Final Year Project

Project Plan

Weakly Supervised Localization with Temporal Information for Assembly Video Understanding

Supervisor: Professor Wenping Wang

Wang Jiarui: 3035332153 Yang Yafei: 3035331501

27 - Sep - 2019

Contents

1 Introduction

Despite the fact that the exploration of human-robot collaboration can be traced back to the moment when robot came out, it is still a promising field especially with the springing up of many deep learning technologies. The rapid development of robotics puts higher request forward the learning ability and efficiency. Instead of guiding robots with explicit human-coded programs, it is more desirable to let them learn the necessary information themselves with easily obtained data.

The real life motivation for our study is to implement a supervisory system to teach, monitor and guide workers on assembly lines. It is a common practice for robot systems to learn from images with category labels and bounding boxes which has been well explored in the work related to supervised object detection [3, 1, 8, 6, 7, 2]. However, it takes a lot of effort to annotate images with precise bounding boxes which is infeasible in the above scenario. Instead, demonstration videos with phase-based labels are more accessible and common in such industry.

In light of that, our work would like to make use of the implicit temporal information in videos instead of focusing on the image level spatial information only. Inspired by some works that use temporal related approaches [8, 6], we would like to feed videos with only phase-based object labels to networks. The intuition behind is to let implicit temporal information serve as free supervision signals [9, 5] to compensate the absence of localization ground truth. Ideally, after trained with simply labeled demonstration videos, the system is able to recognize all assembly phases and localize the objects of interest in each phase so as to guide and monitor assembling work.

2 Objectives

The above application requires the system to temporally identify current assembly phase and spatially locate corresponding objects of interest. So the main focus of our study will be on the evaluation and improvement of classification and localization ability of the system. we will develop a system to achieve localization in weakly supervised manner.

We will apply state-of-the-art networks to get deep features for classification and make use of class activation mapping [9] for visualization. Using the visualized discrimination patch as localization input, we will add temporal information into the network as a free signal for supervision [5]. In this way, we suppose localization performance can be largely improved.

3 Methodology

Class Activation Mapping [9] will be used to visualize deep features learned and also to evaluate localization performance. Temporal coherence and cycle consistency of

time [5] will be applied to further supervise the training and thus improve localization accuracy.

3.1 Visualization of deep features with CAM

Zhou et al. [9] introduce the concept of class activation map (CAM), which is the discriminative image region used by convolutional neural networks (CNNs) to identify the presented image. The most critical step is performing global average pooling (GAP) on the final convolutional feature maps which are then sent to a fully-connected layer to generate output scores. Next step is to map back the weights of the fully-connected layer to the last convolutional layer. Different weights correspond to different colors on heat maps so that the level of importance can be visualized. In this way, we can also have a glance of the localization ability of the network by comparing the discriminative regions with object locations.

3.2 Temporal information

Although it has been proved that the classification network also has localization ability [9], the accuracy is not satisfying enough due to the absence of localization supervision. We will take advantage of implicit temporal information in videos to provide auxiliary supervision signal.

Temporal coherence

It is reasonable to assume that the change in object location is continuous. Thus, it is expected that the position for objects of interest will not vary too much between frames. Based on the theory of temporal coherence, we can penalize the change of deep features of adjacent frames to smooth the localization outcomes. This extra constraint is believed to improve the localization ability.

Temporal Cycle Consistency

Temporal Cycle Consistency (TCC) [5] provides another way to turn temporal information into a supervision signal. The key idea of TCC is to embed patch p and image$I$ into the feature space by encoder ϕ. Forward and then backward the patch to get the cycle consistency loss l_θ. l_θ serves as the loss function to learn embedding network ϕ.

The TCC model is a free supervision technique. We propose to use localization information generated in CAMs as mask for propagation. In this way, connectivity between neighbor frames can be ensured by performing video cycle consistency.

3.3 Data collection

We will train the model on self-collected dataset, which contains video demonstrating the procedure of assembling hard drive, power supply and CD-ROM into a computer case. Demonstration videos will be labelled with the category of objects on phase basis. For example, in the phase of assembling power supply, we will label the corresponding time interval as power supply. We will also collect the videos on different background so as to train our model's generalizability. At the same time, we will provided position information for some testing data to evaluate localization ability.

3.4 Evaluation

We plan to compare the accuracy of localization with models that only make use of CAMs [9, 4]. Moreover, we will also estimate our model with Wang et al. [5]'s work which uses temporal consistency. Models using temporal coherence will be examined as well. At the same time, we propose to test the accuracy in different assembling working scenarios.

4 Schedule and Milestones

Time	Work
Sep, 2019	Review relevant literature. **Milestone 1: Review relevant literature. Finish project plan and build project website.**
Oct, 2019	Review relevant literature. Collect dataset.
Nov, 2019	Test and improve localization with weakly supervised learning in assembly line dataset, namely test CAM model.
Dec, 2019	Test and apply techniques to enforce temporal smoothness of the localization results. **Milestone 2: Provide a system using weakly supervised learning for localization. Submit interim report.**
Feb, 2019	Collect more data and extend our evaluation.
Mar, 2019	**Milestone 3: Evaluate the system. Submit final report and poster design. Prepare presentation.**
Apr, 2019	Final exhibition.

5 Division of Work

Name	Work
Wang	Test and re-implement algorithm aimed at tracking specific part of object in videos.
Yang	Test and re-implement algorithm in object localization in videos.

6 Summary

The ability of our system to learn from demonstration video is a huge progress to the achievement of intelligent processing. Our work aims at implementing a weakly supervised model with accurate localization result under the help of class activation map and temporal information in videos. More details can be viewed at our website: https://i.cs.hku.hk/fyp/2019/fyp19015/.

References

[1] 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.

[2] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *2011 International conference on computer vision*, pages 1995–2002. IEEE, 2011.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[5] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[6] Rong Yan, Jian Zhang, Jie Yang, and Alexander G Hauptmann. A discriminative learning framework with pairwise constraints for video object classification. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):578–593, 2006.

[7] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018.

[8] Lun Zhang, Stan Z Li, Xiaotong Yuan, and Shiming Xiang. Real-time object classification in video surveillance based on appearance learning. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[9] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.