RHODES UNIVERSITY
Where leaders learn

Real Time Detection and Tracking of People in Crowds

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

April 16, 2022

Abstract

Tracking of objects in videos streams is a powerfull tool in the field of computer vision. This research will investigate the recognition and tracking of mulitple faces in settings where multiple faces are concurrently visible. The research will take the form of implementing a system that can detect and track the motion of faces in a video stream after being initialized with minimal input data. The task starts by initializing the tracker with bounding boxes in images that define the faces of the targets. Once the tracker is initialized, the tracker will identify and follow the motion of the targets in a video stream. The tracker will be able to detect if a target face disappears from the video stream. If the target face reappears in the video stream, the tracker will be able to identify and track the face as long as the face remains in the field of view.

1 Introduction

Consider a continuous video stream in which a set of faces appear. The individual faces might move and change orientation in the video stream. Imagine that there is some subset of these faces of interest, and pictures of these faces–these pictures may be unrelated to the video or frames of the video. With these ideas in mind, the goal of this research is to develop a system that can automatically identify and track the target faces in the video stream.

The intial input given to the system is images that have uniquely labelled regions of interest which define individual faces. The input images are required to contain at least one instance of each target face, but there is no maximum limit. This constitutes the initialization of operation, after which the system functions autonomously.

When given an arbitrary video the system detects the presence or absence of any of the target faces within the video as the video progresses with time. Subsequently, the system labels every target face that it detects with the label that was given to the face in the initialization of the system. Once a face is labelled, the system follows the motion of the face, and any other target faces appearing in the current frame, for as long as the face is visible. This constitutes the running phase of the system, where the system identifies and tracks faces in a given video stream.

While the system is running it determines information about the target faces. It can, hence, extract the the number of times each target face appears, the amount of time for which each target appears and the trajectory of each while it is apparent in the video. This is the output stage of the operation, which concludes the operation of the system.

The system is designed to operate with minimal input data given in the initialization stage. With this constraint, it is desirable for the system to use all the data it can get access to. The system, thus, uses the the video in the running phase to learn more about each target face–in this way it can identify and track faces more accurately as the video progresses.

The next section of this proposal gives the formal research statement. This is followed by a section on the research objectives.

Following this, there is a section that introduces works that are related to this research. The afore mentioned section serves as a miniture literature review.

The related works section is followed by a section discussing the approach to research that will be followed. This is followed by a timeline of the deadlines for the project.

Following the timeline section are two sections discussing the practicalites of the research. First is a discussion on the limitations of the research. Second is a brief section discussing further applications of the research.

Finally the conclusion sumarises this proposal.

2 Reseach Statement

- Machine learning and computer vision techniques can be used to design and implement a system that, when given minimal input data, is capable of counting the number and measuring the duration of appearances of multiple target faces in a single video stream.

The central computer vision technique that is tested, by this research, is the Tracking, Learning, and Detection framework–discussed in Section 4. The definition of minimal input data, in the context of this research, is a single image for each face that is required to be tracked, where each image includes a bounding box that defines the location of the face in the image and a label for the face. The goal of this research is to implement a working system that meets the conditions specified by the Research Statement.

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

In 2011 Kalal, Mikolajczyk, and Matas invented the Tracking, Learning and Detection(TLD) framework for the longterm tracking of objects in a video stream. Kalal's original implementation uses a median flow tracker, P-N learning, and a random forrest based detector [2]. These three components enable the tracking, learning and detection capabilites of the system. The three components exchange information as shown in Figure 1 The system requires online(learning as data becomes available) learning in order for the system to work, Kalal developed the P-N Learning paradigm [3], a semi-supervised bootstrapping model [4], tailored to the needs of TLD.

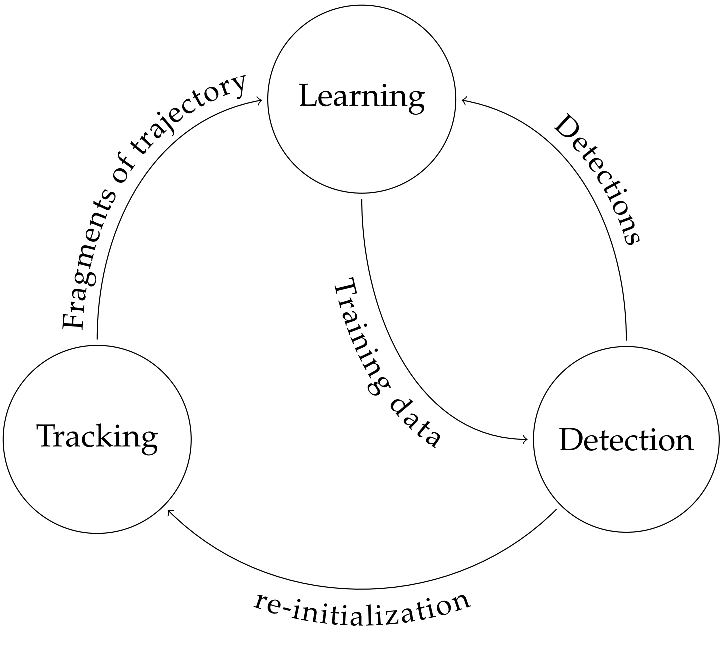

Figure 1: The interaction between tracking, learning and detection in TLD. Figure from [1]

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[5]. 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 [6].

There has also been investigation into the use of Convolutional Neural Networks in tracking [7]. 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 [8] 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 [9] and [10]. This is a good reference, and will be explored further, but not all trackers can satisfy the requirements of this project.

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 [2].

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] 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.

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

[3] 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`.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[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`.