Pattern Recognition Letters 132 (2020) 76–83

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Spatial–temporal multi-task learning for salient region detection

Zhe Chen[a], Ruili Wang[b,*], Ming Yu[c], Hongmin Gao[a], Qi Li[a], Huibin Wang[a]

[a] College of Computer and Information, Hohai University, Nanjing 210098, China
[b] Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand
[c] School of Computer Science and Engineering, Hebei University of Technology, Tianjin, China

ARTICLE INFO

Article history:
Available online 17 October 2018

ABSTRACT

This paper proposes a novel multi-task learning based salient region detection method by fusing spatial and temporal features. Salient region detection has been widely used in various computer vision tasks, being as a general preprocessor to identify interest objects. Despite the recent successes, existing saliency models still lag behind the performance of human when visually perceives dynamic scenes. Most of the existing models largely rely on various spatial features. However, these spatial feature based methods have several deficiencies: (i) they can hardly adapt to the situation where moving objects are included, and (ii) they cannot model the human vision process in dynamic scenes. Recently, some saliency models introduce temporal features in their detecting process, such as the optical flow and stacking frames. The potential of temporal features for saliency optimization has been demonstrated. However, since temporal features in these models are merely used as a compensation to static features, the advantages of temporal features have not yet been fully explored. Aiming to comprehensively address these issues above, our method fuses spatial and temporal features, and learns the mapping relationship from various features to salient regions using our multi-task learning framework. The final salient region is generated by our unified Bayesian framework. The experimental results demonstrated that our proposed approach outperforms previous methods.

1. Introduction

Salient region detection is an important stage for many computer vision tasks, since it can efficiently and initially identify the "important" regions in images and videos. This condensed information is then processed by high-level processors according to various purposes such as object detection [1,2], tracking [3], and image classification [4]. Generally, the aim of salient region detection is to identify the "pop-out" of scenes. This concept is different from the attention point detection which intends to predict the separate fixation point and eye movement while human watches a new scene.

Previously, salient region detection methods are based on various spatial features. For example, color, intensity and orientation features are introduced in the Itti models [5]. This kind of models helps us to understand our visual sensitivity to still natural scenes. Later, various high-level spatial features are employed for salient region detection. For example, the performances of color contrast and contour features have been demonstrated for salient region detection in complicated conditions. However, although significant successes have been achieved by these spatial feature based methods, the disadvantage of them lies in their inadaptability to dynamic scenes. In dynamic scenes, human visual attention mainly focuses on the motion changes which contain more semantic contents, and the corresponding visual saliency consequentially locates on the regions where motions happened. This type of motion saliency is originated from temporal features and cannot be correctly identified by spatial features. Aiming to address this issue, many models introduce the temporal features as a cue to detect salient motions such as the optical flow [6], frame difference [7] and frame stacks [8]. In these methods, temporal and spatial features are jointly inputted into a unified model and the temporal features are considered as a compensation to the spatial/static features. However, the temporal factor has not yet been fully explored for salient region detection.

Another problem for existing salient region detection methods is the large gap between mathematical models and biological vision mechanisms. The primates, as we known, have a strong ability to learn and store experiences in memory. For humans, experience plays a significantly important role for salient region detection, forming an experience-driven saliency [9,10]. For example, we habitually deploy our visual saliency on the texts and the faces appearing in scenes. This kind of salient regions cannot be well identified by any original image information, and the historic experi-

* Corresponding author.
E-mail address: ruili.wang@massey.ac.nz (R. Wang).

https://doi.org/10.1016/j.patrec.2018.10.019
0167-8655/

ence should be learned and considered. Aiming to mathematically establish the mapping relationship between the historic experience and the salient region deployment, various learning method has been proposed. Also, some top-down visual features, such as the object recognition, face detection results, are extracted as a cue for saliency estimation [11,12]. Excellent performance of these learning based methods proves the importance of the experiential cues for salient region detection. However, most of these learning models are trained on the static/spatial features in still images, while the temporal information between frames is overlooked by most of the existing methods [13].

Inspired by the learning based salient region detection strategy and considering the importance of temporal features [14], this paper proposes a novel spatial-temporal multi-task learning based salient region detection method. Two-stream tasks are respectively established with spatial and temporal features. The saliency probabilities driven by spatial and temporal features are jointly calculated under a unified Bayesian framework. The contribution of this paper is two-fold.

(i) Temporal experience learning. In order to fully explore the importance of the temporal factor for salient region detection, various temporal features are introduced in the learning processes, such as the optical flow and the temporal frequency variation. By using these features, our model can learn how motion changes arouse our visual attention and in turn result in the salient region deployment. This strategy can generalize our salient region detection method to complicated dynamic scenes.
(ii) Bayesian framework based multi-task learning framework. Aiming to normalize and fuse the determination results from multiple tasks, a Bayesian framework is proposed to generate the final salient region detection result.

2. Related works

Our approach aims to provide a salient region detection method by jointly learning spatial and temporal features. Thus, spatial and temporal feature based methods are the most related topics to this paper.

Generally, the spatial feature based salient region detection methods are based on local or global variations between pixels. A typical local contrast based method is achieved by the Itti model which defines the saliency by the central-surrounded contrast in multiple scales, and the final results are provided by contrast combination [5]. Later, the fuzzy model is used to modify the local contrast calculation, significantly enhancing the relaxation of the Itti model [15]. The Itti model is further optimized by a contrast normalization method. This method is efficient and can likely cater for the online application after optimization [16].

Moreover, the pyramid structure is used for fusing multi-scaled contrasts, which theoretically can achieve a good performance in depicting both the detail and regional features [17]. Generally, the advantage of the local contrast lies in its sensitivity to point-point changes in local regions, and a good performance for detecting salient edges and contours. However, according to human visual perceptions, the visual system is more sensitive to the global and regional contents. Hence, the local contrast based method in some cases cannot provide satisfactory results. Aiming to solve this problem, the global information is increasingly used in the saliency detection models. Different from the local feature, global feature based methods identify the saliency of a point by its contrast against all the points in the entire image. Recently, the global information is considered as one of the features in the context-aware saliency detection models [18]. In this method, multiple cues, such as the local low-level clues, global considerations, visual organization rules and high-level features are simultaneously modeled to highlight salient objects attached with the scene context. Beside this work, there are many salient region detection models completely built on the global contrast. For example, the pixel-level global saliency can be detected by comparing a pixel to all the pixels including in an image. However, for the case of efficiency, the pixel-level calculation is only achieved on the intensity information thus ignoring many informative cues in other channels [19]. Aiming to solve this problem, the color histogram is used for calculating point-to-point feature contrast. Moreover, regional contrasts in this work are extracted to remain the uniform in salient regions [1]. Also, there are many spectrum frequency based salient region detection methods proposed recently [7,20,21]. In principle, the calculation of spectrum frequency can highlight the global saliency over images. However, according to numerous evaluations on these methods, the local image variations are likely mistaken as the saliency factor if the image is not resized before.

In contrast to vast spatial feature based saliency/salient region detection methods, fewer methods are related to the saliency/salient region detection in dynamic scenes, and most of them are just deemed as the extension to typical spatial feature based methods. For example, the motion features are inputted into the updated Itti model to simulate the motion sensitivity of our vision system [22–24]. Other mathematical methods rather than modeling a visual system just provide results functionally similar to that given by our brain. Some of them are free of features and complex models and only depend on temporal spectrum analysis. For example, frame-difference features are analyzed in the phase spectrum space, establishing a phase quaternion Fourier transform (PQFT) based method [7]. However, the frame-difference feature is susceptible to the motion noises and camera shaking. This problem causes a high false alarm rate in detected salient regions. Moreover, the temporal feature in the PQFT model is just considered as a compensation to the spatial features which mainly control the salient region detection process [7]. By building the time slice, the spectral residual, which is initially used in still images, is operated on the video sequence, making the TSR model for motion saliency detection [25]. However, it is mentioned that the reason for using the amplitude spectrum to detect the saliency is questionable for its lower computational efficiency. By stacking the pixels in the same point of the temporal sequence, the TFT method relates the variation in the phase spectrum with the temporal motion, identifying the region of the motion saliency by the amplitude of the phase spectrum [26].

Moreover, there are many machine learning based salient region detection methods. Various sources, including the low /high-level visual features and bottom-up/top-down features, are extracted for training the saliency model [17,27–30]. Generally, the learning based strategy is functionally closed to our visual perceptions, that the salient region is selected according to our experience. Aiming to define this experience-driven saliency, some well-performed classifiers such as the support vector machine (SVM) and the artificial neural network (ANN) have been trained with a small database [12]. Due to the usage of task-depended cues, these models are specialized to any tasks while their generalization is seriously limited. The deep learning structure has an excellent ability to solve this issue. Various deep networks have been used for salient region detection [28–30]. For these deep learning methods, a large training database is necessary. However, up to date, no labeled database has been generated for salient region identification, which seriously blocks the deep saliency detection.

Theoretically, the spatial/temporal feature based salient region detection methods identify the salient region from two separated aspects: variations between pixels or frames. In some cases, saliency deployment largely depends on the spatial information, while the saliency in other cases mainly relies on the temporal

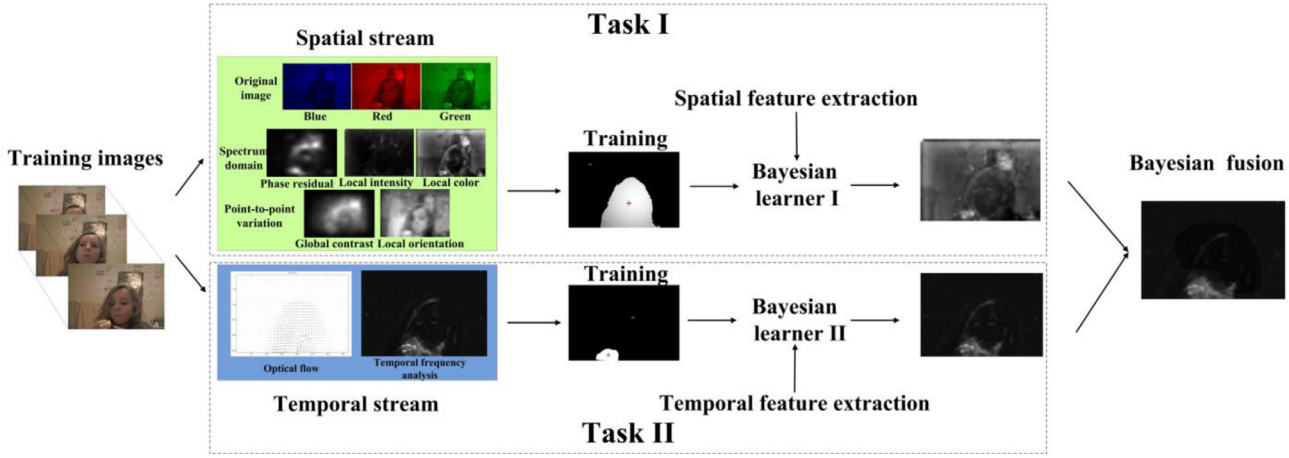

Fig. 1. Flowchart of our salient region detection method. It mainly consists of two streams respectively, corresponding to spatial/temporal tasks. The final comprehensive result is generated by the Bayesian fusion.

information. Aiming to comprehensively take advantage both of these two features, we propose a two-stream learning model for salient region detection. Moreover, in order to improve the model generalization and avoid the overfitting problem, the multi-task mechanism is used here and underlies our spatial–temporal multi-task learning based salient region detection method.

3. Proposed method

3.1. Framework of our proposed method

According to our previous discussions, learning based methods have a better ability to model our visual attention process, especially in dynamic scenes. In this paper, task-independent features are employed for salient region detection. Two Bayesian estimation models are proposed for estimating the probability of the saliency, one is for processing the spatial features and other for the temporal features. A multitask learning mechanism is used to form a two-stream system. The overall method is established on a uniform Bayesian mechanism which jointly fuses the results given by spatial and temporal features. The framework of our method is illustrated in Fig. 1. Various spatial/temporal features are jointly extracted to train and test the Bayesian probabilistic model. Outputs of multi-task processes are fused under a unified Bayesian framework, generating a comprehensive determination result for salient region detection.

3.2. Bayesian learning model for salient region detection

In contrast to the salient region detection in still images, the saliency identification in our method is based on the historical experience. In order to learn a salient region detection model, a Bayesian model is employed here, inspired by the work of [31]. For estimating the saliency probability, the mapping relationship between the saliency/background and their features should be carefully defined. In our method, the features and locations of any pixels are the two determinative factors for saliency determination. Let s be the label of the image saliency at the location $\mathbf{x}=(x, y)$, the corresponding feature in this location is $\mathbf{f}$. Saliency appears in the condition of $s=1$, otherwise default values $s=0$ belong to the visually unattractive pixels. The probability of the pixel $\mathbf{x}$ which is salient can be formed as:

$$P(s|\mathbf{f},\mathbf{x}) = \frac{P(\mathbf{f}|s)P(s|\mathbf{x})}{P(\mathbf{f}|\mathbf{x})} \tag{1}$$

Since the image feature $\mathbf{f}$ is independent of the location $\mathbf{x}$, $P(\mathbf{f}|\mathbf{x})=P(\mathbf{f})$. Hence:

$$\frac{P(\mathbf{f}|s)P(s|\mathbf{x})}{P(\mathbf{f})} = \frac{P(s|\mathbf{f})P(s|\mathbf{x})}{P(s)} \tag{2}$$

Since the previous probabilities for the saliency s are equal, hence:

$$P(s|\mathbf{f},\mathbf{x}) = \frac{P(s|\mathbf{f})P(s|\mathbf{x})}{P(s)} \propto P(s|\mathbf{f})P(s|\mathbf{x}) \tag{3}$$

The function of Eq. (3) keeps the correspondence to the discipline mentioned above, that the saliency is jointly controlled by the pixel feature ($P(s|\mathbf{f})$) and its location ($P(s|\mathbf{x})$).

We use the kernel measurement to calculate the $P(s|\mathbf{f})$, as:

$$P(s|\mathbf{f}) = 1 - P(\mathbf{f}|\psi) = 1 - \frac{1}{N}\sum_{i=1}^{N} K(\mathbf{f}-\psi_i) \tag{4}$$

where ψ_i is the learned historic salient samples (features) which arouse our visual saliency and N is the number of the samples in the vector $\boldsymbol{\psi}$. Here, these samples are given by the labeled training data.

Specially, the Gaussian kernel is applied here for d-dimension image features, as:

$$\begin{aligned} P(s|\mathbf{f}) &= 1 - \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{d} K_{\delta_j}\left(f_j - \psi_{i_j}\right) \\ &= 1 - \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\delta_j^2}} \exp\left(-\frac{1}{2}\frac{(f_j-\psi_{i_j})^2}{\delta_j^2}\right) \end{aligned} \tag{5}$$

where K_{δ_j} is the kernel function with a bandwidth δ_j^2 in the jth dimension of the image feature space.

$P(s|\mathbf{x})$ is scaled by the measurement of Euclidean distance, as:

$$\begin{aligned} P(s|\mathbf{x}) &= 1 - \exp\left(d(\mathbf{x},\mathbf{x_o})\right) \\ &= 1 - \exp\left(\sqrt{(x-x_o)^2 + (y-y_o)^2}\right) \end{aligned} \tag{6}$$

where $d(\mathbf{x}, \mathbf{x_o})$ is the Euclidean distance between the pixel $\mathbf{x}$ and the salient pixel $\mathbf{x_o}$. The initial position of the salient pixel $\mathbf{x_o}$ can be identified with a supervised or unsupervised approach. Using the supervised approach, the start points for the saliency detection are manually selected in a pop-out area in an image, while the unsupervised approach can be realized by automatic image segmentation or clustering methods [32]. This measurement keeps the correspondence to our visual center-surround mechanism that more

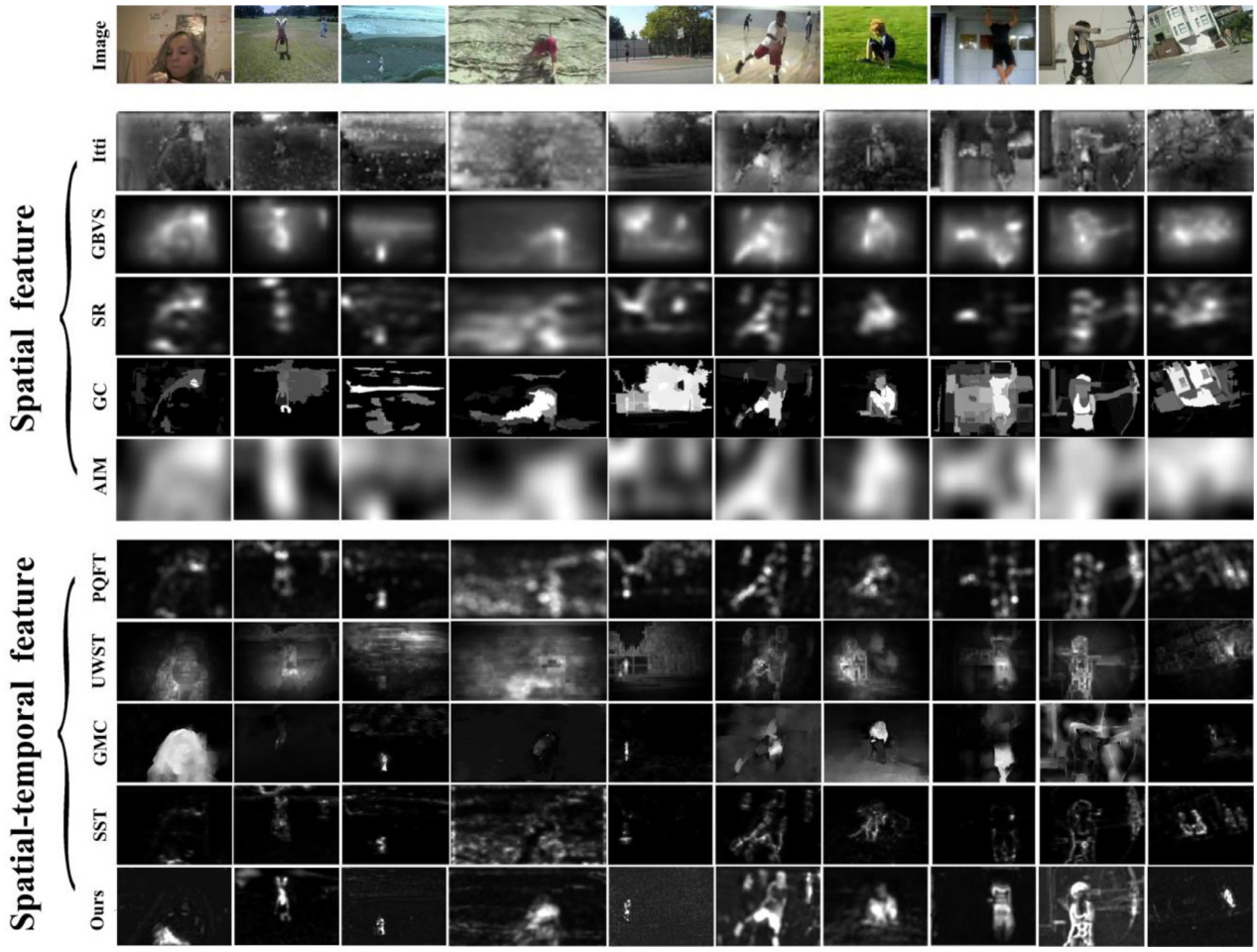

Fig. 2. Qualitative comparison of saliency maps. First row: original images; Second–sixth rows: five typical methods using spatial features; Seventh–last rows: four typical methods jointly using spatial–temporal features and our proposed method.

salient values are likely given to the pixels closer to the salient pixels.

4. Spatial-temporal feature extraction

Various spatial–temporal features are jointly extracted and fused for our salient region detection. Among them, some are given directly from the original image information and some are extracted with variation detectors.

4.1. Spatial feature

The original image information and high-level spatial features are extracted here. Respected to the original image information, the colors (red, green and blue channels), the intensity and the orientation of image pixels are extracted. Moreover, high-level features in our method include the local sensitivity features given by the Itti model [5] and the global contrast features given by the GC model [1]. Also, the global features are also introduced by the phase analysis method [21]. Overall, the spatial features are listed below:

- 1-dimensional intensity and 2-dimensional color channel contrasts (red/green and blue/yellow) are extracted.
- 3-dimensional local variation features are detected using the Itti model, respected to the intensity, color, and the orientation.
- 1-dimensional global contrast is calculated by the pixel-to-pixel distance measurement.
- 1-dimensional spectrum residual feature is calculated by the spectrum analysis method.

This results in 8-dimensional spatial features. Note that this combined spatial feature can identify the spatially salient region from diverse aspects. Here, we do not introduce semantic image features, such as image detection or object recognition features. The reason is that although the semantic features can significantly improve the salient region detection in special tasks, they will largely reduce the model generalization and might transfer our method into a task-dependent type. This does not obey the aim of our method proposed in this paper.

4.2. Temporal feature

Different from the spatial features of still images, temporal features present the variations between frames. Biological discoveries have demonstrated the importance of the temporal variations for attracting our visual saliency [33,34]. These temporal visual variations are perceived in the first level of our vision system, which is also the organ simultaneously processing spatial features. Hence, temporal features theoretically should be simple in structure and feasible for efficient calculations. Generally, temporal features can be classified into two categories, as the short-term and long-term temporal features. For the short-term temporal feature, they are to detect the variations between two consecutive frames, while the long-term features present the changes during a long period (such

as 200 frames). Both of these two features can arouse our visual saliency, controlled by our instant and long memory mechanisms.

Our method jointly extracts these two categories of temporal features. The short-term temporal feature is extracted by the optical flow between two frames, while the temporal spectrum analysis method is used to extract the long-term temporal feature. Overall, the temporal feature we introduced as follows:

- 2-dimensional optical flow vector is extracted, corresponding to the movement in the u and v directions respectively.
- 1-dimensional long-term variation is extracted by the temporal spectrum analysis.

This results in 3-dimensional spatial features. The optical flow vector is automatically extracted by the fast flow model [35]. The long-term variation is calculated by our previously proposed temporal Fourier transform (TFT) method [26]. In the TFT method, the temporal slice for the temporal Fourier transform is built by randomly selected historic 15 frames and the current frame. Hence, a time slice with 16 frames is operated as an element for our temporal spectrum analysis. To avoid the redundancy between the long-term and short-term features, the time slice in TFT method does not contain the frame temporally closed to the current time step.

5. Spatial-temporal feature extraction

In some cases, a redundancy likely exists between the spatial and temporal features. For example, the temporal variations are more significant in textural transmission regions. Hence, if we input the spatial and temporal features into a single model, an overfitting effect will appear during the training process. In order to prevent this problem, a multi-task learning framework is proposed in our method. Two-stream models are respectively trained on spatial and temporal features. The outputs of the two-stream models are fused under a Bayesian framework. The reason for using a Bayesian framework rather than a simple likelihood is that the strong correlation between the spatial and temporal features will lead a strong correlation between the conditional probability with the spatial features $P(s|\mathbf{f}_s, \mathbf{x})$ and the temporal features $P(s|\mathbf{f}_t, \mathbf{x})$. This will leads the likelihood calculation results questionable.

The Bayesian framework mathematically forms the saliency labels under the condition of the pixel location, spatial and temporal features, as:

$$P(s|\mathbf{f}_s, \mathbf{f}_t, \mathbf{x}) = \frac{P(\mathbf{f}_s|s, \mathbf{x})P(s|\mathbf{f}_t, \mathbf{x})}{P(\mathbf{f}_s|\mathbf{f}_t, \mathbf{x}))} \tag{7}$$

where $\mathbf{f}_s$ and $\mathbf{f}_t$ are the spatial and temporal features, $P(s|\mathbf{f}_t, \mathbf{x})$ is given by Eq. (5) combined with the temporal features $\mathbf{f}_t$. The first term of numerator can be calculated, as:

$$P(\mathbf{f}_s|s, \mathbf{x}) = \frac{P(s|\mathbf{f}_s, \mathbf{x})P(\mathbf{f}_s|\mathbf{x})}{P(s|\mathbf{x})} = \frac{P(s|\mathbf{f}_s, \mathbf{x})P(\mathbf{f}_s)}{P(s|\mathbf{x})} \tag{8}$$

where $P(s|\mathbf{f}_s, \mathbf{x})$ is given by Eq. (5) combined with the spatial features $\mathbf{f}_s$, and $P(s|\mathbf{x})$ is given by Eq. (6). $P(\mathbf{f}_s)$ is the probability of the spatial features $\mathbf{f}_s$, which can be conventionally estimated by the feature histogram.

The normalization factor in Eq. (7) can be calculated as:

$$P(\mathbf{f}_s|\mathbf{f}_t, \mathbf{x}) = P(\mathbf{f}_s|\mathbf{f}_t)P(\mathbf{f}_t|\mathbf{x}) = P(\mathbf{f}_s|\mathbf{f}_t)P(\mathbf{f}_t) = P(\mathbf{f}_s, \mathbf{f}_t) \tag{9}$$

where $P(\mathbf{f}_s, \mathbf{f}_t)$ is the probability of occurrence of spatial features $\mathbf{f}_s$ and temporal features $\mathbf{f}_t$.

Using the Eqs. (7) and (8), the probability of the pixel $\mathbf{x}$ being salient can be calculated.

Finally, the salient regions are detected by comprehensively considering the spatial, temporal features and the pixel location, as:

$$P(s|\mathbf{f}_s, \mathbf{f}_t, \mathbf{x}) = \begin{cases} P(s|\mathbf{f}_s, \mathbf{f}_t, \mathbf{x}) & if\ P(s|\mathbf{f}_s, \mathbf{f}_t, \mathbf{x}) > \mathrm{T} \\ 0 & else \end{cases} \tag{10}$$

where T is the threshold which can remove the spatial and temporal noises. In our work, the threshold T is automatically determined by the average of $P(s|\mathbf{f}_s, \mathbf{f}_t, \mathbf{x})$.

6. Experimental results and analysis

6.1. Experimental setup

This section presents a comprehensive evaluation of our method and the performance of our method is demonstrated by comparing to several existing methods. For salient region detection, compared methods include the typical spatial feature based methods, such as the Itti [5], global contrast (GC) [1], graph (GBVS) based methods [16] and the attention detection method based on information maximization (AIM) [36]. Also, the spectral residual based (SR) method being as a typical spectrum analysis based method is selected [21]. Moreover, we compare our method to the methods which jointly extract the spatial and temporal features to identify the saliency, such as the PQFT [7], the spatiotemporal cues and uncertainty weighting based method (UWST) [37], global motion and contrast based method (GMC) [38] and the static and space-time visual saliency based method (SST) [39]. Moreover, we apply our salient region detection method for object segmentation tasks. Our method based segmentation results are quantitatively and qualitatively compared to five typical background learning based methods, such as the spatial–temporal mixture of Gaussian model (ST-MoG) [40], fast principal component pursuit (FPCP) [41], Grassmannian robust adaptive subspace tracking (GRASTA) [42], robust orthonormal subspace learning (ROSL) [43] and motion-assisted matrix restoration (MAMR) [44] methods.

In each of these experiments, we keep the resolution of the input frames as the original resolution of the frames themselves. All of the tests were run using MATLAB 2013a on a Windows platform. The PC was equipped with a core 2.4G and 4G of memory. Since there is no existing published database is available to train and test our salient region detection model, we establish a database for our research. Our video sources are selected from the famous Hmdb-51 database [45], totally 300 frames were labeled by 20 volunteers and the average results were deemed as the ground truth.

6.2. Evaluation metrics

For salient region detection, the ROC curve is used to evaluate the performance of our and compared methods [46]. The abscissa and vertical axes of the ROC maps plot the false positive rate (FPR) and true positive rate (TPR) as follows:

$$\mathrm{FPR} = \frac{f_p}{f_p + t_n}, \qquad \mathrm{TPR} = \frac{t_p}{t_p + f_n} \tag{11}$$

where t_p, t_n, f_p, and f_n denote the numbers of the true positive, true negative, false positive, and false negative, respectively.

For object segmentation, the performance is evaluated with respect to the overlap ratio C and six criteria [47] i.e., the precision (Pr), similarity (Sim), true positive rate (TPR), F-score (FS), false positive rate (FPR), percentage of wrong classifications (PWC).

$$\begin{aligned} C &= \frac{\Omega' \bigcap \Omega}{\Omega' \bigcup \Omega} \\ \mathrm{Pr} &= \frac{t_p}{t_p + f_t}, \quad \mathrm{TPR} = \frac{t_p}{t_p + f_n}, \quad \mathrm{Fs} = 2 \times \frac{\mathrm{Pr} \times \mathrm{TPR}}{\mathrm{Pr} \times \mathrm{TPR}} \\ \mathrm{Sim} &= \frac{t_p}{t_p + f_p + f_n}, \quad \mathrm{FPR} = \frac{f_p}{f_p + t_n}, \\ \mathrm{PWC} &= 100 \times \frac{f_n + f_p}{t_p + t_n + f_p + f_n} \end{aligned} \tag{12}$$

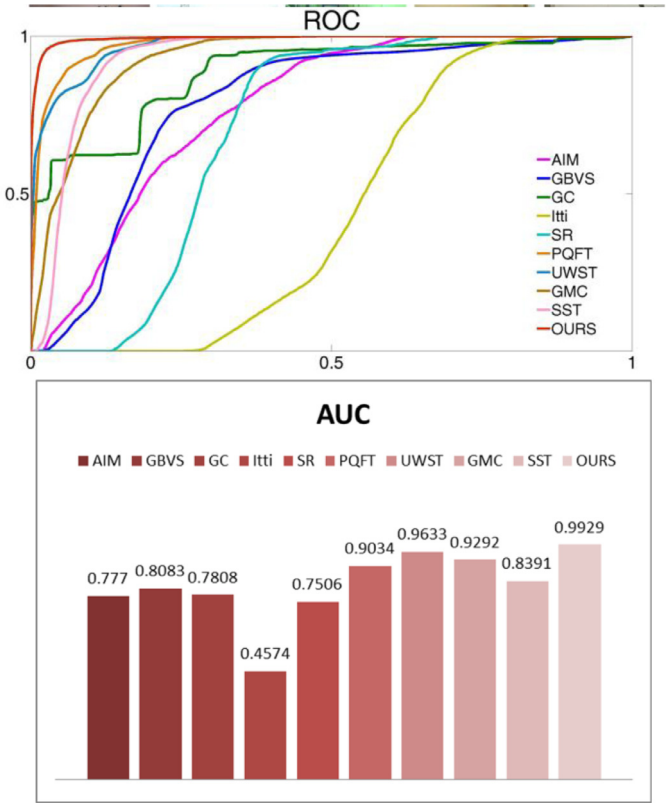

Fig. 3. Quantitative evaluation of saliency maps.

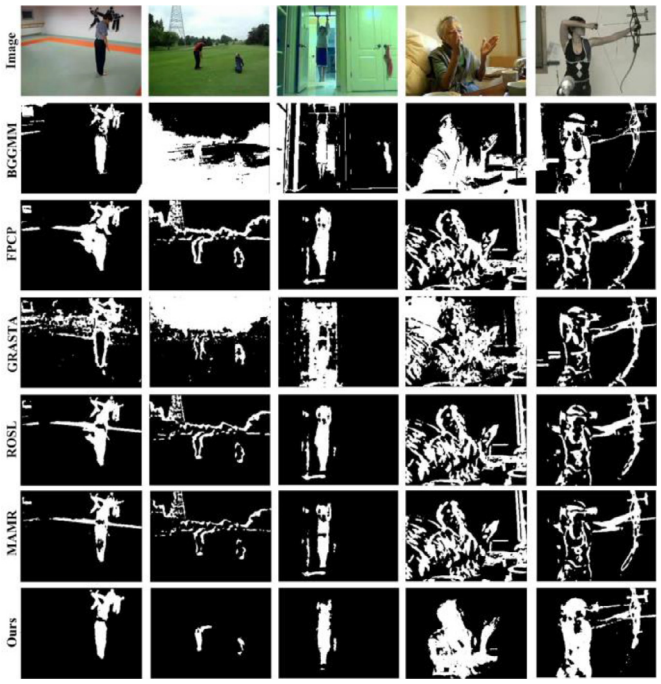

Fig. 4. Qualitative comparison of object segmentation results. First row: original images; Second–fifth rows: five typical background learning based object segmentation methods; last rows: our salient region based object segmentation method.

where Ω' is the detected results, Ω is the ground-truth, Ω'_e is the region of the segmented object area, Ω_o is the ground-truth region of the moving object to be detected, and Ω_b is the background region in the ground truth

6.3. Qualitative performance comparison

Fig. 2 shows the salient region detection results provided by five spatial feature and four spatial-temporal feature based methods and our proposed method. From the second to the sixth rows of Fig. 2, we present the spatial feature based salient region detection results given by five typical methods [5,1,16, 36,21], while the seventh to the tenth rows show four types of spatial–temporal feature based salient region detection results [7,37–39], and our results are presented in the bottom of Fig. 2. All the parameters in compared methods are kept the same to their original edition provided by authors. From these experimental comparisons, our method is able to obtain satisfying salient region detection results. Generally, the performances of spatial-temporal feature based methods are better than those only based on spatial features. The reason is that in the dynamic scenes the motion contents contain more semantic features which will attract our attention, while the contrast in still images will be visually inhibited. By introducing the temporal features, the variations between frames can be extracted and the saliency will focus on the motion changes. This is consistent with our visual perceptions. Moreover, our method also wins over other spatial–temporal feature based methods, achieving the best performance. This is attributed to the usage of the multi-task learning mechanism in our method. By multi-task learning on the training data, the mapping relation between salient regions and image features can be established. This relation indicates the historic experiences and underlies an optimal fusion strategy between the spatial and temporal features. However, the spatial and temporal features, in previous works, are mostly fused with an inflexible form which is somewhat ill-founded and cannot fully explore the contributions of various features.

6.4. Quantitative performance comparison

Fig. 3 shows the ROC curves and AUC of experimental methods over Hmdb-51 database. These results demonstrate that our method consistently outperforms the other methods. Moreover, the importance of the temporal feature can be demonstrated by the higher AUC scores in contrast to the counterparts given by only the spatial features. By the way, we observe a significant step effect of the ROC curves for some saliency models. This effect is caused by the histogram based strategy which was used to accelerate the pixel-to-pixel contrast calculation.

6.5. Application to salient object detection

Practically, salient region detection methods are commonly used as a preprocessor to initially remove the uninteresting and redundant background information, which can increase the efficiency and accuracy of the final results. A typical sample is the salient region detection based object segmentation task.

In order to fairly demonstrate the contribution of our method to salient object segmentation, a simple OTSU method is used here for thresholding our salient region detection [48]. The salient object detection results are shown in Fig. 4. From these results, we can see that the best performance is obtained by our method. Theoretically, all the compared benchmarks all have an ability to remove the background and dynamic noises. However, for the sake of the model scale and computation efficiency, only short-term unlabeled samples are used to train these compared models. As a result, these methods can hardly deal with the issues caused by the dynamic background. Recall that our salient region detection method randomly selects labeled samples in a long-term period. Hence, the background distribution can be better held by our method, underlying our good performance for removing background information. Moreover, this ability is enhanced by our

Table 1
Average performance comparison of ST–MoG, FPCP, GRASTA, MAMR, RPSL and our method.

Method	$\bar{C}$	Pr	TPR	Fs	Sim	FPR	PWC
ST-MoG	0.3936	0.4557	07,713	0.4953	0.3629	0.1612	18.2729
FPCP	0.4760	0.5170	0.7282	0.5796	0.4240	0.0962	12.6125
GRASTA	0.3500	0.4057	0.6247	0.4474	0. 3042	0.1637	19.6115
MAMR	0.5250	0.5974	0.6562	0.6074	0.4256	0.0685	11.1278
RPSL	0.5007	0.5542	0.7065	0.5969	0.4421	0.0821	11.7800
Ours	0.7418	0.8708	0.8070	0.8324	0.7182	0.0168	3.8793

multi-task learning strategy where all pixels in the background are labeled with "−1", no matter whether they are temporally changed.

Table 1 shows the quantitative evaluation of the object segmentation results obtained by compared methods and our method. According to this comparison, our method achieves the best scores in all criteria.

7. Conclusions

Our method learns a two-stream Bayesian model by integrating spatial and temporal features in a unified multi-task learning framework. Our method is able to adapt to various features such as the low-level image information and high-level spatial and temporal semantic features. These features can jointly model the informative cues that arousing our visual attention in dynamic scenes. The usage of the learning strategy can introduce the historic experience into our proposed model, thus generate experience-driven salient region detection results. This contributes to the improvement of salient region detection. The application of our method is demonstrated by the salient object segmentation task wherein our salient region detection method is introduced as the preprocessor.

Theoretically, according to the updating views, a deeper learning strategy may more completely explore the image features. However, the reason for giving up the deep learning framework in our method is two folded. Firstly, a nice deep learning method must depend on a big database. However, up to date, such a big database has not yet been completed. Secondly, a Bayesian pattern can be sufficient to model spatial and temporal features. In our method, all low-level and high-level features are formed as simple normalized maps which can be feasibly held even by economic Bayesian models.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (Nos. 61671201, 51709083, 61501173), the Natural Science Foundation of Jiangsu Province (No. BK20150824), the Fundamental Research Funds for the Central Universities(No. 2017B01914), the Jiangsu Overseas Scholar Program for University Prominent Young & Middle-aged Teachers and Presidents, and the Marsden Fund, New Zealand.

References

[1] M.M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S.M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[2] J. Feng, Y. Wei, L. Tao, C. Zhang, J. Sun, Salient object detection by composition, in: IEEE International Conference on Computer Vision, 2011, pp. 1028–1035.
[3] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, A.V. Hengel, A survey of appearance models in visual object tracking, ACM Trans. Intell. Syst. Technol. 4 (4) (2013) 58–63.
[4] F. Moosmann, E. Nowak, F. Jurie, Randomized clustering forests for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 30 (9) (2008) 1632–1646.
[5] 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.
[6] L. Wixson, Detecting salient motion by accumulating directionally-consistent flow, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 774–780.
[7] C. Guo, Q. Ma, L. Zhang, Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform, in: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 2008, pp. 1–8.
[8] Y. Liu, C.S. Bouganis, P.Y.K. Cheung, A spatiotemporal saliency framework, in: International Conference on Image Processing, Atlanta, 2006, pp. 437–440.
[9] T.S. Lee, C.F. Yang, R.D. Romero, D. Mumford, Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency, Nat. Neurosci. 5 (6) (2002) 589–597.
[10] R. VanRullen, Visual saliency and spike timing in the ventral visual pathway, J. Physiol. Paris 97 (2-3) (2003) 365–377.
[11] J. Zhu, Y. Qiu, R. Zhang, J. Huang, W. Zhang, Top–down saliency detection via contextual pooling, J. Signal Process. Syst. 74 (1) (2014) 33–46.
[12] A. Borji, Boosting bottom-up and top-down visual features for saliency estimation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 438–445.
[13] X. Zhu, S. Zhang, R. Hu, Y. Zhu, J. Song, Local and global structure preservation for robust unsupervised spectral feature selection, IEEE Trans. Knowl. Data Eng. 30 (3) (2018) 517–529.
[14] W. Zheng, X. Zhu, Y. Zhu, R. Hu, C. Lei, Dynamic graph learning for spectral feature selection, multimedia tools and applications, 2017. https://doi.org/10.1007/s11042-017-5272-y.
[15] Y. Ma, H. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: Proceedings of the Eleventh ACM International Conference on Multimedia, 2003, pp. 374–381.
[16] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, Adv. Neural Inf. Process. Syst. (2007) 545–552.
[17] T. Liu, et al., Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[18] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1915–1926.
[19] J.H. Reynolds, R. Desimone, Interacting roles of attention and visual salience in V4, Neuron 37 (5) (2003) 853–863.
[20] 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.
[21] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007, pp. 1–8.
[22] V. Mahadevan, N. Vasconcelos, Biologically inspired object tracking using center-surround saliency mechanisms, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 541–554.
[23] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 171–177.
[24] W. Zheng, X. Zhu, G. Wen, Y. Zhu, H. Yu, J. Gan, Unsupervised feature selection by self-paced learning regularization, 2018. https://doi.org/10.1016/j.patrec.2018.06.029.
[25] X. Cui, D. Liu, D. Metaxas, Temporal spectral residual: fast motion saliency detection, in: Proceedings of the 17th ACM International Conference on Multimedia, 2009, pp. 617–620.
[26] Z. Chen, X. Wang, Z. Sun, Z. Wang, Motion saliency detection using a temporal Fourier transform, Opt. Laser Technol. 1 (80) (2016) 1–10.
[27] 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.
[28] X. Li, L. Zhao, L. Wei, MH. Yang, F. Wu, Y. Zhuang, H. Ling, J. Wang, DeepSaliency: multi-task deep neural network model for salient object detection, IEEE Trans. Image Process. 25 (8) (2016) 3919–3930.
[29] R. Zhao, W. Oyang, X. Wang, Person re-identification by saliency learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 356–370.
[30] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2017) 865–878.

[31] X. Wang, H. Ma, X. Chen, Geodesic weighted Bayesian model for saliency optimization, Pattern Recognit. Lett. 1 (75) (2016) 1–8.
[32] Y. Zheng, B. Jeon, D. Xu, Q.M. Wu, H. Zhang, Image segmentation by generalized hierarchical fuzzy *C*-means algorithm, J. Intell. Fuzzy Syst. 28 (2) (2015) 961–973.
[33] S. Treue, J.C. Trujillo, Feature-based attention influences motion processing gain in macaque visual cortex, Nature 399 (6736) (1999) 575–580.
[34] M.M. Ahrens, D. Veniero, M. Harvey, G. Thut, Spatial extrapolation versus temporal entrainment of reflexive attention by apparent motion stimuli are governed by separate mechanisms, Perception 44 (S1) (2015) 136–148.
[35] A. Ranjan, M.J. Black, Optical flow estimation using a spatial pyramid network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4161–4170.
[36] N.D.B. Bruce, J.K. Tsotsos, Saliency based on information maximization, in: Proceedings of NIPS, 2005, pp. 155–162.
[37] T. Alshawi, Z. Long, G. AlRegib, Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection, IEEE Trans. Image Process. (2018).
[38] Z. Tu, A. Abel, L. Zhang, et al., A new spatio-temporal saliency-based video object segmentation, Cogn. Comput. 4 (2016) 1–19.
[39] H.J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis. 9 (12) (2009) 15–27.
[40] S.D. Babacan, T.N. Pappas, Spatiotemporal algorithm for background subtraction, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07), 2007 I-1065–I-1068.
[41] P. Rodríguez, B. Wohlberg, Fast principal component pursuit via alternating minimization, in: IEEE International Conference on Image Processing, Melbourne, 2013, pp. 69–73.
[42] J. He, L. Balzano, A. Szlam, Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012, pp. 1568–1575.
[43] X. Shu, F. Porikli, N. Ahuja, Robust orthonormal subspace learning: efficient recovery of corrupted low-rank matrices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3874–3881.
[44] X. Ye, J. Yang, X. Sun, K. Li, C. Hou, Y. Wang, Foreground–background separation from video clips via motion-assisted matrix restoration, IEEE Trans. Circuits Syst. Video Technol. 25 (11) (2015) 1721–1734.
[45] Hmdb-51. [online]. Avaliable: http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database.
[46] A.B. Mabrouk, E. Zagrouba, Spatio-temporal feature using optical flow based distribution for violence detection, Pattern Recognit. Lett. 92 (2017) 62–67.
[47] A.W. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking: an experimental survey, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1442–1468.
[48] Z. He, L. Sun, Surface defect detection method for glass substrate using improved Otsu segmentation, Appl. opt. 54 (33) (2015) 9823–9830.