**ISBN** 978-981-14-1684-2
**Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering**
(WCSE 2019 SUMMER)
*Hong Kong, 15-17 June, 2019, pp. 164-170*
doi: 10.18178/wcse.2019.06.024

# Deep Video Object Contour Extraction Using Fully Convolutional Network

Die Li, Murong Jiang, Guocai Du, Chunna Zhao[+], Yinghua Li ,

School of Information Science and Engineering, Yunnan University, Kunming 650501, China
+86 18468015242, +86 15911538327
Annie5242@163.com, jiangmr@ynu.edu.cn

**Abstract.** Complex scene object contour extraction problem has become an important topic in computer vision. To solve the problem, a deep fully convolutional neural network model is established in this paper. The model tackles the task of semi-supervised video contour extract. In our model, the interactive segmentation method and Mask-RCNN algorithm are respectively applied to the first frame of a video to obtain the target semantic information and segmentation mask. Then the binary object mask is processed by the edge detection algorithm to generate a object contour mask. Next the video and the first frame contour mask are input to network of One-Shot Video Object Segmentation algorithm, and contour features and semantic information of the object are learned by the network. Finally the contour semantic information is automatically passed to subsequent frames, and the contours that particular objects in each frame of video are extracted.

Experiments show that this model can detect and locate one or more objects quickly and accurately in a video sequence for various complex scene. Compared with the general edge detection operator, our algorithm does not need to extract redundant background edges and texture details and can be better applied to pose estimation and target recognition.

**Keywords:** Contour extraction; convolutional neural network; interactive segmentation; edge detection; Video Object Segmentation

## 1. Introduction

Shape information can be provided by target contours, which can be used for pose estimation, target recognition, etc., and now contour extraction become a hot research topic[1-3]. With the advent of the era of big data, the number of images and videos that can be obtained every day is numerous, and huge data puts a lot of pressure to storage devices. However, there is often a large amount of redundant information in the video. For example, for a series of videos, we may only need information about some frame containing the same specific target. So it is necessary for us to save the segmented frame or the edge and contour frame of the target.

At present, video-based object extraction algorithms have algorithms based on edge detection, target detection, and Markov random fields. However, the object contours cannot be obtained directly by these algorithms, and subsequent processes are required, such as false alarm removal, edge tracking and edge closure. A contour extraction method based on a complete convolutional neural network is proposed in this paper. A frame contour mask label is required in this algorithm, and the complete contour information of a specific object can be extracted. On the other hand, for the existing video foreground extraction method, whether it is the frame difference method or the background difference method, time consistency is

[+] Corresponding author. Tel.:15391336695; fax: +087165930963.
*E-mail address*: chunnazhao@163.com.

dependent. The problem of how to extract the contour of an object when processing each frame independently is discussed in this paper, which ignores time information and redundancy.

To the best of our knowledge, this paper is the first to apply a deep learning framework to solve video contour extraction. Our model is inspired by One-Shot Video Object segmentation (Referred to as OSVOS)[4] and the interactive object selection[5]. The idea of our algorithm is that the user uses the interactive target selection algorithm to select a specific object in the first frame of the video. The algorithm generates a binary mask of the specific object. Then the edge detection operator is used to obtain the object contour label, and video and contour label are input into the OSVOS network. Finally, all object concour of video are output. Figure 1 shows an overview of the method.

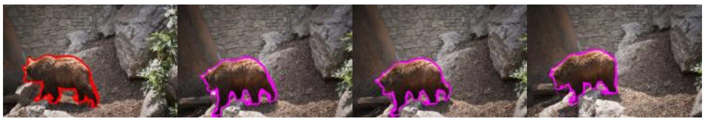

Fig. 1: Contour extraction results show

In Figure 1, the red contour image is a contour label frame, and the rose red contour is a contour extraction result of a subsequent frame.

The main contributions of this paper are summarized as follows:

1) For the foreground target contour extraction problem in video, deep interaction object selection algorithm and one-shot video object segmentation algorithm are combined, and an effective contour extraction model is proposed.
2) For the problem of obtaining the first frame contour mask label, the deep interaction selection algorithm and the edge detection algorithm are jointly applied by us to realize the acquisition of the first frame mask, and the object mask can be obtained only by the user performing a simple click operation.
3) In order to enhance object selection and propagation performance of the model, we apply the instance-aware semantic segmentation algorithm to obtain the location and category information of the object, and the task of multi-object contour extraction is implemented by the improved model.

The rest of the paper is organized as follows. Section 2, the related work is briefly reviewed. Section 3, the proposed model is elaborated. The experimental results are presented and analyzed in section 4 and the paper is summarized in section 5.

## 2. Related Work

The interactive segmentation method is not a new topic in computer vision processing. The contour-based method and the bounding box method are widely used in interactive image segmentation [6-8]. However, a large amount of user interaction is required in these methods, and stroke operations are applied to each frame of video. Furthermore, under images with similar foreground and background appearance, complex textures and appearances, and images under dim lighting conditions, these algorithms have difficulty distinguishing between foreground and background. Therefore, these methods are difficult to achieve the desired results.

At present, semantic segmentation and instance semantic segmentation are increasingly attracting researchers' attention [9-11]. Farabet et al. performed scene marking by using a multi-scale convolutional network trained from raw pixels [12]. The paper[13] extracted regional and foreground features by using the R-CNN framework [11] and regional proposals. Since the task of semantic segmentation is closely related to interactive segmentation, many algorithms have been proposed to combine the two. Ning Xu et al. proposed a simple Interactive semantic segmentation model. Deep learning techniques are used by the model to understand objectivity and semantics. Users only need to make several positive and negative clicks on the target and background. That is, the object and the background in the image can be separated.

In this paper, we propose an interactive video contour extraction model for video contour extraction problem. User interaction and semantic communication are used to extract the contour of video in the model.

Firstly, the interactive selection algorithm is applied to segment the object of interest in the first frame. Then, edge detection algorithm is used to extract contour of segmented image. Next, Semantic segmentation and edge detection algorithms are used to obtain the semantic information of the first frame of image. Finally, video sequence and contour mask are input to the segmentation network model, and the neural network semantically selects and propagates the specific object containing the label, and the object contour is extracted independently in the subsequent frames of the video.

## 3. Proposed Model

The idea of this model is derived from the one-shot video object segmentation algorithm[4]. We train a Fully Convolutional Neural Network to extract the contour of a specific object from a video. We use three successive steps. First, our CNN network is pre-trained in an image dataset containing various objects, to construct a model that is able to distinguish the general concept of a foreground object, that is, "it is an object." Then, during the training phase, the network is trained in a video data set containing contour labels, to construct a model that is able to distinguish the general concept of object contour. Finally, during the testing phase, we fine-tune the network for a bit of iteration on the particular object contour, that is, "this is the contour of a particular object".

### 3.1 Model overview

We hope to use the deep learning network to generate perceptual multi-level functions, capture the contours of specific object, and complete the video object contour extraction tasks. After studying a large number of video image segmentation algorithms, we improved OSVOS algorithm and improved it into our video object contour extraction model. We have made the following improvements: one was to increase the first frame contour extraction process, and the second was to increase and modify the training samples, and the third was to increase the semantic segmentation network extended object segmentation model.

Our model consists of three deep learning networks, an interactive segmentation network, an instance semantic segmentation network, and a video target segmentation network. The first frame image is segmented by the interactive segmentation network, and a specific target mask is output by the segmentation network. The edge detection operator is applied to the segmentation mask, and the contour label is produced by the operator. Another network, Mask-RCNN is used by us to get the location, category, and mask of the object. Then the edge detection operator is applied by us to extract the contour of the mask. The third network, all video frames, the interactive contour mask and semantic contour mask are input to the video segmentation network. After the first frame image is processed by the video segmentation network, the semantic information of the object is obtained, and the semantic information is propagated to the subsequent frame. Then, the network is gradually fine-tuned according to the label and the semantic information network. Finally, the object contour of all the frames is obtained. Our video object contour extraction model architecture is shown in Figure 2.

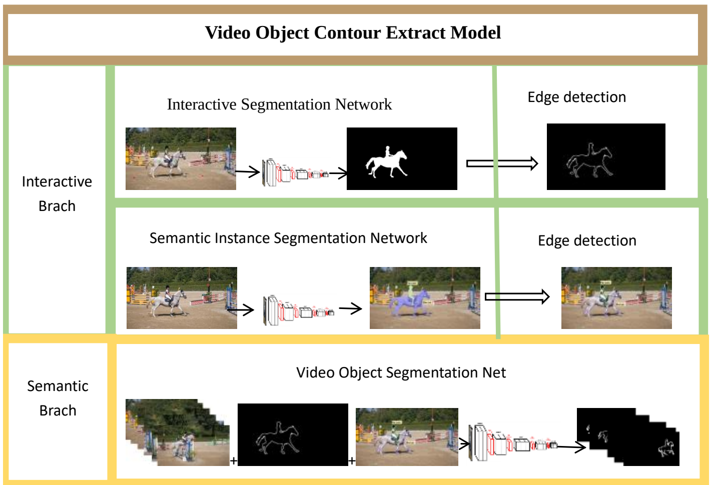
Video Object Contour Extract Model
Interactive Segmentation Network
Edge detection
Interactive
Brach
Semantic Instance Segmentation Network
Edge detection
Semantic
Brach
Video Object Segmentation Net
+
+

Fig. 2: Network architecture

## 3.2 Label acquisition

In order to obtain the object mask label of the first frame image, an interactive segmentation network is added by us. The paper[5] was chosen as our interactive object selection algorithm. This algorithm is different from most interactive selection algorithms that currently exist, and it combines user interaction with deep learning techniques. The user performs a small number of clicks on the object of interest and the background, and the green positive sample point and the red negative sample point are produced on the image, and then the positive and negative sample points are converted into Euclidean distance maps, respectively. The positive and negative Euclidean distance maps are connected to the RGB channels of the image to form (image, user interaction pairs). Image and user interaction pairs are entered into the FCN network. The FCN model randomly samples the generated sample pairs, and then the network is gradually fine-tuned. Combined with Graph Cut optimization algorithm, the target boundary is accurately positioned, and the final target binary mask image is output by the network. The interactive segmentation network architecture is shown in Figure 3.

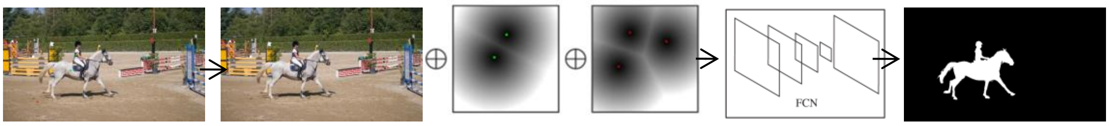

Fig. 3: Interactive segmentation network

For getting the target's contour label, we apply the edge detection operator to process the binary segmentation mask. In this model, we use the Canny edge detection operator for processing. The calculation process of the Canny operator is as follows. First, the gradient intensity and direction of each pixel in the image are calculated by Gaussian filtering. Then, non-maximum suppression is applied to eliminate the spurious response caused by edge detection. Finally, the double threshold is applied to detect real edges and potential edges, and the target contour mask is obtained by suppressing the isolated weak edges.

Meanwhile, for improving the accuracy of video contour extraction, a semantic segmentation network is added to our model. The semantic segmentation algorithm we applied is the Mask-RCNN algorithm. As we all know, the mask branch was added to the Faster-RCNN[14] algorithm to form the Mask-RCNN. Therefore, the binary mask generated by the Mask-RCNN contains the image's positioning and category labels. Then the contour is extracted from the binary mask by the Canny operator. The contour mask at this time includes the positioning, the object category, and the contour information of all objects on the image. For the sake of distinction, we refer to the contour obtained by this process as the semantic contour mask. And the contour mask of a particular target obtained by interactive segmentation is called an interactive contour mask. (Note that the semantic contour mask is not a substitute for the interactive contour mask, the former obtains the contour of all targets in the image is not the contour of a particular object). The sketch of the branch is shown in the green table of Figure 2.

## 3.3 Video object contour acquisition

The OSVOS is an effective video object segmentation algorithm, which is applied to our model to extract video object contour. The VGG16 framework was applied to the OSVOS model. In our model, the filter size was reduced, the number of channels was greatly increased, more information could be extracted, and higher accuracy was achieved. In addition, the fully connected layer for classification was removed in the network model, and the memory and time costs during training and testing were significantly reduced. The improved video object segmentation network mainly includes three successive training processes. The network architecture shown in Figure 4.

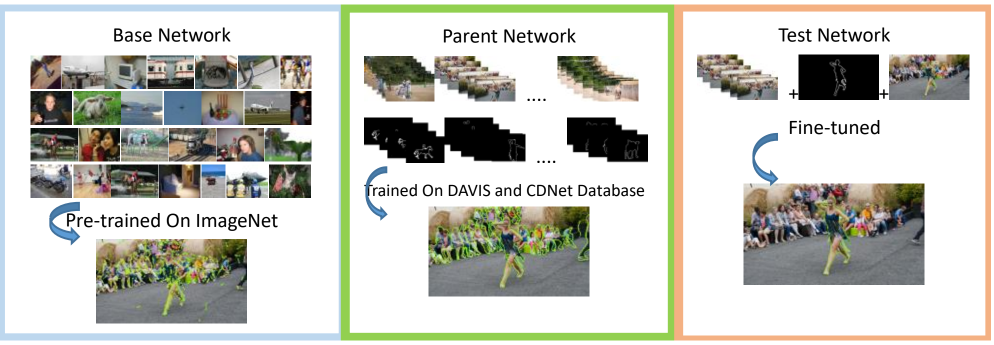

Fig. 4: Overview of our semantic branch

First, our FCNs are pre-trained on ImageNet to learn to mark images. After learning, the target contours are not segmented, while a priori is provided for subsequent contour extraction. During the parent network learning phase, our network is trained on the DAVIS2016 and database2014[15] training sets. All labels used in the parent network are contour labels. And these labels were extracted from segmentation mask of data sets by edge detection. During the test network phase, the video and the first frame of the interactive contour mask and the semantic contour mask are entered, the network is further trained and fine-tuned, and all contours of the particular object are output by the network.

### 3.4 Running process

In general, our video contour extraction model runs as follows. First, any frame of the video is input separately the interactive network and the Mask-RCNN network. Through optimization and edge detection, the interactive contour mask and semantic contour mask are obtained. And the both and a video are simultaneously input to the video object segmentation network, and all frame contours of the video are obtained.

## 4. Experimental Verification

In order to verify the performance of the proposed algorithm, we selected several dozens of natural scene video and surveillance video in complex background for experimental verification. The prospects for these videos include animals, people, cars, and so on. The video includes dynamic background, camera shake, bad weather, intermittent motion of objects, and more. Part of the frame extraction effect is shown below.

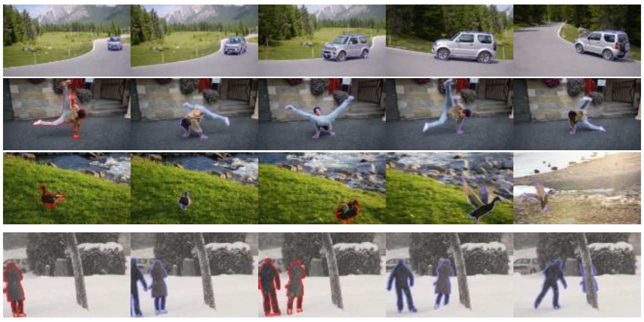

Fig. 5: Edge extraction effect diagram

In Figure 5, we show the contour extraction of the car, animal, single character and double in different scenes. Among them, the contour mask input by this model is represented as a red contour, and the subsequent output frames are represented as a purple contour. The first row and the second row respectively show the contour extraction results of some frames in the car and dancer motion in the scene with little background change. The third row shows video sequence of the swan from still to takeoff in a dynamic scene where the illumination changes. In order to achieve a better contour extraction effect, we used two marker frames, such as the first and third columns of red contour frames in the third row. The fourth line is a skiing scene from single to double in bad weather. We also use two mask frames, such as the first and third columns of the fourth row. Our model was tested on the DAVIS and database2014 test sets. The length of video to be tested varies from tens of seconds to ten minutes, and the foreground contour of a specific target can be effectively and quickly detected in the case where only one frame of the marker frame is given. In the case where two video marker frames are given, the dice coefficient of the extraction result is increased by several percentage points.

At present, the video foreground contour extraction algorithm based on deep learning is the first one. Since the contour extraction and edge detection algorithms both process the boundary of the image, we compare the proposed algorithm with several classic edge detection algorithms, such as Sobel operator[16] and Canny operator [17].

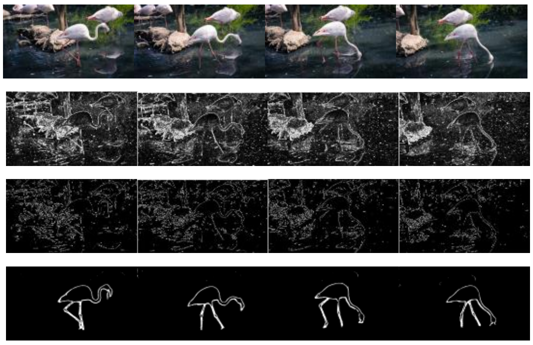

Fig. 6: Comparison of Canny operator, Sobel operator and our operator for multi-target video scenes

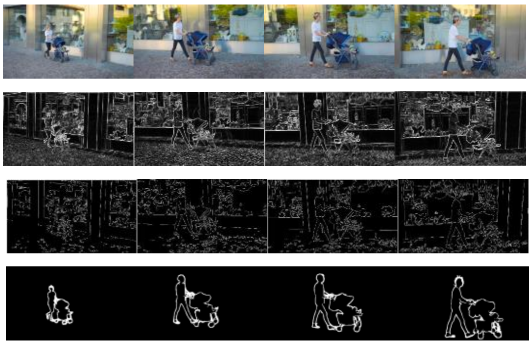

Fig. 7: Comparison of Canny operator, Sobel operator and our operator for videos with multiple objects in the label

In Figure 6 and Figure 7, the first row is a randomly selected experimental video frame, the second row is edge image sequence of Sobel operator corresponding to each frame of the first row, and the third row is edge detection image sequence of Canny operator, and the fourth row is edge extraction image sequence of our model. Through the obvious change of brightness at the edge, the filter operator is used to convolve with the image to extract the edge of the image, which is the characteristic of the Sobel operator and the Canny operator. The edges of the image are extracted from these operators, including the outer contour and texture details of the target object, but also the background contour and details. Background contours and details are not the information we care about, so the extracted contours have a lot of redundancy and noise. The plants and lakes in Figure 6 and the doors and windows and the bottom plate are extracted in Figure 7， while the algorithm proposed in this paper only extracts the object contour. In particular, there are many similar objects in the video image sequence, as shown in Figure 6. When we only need to extract the contour of an object of interest, the algorithm can lock the target well for contour extraction, but the edge detection operators Sobel and Canny cannot.

## 5. Conclusion

In this paper, the one-time learning ability of the machine is used in contour extraction, and a semi-supervised video object contour extraction model based on deep learning is introduced. The video and its one-frame image outline mark are provided by us, and the network can automatically output all the contours of the object in the video. In order to improve the practicability of the model, an interactive object selection algorithm based on deep learning is applied in this paper, and we use it to extract the first frame contour of the video. Several clicks are provided by the user, the foreground target and background will be separated, the Canny edge detection operator is applied to the segmented image, and the target contour information will be extracted through the edge detection operator. The target contour label is input into the single frame mark video segmentation algorithm. The algorithm combines the first frame image and the target contour label to learn the target semantic information. The neural network transmits the semantic information to the subsequent frames, and the contours of all the video frames are extracted. The contour extraction model proposed in this paper has strong robustness. The semantic selection of the first frame provides the appearance and semantic prior to the contour extraction of subsequent frames, which helps the quality of the whole video to be maintained for a long time.

## 6. Acknowledgments

This work was supported by the National Natural Science Foundation of China (61862062) and the National Natural Science Foundation of China Youth Science Fund Project (61104035). We gratefully acknowledge all supporter for this research.

## 7. References

[1] Urata, S., Yasukawa, H.2012. Improvement of contour extraction precision of active contour model with structuring elements. In IEEE Interna- tional Conference on Acoustics, Speech and Signal Processing (ICASSP)

[2] Zhou, X, Wang, P., Wang, C.Q. 2013. GVF Snake Based Target Contour Extraction in SAR Imagery. In Seventh International Conference on Image and Graphics.

[3] Liu, C. Y., Wang,Y., and Wang, S.Y. 2018. Target Detection Based on Saliency Analysis and ContourExtraction for Synthetic Aperture Radar Images. In IEEE International Geoscience and Remote Sensing Symposium.

[4] Caelles, S.,Maninis, K. K., Pont-Tuset, J. et al. 2017.One-shot video object segmentation. In CVPR.

[5] Xu, N., Price, B., Cohen, S., et al. 2016. Deep interactive object selection. IEEE Conference on computer vision and pattern recognition.

[6] Rother, C., Kolmogorov, V. and Blake,A. 2004. “GrabCut” — Interactive Foreground Extraction using Iterated Graph Cuts. SIGGRAPH '04 ACM SIGGRAPH , 309-314

[7] Gulshan, V., Rother, C., Criminisi, A., et al. 2010. Geodesic star convexity for interactive image segmentation. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 3129–3136.

[8] Park, S., Lee, H. S. and Kim, J. 2016. Seed growing for interactive image segmentation with geodesic voting. IEEE International Conference on Image Processing (ICIP).

[9] Ong, H. T. and Ma.K.K. 2011. Semantic image segmentation using oriented pattern analysis. In 8th International Conference on Information, Communications & Signal Processing

[10] He,K., Gkioxari,G. and Doll ́ar, P.,et al. 2017. Mask R-CNN. In ICCV.

[11] Girshick,R., Donahue,J., Darrell, T.,et al. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on computer vision and pattern recognition.

[12] Farabet,C., Couprie,C., Najman, L.,et al. Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(8):1915– 1929, 2013.

[13] Hariharan,B., Arbel ́aez, P., Girshick,R., et al. 2014. Simultaneous detection and segmentation. In ECCV.

[14] Ren, S., He, K., Girshick. 2015. R. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS.

[15] Wang, Y., Jodoin,P. M., Porikli, F., et al. 2014. CDnet2014:An Expanded Change Detection Benchmark Dataset, in Proc.IEEE Workshop on Change Detection(CDW-2014) at CVPR-2014, 387-394.

[16] Sun, L.H., Zhao, E. L., Ma, L., et al. 2014. An edge detection method based on improved sobel operator. Advanced Materials Research. 1529-1532

[17] Canny, J. 1986. A computational approach to edge detection[J]. IEEE Trans. Pattern Anal. Mach. Intell, 679-698.