Real Time Detection and Tracking of People in Crowds

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

April 18, 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 the faces are concurrently visible. The research will take the form of implementing a tracking system that, after being initialized with minimal input data, can detect faces and track their motion in a video stream. The task of tracking starts by initializing the tracker with bounding boxes that define the faces of the targets in images. Once the tracker is initialized, the tracker will be able to detect if a target face is present in or absent from an arbitrary video stream. The tracker will identify the visible target faces and follow their motion in the video stream. If a target face disappears and later reappears in the video stream, the tracker will be able to identify and track the face again as long as the face remains in the field of view.

1 Introduction

Consider a continuous video stream in which a set of faces appears, each face might appear at different times. The individual faces might move and change orientation in the video stream. Imagine that there exists some subset of these faces that is of interest–the target faces. A set of pictures of these faces must exist, 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 or bounding boxes 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 associated with 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 the supplied 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 face 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 supplied 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 with better accuracy 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. This section serves as a miniture literature review.

The related works section is followed by a section discussing the approach to research that the project will follow. 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 long term tracking 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 is the Tracking, Learning, and Detection framework–discussed in Section 4. This framework allows for long-term tracking with minimal input data.

Long-term tracking(LT) is tracking of objects that can undergo partial oclusions, change appearance, and disappear and reappear from the field of view. This is opposed to short-term track-

ing(ST) where the object remains fully in the field of view for the whole duration of tracking. ST can be used as a basis for LT, as is done in the case of TLD.

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 at least one bounding box. The bounding box 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

- Build a system for idenitifying and tracking a single face in a video stream.
- Investigate ways to extend the system so that it can track multiple faces.
- Build a system for idenitifying and tracking multiple faces in crowded scenes using the TLD frame work.
- Enhance the system's long-term tracking capabilities.
- Investigate ways to increase the scalability of the system, so that it can track more faces simlutaneously.
- Test the system's qualitative tracking and detection performance on public data.
- Enable the system to extract and record quantitative data about the targets.
- Test the quantitative performance of the system on public data.

4 Related Works

In 2011 Kalal *et al.* 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 and nearest neighbour based detector [2]. These three components give the respective tracking, learning and detection components of the system.

The learning compenent of TLD forms the backbone of the system, governing the interaction between the detector and tracker. The three components exchange information as shown in Figure 1, this allows the tracker to improve it's performance as time progresses [1]. For the system to operate, it requires online learning–learning as data becomes available. Kalal developed the P-N Learning paradigm [3], a semi-supervised bootstrapping model [4], tailored to the needs of TLD.

Henriques *et al.* [5] propose Kernelized Correlation Filters(KCF) and the novel Dual Correlaion filter(DCF). Both KCF and DCF use circulant matrices and the kernel Trick. The implementation of KCF by Henriques *et al.* uses a Gaussian Kernel, whereas the DCF implementation uses a linear kernel. The calculations involved with the linear kernel are less computationally complex than KCF. DCF can, hence, be processed faster, but, at the cost of some tracking precision.

Work by Galoogahi *et al.* [6] allows KCF and DCF to be applied to modern and useful feature descriptors. Henriques *et al.* show that KCF and DCF be be applied to Histogram of Oriented Gradient(HOG) features to track and detect objects in a video stream with lower computation times and better accuracy. KCF and DCF applied to HOG features are shown to outperfrom many tracking systems Table 1. The results shown by Table 1 are obtained from running the algorithms on a standard four core desktop processor from 2014.

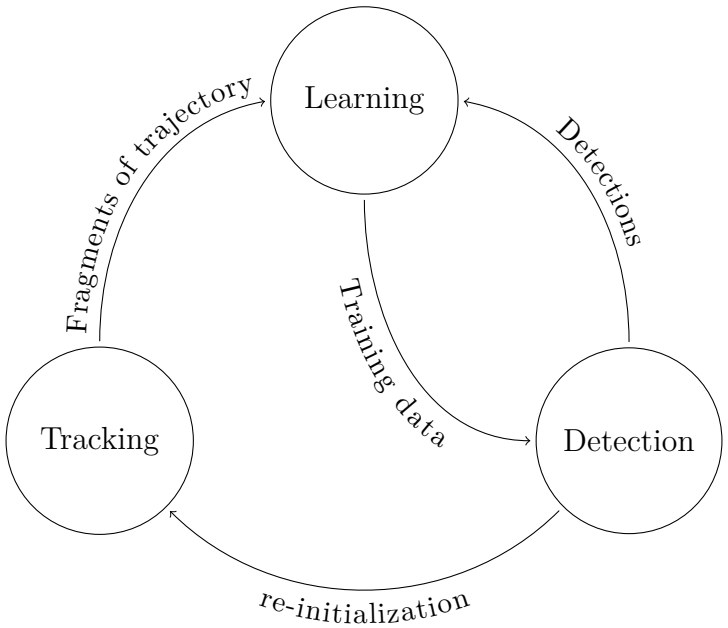

Figure 1: The interaction between tracking, learning and detection in TLD. Figure from [1]

Algorithm	feature	Mean precision	Mean FPS
KCF	HOG	73.2%	172
DCF	HOG	72.8%	292
KCF	Raw pixels	56.0%	154
DCF	Raw pixels	45.1%	278
TLD		60.8%	28
Struck[7]		65.6%	20
MOSSE[8]		43.1%	615

Table 1: Comparison of various trackers, adapted from [5]

The system implemented by Henriques *et al.* does not, however, encorporate a failure recovery mechanism–section 8 of [5]. In other words Henriques *et al.* only explore KCF in the domain of ST. This is in contrast to the original TLD system which provides a failure recovery mechanism in the detection component [1]. The ST using KCF and DCF done by Henriques *et al.* can be used in a TLD framework for LT.

Ma *et al.* [9] investigate the problem of single object LT using correlation tracking. Ma *et al.* use two Gaussian ridge regression [4] models for tracking. One model uses the relative change in background and target as time progresses, the other model tracks by using the target's appearance. The first model is used to track the object's trajectory through fast motion and occlusions, and the second is used for scale change. Using both tracking models they train an online detector that is both flexible(from first tracker model) and stable(from second tracker model).

Ma *et al.* train a random fern classifier [10] [1] online in order to handle tracker failure. This solves the LT problem in a similar way to Kalal.

The Visual Object Tracking Challenge(VOT) is a challenge that benchmarks various trackers every year [11] [12]. VOT investigates both ST and LT. In recent years, VOT has also introduced a real time challenge [12].

There has also been investigation into the use of Convolutional Neural Networks in tracking [13]. The work by Nam *et al.* offers high performance tracking of generic objects. The system implemented by Nam *et al.* requires a large amount of offline training.

Schrittwieser *et al.* [14] propose a new online reinforcement learning method that can be used to train models with minimal input data. Schrittwieser *et al.* describe the *Reanalyse* algorithm. Given a state of a machine learning model the *Reanalyse* algorithm generates training targets for the model from some input data. When the model has improved by training, the *Reanalyse* algorithm generates more training targets based on the new state of the model, the already seen input data, and any new input data. The algorithm allows the available training data to be cycled–this allows the algorithm to extract most of the information from a limited dataset.

5 Approach to Research

The first phase of this research will consist of in-depth reading of literature and further literature reviews. The literature reviews will start by reviewing Kalal's work on TLD. After a thorough review of Kalal's work, works pertaining to other trackers will be reviewed. Following this, there will be a further review of the literature discussing the online training of classifiers. The implementation of the system requires data for testing–a brief phase of data acquisition from public sources will provide data for developmental and testing purposes. This should suffice for the literature review and data collection.

The second stage of this research will relate to the practicalities of the system's implementation. The system will be implemented in C++, this requires proficient understanding of the C++ language. Investigation of the openCV C++ library will be done, and the available utilities will be surveyed.

The third phase consists of a functional reimplementation of the original TLD. Testing of this base system is required at this point, so that problems do not occur in later stages of the full system implementation. Following satisfactory performance of the TLD reimplementation, the next stage of research will commence.

At this point further improvements to the base TLD model will be made. The focus of the improvement will be on the tracking and detection components of the system. This involves either keeping the base model and improving the individual components or restructuring the model to improve model performance. This will constitute the fourth stage of the research.

Following this, an investigation of extending the system to track multiple object simultaneously will be made. There are naive ways to implement a multiple object tracking system–for example creating many different single target trackers to detect and track each object. This stage, therefore, requires significant planning, research and reviewing the current implementation in order to acheive good model performance and efficiency. This will complete the implementation of the system.

The final stage of this research involves three things. First, a full review of the implementation will be carried out. If improvements to the implementation are required and time permits, a return to stage four will be made. Second is a testing phase, where the complete system will be tested with videos obtained from the initial phase of the research. Third, a thesis will give a description of the implementation and specifications of the system. This completes the research project.

6 Timeline

Time	Deliverable
30 March 2022	Seminar 1: Presentation of project
11 April 2022	Draft proposal
19 April 2022	Final proposal
26 April 2022	Functional re-implementation of TLD
29 April 2022	Literature review
30 April 2022	Using KCF as Tracking stage of tracker
6 May 2022	Implementing DCF and comparing to KCF
13 May 2022	Reviewing VOT for better trackers
20 May 2022	Investingating random fern detectors
25 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	Test and make small improvements 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

The first concern of this research is in regards to ethics. Owing to ethical concerns, the system can only be tested on public data, which limits the test cases for the system. An ethical clearance is required to test the system on examples that are closer to real world applications. Time restrictions on the project make doing an ethical clearance impractical.

Further on this point, the second limitation of this research is time. The project is set to take one year, being an honours project. There is only so much literature that can be reviewed and still allow for the implementation of a system. On account of this restiction, the research might not explore certain areas of concern.

This research is only concerned with the tracking of faces. There are many other features that can be used to detect and track humans, for example gait. Sometimes, there are also needs to track objects other than faces. The system implemented by the research cannot guarantee tracking capabilities of objects other than human faces.

There will be an upper limit on the number of faces that can be tracked simultaneously in real time. There are two problematic cases: First, limited compute power, second screen space. The first case occurs if the runtime complexity of the implemented system increases proportionally to the number of faces being tracked. Formally, if system has runtime complexity worse than $O(1)$, the system will eventually fail to track, in real time, as more faces are added. The second problem case occurs in very large, dense crowds of faces where there is not enough space in the field of view for all the faces, or some faces are captured in insufficient resolution.

8 Applications

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

Another application on the larger scale is a security system for monitoring the motion of people. This application requires the system to be extended to a distributed system. This could enhance airport security for example, adding functionality to label suspected terrorists.

The system also has potential to be used as an assistive technology for visually impaired people. If the system can be ported to a mobile device, it could help a visually impaired person identify people at a party or on a television program. This application assumes the system has had the appropriate offline training in it's initialization phase.

9 Conclusion

This research uses the Tracking, Learning and Detection(TLD) framework to implement a system that can track and detect multiple faces simultaneously. TLD is used in the research to develop a long-term tracker that can recognize human faces and follow the trajectory of the faces in a video stream.

The system that the research implements is tested on public data. Testing is done by having the system extract information about faces that appear in a given video stream. The information collected by the implemented system is the number and duration of appearances for each target face.

The research is limited to the tracking of faces, and is not aimed at tracking general objects or other human features. Owing to the time limitations of an honours project, this project has substantial dependence on research done by other people and available tools.

References

[1] Z. Kalal *et al.*, "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 *et al.*, "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 *et al.*, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, pp. 583–596, 3 2014.

[6] H. K. Galoogahi *et al.*, "Multi-channel correlation filters," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013.

[7] S. Hare *et al.*, "Struck: Structured output tracking with kernels," in *2011 International Conference on Computer Vision*, 2011, pp. 263–270. DOI: `10.1109/ICCV.2011.6126251`.

[8] D. S. Bolme *et al.*, "Visual object tracking using adaptive correlation filters," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2544–2550, 2010.

[9] C. Ma *et al.*, "Long-term correlation tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[10] M. Ozuysal *et al.*, "Fast keypoint recognition in ten lines of code," Jun. 2007. DOI: 10.1109/CVPR.2007.383123.

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

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

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

[14] J. Schrittwieser *et al.*, "Online and offline reinforcement learning by planning with a learned model," in *Advances in Neural Information Processing Systems*, M. Ranzato *et al.*, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 27 580–27 591. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/e8258e5140317ff36c7f8225a3bf9590-Paper.pdf.