Tracked Based Semi Supervised Learning

Brian Chung
Stanford University
bpchung@stanford.edu

Jinjian Zhai
Stanford University
jameszjj@stanford.edu

Abstract

In this paper, we tackle the issue of using semi supervised learning in classifying tracking data. Based on the earlier work of Teichman and Thrun, we modify the approach using LIDAR data to extract useful data from a single fixed camera source.

Our tracking data comes from the background subtraction and segmentation of camera video data. Consequently, tracks are labeled and the related information is used in training the object classifier.

The classifier will find a few positive examples, and the resulting objects are further segmented and fed forward to train the classifier. Thus, the classifier and tracking work hand in an EM algorithm style of method.

Future Distribution Permission

The authors of this report give permission for this document to be distributed to Stanford-affiliated students taking future courses.

1. Introduction

Currently, there is a great need for high throughput classification of video data. Rather than manually hand labeling individual video frames, we seek to automate the process through model free segmentation and track classification. In particular, we separate the process by learning into three separate tasks.

After background subtraction and segmentation, a user will hand label a few of those tracks. The set of tracks are fed into a classifier which will also classify further objects as being part of that object class. This new information feeds back into the segmentation/tracking step to glean further class data.

As seen in Teichman's work[1], tracking based semi supervised learning is surprisingly resilient to noise and has an uncanny ability to learn new useful instances of the classes.

Difficulties will arise in optimizing over intersections of tracks.

1.1. Track Identification

Accurate background subtraction will be vital to identify proper tracks. Due to the fixed camera, we can obtain a background model from the data. We will begin with a method of segmentation through bilayer segmentation. Through background subtraction, tracks of objects can be gleaned from each sample. These tracks will be given their appropriate labels.

1.2. Object Classification

The classifier will be an off-the-shelf classifier such as Felzenszwalb[5]. Once new frames of the class are identified, the rest of the track is obtained.

2. Image Segmentation

2.1. Introduction

Despite the multitudes of segmentation methods such as color or texture separation, the subtractions tend to fail when deriving background models from motion based video. These models typically derive a "mean" image based on the set of frames and subtract individual frames by the mean image to obtain the foreground objects[3].

The background subtraction method typically fails for real world data. Items such as trees, birds, and other variant objects can create false positives. In other instances, false negatives arise from collisions or occlusions of the species.

Because of those issues, our group decided to base our work off an implementation of Billayer Segmentation of Live Video by Criminisi, et al.[2]. Crimini's approach uses a probabilistic combination of motion, color, and contrast in a Hidden Markov Model.

2.2. Mathematics

In Criminisi's model, each time step's motion model is augmented in order to favor "coherence" frame to frame and forgoes traditional optical flow estimation:

$$m=(g, \dot{z})$$

, where $\dot{z}=(\dot{z}_1, \dot{z}_2, \cdots, \dot{z}_N)$ is the temporal derivative of pixel array z and $g_i = |\nabla z_i|$ is the spatial gradient of z.

In particular, Criminisi's work utilizes energy minimization, where terms are chosen in order to reduce the fragmentation of objects within the frames:

$$\left(\hat{a}^1, \cdots, \hat{a}^t\right) = \arg\min \sum_{i=1}^{t} E^i$$

, where $E = V^{\mathrm{T}} + V^S + U^C + U^{\mathrm{M}}$. $\left(V^{\mathrm{T}}, V^S, U^C, U^{\mathrm{M}}\right)$ are temporal prior, spatial prior, color likelihood and motion likelihood respectively.

2.3. Implementation

Both back ground subtraction and bilayer segmentation require large amounts of hand labeled training data. Rather, our work uses an implementation of Bilayer Segmentation that only requires initial hand labeled data of the desired species.

The existing implementation only tracks one object throughout the sequence. In order to improve performance, the ability to track multiple objects was added. Furthermore, the implementation did not keep track of an object's trajectory throughout the sequence. This means that an object in one frame did not have a sequence of events in history to compare with.

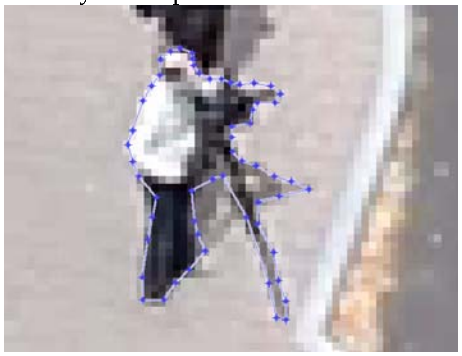


Fig. 1. (a). A pedestrian holding a tripod was masked by a polygon.

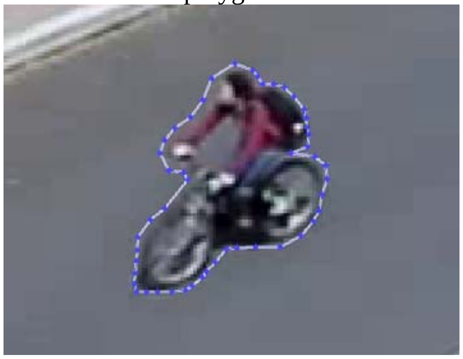


Fig. 1. (b). A bike rider masked by a polygon.

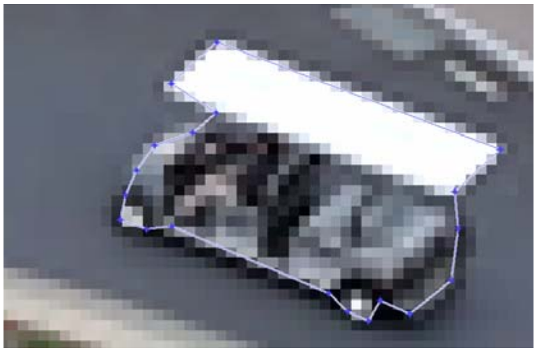


Fig. 1. (c). A golf cart masked by a polygon.

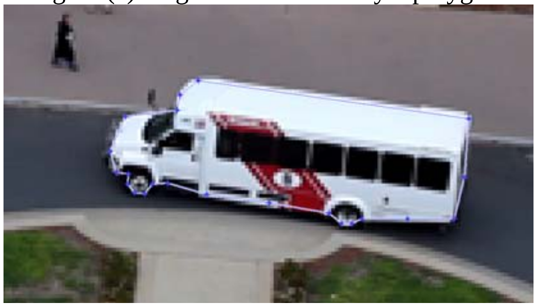


Fig. 1. (d) A bus masked by a polygon.

Our model has the following object classes:
0: pedestrian
1: biker
2: golf cart/mini-van
3: bus/truck

130 minutes of video were fetched from the top of Hoover Tower[4] on Stanford campus. "ff-mpeg"[6] was used to cut the video into movie clips of 300 frames in each batch. The frame per second (fps) is 30 and the clips are 10 seconds each. The video were compressed to 480X240 resolution due to the memory limitation of Matlab. In total 500 clips are used for training and 500 clips are used for testing.

Then matlab code was written to process the video using the energy model of Criminisi. Objects of all four classes are labeled and training data are obtained.

Video processing functions were used in matlab to easily process the avi and mov files. Critical steps are written in cpp file and compiled in matlab using the mex function to save computational time.

Then "linked list" data structure of keeping histories was added, which includes information of the object such as object class, center position, mask information (size, shape, etc.) and times series of locations.

2.4. Segmentation Results

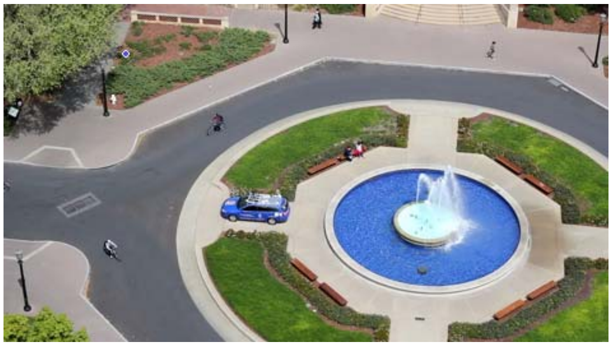
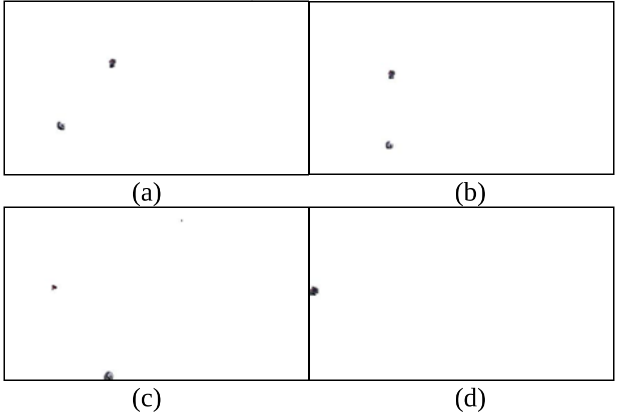


Fig. 2. The original picture of two bikers.

As seen in Fig. 2 and Fig. 3, the objects (bicyclists) are properly tracked.

The segmentation correctly tracks objects despite changes to the frames such as rotation and scale transformations (i.e. bicyclists bike around the circle).



(a) (b)

(c) (d)

Fig. 3. The small circles in the whole picture represent the segmentation and trajectory of two bikers after removing background.

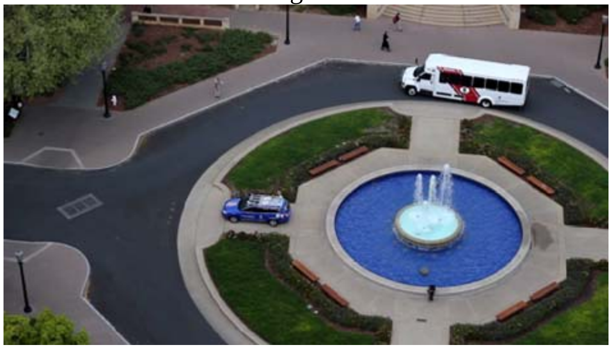
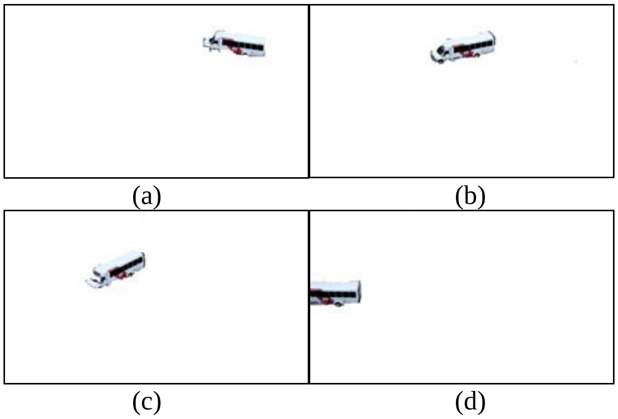


Fig. 4. The scene at 00:06:40.0 of video overhead2-04-07-2011_13-09-21.avi



(a) (b)

(c) (d)

Fig. 5. The trajectory of bus filtered by the mask.

In other cases, the segmentation fails to properly segment the entire object and saves only a portion of it. By adjusting the weights in the energy equation, we will be better able to account for these differences.

3. Object Classification

3.1. Future Work and Evaluation Plan

Consequently, the properly segmented images must be fed into the classifier. Informally, we will have enough (in total of 1000) results for the final paper with classification performance. These metrics will include ability to discover new instances of species (can the classifier find objects in new frames), and feed-forward performance of the track based system (how well does this idea work).

4. References

[1] A. Teichman and S. Thrun. Tracking-based semi-supervised learning. Robotics: Science and Systems (RSS), 2011.

[2] A. Criminisi, G. Cross, A. Blake and V. Kolmogorov. Bilayer Segmentation of Live Video. CVPR, 2006

[3] Y. Wang, P. Perona and C. Fanti. Foreground-Background Segmentation of Video Sequences.

[4] http://robots.stanford.edu/teichman/neovision2/overhead_videos.tar

[5] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, September 2010

[6] http://ffmpeg.org/general.html