Carnegie Mellon University

Improving Video Segmentation Using Region Proposals and FCNs

Anubhav Ashok Sree Harsha Kalli

16720
Computer Vision
Fall 2016 poster presentations

Introduction

Recent years have seen great progress in image segmentation, however the naive application of these algorithms to every frame is expensive and ignores the temporal aspect of a video. Efficient and reliable video segmentation plays a key role in autonomous driving systems.

In this project we combine the powerful object segmentation capabilities of DeepMask[2] with the comprehensive semantic segmentation of clockwork-FCN[1] to produce a system that produces accurate segmentation of videos.

Semantic segmentation

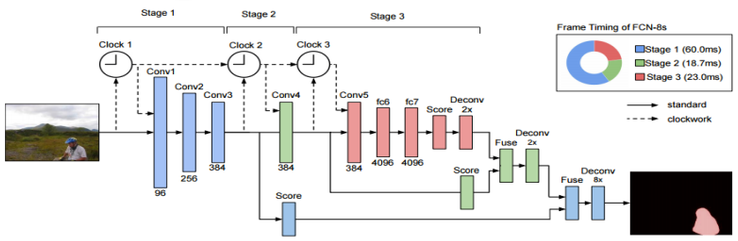

Fig. 2 Clockwork-FCN architecture

The clockwork-FCN[1] framework augments the original FCN[3] for image segmentation method, which segments and image end-to-end, with the clockwork-RNN method to leverage the temporal consistency of a video.

It relies on the assumption that while pixels may change rapidly from frame-to-frame, the semantic content of a scene evolves more slowly.

Object segmentation

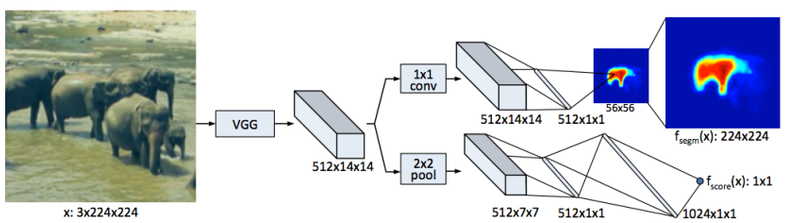

Fig.1 DeepMask architecture

The DeepMask[2] framework uses a convolutional neural network (CNN) to generate object segments in an efficient manner. Multiple region proposals are generated from the input image and then passed through a VGG net to generate features.

These are then passed into the net described in the architecture above to generate a score-mask pair concurrently. The masks are then sorted and thresholded by the scores. The final object segment is created by upsampling the output of the last layer.

Our system

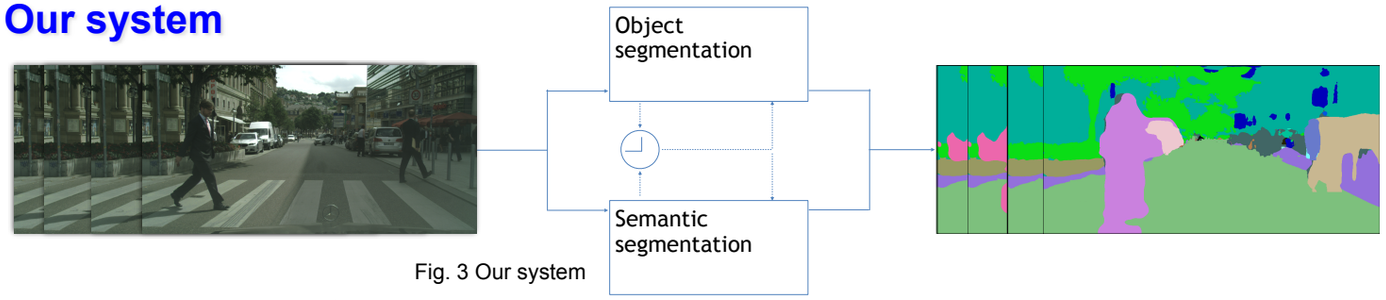

Fig. 3 Our system

While the clockwork-FCN[1] method produced great results for environment classes such as the road, sky and trees, it produced poor results on objects such as cars, people and pets.

On the other hand, DeepMask[2] aims to produce object masks and not segment the entire scene.

Our system retains the best aspects of both clockwork-FCN[1] and DeepMask[2] to produce temporally consistent segments for both objects and scene.

Results

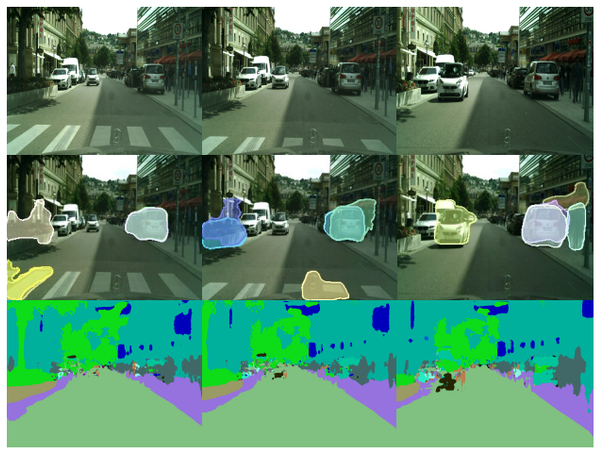

Fig. 4 Standalone output

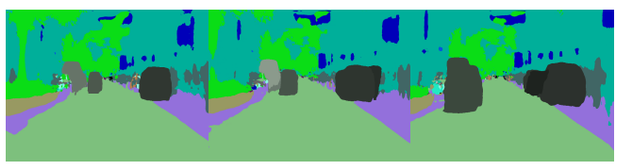

Fig. 5 Output of our system

We compare the results of our system with each of the standalone systems above.
As observed, the combined system segments the road, trees, buildings, people and cars better than each individual one.

Conclusion

In our project we showed the effectiveness of combining semantic and object segmentation modules to produce a video segmentation system.

Some extensions of the system include applying DeepMask selectively across frames and interpolating the intermediate frames to reduce computation or adding a clock to DeepMask that syncs with that of clockwork-FCN to enforce temporal consistency for object segments.

References

1. Shelhamer, Evan, et al. "Clockwork Convnets for Video Semantic Segmentation." 2016.
2. Pinheiro, Pedro O., Ronan Collobert, and Piotr Dollar. "Learning to segment object candidates." 2015.
3. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." 2015.
4. Koutnik, Jan, et al. "A clockwork rnn." 2014.