BeaverDam: Video Annotation Tool for Computer Vision Training Labels

by

Anting Shen

A thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science
in

Engineering - Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Kurt Keutzer, Chair
Professor Only Somewhat Important Guy

Fall 2016

Contents

Chapter 1

Introduction

Deep learning applications in recent years have come to require rapidly growing amounts of labeled training data. Often, accuracies can be boosted by adding data as much as by spending years on algorithmic development. For example, on the VOC07 benchmark, Faster-RCNN [1] with VGG-16 was able to eliminate 27.5% of errors in the much older R-CNN [2] backed by an equally old neural network architecture (mAP improved from 58.5 to 69.9). However, simply by including additional data from VOC12 and COCO, 29.5% of the remaining error was eliminated (mAP improved from 69.9 to 78.8). Therefore, for real-world application development, data can be cheaper and more effective than scientists. While many existing tools support image classification – it is even built into Amazon Mechanical Turk (MTurk) – and some tools support bounding box labeling in images, few tools exist for frame-by-frame labeling in videos. VATIC [3] stands out as being one of the best, as not only does it make high quality annotations one of its main goals, but also cost and scalability.

My work borrows and improves upon many concepts and results from VATIC's user studies, but I focus on an additional goal that is extremely important in creating datasets for real applications. That goal is researcher happiness. Although VATIC extensively tested its "User Interfaces", I argue in chapter 2 that both the annotators and the experimenters are users, and the interfaces should be smooth for both when creating a tool.

Then, in chapter 3, I discuss my take on VATIC's User Interface principles for the annotator, and improvements upon them.

I also release all related code for BeaverDam, my video labeling platform, on Github.[1]

Related Work

Static image annotators

Vatic, LabelMe, etc

Things other people cite

[1] http://github.com/antingshen/beaverdam

Chapter 2

Experimenter Interface

Motivation/Intro

2.1 Interface for researcher

Deployment script

Web based admin interface instead of command line, verification

Availability of highly flexible Python based shell

Video extraction after annotation (exception of H264)

2.2 Decoupled modules

Comparison with microservice

Serving videos

Hosting annotator (ansible)

Crowdsourcing platform independence

Tracking module

2.3 Patterns

Django backend

Patterned frontend (MVC, event-based)

2.4 Dependencies

Few dependencies (sqlite)

Easily restore state

Uses newer technologies

2.5 Security

DB Backup

Authentication enables decoupling, web admin

Other standard procedures, HTTPS & HSTS, CSRF, clickjacking

Chapter 3

Annotator Interface

Intro & our method of user studies (not as good as VATIC though)

3.1 Keyframe scheduling & Multiple object annotation

Keyframe viewer when on custom schedule

3.2 Video playback for maintaining identity

Importance of playback

Video loading issue

HTML5 video advantages

3.3 Reducing clicks

Create by default

Object selector

Keyboard shortcuts

3.4 Handling frame exit/enters

Border padding

Allow out-of-border boxes

3.5 Micro vs Macro tasks

Comparison from existing literature

Proposal advocating for micro-tasks in video labeling

Extensible task structure

3.6 Interpolation & Tracking

Chapter 4

Conclusion

Ha! I’m done!

References

[1] 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 (NIPS)*, 2015.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014.

[3] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, pp. 1–21, 10.1007/s11263-012-0564-1. [Online]. Available: http://dx.doi.org/10.1007/s11263-012-0564-1