Real Time Detection and Tracking of People in Crowds

J L Gouws
Supervisor: Mr. J Connan
Computer Science Department, Rhodes University

April 11, 2022

Abstract

Tracking of objects in videos streams is a powerfull tool in the field of computer vision. It is, however, a relatively unexplored field when compared to other areas of computer vision. That being said, there have been impressive developments in the field in recent years. I propose to use methods of machine learning and image processing to further explore object tracking and detection. The specific focus of my research will be the detection and tracking of people in crowds. Another concern of the research is performance of the implemented system. A large amount of computation power is often needed to analyze videos of crowds, investigation into reducing the required computation power of this system will also be carried out.

1 Introduction

Analysis of videos of crowds of people has many practical applications. Most real world situations that involve people have people that are moving. On top of that most often people are not isolated, but instead form groups of people.

There is huge potential for research in these areas, and a reliable system with high performance would be very useful in industry. For these reasons it makes sense to investigate ways to develop a system that can efficiently and accurately detect and track multiple people in a video stream.

In the past visual tracking has been a largely understudied field, but more recently there has been a significant increase in available tracking systems [1]. Most trackers have niche target areas. A lot of trackers are very fast and accurate, but fail in the long term. Some trackers combat this long term failure [2], but this often comes at the cost of performance [3] [1].

There are ofcourse ethical concerns. As far as research and testing goes only relatively small datasets should be required. This data can be sourced from a group of volunteers, and so ethical concerns should not be a concern for development. But for serious real world applications investigation of larger datasets and an ethical clearance might be required.

2 Reseach Statement

The TLD frame work can be used to develop a system that is capable of detecting and tracking people in a crowd. The system can be designed so as to start with minimal initial offline data and training, and require minimal compute power.

3 Research Objectives

- Using the TLD framework to build a system for idenitifying and tracking people in crowded scenes.
- Investingating ways to improve tracking and classification performance.
- Investigating ways to decrease the computation power required by the system.
- Developing ways to store the system state so that it can maintain it's learned parameters between instances.
- Improving the scalability of the system so that new tracking targets can be added easily.

4 Related Works

4.1 Closely related works

In 2011 Zdenek Kalal invented the Tracking, Learning and Detection(TLD) framework for the longterm tracking of objects in a video stream [2] [4]. Kalal's original implementation uses a median flow tracking stage, P-N learning, and a random forrest based detector. The system requires online(learning as data becomes available) learning in order for the system to work, Kalal developed the P-N Learning paradigm [5], a semi-supervised bootstrapping model [6], tailored to the needs of TLD.

There have been improvements in tracking methods since his development of TLD. Most notably Kernelized Correlation Filters(KCF) being applied on Histogram of Oriented Gradient features[3]. KCFs, however, do not have the ability to detect, and so if they fail they are unlikely to

recover. KCFs also require a stage that will start them tracking. KCFs have great potential to be used as the tracking stage for a TLD tracker.

Kalal uses Random Forest for detection in his implementation of TLD. A more modern approach would be to use Extreme Gradient Boosting(XGB) for the detector. XGB generally offers similar if not better classification performance than Random Forest and requires significantly less time to train [7].

There has also been investigation into the use of Convolutional Neural Networks in tracking [8]. This offers high performance tracking of generic objects. It will be investigated for extension to tracking and detecting people in crowds. It will not be a main point of research unless it turns out to be fruitful–seeing as CNNs tend to require large amounts of training time, data, and memory space.

There has also been some quite old research that involves localizing multiple targets. The research of Taylor and Drummond [9] offer this at high FPS even on low powered devices.

Currently there are vary many trackers availble all with varying degrees of performance. These have been benchmarked in [1] and [10]. This is a good reference, and will be explored further, but not all trackers can satisfy the requirements of this project.

4.2 Less related work

There has also been relevant work done on correlation tracking by [11]. Ma. et al investigated the problem of long term tracking where the target undergoes abrupt motion, heavy occlusions and disapearing from view. There work will probably integrate well with Kernelized Correlation filters, and it has very practical advantages. It is not the forefront of this research to handle occlusions, but it is a possible extension.

The P-N learning system is not completely unrelated to Reinforcement Learning Techniques. There have been promising recent results in online reiforcement learning [12]. The work of [12] allows for variable data budgets, which should integrate well with a constant flux of input from a video stream. TLD may not be compatible with the methods proposed in [12], but the feaibility of the ideas will be explored in the paper.

5 Approach to Research

Initially I will start by reading literature on tracking and facial recognition. This will mainly start by reviewing Kalal's work in TLD, namely his Doctoral thesis [4].

Next I will go into a phase of re-implementing predator in C++, with help from the OpenCV libraries. Once this is stable and running with decent performance, I will move onto the next stage.

Next I will be looking at ways to improve my implementation of the TLD system. I will mainly focus on looking at improving the tracking and detection stages of the system. I will then spend a small amount of time looking at the learning component, but do not intend to completely change it.

Next I will extend the tracking and detection system so that it is able to identify and track multiple targets in a video stream. This will probably require a lot of work and testing. It is important that the system is both reasonably accurate. It is particularly that the system does not get confused and misidentify objects in the start up phase. There is always high potential for confusion in the start up of system, where there are potentially similar looking faces.

Finally I will be looking at ways to enhance the performance of tracking multiple objects. This will come later and mainly be considered and extension. It would be useful to have the system

run on mobile devices or micro-computers, with the potential to be extended to a distributed system.

6 Timeline

Time	Deliverable
30 March 2022	Seminar 1: Presentation of project
11 April 2022	Draft proposal
14 April 2022	Final proposal
18 April 2022	Functional re-implementation of TLD
25 April 2022	Using KCF as Tracking stage of tracker
29 April 2022	Literature review
6 May 2022	Using and XGB classifier for detection
13 May 2022	Reviewing VOT for better trackers
23 May 2022	Final decision on Tracking and Detection stages.
11-13 July 2022	Seminar 2: Progress Presentation
12 August 2022	Extension of system to multiple targets
19 August 2022	Progress Report
26 August 2022	Testing and tweaking the system
3 October 2022	First Draft of thesis
10 Octorber 2022	Completion of implementation
14 October 2022	Short ACM-style paper
17-19 October 2022	Seminar 3: Final Oral presentation
28 October 2022	Final project submission

7 Limitations

It is difficult to accurately identify people from the rear. In this case it might be good to try and prevent the system from learning the appearance of the back of a person's head. A solution to this problem is also developing a system that can be distributed so as to be able to track people from a surrounded view.

It is possible to track semi-ocluded targets, but detecting them is difficult. This will cause problems for the initialization and re-initialization of the tracker. It might be possible to impute some facial features before inputing the stream into the detection system. If a face is fully ocluded, however, there is not much that the system will be able to do, but a human will not be able to do very much.

The system will require relatively good resolution cameras in order to identify people accurately from a long distance. In most situations involving crowds, a camera capturing a video stream will need to be placed a significant distance from the targets. Hence, the best we can do in software is try to use image processing techniques to artificially enlarge the target, or work on detectors that can work on minimal resolution.

There is a problem with access to data that can be trained on, and ethics. Most crowd video data will be impractical to use for testing, owing to resolution and crowd size. There are also ethical issues to be taken into consideration. For testing purposes, I should be able to source my own data, but it will not necessarily resemble real world data.

8 Applications

This reaseach has multiple real world applications. One includes an automated class register that can be used in schools.

Another approach on the larger scale is a security system for monitoring the motion of people. This could enhance airport security for example, maybe being able to label suspected terrorists.

It also has potential to be used as an assistive technology for visually impaired people. If it can run on a mobile device, it might be able to help a visually impaired person identify people at a party or on a television program. This is assuming the system has had the offline training to be able to do so.

References

[1] M. Kristan, A. Leonardis, J. Matas, *et al.*, "The visual object tracking vot2017 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct. 2017.

[2] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, pp. 1409–1422, 7 2011.

[3] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, pp. 583–596, 3 2014.

[4] Z. Kalal. "Tracking learning detection." (2011), [Online]. Available: `http://www.ee.surrey.ac.uk/CVSSP/Publications/papers/Kalal-PhD_Thesis-2011.pdf`.

[5] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-n learning: Bootstrapping binary classifiers by structural constraints," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 49–56. DOI: `10.1109/CVPR.2010.5540231`.

[6] K. Murphy, *Machine Learning: A Probabilistic Perspective*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2012, ISBN: 9780262018029.

[7] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021.

[8] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[9] S. Taylor and T. Drummond, "Multiple target localisation at over 100 fps," Jan. 2009. DOI: `10.5244/C.23.58`.

[10] M. Kristan, A. Leonardis, J. Matas, *et al.*, "The eighth visual object tracking vot2020 challenge results," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds., Cham: Springer International Publishing, 2020, pp. 547–601.

[11] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[12] J. Schrittwieser, T. Hubert, A. Mandhane, M. Barekatain, I. Antonoglou, and D. Silver, "Online and offline reinforcement learning by planning with a learned model," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 27 580–27 591. [Online]. Available: `https://proceedings.neurips.cc/paper/2021/file/e8258e5140317ff36c7f8225a3bf9590-Paper.pdf`.