Instance Level Object Segmentation in Videos
Project Proposal for Visual Computing Course, Ecole CentraleSupelec, Paris - Spring 2019

Ayush K. RAI
Ecole CentraleSupelec, Paris
ayush.rai2512student-cs.fr

Kai-Wei TSOU
Ecole CentraleSupele, Paris
kai-wei.tsou@supelec.fr

Motivation and Problem Definition

Visual Understanding of complex urban street scenes at an Instance level is an open research problem in computer vision. In the recent times the Instance Level Object Segmentation task has received huge amount of attention not only from the computer vision research community but also from research and development departments of many giant tech firms working in the field of Autonomus Driving. Infact many large scale publicly available datasets like CityScapes[2] and Apolloscape [6] have been released to help the researchers address this problem. Last year a Kaggle Challenge on Video Segmentation was organized as a part of 2018 CVPR Autonomous Driving Workshop. We strongly believe that this task of Instance Level Object Segmentation holds a high academic and research value. ***Therefore we propose to work on this problem as a part of our course project in Introduction to Visual Computing Course offered at Ecole CentraleSupelec, Paris - Spring 2019.***

1. Methodology

Alot of work has been done in the field of Instance Segmentation in images involving graph based image segmentation methods like normalized cut, applying exact and approximate inference techniques by modeling the images as Markov Random Fields (MRF) and Conditional Random Fields (CRF) and deep learning based methods especially in the recent times. However Instance segmentation in videos is an evolving problem and has been less studied as compared to Instance Segmentation in Images. In our project we plan to begin with Instance Segmentation based methods on Images and extend it to sequence of frames i.e videos. Based on our literature survey we propose following methods:

1.1. Graph Based Segmentation

The first approach we consider is to apply normalized cut [10]. The advantage of normalized cut is that it considers not only the cut cost but also the total edge connections to all the nodes within a group. In addition, compared with neural network, the computation of normalized cut is relatively cheap. The objective function of two-class normalized cut is shown in equation 1:

$$\begin{aligned} N_{cut}(A,B) &= \frac{cut(A,B)}{asso(A,V)} + \frac{cut(A,B)}{asso(B,V)} \\ &= 2 - \left(\frac{asso(A,A)}{asso(A,V)} + \frac{asso(B,B)}{asso(B,V)}\right) \end{aligned} \tag{1}$$

where A, B are two partitions set, V is a set of all nodes and:

$$cut(A,B) = \sum_{u \in A, v \in B} w(u,v) \tag{2}$$

$$asso(A,V) = \sum_{u \in A, t \in V} w(u,t) \tag{3}$$

2. Deep Learning Based Methods

2.1. U-Net

The U-net architecture [9] has achieved very good performance on image segmentation applications. So we decide to adopt U-Net to solve video segmentation problem. The original U-Net architecture is similar to the convolutional auto-encoder architecture. The encoder contains repeated application of 3×3 convolution layer followed by ReLU, and 2×2 max pooling layer. After each convolution layer, U-Net double the channel size. Then The decoder up-sample the obtained encoded map with repeated 3×3 deconvolution layer and 3×3 convolution layer. Note that the U-Net also concatenate the corresponding feature maps from the encoder layer with up-sampled feature maps at each deconvolution operation in order to better localize and learn representations. The overall architecture of the U-Net is shown in figure 1.

2.2. Region Proposal Based Networks

Another model we consider is R-CNN [4] and its variants [3, 8, 5]. R-CNN contains several stages of process.

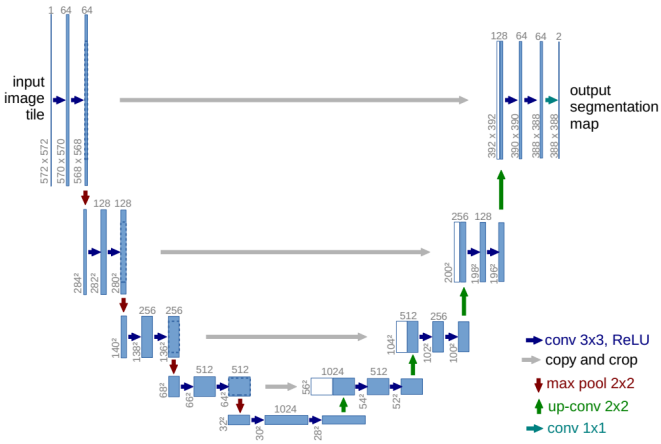

Figure 1. The architecture of U-Net.

First, the R-CNN would select several possible regions (about 2000 regions), and then use pre-trained CNN, such as AlexNet, to extract features from the selected regions. Next, a SVM is used to classify the region based on the extracted features. Finally, we adopt another regression model to detect the location of bounding box. The framework of R-CNN is shown in figure 2. We might also try to adopt variant models of R-CNN, such as faster R-CNN and mask R-CNN.

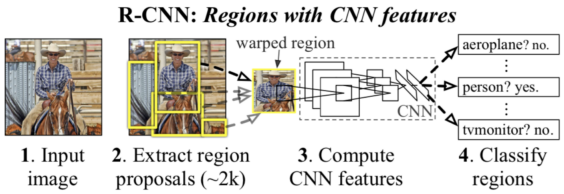

Figure 2. The framework of R-CNN.

In our project, we will mostly focus on deep learning based methods which are currently the state of art on standard datasets. If we have enough time then we will also dive into strategies to use the temporal information in the subsequent frames of the video to achieve better results like [1]

3. Evaluation

In the Kaggle Competition, the task is to predict segmentations of different movable objects appearing in the view of a car camera. The **competition dataset (provided by Baidu Inc.)** contains a large number of segmented and original driving images and there are multiple object instances. However in this competition, only seven different instance-level annotations **car, motorcycle, bicycle, pedestrian, truck, bus, and tricycle** are evaluated. The corresponding groups, such as car group and bicycle group are not evaluated currently.

3.1. Evaluation Metric

Interpolated average precision (AP) is used as the metric for object segmentation as mentioned in [7]. The mean AP (mAP) is computed for all the video clips and all the classes for evaluation at different intersection-over-union (IoU) thresholds. The IoU between a predicted instance A and a ground truth instance B is computed by:

$$IoU(A, B) = \frac{A \cap B}{A \cup B}$$

To obtain the Precision-Recall curve, ten IoU thresholds are chosen in range [0.5, 1.0) with step size of 0.05. After this the ground truth instances are matched with predicted instances at different IoU thresholds. For example, given an IoU threshold 0.5, a predicted instance is considered as matched if the IoU with a ground truth instance is greater than 0.5. If there are multiple predicted instances matched to a ground truth instance, the predicted instance that is larger than the IoU threshold and has the largest confidence is considered as the true positive, and remaining predicted instances are false positives. The predicted instances that are not matched with any ground truth instances are counted as false positives. If IoU between a predicted instance and ignoring labels is larger than the IoU threshold, this predicted instance is removed from the evaluation.

References

[1] K. Chen, J. Wang, S. Yang, X. Zhang, Y. Xiong, C. C. Loy, and D. Lin. Optimizing video object detection via a scale-time lattice. *arXiv preprint arXiv:1804.05472*, 2018. 2

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[3] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 1

[6] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. *arXiv preprint arXiv:1803.06184*, 2018. 1

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[10] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 1