COMP4801 Final Year Project Plan

ZHANG Qiping, CAI Jinyu

September 28, 2018

1 Objectives

This project aims to design an efficient approach for semantic video segmentation, which allows researchers to accurately segment all objects in the image scenes of consecutive video frames and classify them based on their concept/meaning.

The intermediate goal for this project, is to apply deep learning techniques to perform semantic segmentation of videos captured from in-car cameras based on the state-of-art models and algorithms. Specifically, we will first select a suitable semantic video segmentation dataset to perform training and testing of machine learning models in our further work, and then investigate the high-performance approaches recently developed by other researchers. An evaluation will be made on these models in terms of their efficiency and accuracy, and comparison will be made to check the difference made by their distinctive models, algorithms and system architectures. In this way, advantages and deficiencies of the existing methods could be exposed and verified, and possible ways of improvement and modification could be discovered for further experiment.

The ultimate goal for this project, is to design and train a deep neural network (e.g. a convolutional neural network (CNN)) that can produce temporally coherent semantic segmentation of a video. Different from the general image segmentation tasks, our desired model for semantic video segmentation should also be able to maintain a consistency along time-dimension, besides the spatial correctness and meaningfulness required by object classification in a single 2D image. In other words, a qualified and ideal model for semantic video segmentation should not only correctly classify the pixels of each image frame into real-world objects, but the classification of nearby frames should not undergo an obvious fluctuation. The model to be constructed in this project is expected to outperform the previous approaches in at least one aspect, for example, leading to less computational power consumption or higher segmentation accuracy.

2 Problem Statement

As one of the most traditional and basic fields in computer vision research, semantic segmentation intuitively refers to the process of assigning a class label (e.g., road, tree, sky, pedestrian, car, ...) to each pixel of an image. Great importance is continuously attached to this area since it usually performs as the substantial preprocessing stage in many vision tasks, such as scene parsing and understanding. Generally, semantic segmentation consists of three basic steps: object detection, shape recognition and classification, each of which contains space of further

efficiency improvements. However, when the problem domain generalizes to video instead of a single image, researchers are confronted with more challenges: in brief, as a concatenation of consecutive frames, video segmentation is a more difficult problem because consistency should be maintained between a certain number of neighborhood images, i.e. the nearby frames should not contain greatly distinctive classification patterns with each other. This inherent characteristic has contributed to the obstacles in the design of semantic video segmentation algorithms.

According to most reviews and articles, there are several common problems along with the prosperousness of research in semantic video segmentation. A typical problem that have existed since the involvement of machine learning is the lack of large, complete and representative datasets. Although this problem has been attenuated to some degree with the recent development of reference datasets, which has a set of standardized training and testing method, deficiency remains in terms of their scale and abundance. But these existing improvements have already been able to direct a large number of computer vision scientists into the field of semantic video segmentation. Meanwhile, another emerging problem is that, it is quite time-consuming to construct a dataset with pixel-level labeling and high accuracy (low mis-judgement rate by human labeler) simultaneously. This phenomenon causes semi-supervised and weakly supervised methods to be conceived by researchers and occupy their positions in in this field. On the other hand, for some video-level labeling datasets, location detection accuracy has become a dominant problem, which consequently requires to apply other approaches to preserve system performance, leading to greater cost in system design and development for this kind of dataset. This trade-off basically prevents several graphical models like Markov Random Field (requiring high labeling accuracy) from being widely used for datasets belonging to this category.

Furthermore, even though it has been several years since semantic video segmentation began taking an important position in machine vision study, changes in the appearance of objects such as angle and direction, size and scale, blurring and reduced quality, camouflage of objects in the environment, objects overlap, etc. are still among the main difficulties in this area. Nevertheless, this problem also got solved to some certain extent with the development of deep neural networks (DNNs), which has helped boost system reliability to a great deal in recent researches. Also, with the increase of computational power, which leads to more layers in current DNNs, a further enhance in segmentation accuracy could be achieved. With the approaches and tools developed in DNNs, convolutional neural networks (CNNs) accomplished superior performance over other methods in feature extraction, resulting in its important role in semantic video segmentation in view of present researchers. Several derivative models from CNNs, such as fully convolutional network (FCN) have also been introduced, to obtain a large map of the labels by tagging every small part of the image.

Nowadays, although the problems described above have found several solutions that gained great success in semantic video segmentation, there is still a certain distance from this research area to complete maturity. A series of problems still remain in this field of study, for example, the difficulty in segmenting the blurred objects in the video due to the wrong focus of the camera, and the vulnerability encountered while trying to classify an object in a different point of view with that of the training dataset. Considering all these problems in semantic video segmentation, our project aims to compare and modify the previous state-of-art models, and develop a new approach that could mitigate their deficiency and obtain a better overall performance.

3 Background

The basic background for this project could be briefly summarized into 3 perspectives: computer vision, machine learning, and deep neural network. As an interdisciplinary field that studies how machines can be programmed to obtain high-level comprehension from digital images or videos, computer vision is concerned with the theory behind artificial systems that extract information from images, where video sequences are an extremely important source of data. Due to this reason, semantic video segmentation naturally becomes an essential preprocessing stage in many computer vision tasks, to classify and identify the significant objects in continuous video frames. Therefore, in order to fully understand the target and the methodology of this project, the fundamental concepts and knowledge regarding to computer vision will be the prerequisites of our work. Moreover, as is elaborated in the objective and problem statement section, the introduction of machine learning approaches to image segmentation tasks has efficiently and successfully contributed to the enhance of system in balanced manner through training. This explains why most of the latest research on semantic video segmentation are based on deep neural networks, convolutional neural networks and their generalization. As a consequence, background knowledge about machine learning and DNNs/CNNs will also play a non-negligible role in this project.

For the technical background of this project, several most popular deep learning frameworks, such as Caffe and PyTorch will be applied to finish the training and testing task of deep neural network. As these deep learning platforms enable fast, flexible experimentation and efficient production through a distributed training, and ecosystem of tools and libraries, a great number of latest research works are based on these frameworks. Since one of our tasks in the early stage of this project is to reproduce the outcomes of these previous researches and make performance assessment, applying these machine learning platforms will bring about convenience to both the intermediate study and the modification later on. With plenty of open-source packages and online tutorials/documentations, it is hopeful to quickly get started with our coding and experiments. Meanwhile, due to the inherent features of these deep learning frameworks, Python and C++ will be the major programming languages used in this project, where Linux will become our main developing environment accordingly. The GPU server (1080 Ti/Titan) provided by supervisor will give the necessary computational power needed in the training/testing of neural networks, and other data processing tools, such as Matlab, will also be utilized based on actual requirement.

4 Literature Review

According to [22], the latest researches on semantic video segmentation mainly focus on three aspects: (1) Input of Semantic Segmentation Systems; (2) Feature Extraction; (3) Modeling and Classification.

For system inputs of semantic video segmentation, [5, 4] make use of binary inputs in their models. On the other hand, systems presented in [15, 27] are established on basis of multi-class inputs. In this case, the quality and accuracy of training mainly depend on the number of existing categories provided in training dataset. Meanwhile, as most researchers have been applying common RGB videos for training [12, 15], there are still a certain number of them including geographical coordinates to obtain a higher segmentation precision [18, 1].

In terms of feature extraction, [15, 26] applies the method called super-voxels, in which their

algorithms firstly try to detect smallest component of a video able to be considered as a 3D structure of images, and then extract the defined features from these detected super-voxels. More traditionally, a majority of research works make use of fundamental hand-craft features, such as pixel color features of video frames [14, 18], histogram of oriented gradient (HOG) [12, 25], appearance-based features (e.g. texture features) [21, 13] and 3D optical flow features [18]. On the other hand, with the recent development of deep learning approaches, application of pre-trained models on CNNs has also become popular method for extraction of automatic features from input dataset [29, 11].

Lastly, for modeling and classification, the extant approaches could mainly be divided into the following categories: a. Unsupervised Methods, such as clustering algorithms [16], graph-based algorithms [8, 28] and random walk algorithms [1, 3]; b. Support Vector Machine (SVM) [10, 13]; c. Random Decision Forest (RDF) [7, 21]; d. Markov Random Field (MRF) [23, 17]; e. Conditional Random Field (CRF) [6, 14]; f. Neural Networks, including traditional neural network [24, 20] and deep neural networks (DNNs, which is further composed of several generalized models like CNNs, RNNs and FCNs [9, 29, 11, 2]). All the models and methods described above consist of numerous subdivisions, which is beyond the scope of literature review in this project plan.

5 Approach and Methodology

This project will use the mainstream machine learning platform PyTorch developed by Facebook's AI research group to carry out model training and evaluation experiments. Because of PyTorch's flexibility in modifying neural networks architecture, concise code for manipulating training strategy and popularity among the machine learning community, it allows us to run existing examples proposed by other researchers which are mainly open-source PyTorch projects. In this way, we can identify key drawbacks of semantic segmentation and make improvements to tackle some particular obstacles. Besides PyTorch, some open source computer vision libraries like OpenCV will be leveraged for the purpose of processing dataset. These libraries provide standard and performance-optimized codes for commonly used features, for example, reading and transforming images. In terms of hardware, this project will take the advantages of GPUs for training models. GPUs pipelines work parallelly and thus make the training process efficient enough for complex model architecture and large dataset.

6 Deliverable

Dataset is a crucial part of a machine learning project as machine learning is mainly about learning the patterns of a given dataset. However, this project will not include collecting its own dataset, but employ existing open source benchmark dataset. One reason is that collecting data and annotating the ground truth require considerable human labor work, which is hard for a final year project. Another reason is that, open source datasets are well organized and accurately annotated. Many research teams work on them and thus they can provide benchmarks for evaluating model performance.

As machine learning is data-oriented, the final output model will only work with similar types of data as the training data. For example, if we train the model to segment a road condition video, it will not work if the input videos are of other scenarios such as a medical detection video inside

a human body. Thus, this project will avoid emphasizing on the final output of the model, but focusing on the techniques our model uses to manipulate the video and the network architecture of the model. Particularly, this project will try to find techniques to achieve low-latency and at the same time maintain satisfactory accuracy level. These innovations will serve for general semantic segmentation problems and thus will be a highlight of this project.

7 Feasibility Assessment

With the emerging of deep neural network and machine learning, recent years have witnessed fast improvement on semantic segmentation. However, challenges still remain for real-time video segmentation as it requires far more computational resource than a single picture. One advantage we can take of is the similarity of the neighbouring frames, which enables us to use a shared feature to predict multiple neighbouring frames at the same time. However, it is also a challenge to handle the subtle differences between them. Another possible risk of this project is the limited computational resource. [19] spent 6 days training a model on a single Nvidia GTX Titan X GPU with 12G memory. Typical training requires multiple GPUs to accelerate the process. However, this project, together with other three final year projects, shares only four GPUs. Considering it takes multiple trials to verify a proposed techniques, the limited computational resource and time may be a potential risk of this project.

8 Schedule

This project is cooperated by Zhang Qiping and Cai Jinyu. Both will participate in brainstorming and carrying out experiments and will share a fair part of this project. However, Cai will focus on the implementation of ideas as he is familiar with the tools. Zhang will focus on setting up different experiment strategies as he has more research experience. As is shown by Fig.1 This project starts in early September, with the project plan proposed by the 30th of September, 2018. Then one month will be spent on literature review, research tools and data preparation. And between November 1st, 2018 and March 30th, 2019, there will be two phases of experiments and analysis, each taking around two and a half months. Phase II experiments will be modifying and optimizing our models based on the progress and results of phase I experiments and feedback from the supervisor. An interim report will be submitted after finishing phase I experiments. And after finishing phase II experiments, one month will be left for finalizing the final report and presentation on April 30th, 2019.

9 Conclusion

As a fundamental field in computer vision, semantic segmentation has been greatly improved with the help of machine learning and neural network. However, video-based semantic segmentation still remains challenging because of its nature of high throughput and demand for coherence. This project will rely on commonly used machine learning framework PyTorch to improve the efficiency and accuracy of video semantic segmentation. Model training and performance evaluation will be carried out on benchmark data set for its accurate annotation and convenient packaging. Although related studies have shown potential in decreasing latency without sacrificing accuracy,

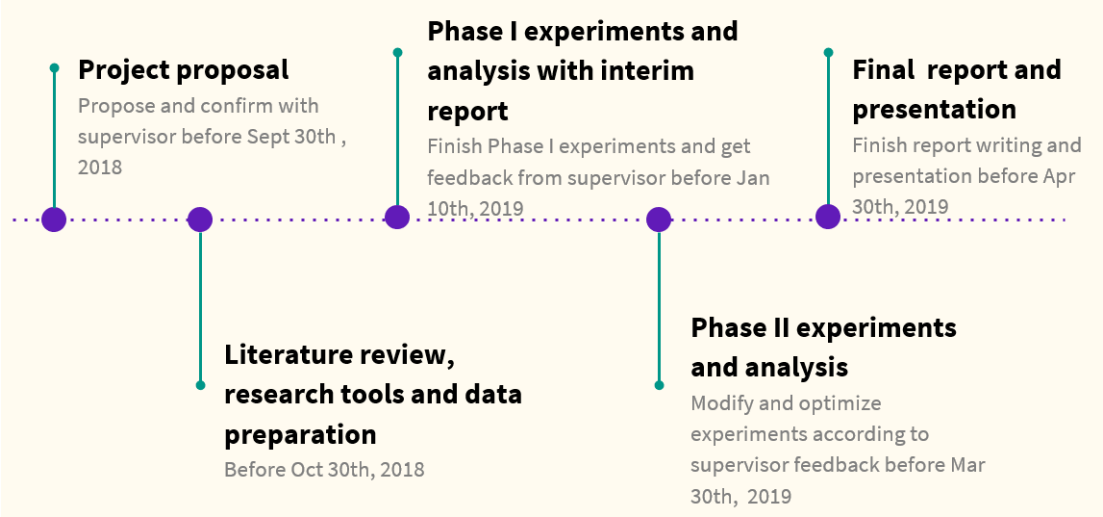

Figure 1: Project Timeline

risks still exist from the limitation of computational resources allocated to this project. This project will be two-phased and is expected to finish by late April.

References

[1] Shervin Ardeshir, Kofi Malcolm Collins-Sibley, and Mubarak Shah. Geo-semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2792–2799, 2015.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[3] Gedas Bertasius, Lorenzo Torresani, X Yu Stella, and Jianbo Shi. Convolutional random walk networks for semantic image segmentation. In *CVPR*, pages 6137–6145, 2017.

[4] Sebastian Bittel, Vitali Kaiser, Marvin Teichmann, and Martin Thoma. Pixel-wise segmentation of street with neural networks. *arXiv preprint arXiv:1511.00513*, 2015.

[5] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3241–3248. IEEE, 2010.

[6] Feng-Ju Chang, Yen-Yu Lin, and Kuang-Jui Hsu. Multiple structured-instance learning for semantic segmentation with uncertain training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–367, 2014.

[7] Björn Fröhlich, Erik Rodner, Michael Kemmler, and Joachim Denzler. Large-scale gaussian process multi-class classification for semantic segmentation and facade recognition. *Machine vision and applications*, 24(5):1043–1053, 2013.

[8] Fabio Galasso, Margret Keuper, Thomas Brox, and Bernt Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–56, 2014.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra 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.

[10] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[11] Yang He, Wei-Chen Chiu, Margret Keuper, and Mario Fritz. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Anna Khoreva, Fabio Galasso, Matthias Hein, and Bernt Schiele. Classifier based graph construction for video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–960, 2015.

[13] M Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller. Parameter estimation and energy minimization for region-based semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1373–1386, 2015.

[14] Buyu Liu and Xuming He. Multiclass semantic video segmentation with object-level active inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4286–4294, 2015.

[15] Xiao Liu, Dacheng Tao, Mingli Song, Ying Ruan, Chun Chen, and Jiajun Bu. Weakly supervised multiclass video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 57–64, 2014.

[16] Yang Liu, Jing Liu, Zechao Li, Jinhui Tang, and Hanqing Lu. Weakly-supervised dual clustering for image semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2075–2082, 2013.

[17] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015.

[18] Andelo Martinovic, Jan Knopp, Hayko Riemenschneider, and Luc Van Gool. 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2015.

[19] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[20] Soo Beom Park, Jae Won Lee, and Sang Kyoon Kim. Content-based image classification using a neural network. *Pattern Recognition Letters*, 25(3):287–300, 2004.

[21] Daniele Ravì, Miroslaw Bober, Giovanni Maria Farinella, Mirko Guarnera, and Sebastiano Battiato. Semantic segmentation of images exploiting dct based features and random forest. *Pattern Recognition*, 52:260–273, 2016.

[22] Mohammad Hajizadeh Saffar, Mohsen Fayyaz, Mohammad Sabokrou, and Mahmood Fathy. Semantic video segmentation: A review on recent approaches. *arXiv preprint arXiv:1806.06172*, 2018.

[23] Abhishek Sharma, Oncel Tuzel, and David W Jacobs. Deep hierarchical parsing for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 530–538, 2015.

[24] Chih-Fong Tsai, Ken McGarry, and John Tait. Image classification using hybrid neural networks. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 431–432. ACM, 2003.

[25] David Varas, Mónica Alfaro, and Ferran Marques. Multiresolution hierarchy co-clustering for semantic segmentation in sequences with small variations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4579–4587, 2015.

[26] Chenliang Xu and Jason J Corso. Evaluation of super-voxel methods for early video processing. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1202–1209. IEEE, 2012.

[27] Yi Yang, Sam Hallman, Deva Ramanan, and Charless C Fowlkes. Layered object models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1731–1743, 2012.

[28] Luming Zhang, Mingli Song, Zicheng Liu, Xiao Liu, Jiajun Bu, and Chun Chen. Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1908–1915, 2013.

[29] Yukun Zhu, Raquel Urtasun, Ruslan Salakhutdinov, and Sanja Fidler. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4703–4711, 2015.