Video Saliency Detection Using Deep Convolutional Neural Networks

Xiaofei Zhou[1,2,3], Zhi Liu[2,3(✉)], Chen Gong[4], Gongyang Li[2,3], and Mengke Huang[2,3]

[1] Institute of Information and Control, Hangzhou Dianzi University, Hangzhou, China
zxforchid@outlook.com
[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
liuzhisjtu@163.com, lllgongyang@gmail.com, mengkehuang@gmail.com
[3] School of Communication and Information Engineering, Shanghai University, Shanghai, China
[4] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
chen.gong@njust.edu.cn

Abstract. Numerous deep learning based efforts have been done for image saliency detection, and thus, it is a natural idea that we can construct video saliency model on basis of these image saliency models in an effective way. Besides, as for the limited number of training videos, existing video saliency model is trained with large-scale synthetic video data. In this paper, we construct video saliency model based on existing image saliency model and perform training on the limited video data. Concretely, our video saliency model consists of three steps including feature extraction, feature aggregation and spatial refinement. Firstly, the concatenation of current frame and its optical flow image is fed into the feature extraction network, yielding feature maps. Then, a tensor, which consists of the generated feature maps and the original information including the current frame and the optical flow image, is passed to the aggregation network, in which the original information can provide complementary information for aggregation. Finally, in order to obtain a high-quality saliency map with well-defined boundaries, the output of aggregation network and the current frame are used to perform spatial refinement, yielding the final saliency map for the current frame. The extensive qualitative and quantitative experiments on two challenging video datasets show that the proposed model consistently outperforms the state-of-the-art saliency models for detecting salient objects in videos.

This work was supported by the National Natural Science Foundation of China under Grants 61771301 and 61602246, the Natural Science Foundation of Jiangsu Province under Grant BK20171430, the Fundamental Research Funds for the Central Universities under Grant 30918011319 and the Summit of the Six Top Talents Program under Grant DZXX-027.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 308–319, 2018.
https://doi.org/10.1007/978-3-030-03335-4_27

Keywords: Video saliency · Convolutional neural networks
Feature aggregation

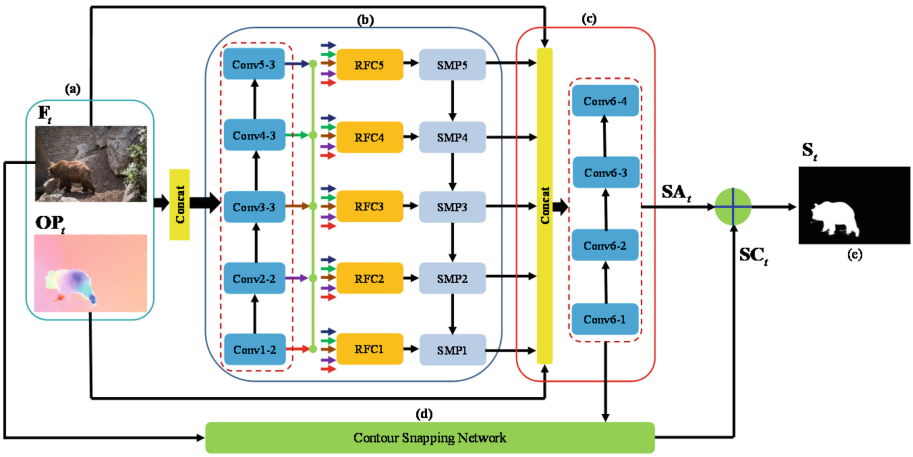

Fig. 1. The main flowchart of the proposed video saliency model. Given (a) the current frame $\mathbf{F}_t$ and its optical flow image $\mathbf{OP}_t$, we obtain five feature maps $\{\mathbf{SMP}_i, i = 1, \ldots, 5\}$ via (b) the feature extraction network. Then these feature maps and the original information including $\mathbf{F}_t$ and $\mathbf{OP}_t$ are concatenated and passed to (c) the aggregation network. Besides, the current frame $\mathbf{F}_t$ and the output of aggregation network $\mathbf{SA}_t$ are passed to (d) the contour snapping network, to perform spatial refinement. (e) Shows the saliency map $\mathbf{S}_t$ of current frame, which is the summation of the outputs of aggregation network and contour snapping network, *i.e.* $\mathbf{SA}_t$ and $\mathbf{SC}_t$

1 Introduction

Saliency detection aims to identify the salient object regions in images or videos, which plays an important role as a preprocessing step in many computer vision applications such as object detection and segmentation [7,11,21,29,33], content-aware image/video retargeting [8,28], and content-based image/video compression [12,13]. According the input of the visual system, saliency detection can be categorized into two classes including image saliency models and video saliency models. Up to now, numerous efforts have been devoted to the saliency detection for still images, but the research on video saliency has received relatively few attention. In this paper, we focus on the video saliency detection.

Video saliency detection is different from image saliency detection, since it takes into account both spatial and temporal information of the video sequences simultaneously. In order to deal with both cues in videos and pop-out the

prominent objects from videos, many priors efforts have been done from various aspects such as the center-surround scheme [15,22], information theory [18], machine learning [16,25], information fusion [5,9], and regional saliency assessment [19,20,26,32]. The above saliency models can obtain satisfactory results to some degree, however their performances degrade in dealing with complicated motion and complex scenes such as fast motion, dynamic background, nonlinear deformation, and occlusion, etc. Fortunately, convolutional neural networks (CNNs) have been successfully applied to many areas in computer vision such as object detection and semantic segmentation [6,10]. Further, it also pushes forward the progress of saliency detection in still images such as the multi-context deep learning framework [31] and the aggregation of multi-level convolutional feature framework [30]. Obviously, it is a natural idea that we can construct video saliency model based on existing deep learning based image saliency models. Unfortunately, we can see that the temporal information over frames is not incorporated by these deep saliency models, thus, it is not appropriate to conduct video saliency detection on each frame by using existing deep saliency models directly. Recently, deep learning is also applied in video saliency detection such as the two cascade modules based deep saliency network in [27]. However, due to the limited number of annotated training videos, this model is trained on large-scale synthetic video data.

Motivated by this, we propose a video saliency model based on existing image saliency model and train it with limited video data only. Concretely, our model consists of three steps including feature extraction, feature aggregation and spatial refinement. The current frame and its optical flow image are first concatenated and fed into the feature extraction network, generating the corresponding feature maps. Notably, we employ an off-the-shelf convolutional neural networks (CNNs) based image saliency model [30] as our feature extraction network. Then, the obtained feature maps, the current frame and its optical image are combined and passed to the aggregation network, which is used to perform feature integration. Finally, a contour snapping network based spatial refinement network is deployed to the output of aggregation network and generates the final saliency map.

The advantages of our model are threefold. Firstly, the input of the feature extraction network is the concatenation of the current frame and its optical flow image, which gives a strong prior for the salient objects in videos. Secondly, the aggregation network not only incorporates the feature maps generated by the feature extraction network, but also aggregates the original information including the current frame and the optical flow image. The original information can provide complementary information for the aggregation of feature maps. Thirdly, a contour snapping based spatial refinement is introduced to improve the quality of spatiotemporal saliency maps, which not only highlight salient objects effectively, but also be with well-defined boundaries. Overall, our main contributions are summarized as follows:

1. Based on existing image saliency models, we propose a deep convolutional neural network based video saliency model, which consists of three steps

including feature extraction, feature aggregation and spatial refinement. Specifically, the three steps correspond to three sub-networks including feature extraction network, feature aggregation network and contour snapping network.
2. In order to obtain complementary information for the aggregation of feature maps, we incorporated the original information including the current frame and its optical flow image into the aggregation network. Concretely, a tensor that consists of feature maps and original information is fed into the aggregation network.
3. We compare our model with several state-of-the-art saliency models on two public video datasets, and the experimental results firmly demonstrate the effectiveness and superiority of the proposed model.

2 Our Approach

Figure 1 shows an overview of the proposed video saliency model. Concretely, in the first step, *i.e.* feature extraction, the input is the concatenation of the current frame $\mathbf{F}_t$ and its corresponding optical flow image $\mathbf{OP}_t$. Then, we obtain the feature maps originated from different layers, as shown in Fig. 1(b) and denoted as $\{\mathbf{SMP}_i, i = 1, 2, 3, 4, 5\}$. Successively, these feature maps and the original information including $\mathbf{F}_t$ and $\mathbf{OP}_t$ are concatenated and passed to the aggregation network as shown in Fig. 1(c). Further, a contour snapping network [4] based spatial refinement shown in Fig. 1(d), is deployed in our model. The contour snapping network incorporates the current frame $\mathbf{F}_t$ and the output of aggregation network $\mathbf{SA}_t$ together. Finally, the saliency map $\mathbf{S}_t$, as shown in Fig. 1(e), is computed as the summation of the output of the aggregation network and the contour snapping network, *i.e.* $\mathbf{SA}_t$ and $\mathbf{SC}_t$.

2.1 Feature Extraction

In order to obtain an appropriate representation for salient objects in videos, we employ the feature extraction network in [30], which achieves a superiority performance in saliency detection for still images, to extract feature maps. We should note that one of the difference between our model and [30] is the input. Specifically, the input of feature extraction in our model is the concatenation of the current frame and its corresponding optical flow image, which is a strong prior for salient objects in videos. Differently, [30] focus on image saliency, thus, its input is the static image only.

As aforementioned, the input of feature extraction is the concatenation of the current frame $\mathbf{F}_t$ and its corresponding optical flow image $\mathbf{OP}_t$, which is generated using the method of large displacement optical flow (LDOF) [3] and then converted to a 3-channel (RGB) color coded optical flow image [2]. Besides, we should note that we concatenate $\{\mathbf{F}_t, \mathbf{OP}_t\}$ in the channel direction, thus generating a tensor with the size of $h \times w \times 6$, in which h and w refer to the height and width of the scaled current frame/optical flow image, respectively.

Here, we set h and w as 256. Then, the generated tensor is fed into the feature extraction network shown in Fig. 1(b), in which $\{\mathbf{RFC}_i, i = 1, 2, 3, 4, 5\}$ refer to the resolution-based feature combination structure as detailed in [30]. The output of feature extraction, namely $\{\mathbf{SMP}_1, \mathbf{SMP}_2, \mathbf{SMP}_3, \mathbf{SMP}_4, \mathbf{SMP}_5\}$ are all with the size of $256 \times 256 \times 2$, which are two channel feature maps consistent with [30].

2.2 Feature Aggregation

With the output of the feature extraction network, namely feature maps $\{\mathbf{SMP}_i, i = 1, 2, 3, 4, 5\}$, we design an aggregation network to effectively aggregate these feature maps and generate the coarse result of video saliency detection, *i.e.* $\mathbf{SA}_t$. To provide complementary information for convolutional features originated from feature maps, the original information including the current frame $\mathbf{F}_t$ and its optical flow image $\mathbf{OP}_t$ is also incorporated to the aggregation operation.

Specifically, we first concatenate these feature maps, the current frame $\mathbf{F}_t$ and its optical flow image $\mathbf{OP}_t$ in the channel direction, yielding a tensor with the size of $256 \times 256 \times 16$. Secondly, the generated tensor is fed into a series of convolutional layers including $\{Conv6-1, Conv6-2, Conv6-3\}$, each of them is a convolutional layer with 3×3 kernel size. Successively, there is a layer denoted as $Conv6-4$, which is a 1×1 convolutional filter. Finally, the output of aggregation network, *i.e.* $\mathbf{SA}_t$, is generated via a softmax layer. Besides, a batch normalization layer [1] and a ReLU layer are deployed between $Conv6-1$ and $Conv6-2$, as well as between $Conv6-2$ and $Conv6-3$.

In our model, the feature extraction network and the aggregation network are jointly trained in an end-to-end manner. Given the training dataset $\mathbf{D}_{train} = \{(\mathbf{F}_n, \mathbf{OP}_n, \mathbf{Y}_n)\}_{n=1}^{N}$ with N training samples, in which $\mathbf{F}_n = \{\mathbf{F}_n^j, j = 1, ..., N_p\}$, $\mathbf{OP}_n = \{\mathbf{OP}_n^j, j = 1, ..., N_p\}$ and $\mathbf{Y}_n = \{\mathbf{Y}_n^j, j = 1, ..., N_p\}$ denote the input current frame, its optical flow image and the binary ground-truth with N_p pixels, respectively. Besides, $\mathbf{Y}_n^j = 1$ indicates the salient object pixel and $\mathbf{Y}_n^j = 0$ represents the background pixel. For simplicity, we drop the subscript n and consider $\{\mathbf{F}, \mathbf{OP}\}$ for each frame independently. Thus, the loss function can be defined as:

$$\begin{aligned}\mathcal{L}(\mathbf{W}, \boldsymbol{b}) = &-\beta \sum_{j \in \mathbf{Y}_+} \log P\left(\mathbf{Y}^j = 1 | \mathbf{F}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right) \\ &-(1-\beta) \sum_{j \in \mathbf{Y}_-} \log P\left(\mathbf{Y}^j = 0 | \mathbf{F}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right),\end{aligned} \tag{1}$$

where $\mathbf{W}$ and $\boldsymbol{b}$ are denoted as kernel weights and bias of convolutional layers, and $\mathbf{Y}_+$ and $\mathbf{Y}_-$ indicate the label sets for salient objects and background, respectively. β refers to the ratio of salient objects pixels in the ground truth $\mathbf{G}$, *i.e.* $\beta = |\mathbf{Y}_+|/|\mathbf{Y}_-|$. $P\left(\mathbf{Y}^j = 1 | \mathbf{F}, \mathbf{OP}; \mathbf{W}, \boldsymbol{b}\right)$ denotes the probability of the pixel belonging to salient objects. Besides, the loss function is also the difference between our model and [30]. Concretely, the loss function in [30] consists of the fusion loss and the layer loss of other five layers. Differently, the loss function in our model is the aggregation loss.

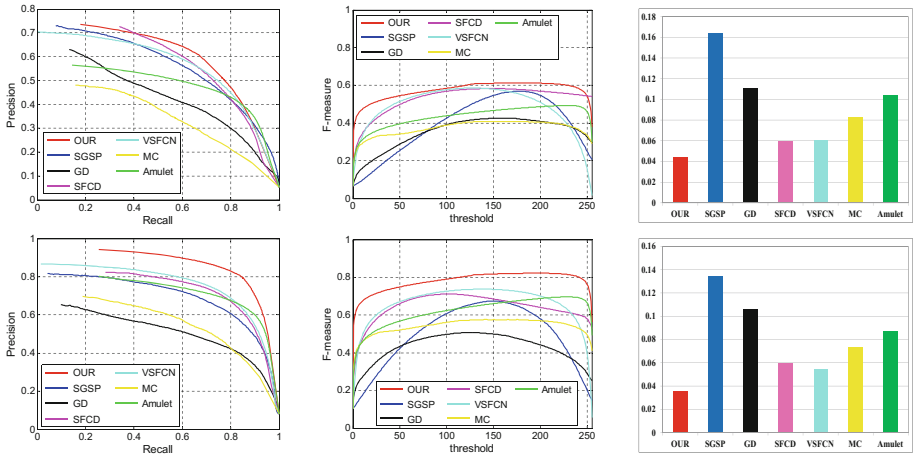

Fig. 2. (better viewed in color) Quantitative evaluation of different saliency models: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE. From top to down, each row shows the results on the UVSD dataset and the DAVIS dataset, respectively.

2.3 Spatial Refinement

To further improve the detection accuracy, we introduce a contour snapping network into our method to perform spatial refinement. The contour snapping network [4] is trained offline and used to detect object contours. Here, we exploit the contour snapping network without training or fine-tuning. In our implementation, we first train the aforementioned networks including feature extraction network and aggregation network shown in Fig. 1(a, b, c) in an end-to-end manner. Then, in the test phase, we add a second branch, *i.e.* the contour snapping network, into our model. Concretely, the current frame $\mathbf{F}_t$ and the output of aggregation network $\mathbf{SA}_t$ are first passed to the contour snapping network shown in Fig. 1(d), and the output is denoted as $\mathbf{SC}_t$. Then, the outputs of contour snapping network and aggregation network are combined via linear summation. Finally, we obtain the final saliency map $\mathbf{S}_t$ for the current frame:

$$\mathbf{S}_t = \mathrm{Norm}\left[\mathbf{SA}_\mathrm{t} + \mathbf{SC}_\mathrm{t}\right], \tag{2}$$

where the operation Norm normalizes the saliency map into the range of $[0, 1]$.

3 Experimental Results

3.1 Experimental Setup

Datasets and Metrics: The datasets in training and test phases consist of three public challenging datasets. Concretely, SegTrackV2 [17] consists of 14

videos with challenging circumstances such as appearance change, motion blur, occlusion, complex deformation and so on. UVSD [19] contains a total of 18 challenging videos with complicated motions and complex scenes. DAVIS [24] is a recent dataset for video object segmentation, which contains 50 high-quality videos with different motions of human, animal and vehicle in challenging circumstances. Similar to [27], we train our model on the binary masks of SegTrackV2 and the training set of DAVIS. When testing, we evaluate the performance of the proposed model over two datasets including UVSD and the test set of DAVIS. Besides, following the evaluation measures used in [27], we evaluate the video saliency detection performance using the precision-recall (PR) curve, F-measure curve by setting its β^2 to 0.3, and mean absolute error (MAE) values.

Implementation Details: To avoid over-fitting, some prior works [23,34] have been done from the perspective of utilizing training-free features. Differently, to reduce the effect of over-fitting and improve the generalization of neural network, we simply augment these two datasets by mirror reflection and rotation techniques ($0°, 90°, 180°, 270°$). In the training phase, we use Stochastic Gradient Descent (SGD) with momentum 0.9 for 22000 iterations with base learning rate 10^{-8}, mini-batch size 32 and a weight decay 0.0001. Besides, the parameters of multi-level feature extraction layers are initialized from the model [30]. For other convolutional layers, we initialize the weights by the "msra" method [14].

3.2 Performance Comparison with State-of-the-art

We compare our model with state-of-the-art saliency models including SGSP [19], GD [26], SFCD [5], VSFCN [27], MC [31] and Amulet [30]. The former three models aim at video saliency while the latter two are deep learning based image saliency models. Besides, our model is denoted as "OUR". In the following, quantitative and qualitative comparisons are performed successively.

A quantitative comparison among OUR, SGSP, GD, SFCD, VSFCN, MC and Amulet is shown in Fig. 2. We can see that our model achieves the best performance in terms of PR curves, F-measure curves and MAE values on UVSD and DAVIS datasets. It clearly demonstrates the effectiveness of our model. Figures 3 and 4 provide the qualitative evaluation for our model and the state-of-the-art saliency models on UVSD and DAVIS, respectively. All these videos exhibits various challenges such as shape complexity, occlusion and non-rigid deformation and motion blur and so on. Thus, it is a challenging task for video saliency detection. Compared with other models, we can see that our model achieves the best performance with completely highlighted salient objects and effectively suppressed background regions, as shown in Figs. 3(c) and 4(c). For the results of MC and Amulet shown in Figs. 3(h, i) and 4(h, i), some background regions are also highlighted due to the lack of temporal information in these two models. For other three models including SGSP, GD and SFCD, their results can pop-out the main parts of salient objects and also highlight some background regions around salient objects. The reason behind this lies in that the features in these models are not discriminative enough. Thus, it is incapable of differentiating the

salient objects and background regions effectively. From the results of VSFCN, as shown in Figs. 3(g) and 4(g), we can see that some background regions are also popped out for videos with fast motion and non-rigid deformation.

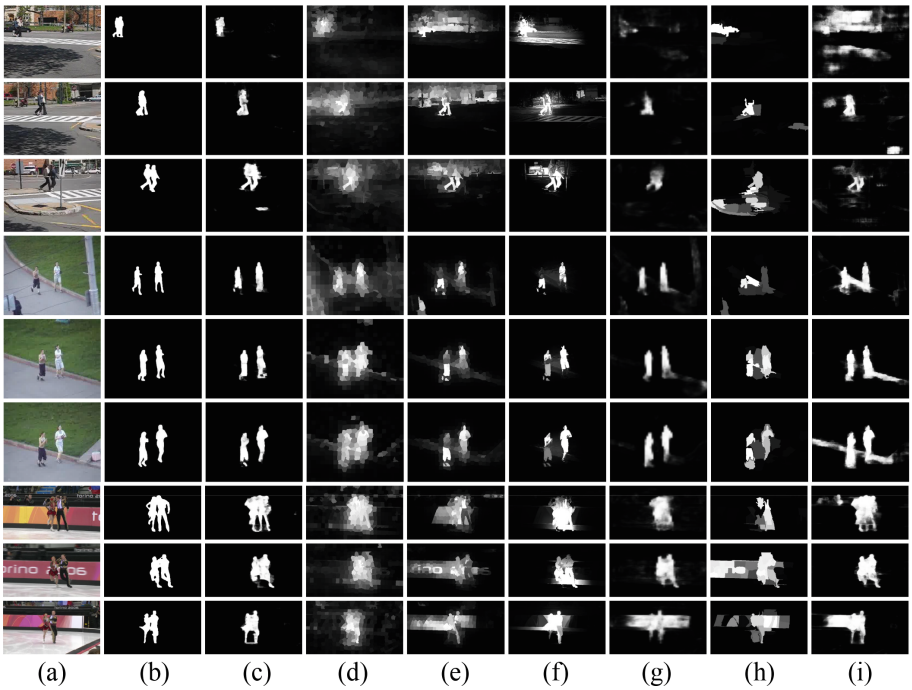

Fig. 3. Examples of spatiotemporal saliency maps for some videos in the UVSD dataset. (a): Input video frames, (b): binary ground truths, (c): OUR, (d): SGSP, (e): GD, (f): SFCD, (g): VSFCN, (h): MC, (i): Amulet.

3.3 Analysis of the Proposed Model

To investigate the effectiveness of feature aggregation and spatial refinement, we conduct ablation experiments on UVSD dataset, and the results are shown in Fig. 5. In these experiments, our model without contour snapping is denoted as "woCS", and on the basis of "woCS", the input of aggregation network without current frame, optical flow image and the previous two are denoted as "woRGB","woOP" and "woRGBOP", respectively. Concretely, firstly,"woCS" achieves the second best performance, and with the help of contour snapping, "OUR" performs best compared to variants of the proposed model. It clearly demonstrates the effectiveness of spatial refinement. Secondly, the performance of "woRGB","woOP" and "woRGBOP" is worse than "woCS", it demonstrates

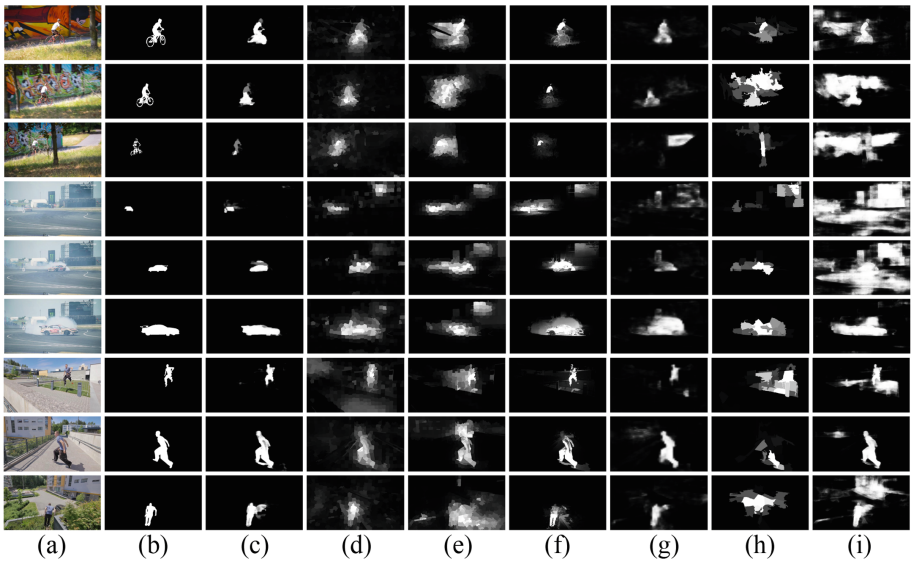

Fig. 4. Examples of spatiotemporal saliency maps for some videos in the DAVIS dataset. (a): Input video frames, (b): binary ground truths, (c): OUR, (d): SGSP, (e): GD, (f): SFCD, (g): VSFCN, (h): MC, (i): Amulet.

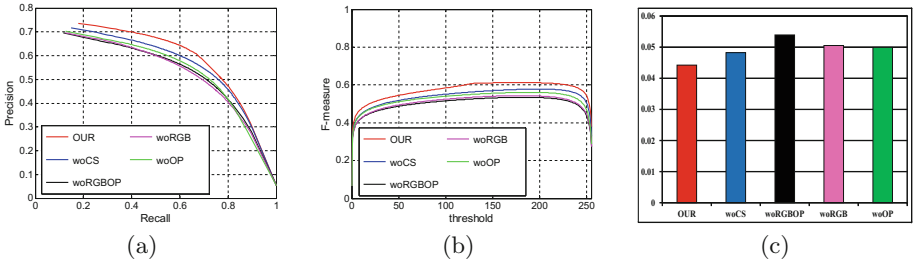

Fig. 5. (better viewed in color) Quantitative evaluation for the model analysis: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE.

the effectiveness and rationality of feature aggregation, which needs the complementary information originated from current frame and optical flow image. Lastly, from the perspective of PR curves, F-measure curves and MAE values, we can see that "woRGB" and "woOP" perform better than "woRGBOP", and it indicates that the complementary information originated from current frame and optical flow image is crucial for aggregation network. Generally speaking, the ablation study shown in Fig. 5 demonstrates the effectiveness and rationality of feature aggregation and spatial refinement in the proposed model.

4 Conclusion

Based on the existing image saliency model, we propose a novel video saliency model, in which feature extraction, feature aggregation and spatial refinement are integrated in a unified architecture. Firstly, the concatenation of the current frame and its optical flow image is fed into the feature extraction network, which defines an appropriate representation for salient objects in videos. Then, the aggregation network is used to aggregate the generated feature maps and the original information. The novelty lies in the introduction of the original information, which provides complementary information for the aggregation of feature maps. Finally, the contour snapping network is introduced to perform spatial refinement, yielding a high-quality saliency map with well-defined boundaries. The experimental results on two public datasets show the effectiveness of the proposed model.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell **39**, 2481–2495 (2017)
2. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
3. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(3), 500–513 (2011)
4. Caelles, S., Maninis, K., Ponttuset, J., Lealtaixe, L., Cremers, D., Van Gool, L.: One-shot video object segmentation, pp. 221–230, June 2016
5. Chen, C., Li, S., Wang, Y., Qin, H., Hao, A.: Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans. Image Process. **26**(7), 3156–3170 (2017)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
7. Chen, L., Shen, J., Wang, W., Ni, B.: Video object segmentation via dense trajectories. IEEE Trans. Multimed. **17**(12), 2225–2234 (2015)
8. Du, H., Liu, Z., Jiang, J., Shen, L.: Stretchability-aware block scaling for image retargeting. J. Vis. Commun. Image Represent. **24**(4), 499–508 (2013)
9. Fang, Y., Wang, Z., Lin, W., Fang, Z.: Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Trans. Image Process. **23**(9), 3910–3921 (2014)
10. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448. IEEE (2015)
11. Gong, C., et al.: Saliency propagation from simple to difficult. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2531–2539. IEEE, June 2015

12. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process. **19**(1), 185–198 (2010)
13. Guo, J., Song, B., Du, X.: Significance evaluation of video data over media cloud based on compressed sensing. IEEE Trans. Multimed. **18**(7), 1297–1304 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imageNet classification. In: The IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034. IEEE (2015)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
16. Lee, W.F., Huang, T.H., Yeh, S.L., Chen, H.H.: Learning-based prediction of visual attention for video signals. IEEE Trans. Image Process. **20**(11), 3028–3038 (2011)
17. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: IEEE International Conference on Computer Vision (ICCV), pp. 2192–2199. IEEE (2013)
18. Liu, C., Yuen, P.C., Qiu, G.: Object motion detection using information theoretic spatio-temporal saliency. Pattern Recognit. **42**(11), 2897–2906 (2009)
19. Liu, Z., Li, J., Ye, L., Sun, G., Shen, L.: Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Trans. Circuits Syst. Video Technol. **27**(12), 2527–2542 (2017)
20. Liu, Z., Zhang, X., Luo, S., Le Meur, O.: Superpixel-based spatiotemporal saliency detection. IEEE Trans. Circuits Syst. Video Technol. **24**(9), 1522–1540 (2014)
21. Liu, Z., Zou, W., Le Meur, O.: Saliency tree: a novel saliency detection framework. IEEE Trans. Image Process. **23**(5), 1937–1952 (2014)
22. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 171–177 (2010)
23. Mahapatra, D., Winkler, S., Yen, S.C.: Motion saliency outweighs other low-level features while watching videos. In: Human Vision and Electronic Imaging XIII, vol. 6806, p. 68060P. International Society for Optics and Photonics (2008)
24. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 724–732. IEEE (2016)
25. Vig, E., Dorr, M., Martinetz, T., Barth, E.: Intrinsic dimensionality predicts the saliency of natural dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **34**(6), 1080–1091 (2012)
26. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3395–3402. IEEE (2015)
27. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. **27**(1), 38–49 (2018)
28. Yan, B., Yuan, B., Yang, B.: Effective video retargeting with jittery assessment. IEEE Trans. Multimed. **16**(1), 272–277 (2014)
29. Ye, L., Liu, Z., Li, L., Shen, L., Bai, C., Wang, Y.: Salient object segmentation via effective integration of saliency and objectness. IEEE Trans. Multimed. **19**(8), 1742–1756 (2017)
30. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: aggregating multi-level convolutional features for salient object detection. In: The IEEE International Conference on Computer Vision (ICCV), pp. 202–211. IEEE, October 2017

31. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1265–1274. IEEE, June 2015
32. Zhou, X., Liu, Z., Gong, C., Liu, W.: Improving video saliency detection via localized estimation and spatiotemporal refinement. IEEE Trans. Multimed. (2018). https://doi.org/10.1109/TMM.2018.2829605
33. Zhou, X., Liu, Z., Sun, G., Ye, L., Wang, X.: Improving saliency detection via multiple kernel boosting and adaptive fusion. IEEE Signal Process. Lett. **23**(4), 517–521 (2016)
34. Zhu, Z., et al.: An adaptive hybrid pattern for noise-robust texture analysis. Pattern Recognit. **48**(8), 2592–2608 (2015)