UNIVERSITY OF CALIFORNIA
RIVERSIDE

Semi-Supervised Tracking of Multiple Identical Objects with Occlusions

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Masters of Science

in

Computer Science

by

Colin Lee

June 2021

Thesis Committee:

Professor Christian R. Shelton, Co-Chairperson
Professor Amit K. Roy-Chowdhury, Co-Chairperson
Professor Samet Oymak

The Thesis of Colin Lee is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

Acknowledgments

ABSTRACT OF THE THESIS

Semi-Supervised Tracking of Multiple Identical Objects with Occlusions

by

Colin Lee

Masters of Science, Graduate Program in Computer Science
University of California, Riverside, June 2021
Professor Christian R. Shelton, Co-Chairperson
Professor Amit K. Roy-Chowdhury, Co-Chairperson

Recent advances in multiple object tracking (MOT) rely primarily on visual appearance features to reconnect tracks lost due to occlusions. However, appearance features cannot be relied on to discriminate between objects that are visually similar or identical, such as animals, people in uniform, or mass-produced items. We propose a new model that relies on spatio-temporal motion features rather than appearance features for such videos. Furthermore, training an MOT method often relies on expensive hand labeling of bounding boxes or segmentation masks with ground truth tracks. By contrast, our videos are labeled only with fixed bounding boxes (effectively only positional information). We train our model in a semi-supervised manner using iterative pseudo-labeling (IPL), a technique often used in natural language processing, but not common to computer vision tasks. We show that appearance features are insufficient for reconnecting tracklets in videos of bee foraging, and that our motion-based IPL method offers an improvement over appearance feature methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In modern multi-object tracking (MOT), visual appearance features are the main method of distinguishing between different objects whose tracks must be reconnected due to occlusion. Such appearance features rely on objects being visually distinct in the first place, but this is not always the case. There are many scenarios in which one might desire to track objects that are visually similar, sometimes to the point of being completely indistinguishable from one another. For instance, many animals, such as bees and ants, are visually identical in appearance and are difficult to distinguish between even for humans (see Figure 1.1). In the common MOT task of pedestrian tracking, these appearance cues often rely on the clothes that people are wearing. However, there are many situations where people dress in uniforms, such as in the military or sports. In such cases, only less ostentatious details could be used differentiate appearances, such as faces or player numbers, but these could often be out of view due to the human's orientation towards the camera. Inanimate objects are perhaps even more likely to suffer from this problem, as many of

Figure 1.1: Two bees: Visually similar objects, nearly identical in appearance and difficult for even humans to differentiate.

them are mass-produced. While this is not usually an extreme problem in the usual task of tracking cars in everyday traffic, certain situations, such as tracking military or government agency vehicles would fall into this category, as they are often of singular make, model, and color.

In ideal conditions with quality video and no occlusions, tracking multiple identical objects is not in of itself a difficult task. Modern object detectors are certainly capable of accurately detecting most objects, and multiple objects having identical appearances would only serve to make the task simpler by reducing the variations and complexity needed to create a reliable detector. Following the detection step, MOT methods generally perform a data association step to link the per-frame detections into sequences of detections(tracks). With no occlusions between objects or the background, tracking is hardly more complex than stitching together reliable detections and thus the data association step is very simple. With the addition of long-term occlusion, this task becomes significantly more difficult due

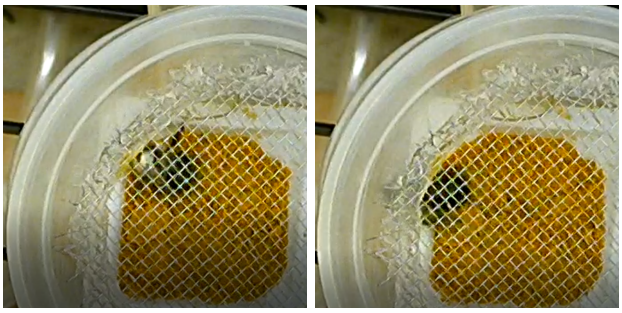

Figure 1.2: Long term occlusions: Bees can often disappear under occluding parts of the experimental apparatus for long periods and reappear elsewhere.

to the need to reconnect lost tracks, but the usual method of reconnecting them via models trained for the visual re-identification (reID) is no longer applicable.

In this work, we will investigate the intersection of visually similar objects and occlusion using videos of bees in an experimental apparatus which often occludes them, using only positional information as ground-truth for training our object detector. First, we will establish the need for new motion features by investigating the effectiveness of visual features on these videos. Then, using sequences of detections provided by an object detector, we will reconstruct full object trajectories (tracks) without the use of visual features, instead relying on motion features provided by a novel spatio-temporal model.

Chapter 2

Related Work

2.1 Appearance-based Tracking-by-Detection

The object detection step in MOT has advanced to a point where it is possible to adequately track objects when there are little or no missing detections due to occlusions or low frame rates. Recent tracking methods, such as SORT[3], IOUTracker[4], and Tracktor[1], almost forgo the data association step entirely, relying solely on the bounding box information from the detection step to string detections into tracks. Despite their very simple nature they are able to achieve state of the art performance.

Accurate object detectors then, in conjunction with high quality video, would have almost obviated the data association step were it not for the simple fact that most real-world video *is not* ideal and includes occlusions. The principle purpose of the modern data association step then, is in dealing with the missing detections. This is often referred to as the re-identification or reID task. Recent work utilizes the discriminitive abilities of modern deep networks to learn appearance features that can be used to identify when the

same object is detected after reappearing from behind an occlusion (or on another camera in a multi-camera setup). While some form of motion prediction is often used in conjunction with these appearance features, the main focus of recent scholarship has been on using appearance features for the reID task. This is evident in extensions of the aforementioned tracking methods: DeepSORT[25], VIOUTracker[5], and Tracktor++[1]. The DeepSORT and Tracktor++ extensions rely primarily on appearance features generated from separately trained reID networks in order to reconnect long range tracklets. VIOUTracker does not use an a reID network, but still uses visual single object trackers(SOT) to reconnect over occlusions. This reliance on appearance features poses a problem for different objects with similar appearances.

A popular method of training these reID networks is to train Siamese CNNs [11, 23, 27, 6] and these are often optimized with a triplet loss function [26]. Originally used for the person re-identification task by [21] in 2015, it has become one of the de facto methods to train Siamese networks for reID tasks in general. Triplet loss training is performed by forming triplets from three images: an anchor image, a positive image, and a negative image. The anchor and the positive images come from the same object, which are usually from the same tracklet in this context, and help learn similarities between two modalities of the same object. The negative image, on the other hand, helps distinguish between the target object and the rest of the targets in the image.

As we will show, these appearance features will prove insufficient when working with visually near-identical objects such as bees.

2.2 Motion Features

In the absence of effective appearance features to distinguish between objects, the most obvious substitute would be spatio-temporal information of the tracklets. The usage of motion features is not a new idea, having been focus of many methods before the introduction of CNNs and the advantages they provided in visually distinguishing between objects. Previous work on tracking multiple identical objects has not been the subject of much study in recent years. Kalman filters [9] and particle filters [8] are a natural choice for creating a motion model from the positional information provided by object detectors. A simple affinity can then be calculated based on the distance from the predicted position to any nearby tracklets or detections. However, these methods require the user to assume a motion model a priori, and thus objects with complex or unpredictable dynamics can cause these to fail.

[7] in 2013 specifically addressed the problem of multiple similar objects with occlusions in 2013. To circumvent the problems with a priori models they used the order of linear regressors to represent the dynamics of a tracklet. Even so, an assumption must be made that the trajectory underlying the tracklets can be approximated by a similar regressor throughout. Whether using an a priori motion model or not, assumptions about the motion of objects typically only work well when object motion is somewhat predictable, such as with cars or pedestrians walking across a limited field of view.

While motion features are not currently as prominent, they are still used in conjunction with appearance features to further refine and improve results. The most visible example of this is the Kalman filter in the SORT [3] and DeepSORT [25] algorithms. How-

ever, modern work featuring motion features usually involves using *in conjunction with* appearance features and not *in the absence of* useful appearance features. [16, 20, 26, 24] This work will investigate the importance of appearance features for long-range tracklet reconnection, in addition to investigating how effective motion features can be for discrimination in lieu of useful appearance features.

In spite of the recent focus on appearance features, there still is some research focusing prominently on the usage of motion features. However, none of these papers are written with a particular focus on visually identical objects and ineffective appearance features, or they require extra ground truth labeling. For instance, [22] utilizes sequences of bounding boxes that they call "anchor tubes", but these rely on properly ground-truthed bounding boxes for the object detector to learn. We will used fixed bounding box sizes in this work, effectively meaning that we will rely solely on a sequence of positional data instead of bounding boxes, object segmentations, or labeled object angles.

While motion features are often acknowledged in many works as useful in the cases of objects with similar appearances, these acknowledgements tend to be made in passing and do not formally treat the problem of reconnecting identical objects over occlusions. To our knowledge, such work is very sparse.

2.2.1 Unsupervised and Semi-supervised MOT

Currently, one of the biggest weaknesses of the modern use of deep learning methods is its high reliance on accurate and well-labeled data. Improving upon this weakness continues to be a very active area of research and the MOT field has been no exception. Well-labeled datasets are expensive to create due to the need for time-consuming human

labeling and are vulnerable to any biases introduced into the dataset by the human labelers.

There are approaches to alleviate the need for strongly-labeled datasets, with the most desirable being completely unsupervised MOT where no labeling at all is required. In the past, this has been covered by methods such as background subtraction. Modern attempts to tackle this task do exist. However, thus far they have been relegated to simpler MOT tasks such as tracking a small number of MNIST numbers without external occlusions, such as in [12] .

Weaker labeling has also been a significant area of interest in single object tracking (SOT), often referred to as the video object tracking (VOT) task. In this field, it is common to initialize the tracker with the first (or first several frames) frame of the video with the single object marked[14]. After these first frames, no other frames of the video are labeled and the tracker is expected to find the object in question [2]. Some MOT approaches attempt to make use of multiple SOT trackers, such as the aforementioned VIOU Tracker, and can perform adequately on simpler high quality videos [5].

In this work, we will attempt to use high confidence tracklets as weak labels. Our approach to this problem is to use iterative psudeo-labeling (IPL), which we use to produce iteratively more confident inter-tracklet strong labels. IPL is currently used primarily in the field of natural language processing, and to our knowledege, has not yet been used in this context. In one of our baseline experiments, we investigate the effects of strongly-labeled object trajectories on a modern reID method versus weakly-labeled tracklets. Semi-supervised learning has been used in MOT before, such as in [19], but not for tracklet linking in the manner presented here.

Chapter 3

Problem

3.1 Formulation

We modify the general problem formulation provided in [17]. We formulate the MOT problem as a multi-variable estimation problem and as follows:

Given a video as a sequence of T frames with M objects, each with its own ground truth trajectory, we denote the state of the mth object at time t by its state $\mathbf{s}_t^m$, which can be represented by things such as its coordinates, bounding box, appearance, or segmentation mask, e.g. $\mathbf{s}_t^m = (x, y, appearance)$. Then the entire ground truth trajectory of the mth object can be written as $\mathbf{S}_{1:T}^m = (\mathbf{s}_1^m, \mathbf{s}_2^m, ..., \mathbf{s}_T^m)$.

To represent the detections from our object detection step, we denote an observation of the mth object at time t in the same manner as its state, $\mathbf{o}_t^m = (x, y, appearance)$. Correspondingly, $\mathbf{O}_{1:T}^m = (\mathbf{o}_1^m, \mathbf{o}_2^m, ..., \mathbf{o}_T^m)$ represents the observed trajectory of the mth object through the video sequence.

Our objective in the data association step of MOT task then, is to find the set of

most probable trajectory from the given observations:

$$\hat{\mathbf{S}}^m_{1:T} = argmax_{\mathbf{S}^m_{1:T}} P(\mathbf{S}^m_{1:T} | \mathbf{O}^m_{1:T}) \tag{3.1}$$

The question from this step is how to determine which of the observations provided belong to the mth object. That is, we must estimate the probability that each observation belongs to the same object:

$$P(m_a = m_b | \mathbf{o}^{m_a}_{t_a}, \mathbf{o}^{m_b}_{t_b}) \tag{3.2}$$

This value will be approximated by our *detection affinity* and based on the results of the simpler MOT methods such as IOUTracker and Tracktor, we can safely make the follow assumption:

If $\mathbf{o}^{m_a}_{t_a} \approx \mathbf{o}^{m_b}_{t_b}$, $t_a \approx t_b$ and $\mathbf{o}^{m_a}_{t_a} \not\approx \mathbf{o}^{m_c}_{t_c}$, $\mathbf{o}^{m_b}_{t_b} \not\approx \mathbf{o}^{m_c}_{t_c} \forall$ other observations $\mathbf{o}^{m_c}_{t_c}$, then

$$P(m_a = m_b | \mathbf{o}^{m_a}_{t_a}, \mathbf{o}^{m_b}_{t_b}) \approx 1$$

That is, if there are two detections very close in both space and time, with no other detections to confuse the situation, then accurate modern object detectors combined with high quality video allow us to assume that the two detections belong to the same object. This allows us to form a high confidence tracklet $\hat{\mathbf{S}}^m_{t_1:t_2}$.

We can then continue by estimating the probability that two high confidence tracklets belong to the same object:

$$P(m_a = m_b | \hat{\mathbf{S}}^{m_a}_{t_1:t_2}, \hat{\mathbf{S}}^{m_b}_{t_3:t_4}) \tag{3.3}$$

Which we can approximate with a *tracklet affinity*. Such an affinity scoring can then be learned using an affinity model as presented below. Once we have affinity scores for each possible pair of tracklets, we must note that one tracklet can only paired with one other tracklet before and after it. This leads to a bipartite matching problem which we can solve using methods such as the Hungarian algorithm, giving us an optimal assignment of the most likely tracklet linkages.

3.2 Method

We build our affinity model by starting similarly to the reID networks used to determine appearance based affinity. Given two tracklets, we determine an embedding for each of them to factor into an affinity scoring. However, given that appearance features between identical objects cannot be relied upon, we instead start by learning an embedding from motion features.

3.2.1 Input Motion Features

In order to assign affinity scores to a pair of tracklets without using appearance features, we begin with the following inputs: the tracklets themselves (as two sequences of positional detection coordinates), the time gap between the two tracklets (in number of frames), and any other tracklets that co-occur with either tracklet in the pair. Production of these high confidence tracklets is illustrated in Figure 3.1.

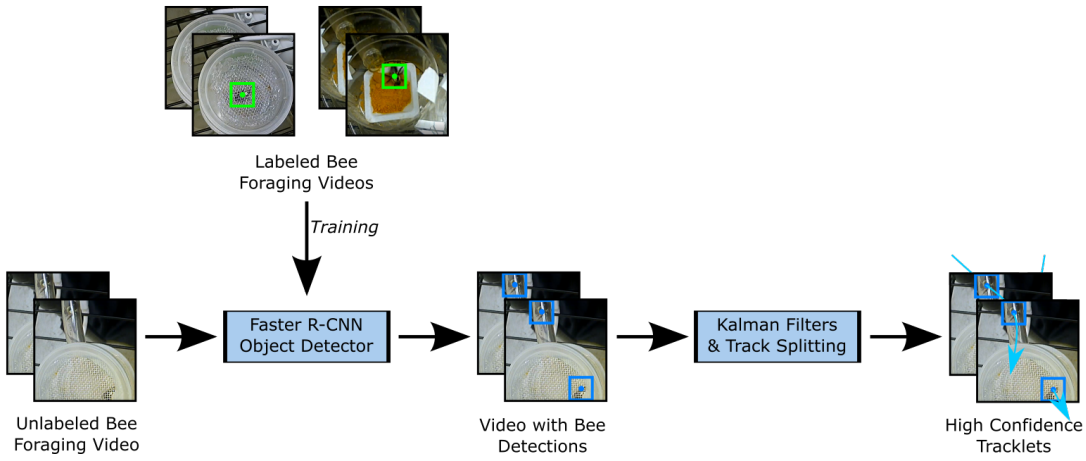

Figure 3.1: High confidence tracklet production

3.2.2 Model

To produce an affinity score from our input features, we utilize a Siamese gated recurrent unit (GRU) network to produce an embedding between the pair of tracklets. The co-occurring tracklets are then zero padded and summed together to collapse them into a single sequence in order to account for varying numbers of co-occurring tracklets between pairs. The summed co-occurring tracklet sequence is then fed into separate GRU network to learn an embedding for the co-occurring tracklets. The two embeddings are then concatenated along with the time gap and fed into a pair of fully connected layers, producing an affinity score.

Siamese networks have long been used to produce affinity scores, so we continue to use them here. We additionally combine them with the GRU networks commonly used on sequential data such as these tracklets. Since we are not using visual appearance data

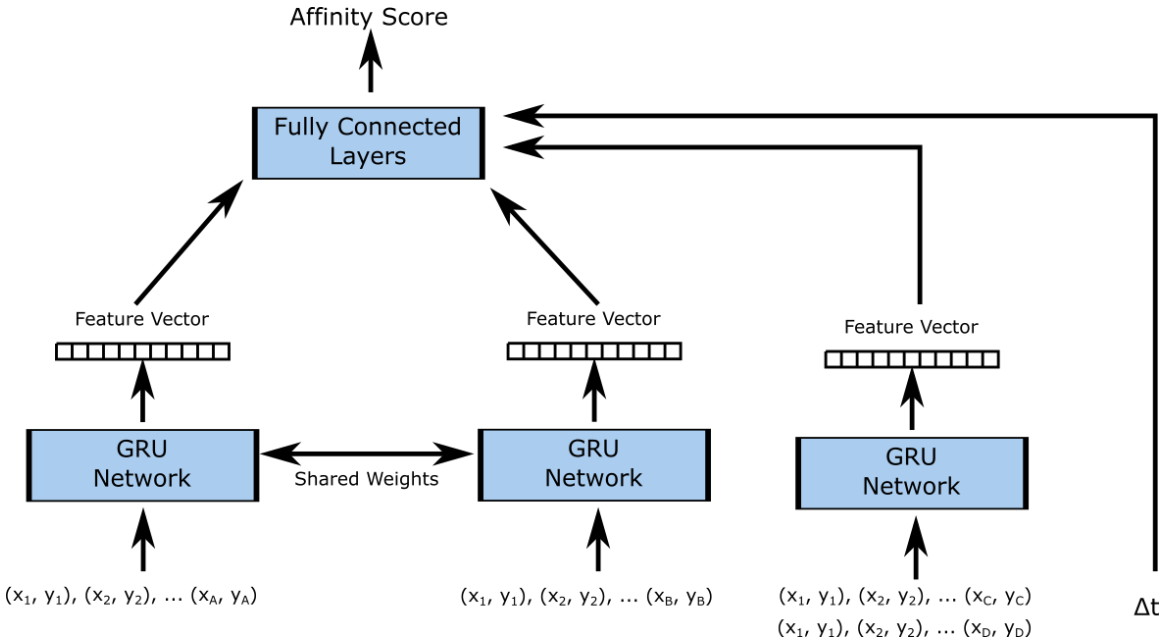

Figure 3.2: Motion affinity model with inputs

as input, CNNs offer comparatively little advantage. The model is illustrated in Figure 3.2.

3.2.3 Training with Iterative Pseudo-Labeling

In order to train our affinity model for long-range reconnection of tracklets without the inter-tracklet ground-truth, we us iterative pseudo-labeling (IPL). In the semi-supervised IPL setting, training begins with a small set of labeled data, and another set of unlabeled data. In this context, the high confidence linkages between detections in the same tracklet will serve as our labeled data, while the unknown linkages between the tracklets will serve as the unlabeled data.

To begin, we must first train our initial affinity model on the "labeled" dataset. To do this we, simply simulate track losses within each tracklet, randomly splitting each

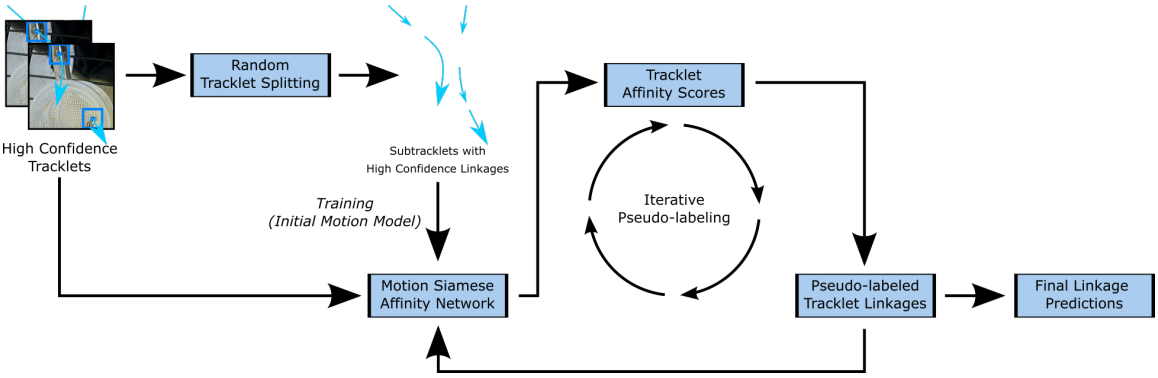

Figure 3.3: Motion IPL methodology

tracklet into a random number of smaller sub-tracklets. Sub-tracklets from within the same tracklet are then randomly paired together to form positive examples, while sub-tracklets from different tracklets are paired to form negative examples. Once the initial affinity model has been trained to convergence, the iterative pseudo-labeling can begin.

First a small sample of the "unlabeled" tracklet pairs are chosen the for next round of pseudo-labeling, based on predetermined sample rate hyperparameter. Next, a forward pass is performed on the chosen tracklet pairs, calculating an affinity score for each of them. A score close to 0 or 1 indicates a higher confidence affinity, so these are marked with negative and positive pseudo-labels respectively. However, any affinity scores that are above a predetermined lower threshold or below an upper threshold are deemed to have too little confidence to be pseudo-labeled, and are thus returned to the unlabeled set.

The model is then trained again for a few iterations on both the "labeled" sub-tracklet dataset and the newly pseudo-labeled dataset using a combined loss between the two datasets, with a weight hyperparameter on the pseudo-label loss.

The process then repeats until all of the unlabeled dataset has received pseudo-

labels and the model has converged. After the IPL training of the model has been completed, it is then used to produce affinity scores as normal for the tracking-by-detection paradigm, and the tracklet linkages can be then optimally solved using a bipartite graph matching method such as the Hungarian algorithm.

Chapter 4

Experiments

Our experimental goals are as follows: (i) establish and assess the limitations of appearance based features for tracking identical objects with challenging occlusions, and (ii) investigate the efficacy of using motion features in lieu of appearance features.

4.1 Data

To test our method, we utilize videos of bees foraging for pollen and nectar in a pair of plastic arenas, provided by the Woodard Lab at UC Riverside's Department of Entomology. Bees are extremely similar in visual appearance, and are difficult even for humans to differentiate between. The foraging arena environment complicates the MOT task in these videos further by occluding the bees at various points throughout the on-camera trajectories of the bees. Occlusions include the opaque edges of the arena, as well as the objects inside the arenas, such as the pollen tray. Additionally, a wire mesh over the tops of the arenas adds some noise to visual signal of the bees. Ground truth labeling

for these videos consists of points centered on the locations of the bees while they are on camera and at least partially unoccluded. These labels were obtained by hand through a human labeler following bees through the video using a mouse pointer.

Our tracklets are produced from detections provided by a Faster-RCNN detector trained on bee foraging videos with fixed-size bounding boxes added to the coordinate ground truth labeling (i.e. only positional information). Detections are attributed to new or existing tracks based on per-track Kalman filters. The tracks from the Kalman filters are then split whenever there exists a gap of more than 1 frame, i.e. anytime the track was lost and picked back up. These tracklets are recorded as a sequence of positions, images associated with those positions, and the starting and ending frames of these sequences (since they are consecutive).

4.2 Metric

Results are measured by identity switches, the number of times that a predicted identity changes in a ground truth trajectory. However, this alone is not sufficient as a the data association method can easily stitch together an arbitrary number of tracklets, so we additionally report the total number of predict trajectories as well.

Ideally, ID switches should be as close to zero as possible, where as the number of predicted trajectories should be as close to the number of ground truth trajectories.

4.3 Baseline

We begin our experiments by conducting a series of baseline tests using different models of tracklet affinity. As these methods were generally quicker to run, we have performed a fine grid search on many of the hyperparameters for them as in 4.1. We then select the best of these runs using based on the lowest number of ID switches while not exceeding more than 5 too few or too many predicted trajectories. These best results are reported in each of the tables below.

4.3.1 Simple Baseline

In order to mark baseline performance on our dataset, we first apply the tracking-by-detection paradigm with a very simple affinity score based on distance between tracklets, d, and the time gap between them, Δt. Time gaps are naturally measured in terms of frames between tracklets and the distances are measured as the Euclidean distance between the ending position of the pre-occurring tracklet and the starting position of the post-occurring tracklet. The time gap is divided by the frames per second and the distances are divided by the length of the camera diagonal, which is the longest possible Euclidean distance for the video. Values are then normalized according to the mean and variance.

$$\text{Affinity} = d + w_1 \Delta t$$

These affinities are stored in an affinity matrix representing a bipartite graph, for which an optimal solution is then found using the Hungarian algorithm. The results for

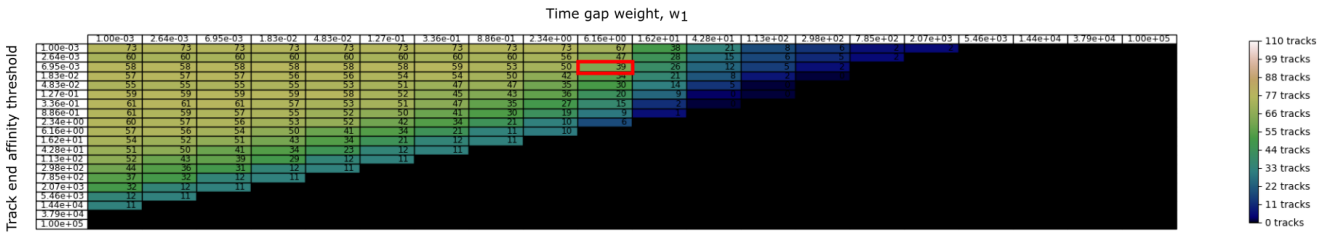

Figure 4.1: Simple baseline: Best ID switches marked in red

	Simple
ID Switches	39
Predicted Tracks	56

Table 4.1: Simple baseline: Best ID switches

this simple baseline are presented in Table 4.1.

4.3.2 Simple Visual Features

Next, we assess the effect of adding simple visual features to the affinity measure. We try both histogram of gradient (HOG) and color histogram features. For both, the features are calculated for each image in each tracklet, then averaged across the tracklet. The color histogram feature consists of three (RGB) histogram vectors, each averaged across the tracklet. Distances are calculated between the two tracklets' features and than averaged across the three channels, α. For the HOG features, distances are simply the distance between the two tracklets' feature vectors.

$$Affinity = d + w_1 \Delta t + w_2 \alpha$$

As before, optimal trajectories are then found from the affinity matrix using the

	HOG	Color hist
ID Switches	34	40
Predicted Tracks	54	57

Table 4.2: Simple appearance baselines: Best ID switches

Hungarian algorithm. The results are displayed in Table 4.2.

The HOG features failed to perform well, as a bee is captured at many different rotations which blurs the values of the different angle bins. The color histograms did not seem to perform much better either, likely due to the visual similtarity of the bees.

4.3.3 Triplet Loss

Here we test a more modern deep learning approach, training a Siamese CNN to produce a 256-dimensional embedding from the previously mentioned tracklet information using triplet loss. Triplets are produced by taking an anchor image from one position in a tracklet (chosen randomly), a positive image from another point in the same tracklet, and a negative image either from a co-occurring tracklet (which by default, means it cannot be the same object), or an image of empty background.

In order to stick with the semi-supervised setting, we form our triplets without using our knowledge of ground-truth inter-tracklet linkages. This is accomplished by finding all pairs of tracklets that co-occur in time. These two tracklets cannot be linked, as an object cannot occupy more than one position at the same time. The anchor and positive images are randomly selected from one of the tracklets, and the negative image is selected from the other. From these we form 3027 triplets and train our a Siamese network for 100 epochs.

	Triplet	Triplet w/ GT
ID Switches	38	39
Predicted Tracks	57	56

Table 4.3: Siamese reID appearance baselines: Best ID switches

The Siamese network trained by triplet loss produces a 256-dimensional appearance vector for each tracklet, which is then combined with the time gap and distance scores to produce an affinity matrix which is optimally solved as above. The ID switches for this are presented in Table 4.3.

The results indicate no significant improvement from the simple appearance features tested above, indicating the ineffectiveness of appearance features on identical objects.

One possible reason was the relatively small amount of tracklet pairs used to form these triplets, which perhaps excluded more useful appearance features in the tracklets that did not occur alongside any other tracklets. To test this, we produced triplets using the inter-tracklet ground truth, instead forming triplets from whole ground-truthed trajectories rather than simple tracklets. The results are shown in Table 4.3 as well. Even with the added benefit of the ground truth, the learned appearance features were still not useful in reconnecting long range track losses.

4.4 Motion IPL

Here we test our Motion IPL method. Our initial affinity model is trained by randomly splitting our each of our high confidence tracklets into an random number of smaller sub-tracklets, then training our Motion IPL network on these artificially disconnected sub-

	ID Switches	Predicted Tracks
Simple	39	56
HOG	34	54
Color hist	40	57
Triplet	38	57
Triplet w/GT	39	56
Motion IPL	**12**	**59**

Table 4.4: Best ID Switches for each method

tracklets. The initial affinity model is trained for 10k iterations.

After initial training of the affinity model on the high confidence linkages between the sub-tracklets, we predict an affinity score on a sample of the original tracklet pairs (using a 20% sampling rate), and assign pseudo labels based to pairs with scores high enough to warrant a positive affinity pseudo-label or low enough for a negative pseudo-label. Here we use a positive pseudo-label threshold of 0.8 and a negative pseudo-label threshold of 0.3. The model is than trained for another 5 iterations and the pseudo-labeling process is repeated until all tracklet pairs have been assigned pseudo-labels. The network is then used to produce affinity scores for each of the original tracklet pairs, and then the Hungarian algorithm is used to assign tracklet linkages based on these affinities as in the previous experiments. Our results are summarized in Table 4.4.

ID switches were reduced to 12, indicating that the usage of more complex motion features can help us effective link tracklets in the absence of useful appearance features. We also note that we end up with 59 predicted trajectories versus the 55 ground-truth trajectories, indicating that there were that there were still several track fragmentations beyond the ID switches.

Chapter 5

Conclusions

We have investigated the inefficacy of the appearance features widely used for modern long-range tracklet reconnection in situations with visually identical objects. In multiple experiments, we have shown that both classical and modern appearance features have limited capability to perform this task semi-supervised. Even with the additional benefit of ground-truth knowledge between tracklets, long-range reconnection ability of a Siamese CNN trained with triplet loss seemed ineffective.

We then proposed a Motion IPL method for reconnecting lost tracks using no appearance features and instead focusing on motion features and **the interactions between tracklets relative to each other throughout time.**

5.1 Limitations

The conclusions of this work are limited by the data it has been tested on. Thus far it has only been tested on a single video of bee foraging. More work will need to be

done to further test and verify these results on not only more videos of the same domain, but hopefully also other challenging domains involving greater quantities of objects. It is possible the success of our method on this video has been due to the minimal number of objects and rather few pathways for movement available in the bee foraging arenas.

With that said, we do note that this domain has also provided some difficult challenges already with frequent occlusions of the objects by background objects, as well as the comparatively erratic and unpredictable movements of bees compared to other commonly examined objects such as cars or pedestrians.

Bibliography

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[4] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[5] Erik Bochinski, Tobias Senst, and Thomas Sikora. Extending IOU based multi-object tracking by visual information. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6:737–744, 1993.

[7] Caglayan Dicle, Octavia I. Camps, and Mario Sznaier. The way they move: Tracking multiple targets with similar appearance. In *2013 IEEE International Conference on Computer Vision*. IEEE, December 2013.

[8] Michael Isard and Andrew Blake. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[9] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *TRANS. ASME, SER. D, J. BASIC ENG*, page 109, 1961.

[10] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoue, Thomas Brox, and Bjorn Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[11] Minyoung Kim, Stefano Alletto, and Luca Rigazio. Similarity mapping with enhanced siamese network for multi-object tracking. *Machine Learning for Intelligent Transportation Systems (MLITS), 2016 NIPS Workshop*, 2016.

[12] Adam R Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential attend, infer, repeat: generative modelling of moving objects. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8615–8625, 2018.

[13] Adam R. Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8615–8625, Red Hook, NY, USA, 2018. Curran Associates Inc.

[14] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Gustav Hager, Alan Lukezic, Abdelrahman Eldesokey, et al. The visual object tracking vot2017 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1949–1972, 2017.

[15] Sangyun Lee and Euntai Kim. Multiple object tracking via feature pyramid siamese networks. *IEEE Access*, 7:8181–8194, 2019.

[16] Weiqiang Li, Jiatong Mu, and Guizhong Liu. Multiple object tracking with motion and appearance cues. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, October 2019.

[17] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, page 103448, 2020.

[18] Liqian Ma, Siyu Tang, Michael J. Black, and Luc Van Gool. Customized multi-person tracker. In *Computer Vision – ACCV 2018*. Springer International Publishing, December 2018.

[19] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3602, 2015.

[20] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[22] ShiJie Sun, Naveed Akhtar, XiangYu Song, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Simultaneous detection and tracking with motion modelling for multiple object tracking. In *European Conference on Computer Vision*, pages 626–643. Springer, 2020.

[23] Bing Wang, Li Wang, Bing Shuai, Zhen Zuo, Ting Liu, Kap Luk Chan, and Gang Wang. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[24] Fan Wang, En Zhu, Lei Luo, and Jun Long. Multi-object tracking combines motion and visual information. In Vicenç Torra, Yasuo Narukawa, Jordi Nin, and Núria Agell, editors, *Modeling Decisions for Artificial Intelligence*, pages 166–178, Cham, 2020. Springer International Publishing.

[25] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[26] Junbo Yin, Wenguan Wang, Qinghao Meng, Ruigang Yang, and Jianbing Shen. A unified object motion and affinity model for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] Shun Zhang, Yihong Gong, Jia-Bin Huang, Jongwoo Lim, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. Tracking persons-of-interest via adaptive discriminative features. In *European conference on computer vision*, pages 415–433. Springer, 2016.

[28] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020.