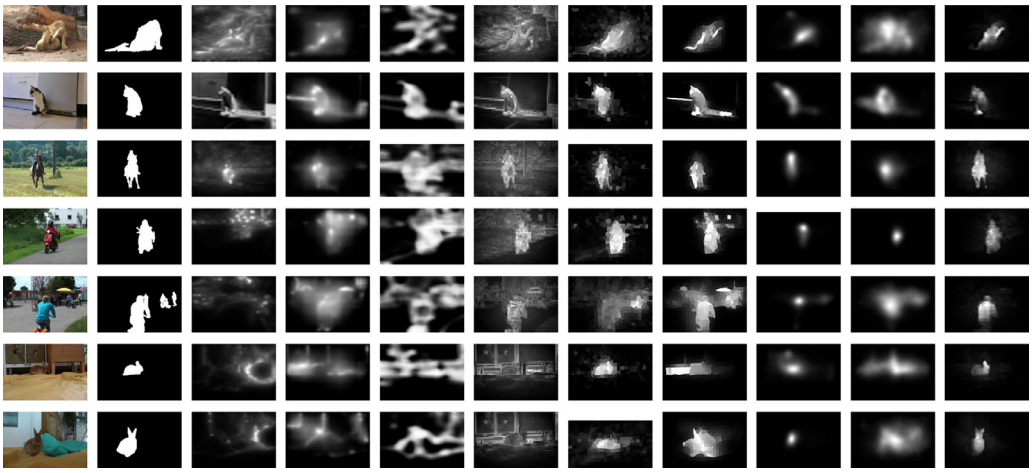

Fig. 6. Visual comparisons of output saliency results generated by different video saliency detection models on FBMS database. First column to the final column: original video frames, the ground truth maps, saliency maps from RWRV [41], CE [65], Seo [68], Fang [56], SGSP [38], LGGR [46], DeepGaze II [47], ICF [47] and the proposed method.

Fig. 7. Visual comparisons of output saliency results generated by different video saliency detection models on DAVIS database. First column to the final column: original images, the ground truth maps, saliency maps from RWRV [41], CE [65], Seo [68], Fang [56], SGSP [38], LGGR [46], DeepGaze [47] and ICF [47] and the proposed method.

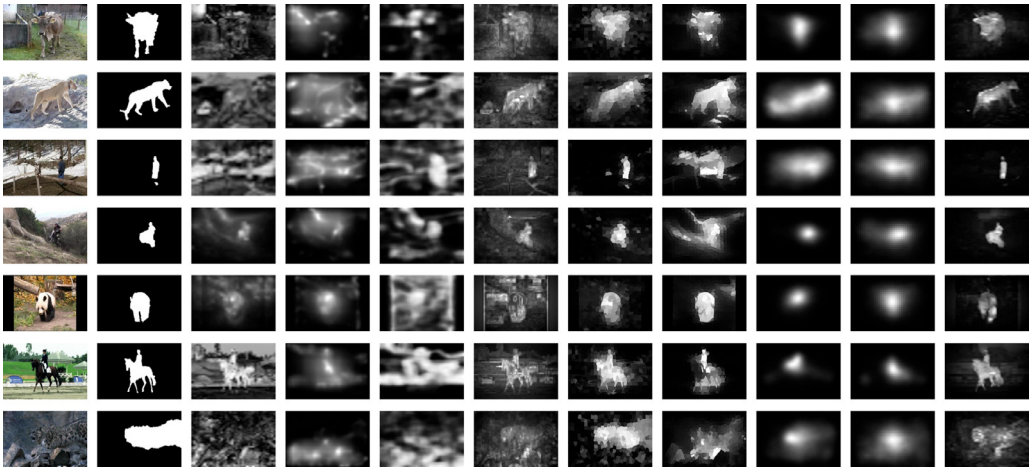

Fig. 8. Visual comparisons of output saliency results generated by different video saliency detection models on Visal database. First column to the final column: original video frames, the ground truth maps, saliency maps from RWRV [41], CE [65], Seo [68], Fang [56], SGSP [38], LGGR [46], DeepGaze II [47], ICF [47] and the proposed method.

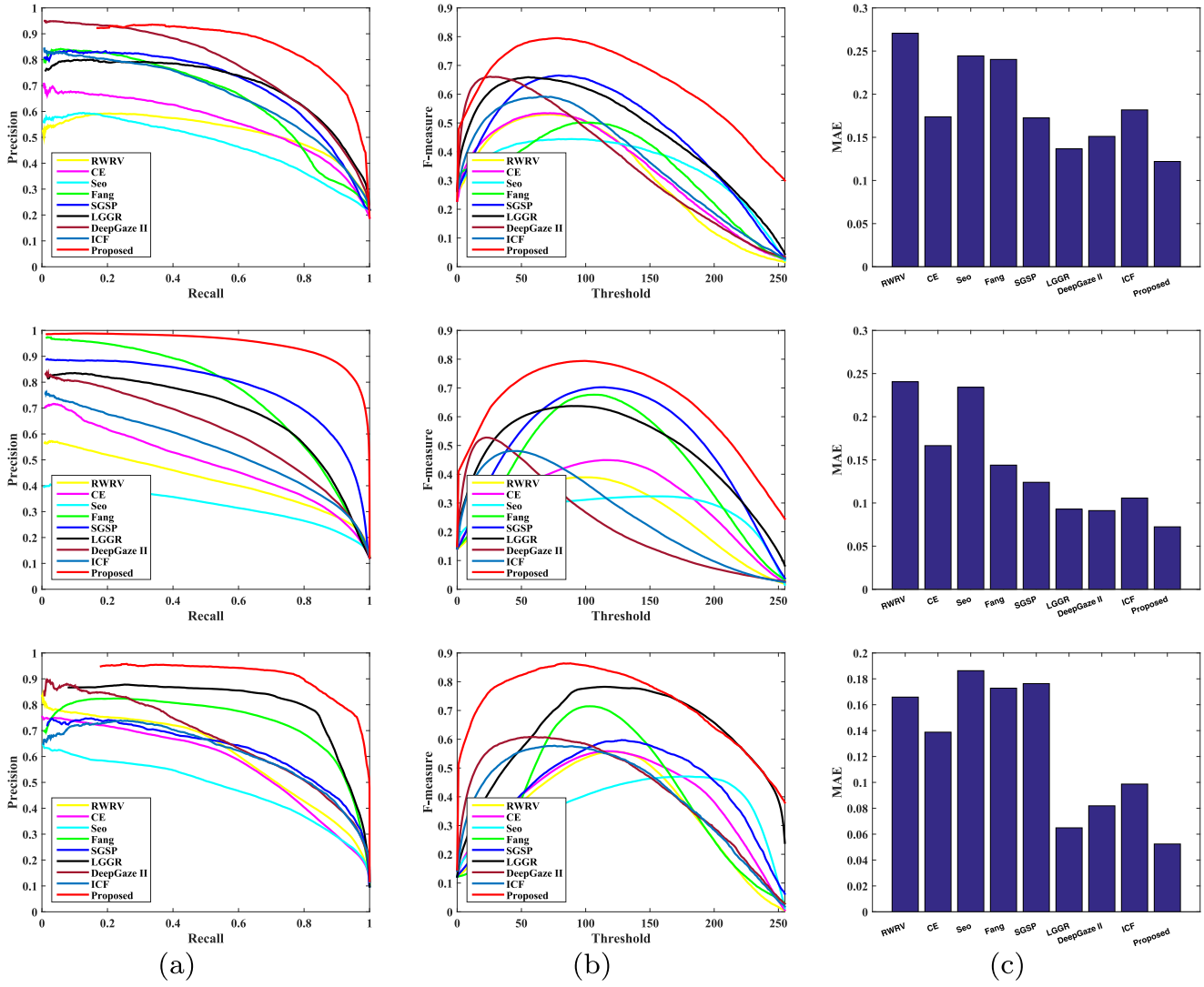

Fig. 9. Experimental results by different saliency detection methods on FBMS [66] (**top**), DAVIS [67] (**mid**), and ViSal [46] (**bottom**): (a) Precision-Recall curves; (b) F-measure; (c) average MAE.

SGSP cannot suppress residual saliency in background regions, as shown in the fifth and sixth rows of Fig. 7. LGGR combines local and global saliency cues linearly for saliency region detection. As shown in the third and fifth rows of Fig. 7, LGGR cannot accurately detect salient regions. DeepGaze II and ICF only extract static features and ignore motion feature. Compared with these existing saliency detection models, the proposed model can detect more accurate saliency results, as shown by saliency results in the last column of Fig. 7. These experimental results also demonstrate that the uncertainty weighting algorithm based on Gestalt theory can be used to highlight saliency of foreground regions and suppress saliency of background regions by fusing spatial and temporal saliency results. Some more comparison samples on Visal are given in Fig. 8. From this figure, we can also see the superiority of the proposed model over other existing saliency detection models.

3.2.2. Quantitative comparison

We provide quantitative results in Fig. 9, where PR curve, F-measure, and MAE [49] values of compared saliency detection models are provided. In this figure, PR curve, F-measure, and MAE values are all averaged over 30 test video sequences in FBMS, 50 video sequences in DAVIS, and 17 video sequences in ViSal. From these experimental results, it can be seen that the proposed method consistently outperforms other existing methods across different metrics. Please note that the computational complexity of the proposed video saliency detection method is higher than some existing related methods, since we have to extract the motion features by optical flow in the proposed method. The proposed method cannot run in real-time for common video sequences.

3.2.3. Fusion by uncertainty weighting

In the section, we conducted a comparison experiment on spatial, temporal, and spatiotemporal saliency with quantitative results, as shown in Fig. 10. From these experimental results, we can see that temporal saliency can obtain better performance than spatial saliency in most cases, while the spatiotemporal saliency consistently outperforms spatial and temporal saliency across different metrics. This demonstrates that the adaptive uncertainty weighting by the two laws of similarity and common fate in Gestalt theory can effectively fuse spatial and temporal saliency for spatiotemporal saliency prediction.

3.3. Parameter choice

In this experiment, we conducted a comparison experiment for performance evaluation with the choice of various parameters used

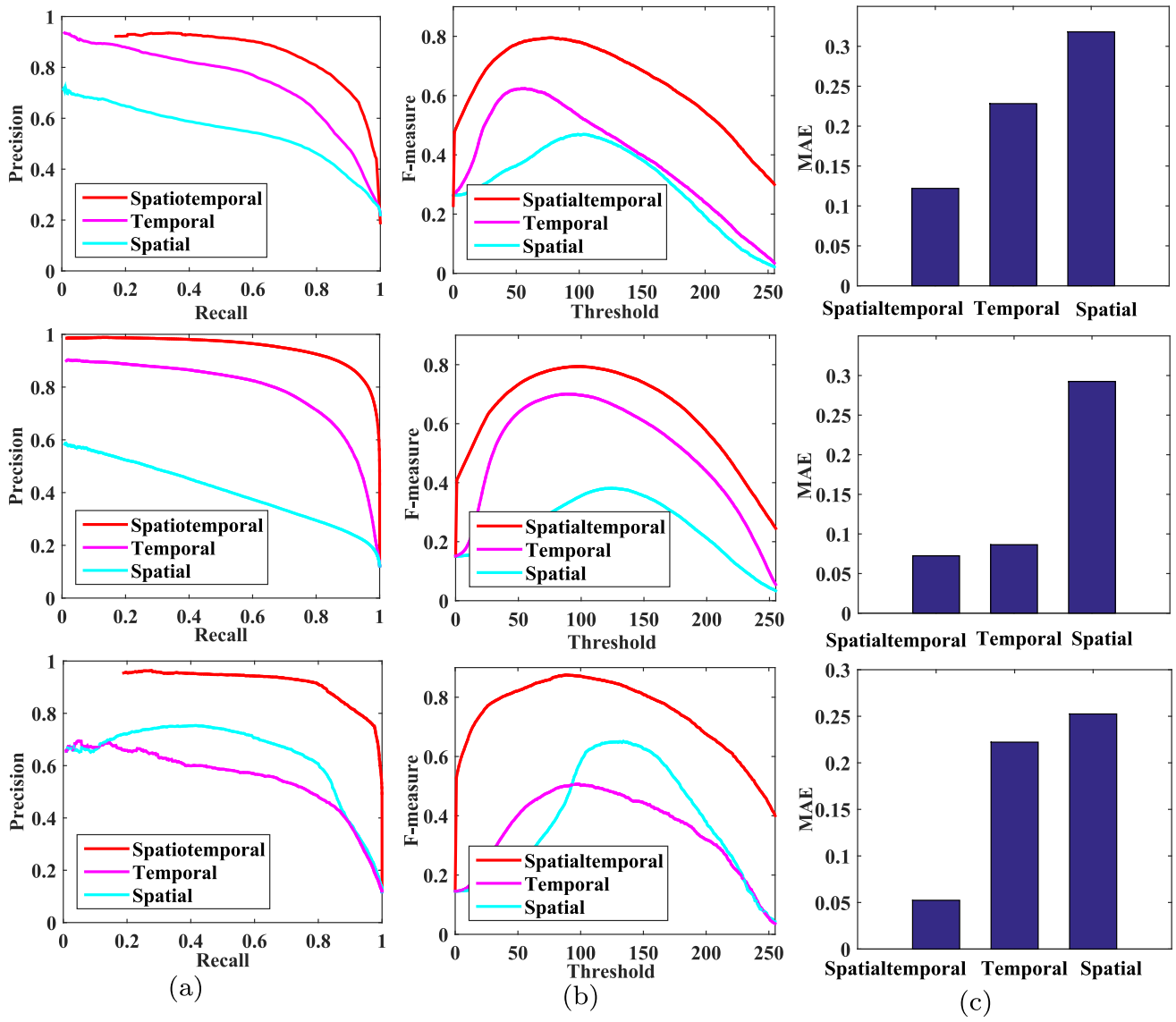

Fig. 10. Quantitative results of spatial, temporal and spatiotemporal saliency maps on FBMS (**top**), DAVIS (**mid**), and ViSal (**bottom**). (a) Precision-Recall curves; (b) F-measure; (c) average MAE.

Table 1
Experimental results on FBMS, DAVIS, and ViSal databases in terms of average MAE.

Databases	$\alpha\backslash\sigma_s$	1	3	5	7	10
FBMS	40	0.1703	0.1590	0.1450	0.1666	0.1787
	50	0.1742	0.1460	0.1343	0.1538	0.1731
	60	0.1561	0.1332	**0.1220**	0.1429	0.1698
	70	0.1595	0.1403	0.1489	0.1530	0.1601
	80	0.1756	0.1509	0.1530	0.1688	0.1704
DAVIS	40	0.1340	0.1276	0.0823	0.1153	0.1200
	50	0.1267	0.1154	0.0851	0.1089	0.1061
	60	0.1123	0.0950	**0.0724**	0.0908	0.1088
	70	0.1259	0.1062	0.0820	0.1040	0.1163
	80	0.1403	0.1284	0.1054	0.1209	0.1365
ViSal	40	0.1033	0.0850	0.0758	0.0899	0.0958
	50	0.0947	0.0844	0.0750	0.0763	0.0978
	60	0.0921	0.0676	**0.0524**	0.0690	0.0864
	70	0.1046	0.0799	0.0574	0.0764	0.0960
	80	0.1150	0.0973	0.0761	0.0942	0.1084

Table 2
Experimental results on FBMS, DAVIS, and ViSal databases in terms of average F-measure.

Databases	$\alpha\backslash\sigma_s$	1	3	5	7	10
FBMS	40	0.6031	0.6058	0.6150	0.6168	0.5920
	50	0.6105	0.6187	0.6238	0.6120	0.6066
	60	0.6150	0.6208	**0.6398**	0.6252	0.6087
	70	0.6079	0.6155	0.6299	0.6174	0.6068
	80	0.5984	0.6001	0.6148	0.6077	0.5938
DAVIS	40	0.6030	0.6160	0.6274	0.6108	0.6166
	50	0.6003	0.6246	0.6322	0.6284	0.6148
	60	0.6257	0.6329	**0.6400**	0.6309	0.6244
	70	0.6162	0.6351	0.6378	0.6252	0.6177
	80	0.6057	0.6208	0.6280	0.6154	0.6040
ViSal	40	0.7063	0.7138	0.7248	0.7222	0.7088
	50	0.7060	0.7299	0.7330	0.7290	0.7154
	60	0.7155	0.7284	**0.7441**	0.7366	0.7278
	70	0.7086	0.7168	0.7405	0.7280	0.7100
	80	0.7041	0.7073	0.7244	0.7156	0.7079

in the proposed video saliency detection model. The comparison experiment is conducted by considering two parameters including σ_s in Eq. (1) and α in Eq. (13). Experimental results are shown in Tables 1 and 2 with different σ_s and α on three databases. From these tables, we can observe that the proposed method with $\sigma_s = 5$ and $\alpha = 60$ can obtain the best performance among different parameter settings.

The computation cost of the proposed method is much higher than some existing methods due to the motion estimation processing (optical flow algorithm) involved. The most time-consuming part is the motion feature extraction by optical flow in the proposed model. We will further investigate to reduce the computational complexity.

4. Conclusion

In this work, we introduce a novel framework of video saliency detection based on spatiotemporal cues and Gestalt theory. For spatial saliency prediction, we extract spatial features including luminance, color, and texture features to compute the feature contrast. For temporal saliency prediction, we extract motion features calculated by optical flow. The spatiotemporal saliency map is calculated by fusing spatial and temporal saliency with uncertainty weighting across modalities estimated by the two laws of similarity and common fate in Gestalt theory. Experimental results show the superiority of the proposed video saliency detection model over other existing ones on three public databases.

In the future, we will improve our algorithm by extracting high-level features based on artificial neural networks. Currently, our algorithm predicts spatiotemporal saliency by extracting low-level features in video sequences. However, it is known that saliency is influenced by high-level features from top-down knowledge. We can extract low-level and high-level features (such as semantic object information like human) simultaneously to design a more effective video saliency detection model in future.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61571212 and 61822109, the Natural Science Foundation of Jiangxi Province under Grant 20181BBH80002, and the Fok Ying-Tong Education Foundation of China under Grant 161061.

References

[1] M. Carrasco, Visual attention: the past 25 years, Vision Res. 51 (13) (2011). 1484–525.
[2] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, Salient object detection in the deep learning era: an in-depth survey, CoRR abs/1904.09146 (2019).
[3] 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.
[4] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: International Conference on Neural Information Processing Systems, 2006, pp. 545–552.
[5] P. Zhang, T. Zhuo, W. Huang, K. Chen, M. Kankanhalli, Online object tracking based on CNN with spatial-temporal saliency guided sampling, Neurocomputing 257 (2017) 115–127.
[6] Y. Fang, Z. Fang, F. Yuan, Y. Yang, S. Yang, N. Xiong, Optimized multioperator image retargeting based on perceptual similarity measure, IEEE Trans. Syst. Man Cybern. 47 (11) (2017) 2956–2966.
[7] C. Aytekin, A. Losifidis, M. Gabbouj, Probabilistic saliency estimation, Pattern Recognit. 74 (1) (2018) 359–372.
[8] J. Zhang, K.A. Ehinger, H. Wei, K. Zhang, J. Yang, A novel graph-based optimization framework for salient object detection, Pattern Recognit. 64 (1) (2017) 39–50.
[9] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 1 (1) (2018). 1–1.
[10] E. Macaluso, C.D. Frith, J. Driver, Directing attention to locations and to sensory modalities: multiple levels of selective processing revealed with PET, Cerebral Cortex 12 (4) (2002) 357–368.
[11] T.S. Lee, D. Mumford, Hierarchical bayesian inference in the visual cortex, JOSA A 20 (7) (2003) 1434–1448.
[12] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1155–1162.
[13] M.S. Livingstone, D.H. Hubel, Anatomy and physiology of a color system in the primate visual cortex, J. Neurosci. Official J. Soc. Neurosci. 4 (1) (1984) 309–356.
[14] C. Yang, L. Zhang, H. Lu, X. Ruan, M. Yang, Saliency detection via graph-based manifold ranking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
[15] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2814–2821.
[16] M.M. Cheng, N.J. Mitra, X.L. Huang, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[17] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2376–2383.
[18] V. Gopalakrishnan, Y. Hu, D. Rajan, Random walks on graphs for salient object detection in images, IEEE Trans. Image Process. 19 (12) (2010) 3232–3242.
[19] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[20] 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.
[21] J. Li, M.D. Levine, X. An, X. Xu, H. He, Visual saliency based on scale-space analysis in the frequency domain, IEEE Trans. Pattern Anal. Mach. Intell. 35 (4) (2013) 996–1010.
[22] J. Li, L.Y. Duan, X. Chen, T. Huang, Y. Tian, Finding the secret of image saliency in the frequency domain, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2428–2440.
[23] C. Guo, L. Zhang, A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (1) (2010) 185–198.
[24] B. Schauerte, R. Stiefelhagen, Quaternion-based spectral saliency detection for eye fixation prediction, in: European Conference on Computer Vision, 2012, pp. 116–129.
[25] Z. Liu, W. Zou, O.L. Meur, Saliency tree: a novel saliency detection framework, IEEE Trans. Image Process. 23 (5) (2014) 1937–1952.
[26] J. Lei, B. Wang, Y. Fang, W. Lin, P.L. Callet, N. Ling, C. Hou, A universal framework for salient object detection, IEEE Trans. Multimed. 18 (9) (2016) 1783–1795.
[27] F. Perazzi, P. Krhenbhl, Y. Pritch, Saliency filters: contrast based filtering for salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.
[28] C. Li, Y. Yuan, W. Cai, Y. Xia, D. Feng, Robust saliency detection via regularized random walks ranking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2710–2717.
[29] S. Pei, W. Chang, C. Shen, Saliency detection using superpixel belief propagation, in: IEEE International Conference on Image Processing, 2014, pp. 1135–1139.
[30] D. Zhang, J. Han, C. Li, J. Wang, Co-saliency detection via looking deep and wide, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2994–3002.
[31] J. Li, Y. Tian, T. Huang, Visual saliency with statistical priors, Int. J. Comput. Vis. 107 (3) (2014) 239–253.
[32] D. Zhang, D. Meng, C. Li, A self-paced multiple-instance learning framework for co-saliency detection, in: IEEE International Conference on Computer Vision, 2015, pp. 594–602.
[33] M. Sun, A. Farhadi, B. Taskar, S. Seitz, Summarizing unconstrained videos using salient montages, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2017) 2256–2269.
[34] T. Xi, W. Zhao, H. Wang, W. Lin, Salient object detection with spatiotemporal background priors for video, IEEE Trans. Image Process. 26 (7) (2017) 3425–3436.
[35] O. Le Meur, P. Le Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Res. 47 (19) (2007) 2483–2498.
[36] Z. Liu, X. Zhang, S. Luo, O.L. Meur, Superpixel-based spatiotemporal saliency detection, IEEE Trans. Circuits Syst. Video Technol. 24 (9) (2014) 1522–1540.
[37] W. Wang, J. Shen, F. Porikli, Saliency-aware geodesic video object segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3395–3402.
[38] Z. Liu, J. Li, L. Ye, G. Sun, L. Shen, Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation, IEEE Trans. Circuits Syst. Video Technol. 27 (12) (2017) 2527–2542.
[39] V. Leboran, A. Garcia-Diaz, X.R. Fdez-Vidal, X.M. Pardo, Dynamic whitening saliency, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2017) 893–907.
[40] S.H. Lee, J.H. Kim, K.P. Choi, Video saliency detection based on spatiotemporal feature learning, in: IEEE International Conference on Image Processing, 2014, pp. 1120–1124.
[41] H. Kim, Y. Kim, J.Y. Sim, C.S. Kim, Spatiotemporal saliency detection for video sequences based on random walk with restart, IEEE Trans. Image Process. 24 (8) (2015) 2552–2564.
[42] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognit. Psychol. 12 (1) (1980) 97–136.
[43] O. Le Meur, P. Le Callet, D. Barba, A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (5) (2006) 802–817.
[44] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.
[45] C. Chen, S. Li, Y. Wang, Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion, IEEE Trans. Image Process. PP (99) (2017). 1–1.
[46] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Trans. Image Process. 24 (11) (2015) 4185–4196.

[47] M. Kummerer, T.S.A. Wallis, L.A. Gatys, M. Bethge, Understanding low- and high-level contributions to fixation prediction, in: IEEE International Conference on Computer Vision, 2017.
[48] S.S.S. Kruthiventi, K. Ayush, R.V. Babu, Deepfix: a fully convolutional neural network for predicting human eye fixations, IEEE Trans. Image Process. 26 (9) (2017) 4446–4456.
[49] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convonlutional networks, IEEE Trans. Image Process. 27 (1) (2018) 38–49.
[50] Y. Yuan, C. Li, J. Kim, W. Cai, D.D. Feng, Dense and sparse labeling with multidimensional features for saliency detection, IEEE Trans. Circuits Syst. Video Technol. 28 (5) (2018) 1130–1143.
[51] W. Wang, J. Shen, F. Guo, M.-M. Cheng, A. Borji, Revisiting video saliency: a large-scale benchmark and a new model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4894–4903.
[52] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (5) (2018) 2368–2378.
[53] W. Wang, J. Shen, X. Dong, A. Borji, Salient object detection driven by fixation prediction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720.
[54] J.C. Banerjee, Gestalt theory of perception, Encyclopaedic Dictionary of Psychological Terms, 1994.
[55] H. Stevenson, Emergence: the gestalt approach to change, Unleashing Executive Org. Potential 7 (2012).
[56] Y. Fang, Z. Wang, W. Lin, Z. Fang, Video saliency incorporating spatiotemporal cues and uncertainty weighting, IEEE Trans. Image Process. 23 (9) (2014) 3910–3921.
[57] A.A. Stocker, E.P. Simoncelli, Noise characteristics and prior expectations in human visual speed perception, Nature Neurosci. 9 (4) (2006) 578–585.
[58] Y. Zhong, A.K. Jain, Object localization using color, texture and shape, Pattern Recognit. 33 (4) (2000) 671–684.
[59] J. Wei, Color object indexing and retrieval in digital libraries, IEEE Trans. Image Process. 11 (8) (2002) 912–922.
[60] D. Sun, S. Roth, M.J. Black, Secrets of optical flow estimation and their principles, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2432–2439.
[61] Z. Wang, Q. Li, Video quality assessment using a statistical model of human visual speed perception, J. Opt. Soc. Am. A Opt. Image Sci. Vis. 24 (12) (2007) B61–B69.
[62] D.A. Poggel, H. Strasburger, M. MacKeben, Cueing attention by relative motion in the periphery of the visual field, Perception 36 (7) (2007) 955–970.
[63] B.A. Wandell, Foundations of Vision, 1995.
[64] R. Achanta, A. Shaji, K. Smith, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[65] Y. Li, Y. Zhou, J. Yan, Z. Niu, J. Yang, Visual saliency based on conditional entropy, in: Asian Conference on Computer Vision, 2009, pp. 246–257.
[66] T. Brox, J. Malik, Object segmentation by long term analysis of point trajectories, in: European Conference on Computer Vision, Springer, 2010, pp. 282–295.
[67] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 724–732.
[68] H.J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis. 9 (12) (2009) 1–27.

Yuming Fang received his Ph.D. degree from Nanyang Technological University in Singapore, M.S. degree from Beijing University of Technology in Beijing, China, and B.E. degree from Sichuan University in Chengdu, China. Currently, he is a Professor in the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He serves as an Associate Editor of IEEE Access and is on the editorial board of Signal Processing : Image Communication. He have authored and co-authored more than 100 academic papers in international journals and conferences in the areas of multimedia processing. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, 3D image/video processing.

Xiaoqiang Zhang received B.Eng. degree in computer engineering from Nanchang Institute of Technology, Nanchang, China, 2016. He is currently a Master student in the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. His reserach interests include computer vision and machine learning.

Feiniu Yuan received B.Eng. and M.E. degrees in mechanical engineering from the Hefei University of Technology, Hefei, China, in 1998 and 2001, respectively, and a Ph.D. degree in pattern recognition and intelligence system from the University of Science and Technology of China (USTC), Hefei, in 2004. From 2004 to 2006, he worked as a post-doctorate with USTC. From 2010 to 2012, he was a Senior Research Fellow with Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore. He is currently a full professor at the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China. His research interests include 3D modeling, image processing and pattern recognition.

Nevrez Imamoglu received the Ph.D. degree in Chiba University, Japan, in 2015. He is currently a researcher in National Institute of Advanced Industrial Science and Technology (AIST), Japan. Previously, he was a JSPS foreign Postdoctoral Fellow in RIKEN Brain Science Institute, Japan. He published over 20 adacemic papers in international journals and conferences in the areas of multimedia processing. His research interests includes image processing and pattern recognition.

Haiwen Liu received the B.S. degree in electronic system and M.S. degree in radio physics from Wuhan University, Wuhan, China, in 1997 and 2000, respectively, and the Ph.D. degree in microwave engineering from Shanghai Jiao Tong University, Shanghai, China, in 2004. From 2004 to 2006, he was a Research Assistant Professor with Waseda University, Kitakyushu, Japan. From 2006 to 2007, he was a Research Fellow with Kiel University, Kiel, Germany, where he was granted the Alexandervon Humboldt Research Fellowship. From 2007 to 2008, he was a Professor with the Institute of Optics and Electronics, Chengdu, China, where he was supported by the 100 Talents Program of Chinese Academy of Sciences, Beijing, China. Currently, he is a full Professor with Xi'an Jiaotong University, Xi'an, China. He has authored more than 100 papers in international and domestic journals and conferences. Dr. Liu is the Associate Editor of the International Journal of RF and Microwave Computer-Aided Engineering and Leading Guest Editor of the International Journal of Antennas and Propagation. He has served as a Technical Program Committee member for many international conferences.