Name: Sean Goh Ann Ray
Matric No: G2202190G

1) What is the work about?

This paper [1] talks about long term real-time tracking of unknown objects in a video stream. Many algorithms fail when the unknown object disappears from the camera's view and reappears again in a different appearance due to geometric transformation, and those which overcome this problem would require offline training to recognise the possible appearances. This paper proposes the use of a novel tracking-learning-detection (TLD) framework for long term tracking, works even when the unknown object moves in and out of the frame, and performs the long term tracking at the video stream's frame rate. The tracking is performed by tracking the object from frame to frame, the detector learns the observed appearances of the object and corrects the tracker, and the learning estimates the detection errors and updates the detector.

2) What are gaps of prior research works?

Prior research works involve the use of tracking, detection and Machine Learning (ML) algorithms, or some combination of the 3 algorithms.

Tracking algorithms work by describing an object's bounding box, and attempts to find a similar bounding box in a similar location in the next frame, thereby allowing the estimation of the object's trajectory over time. However, limitations include the changing appearance of an object (due to geometric transformation of the object and lighting), the changing appearance of the background, and also accumulate errors over time. Additionally, the algorithm usually fail when the object is partially blocked by another object, or when it goes out of frame and reappears again in the video stream, especially when there's a different appearance.

Detection algorithms work by detecting an object's feature within a frame, but it requires beforehand training of the object's geometric shape. Such algorithms also require large amounts of training datasets to accurately detect an object from the background, and therefore cannot be used to perform real-time tracking to detect unknown objects.

ML algorithms used for the training of detection algorithms include Expectation-Maximization (EM), Self-learning, and Cotraining. However, these require a large pool of labelled or semi-labelled training datasets, and is therefore not suitable for this work as the video stream provides only one labelled example to learn from.

Prior works which involved TLD frameworks include pre-training a detector to output to the tracker, while others were successfully implemented but only on low frame rate video streams [2]. However, these TLD frameworks require pre-training and does not perform well when the object changes its appearance. Other adaptive TLD frameworks uses real-time detector which detects changes in the background, thereby enabling tracking. However, this means that the tracking and detection is done by a single process, which may not perform as well as when the tracking and detection algorithm work together.

3) What are motivations of the performed search?

This paper suggests the use of both tracking and detection algorithms at the same time, where the tracking algorithm provides a small set of training data for the detection algorithm, which then undergoes ML to learn new appearances of the object, and outputs the results back to the detection and tracking algorithms to perform the long term tracking. This means that the algorithms supports each other in order to achieve long term tracking of unknown objects.

4) How does the proposed technique address the gaps?

The proposed technique involves a ML stage between the tracking and detection algorithm called P-N learning. It obtains information from both the tracking and detection algorithms and estimates the detector's errors, which then generates training datasets to retrain the detector, which would then update the tracker.

P-N learning consists of 4 blocks, a classifier, training set, supervised training, and P-N experts. The initial training process starts with a labelled training set (the labelled frame) which undergoes supervised learning to train a classifier. The labelled training set also includes additional synthetic positive examples which were produced using geometric transformations of the object of interest. The P-N experts then estimates the errors of the classification, which then adjusts the labels of the training set. The classifier then undergoes another iteration of supervised learning, and this process repeats until convergence or other stopping criterion is met.

P-N learning uses 2 individual experts to identify the detector's errors, the P-expert for false negatives and N-expert for false positives. The separation of the classification errors allows the P-expert to estimate the false negative and adjusts their class labels to positive before adding them to the training set, while the N-expert estimates the false positive and adjusts their class labels to negative before adding them to the training set.

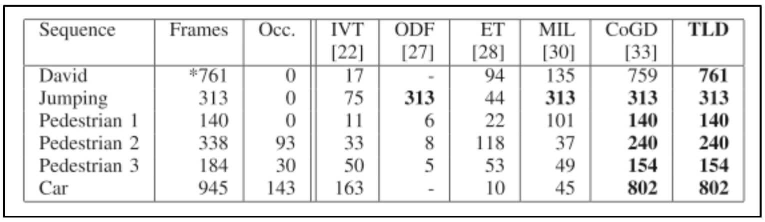
The P-expert assumes the object moves along a trajectory, where it remembers the location of the object within the previous frame, and estimates the new location in the current frame using the tracker. If the detector outputs a negative at a current location, the P-expert relabels that as positive. This new positive example provides the classifier a new appearance of the object, which then updates the detection algorithm.

The N-expert assumes the object can only appear at a single location within the frame. It receives the response of the detector and tracker to select a patch in the frame that it is most confident in, where patches that are not overlapping with the most confident patch are labelled as negative, and the most confident patch updates the tracker with the new location of the object.

This process repeats from frame to frame, where frames after the labelled frame have unlabelled data and will be labelled based on the classifier. The P-N experts then estimates the errors of the classification again, thereby producing new training sets to update the classifier after every frame as described above. This allows the real-time tracking of the object in an unlabelled video stream.

5) What evaluation metrics were adopted to validate the designs?

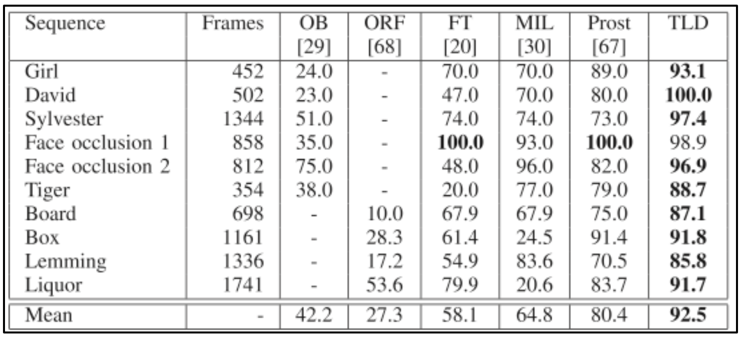
There were multiple evaluation metrics used. In another report [2] which used 5 different trackers on video sequences where the object of interest was fully blocked and even disappeared from the camera's view. The Cotrained Generative-Discriminative tracking (CoGD) proved to be the most effective tracker in that report by a large margin. In summary, CoGD works by combining two tracking methods, which are generative and discriminative. Generative tracking works by learning the closest appearance of an object in a new frame and updates the tracker, while discriminative tracking finds the best decision boundary of an online support vector machine to differentiate the object from the background. Table 1 below shows the performance of the 5 different trackers in comparison to TLD when used on the same sequences. While CoGD and TLD have comparable results, it was noted that CoGD required 10 labelled frames for the initial training set and could only perform the tracking at 2 frames per second while TLD only required one labelled frame for the initial training set and could perform at 20 frames per second.



Sequence	Frames	Occ.	IVT [22]	ODF [27]	ET [28]	MIL [30]	CoGD [33]	**TLD**
David	*761	0	17	-	94	135	759	**761**
Jumping	313	0	75	**313**	44	**313**	**313**	**313**
Pedestrian 1	140	0	11	6	22	101	**140**	**140**
Pedestrian 2	338	93	33	8	118	37	**240**	**240**
Pedestrian 3	184	30	50	5	53	49	**154**	**154**
Car	945	143	163	-	10	45	**802**	**802**

Table 1 Number of Successfully Tracked Frames. Source [1]

In another report [3], 5 different algorithms were used on video sequences where the object of interest was partially blocked from the camera's view. The Parallel Robust Online Simple Tracking (PROST) performed the best among the algorithms used in that report. Prost consists of 3 different algorithms, a simple non-adaptive tracker, an optical flow based mean shift adaptive tracker, and an online random forest to learn new appearances. Table 2 below shows the performance measured by recall in percentage, where TLD performs about 12% better than Prost on average.



Sequence	Frames	OB [29]	ORF [68]	FT [20]	MIL [30]	Prost [67]	TLD
Girl	452	24.0	-	70.0	70.0	89.0	**93.1**
David	502	23.0	-	47.0	70.0	80.0	**100.0**
Sylvester	1344	51.0	-	74.0	74.0	73.0	**97.4**
Face occlusion 1	858	35.0	-	**100.0**	93.0	**100.0**	98.9
Face occlusion 2	812	75.0	-	48.0	96.0	82.0	**96.9**
Tiger	354	38.0	-	20.0	77.0	79.0	**88.7**
Board	698	-	10.0	67.9	67.9	75.0	**87.1**
Box	1161	-	28.3	61.4	24.5	91.4	**91.8**
Lemming	1336	-	17.2	54.9	83.6	70.5	**85.8**
Liquor	1741	-	53.6	79.9	20.6	83.7	**91.7**
Mean	-	42.2	27.3	58.1	64.8	80.4	**92.5**

Table 2 Performance measured by Recall. Source [1]

6) What are the constraints of the proposed technique?

The proposed TLD framework performs badly when there is full out-of-plane rotation, where the tracker will not be able to identify the object of interest unless the appearance was observed before.

The proposed TLD framework only trains the detection algorithm and not the tracking algorithm, meaning that the tracker always make the same errors.

It also does not perform well when the object changes its appearance drastically and in a wide variety of ways, such as a human who is dancing.

7) What are possible future works?

One of the possible future works would be to perform multi-object tracking, as the TLD framework proposed in the paper only tracks a single target.

Another possible improvement is to include allow the tracking algorithm to differentiate the object from the background.

References

[1] Z. Kalal, K. Mikolajczyk and J. Matas, "Tracking-Learning-Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7, July 2012.

[2] Q. Yu, T.B. Dinh, and G. Medioni, “Online Tracking and Reacquisition Using Co-Trained Generative and Discriminative Trackers,” Proc. 10th European Conf. Computer Vision, 2008.

[3] J. Santner, C. Leistner, A. Saffari, T. Pock and H. Bischof, "PROST: Parallel robust online simple tracking," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.