Video saliency detection based on temporal difference and pixel gradient

1st Xiangwei Lu
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
E-mail: 1076604618@qq.com

2nd Muwei Jian
School of Computer Science and Technology
Shandong University of Finance,Economics
Jinan, China
*(Correspondence author) E-mail: jianmuweihk@163.com

3rd Rui Wang
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
E-mail: capricorn.orz@gmail.com

4th Zhichao Yun
ShanDong Vocational College of Science and Technology
WeiFang, China
*(Co-correspondence author) E-mail: 21102926@qq.com

5th Peiguang Lin
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
E-mail: llpwgh@163.com

6th Hui Yu
School of Creative Technologies
University of Portsmouth
Portsmouth, UK
E-mail: hui.yu@port.ac.uk

***Abstract*—Even though temporal information matters for the quality of video saliency detection, many problems such as bad performance in time-space coherence and edge continuity still face present network frameworks. In response to these problems, this paper designs a full convolutional neural network, which integrates temporal differential and pixel gradient to fine tune the edges of targets. Meanwhile, the changes of pixel gradients of original images are used to recursively improve the continuity of target edges and details of central areas. The method presented in the paper has been tested with two available public datasets and its effectiveness been proved after it being compared with 6 other widely accepted methods.**

***Index Terms*—Video saliency detection, Temporal difference, Pixels gradient, Edge refinement, Co-Attention**

I. Introduction

Video saliency detection aims to recognize interesting zones in dynamic scenes by simulating the attention mechanism of human's eyes. It has been widely applied in video compression, video target tracking [1], video quality assessment, video summarization [2], scene understanding, etc.. Prevailing video saliency detection methods generally provide macroscopic perspectives and use optical flow and LSTM convolution models to extract temporal characteristics. For example, Li et al. [3] proposed a flow-guided relapse neuroencoder, which can forward coding by applying the motion information obtained via optical flow and sequence signatures derived from the LSTM net, and enhance the temporal consistency of features of each frame. Song et al. [4] proposed a deep bilateral ConvLSTM structure for learning temporal characteristics by way of a cascading, deep method. However, these methods often overlook extracting and integrating interframe details and result in poor continuity of differential features between consecutive frames. To solve the above mentioned problem, we designed a static feature extraction network with several static saliency networks working in parallel to extract the information features of video frames and obtain initial temporal information. Interframe differential nformation between consecutive frames is used to set up a co-attention mechanism. And the temporal relationships between differential information of consecutive frames were used in assisting learning the temporal and spatial continuity of interframe movements. The detection results have been improved by combining the co-attention module and the pixel gradient-based refinement module.

II. Proposed Method

In this paper, a new method is proposed to improve the continuity and center detail of the video detection edge. Our research is composed of three parts: the joint static feature extraction network (SFN), temporal difference co-attention model (TCM), and pixel gradient-based refinement processing module (GRM). Fig.1 is a diagram of the method. SFN aims to extract temporal features on the basis of an image saliency detection method. After SFN extraction of informa-

This work was supported by National Natural Science Foundation of China (NSFC) (61976123, 61601427, 61876098); the Taishan Young Scholars Program of Shandong Province; Royal Society - K. C. Wong International Fellowship (NIF180909); and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

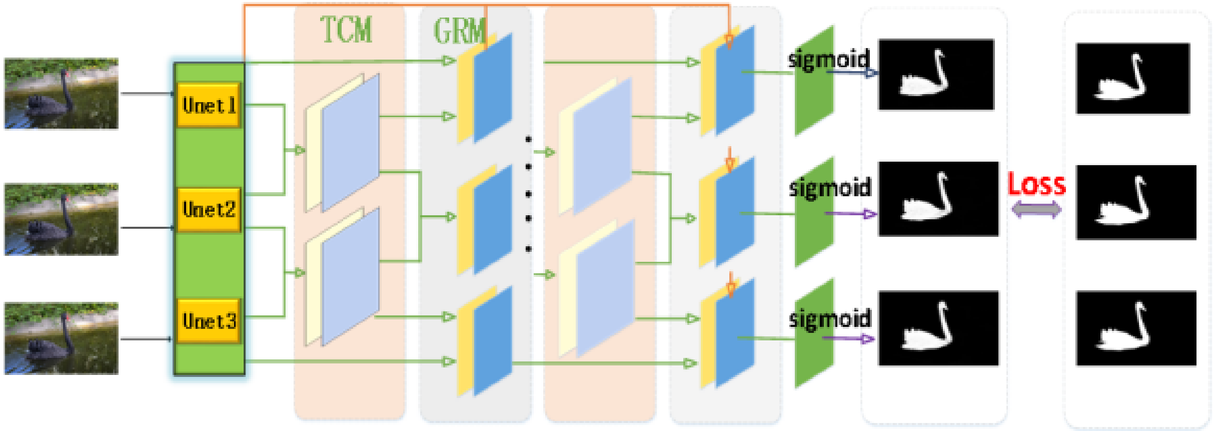


Fig. 1. Diagram of network framework.

tion, TCM integrates temporal difference information via a co-attention mechanism, enhance spatiotemporal consistency between video consecutive frames, and draw attention to both edge and central areas. Then GRM improves details in the edge and central areas by capturing the pixel gradients of initial images.

A. Joint static feature extraction network

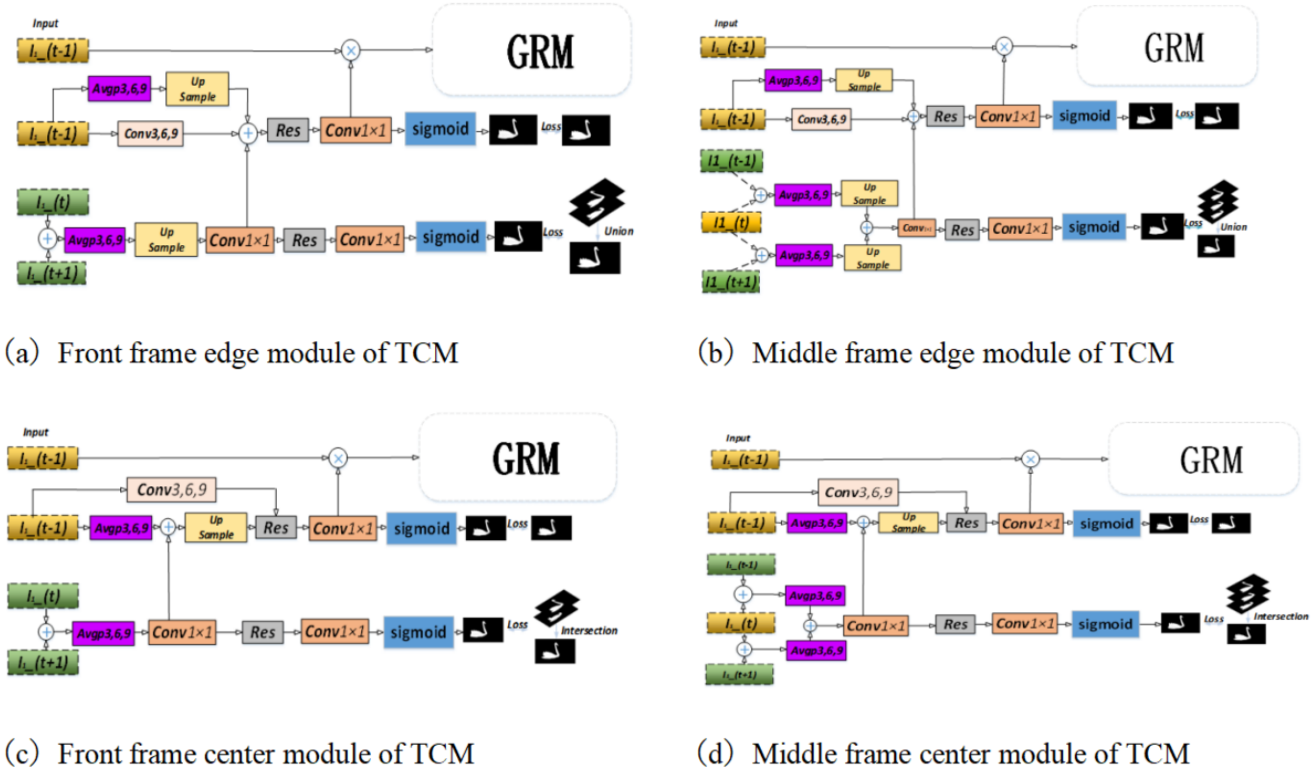
In view of the excellent feature extraction of VGG16, this paper used U-Net with a VGG16 encoder as a subnet of the joint static feature extraction network to extract initial features of video frames. Because of lacking a sufficiently large annotated database for video saliency training, this paper uses the database of DUTS-TR to train the U-Net model and copy the weights obtained from training to the subnets for extracting features of the former, middle, and latter frames. It can ensure the highest relevancy of information output at the beginning of overall dynamic network training.

B. Temporal differential collaborative attention module

Compared to image saliency detection, Saliency detection on video has its biggest advantage: clues of sequential movements. Because of it, the effectiveness of saliency detection can be improved. Therefore, we designed a temporal differential co-attention edge module and central module aiming at the edge and center zones of targets, respectively. As shown in Fig.2 (a), the co-attention edge module of former frames integrates saliency features of SFN outputs through coordinating inter-frame relationships and then assigns weights spatially. This module is composed of two parts: the former frame and the rest. The former frame is trained with annotated GT maps and the rest trained with the union of GT maps to obtain as much edge information as possible. Multi-scale pooling and multi-scale convolution are used for former frames to integrate their information. For the other part, multi-scale pooling and upsampling are used for dimension reduction and integrated information is added to the former frames. Overlapping information is deeply processed and reduced to one dimension via 1×1 convolution blocks, and then is multiplied by I1(t-1) to put weight on initial data. Similar to Fig.2(a), Fig.2 (b) puts together temporal differences between middle frames and former and latter frames and add them to middle frames. After dimension reduction, information of middle frames is updated and more attention can be paid to their edge zones. The TCM central module is designed following the similar principle except that the intersection of GT maps of other parts is used rather than their union, and convolution and pooling are conducted before upsampling.

C. Pixel gradient-based refinement processing module

As in original images pixel values are approximate where salient objects locate but are distinct from those of the background along their edges, pixel gradient can be properly used to set up a barrier to separate the inside from the outside. With the profiles of salient objects known, pixel gradient can act as a powerful tool for capturing and utilizing the details of salient objects. In this paper, we use the TCM modules to integrate the location information of edge zones and central zones to support the gradient-based optimization module. As shown in Fig.3, based on information provided by the TCM edge module, we define the edges of objects through shrinkage first and then expansion. It can help edge optimization get rid of interference by pixel information outside the edge. Meanwhile, as the size of salient objects varies, convolution blocks are designed specifically for multiple-scale pixel expansion. A more balanced approach is used to optimize the information output by TCM central modules. On the basis of shrinkage and expansion, the approach of expansion followed by shrinkage is also used for optimization because besides the problem of hollows, the integrity of edge information needs to be considered when dealing with central information. Shrinkage accomplished by expansion after reversing the colors of saliency maps. (as shown in Formula.1). In addition, to give



(a) Front frame edge module of TCM

(b) Middle frame edge module of TCM

(c) Front frame center module of TCM

(d) Middle frame center module of TCM

Fig. 2. Temporal difference co-attention edge module, I1(t-1), I1(t) and I1(t+1) represent the input information of the former, middle, and latter frames the SFN,Avgp3,6,9 and Conv3,6,9 represents the 3×3, 6×6, and 9×9 multiscale pooling operations and multiscale convolution operations, Res represents residual block, conv1×1 represents 1×1 convolution block, represents element-by-element multiplication, and represents channel concatenate.

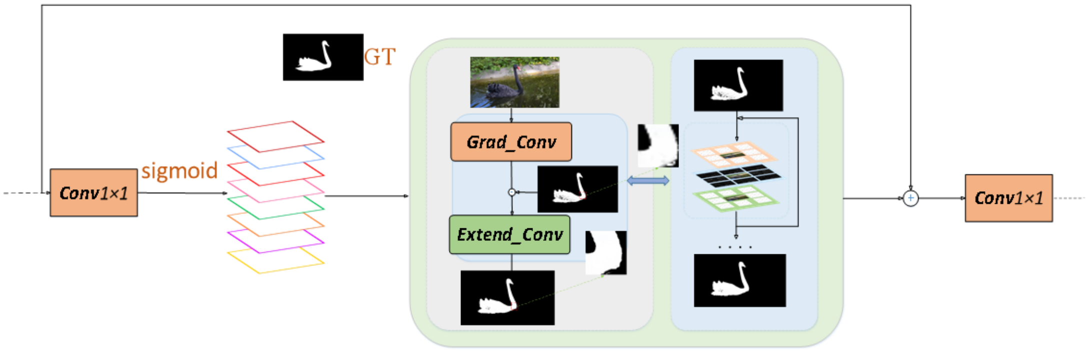


Fig. 3. Pixel gradient refinement module.

full play the combined effects of the pixel gradient module and TCM, they are jointly used to recurrently optimize the network structure.

$$S_g = max(S) + min(S) - S_g \tag{1}$$

S_g represents a gray value, max(S) represents the maximum value in the graph, and max(S) represents the minimum value in the image.

III. Experiment

A. Datasets

Two commonly used benchmark datasets are used for evaluation in this paper: FBMS and DAVIS 2017. The dataset of FBMS (Freiburg-Berkeley Motion Segmentation) is composed of 59 motion video sequences showing obvious motion displacements of salient objects, and hence, is of great value for evaluating how well motion laws are captured. DAVIS 2017 (Densely Annotated Video Segmentation 2017) is composed

TABLE I
COMPARISON OF MAE AND MAXIMUM F-MEASURE OF DIFFERENT METHODS

Method	DAVIS 2017		FBMS	
	maxF	MAE	maxF	MAE
FGRNE [5]]	0.7970	0.0429	0.8066	0.0822
PDB [6]	0.8645	0.0291	0.8374	0.0695
DLVSD [7]	0.7502	0.0558	0.7683	0.1030
SCOM [8]	0.7895	0.0545	0.8020	0.0891
SFLR [9]	0.7496	0.0574	0.6946	0.1305
SGSP [10]	0.7127	0.1274	0.6696	0.1837
Ours	0.8692	0.0293	0.8385	0.0580

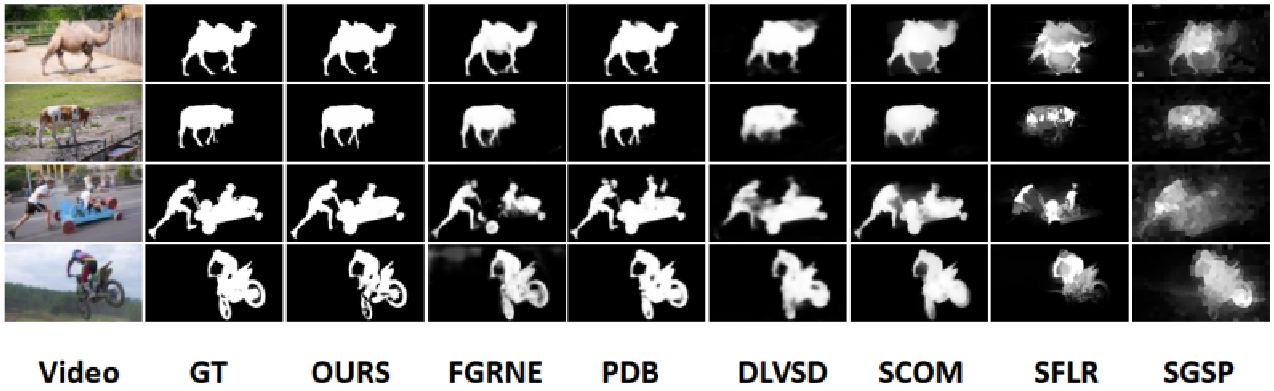


Fig. 4. Comparison of the results of different video saliency detection methods

of 90 high-quality, full HD video sequences, 6242 frames in total.

B. Evaluation criteria

To comprehensively measure the effect of this method, we compare it with 6 popular methods by using two indicators, which are maximum F-measure [10], and mean absolute error (MAE).

$$F_\beta = \frac{(1+\beta^{2}) \times Precision \times Recall}{\beta^2 \times Precision \times Recall} \qquad (2)$$

β^2 is used to put weight on precision and recall to reflect their importance, where β^2 is 0.3. F-Measure is used to measure how perfectly saliency maps match GT after binarization. MAE can measure the level of similarity between the primary saliency map and GT by the average absolute difference in pixel values.

C. Performance Comparison

The method in this paper is compared with 6 other advanced detection methods, which are FGRNE [5], PDB [6], DLVSD [7], SCOM [8], SFLR [9], and SGSP [10]. The superiority of our network model has been highlighted in Fig.4, and Table 1. For the sake of fairness, saliency maps used in the paper come from their authors.

- Quantitative Analysis

In order to quantitatively analyze the effectiveness of our method, two indicators including maxF$_\beta$, and MAE are used in the section for comparing it with other mainstream video saliency detection methods. As indicated by PR curves in Fig.4, our method has sawn an excellent result with its precision and recall being the best among these methods as tested with all two datasets. As shown in Table 1 listing maxF$_\beta$ and MAE of all methods, ours is 0.47%, 0.09% higher than the second best method in terms of maxF$_\beta$ when being tested with datasets of DAVIS 2017 and FBMS, respectively; when it comes to MAE, our method is 0.03%, 1.54% lower. Our method has been proved to be effective as it leads other methods by wide margin in tests with all two different kinds of datasets.

- Qualitative Analysis

Fig.4 shows comparison of visual results from our network model and other network models. According to this comparison, the images based on our method have more clear-cut edges and more details in its center and are free from central hollows, achieving better visual effects than other methods. Our method has secured satisfactory effects for different kinds of primary images, which either have low contrast between the background or object (the first group), high contrast of the object itself (the second group), complicated background (the third group), or many details in the object (the fourth group).

IV. CONCLUSION

In this paper, a new detection model is proposed for video saliency object based on temporal difference and pixel gradient. The detection model is mainly composed of temporal differential co-attention module and pixel gradient refinement module. And the recurrent optimization strategy is applied to improve the precision of spatiotemporal saliency results. Experiments with several available public datasets show that the method in the paper can provide better indicators and visual effects.

References

[1] Wu H, Li G, Luo X. Weighted attentional blocks for probabilistic object tracking[J]. The Visual Computer, 2014, 30(2): 229-243.

[2] GGötze, N., et al. ”Multistage recognition of complex objects with the active vision system NAVIS.” (1996).

[3] Li, Guanbin, et al. ”Flow guided recurrent neural encoder for video salient object detection.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] Song, Hongmei, et al. ”Pyramid dilated deeper convlstm for video salient object detection.” Proceedings of the European conference on computer vision (ECCV). 2018.

[5] Li, Guanbin, et al. ”Flow guided recurrent neural encoder for video salient object detection.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Song, Hongmei, et al. ”Pyramid dilated deeper convlstm for video salient object detection.” Proceedings of the European conference on computer vision (ECCV). 2018.

[7] Wang W, Shen J, Shao L. Video salient object detection via fully convolutional networks[J]. IEEE Transactions on Image Processing, 2017, 27(1): 38-49.

[8] Chen Y, Zou W, Tang Y, et al. SCOM: Spatiotemporal constrained optimization for salient object detection[J]. IEEE Transactions on Image Processing, 2018, 27(7): 3345-3357.

[9] Chen, Yuhuan, et al. ”SCOM: Spatiotemporal constrained optimization for salient object detection.” IEEE Transactions on Image Processing 27.7 (2018): 3345-3357.

[10] Liu, Zhi, et al. ”Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation.” IEEE transactions on circuits and systems for video technology 27.12 (2016): 2527-2542.