POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Thesis

Enhanced Road User Tracking in Cluttered Urban Environments through Multiple Hypothesis Tracking and Probabilistic Object Discrimination

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Francesco Romeo

Advisor: Prof. Simone Vantini

Co-advisor: Yury Tarakanov

Academic year: 2021-2022

1. Introduction

Object tracking is the process of following moving entities in subsequent observations coming from one or multiple sensors, with the aim of estimating and predicting their trajectories. Accurate tracking in complex urban scenarios is crucial for safety applications such as accident mitigation, predictive traffic control and design of safer infrastructure. In this work, a tracking system based on Multiple Hypothesis Tracking (MHT) will be implemented and tested in complex urban scenarios. Together with developing a tracking system, the aim of this thesis is to address the problem of different road users that are merged into a single track due to their proximity. To tackle this issue, a strategy that uses the information contained in bounding boxes describing the objects in the scenario has been defined and tested. The resulting tracker will be tested on a given dataset that comprises a collection of frames coming from one stereovision sensor designed by Viscando AB, a Swedish company specializing in traffic data collection and analysis for safe and smart mobility applications, in collaboration with which this work has been produced.

The assumption of this work is that the combination of Multiple Hypothesis Tracking and the probabilistic approach to discriminate merged objects will generate an algorithm that is able to outperform the conventional single hypothesis tracker in solving the common problems that a tracking system faces.

2. Multiple Hypothesis Tracking

The tracking system that will be implemented uses a **Kalman Filter** (KF) to build trajectories of the road users. In particular, the motion model that is at the basis of the KF is the **constant speed model**. Although this type of motion model does not account for variation in speed and turning rate, it is nonetheless widely used given its simplicity and its linearity, which guarantees the convergence of the KF. Another possible motion model that has the same properties is the constant acceleration one, which will be compared with the constant speed one later on.

A KF allows every type of tracking system to perform two important steps: prediction of the next position of an object, based on the previous positions and by means of the motion model and, if the position is actually available, filtering of the noise contained in the datum. These two operations provide the tracking system with an estimation of the position of the object in the current frame. The second step, i.e., the filtering step, is not possible if a position for the object is not available in a given frame. In this case, the position predicted by the KF will be considered the actual one for the object.

The main difference between a Multiple Hypothesis Tracking (MHT) system and a single hypothesis tracker is that the former allows for multiple trajectories for the same target and eventually chooses the one that is most likely to represent the real target's trajectory based on motion and appearance features and compatibility with the other existing tracks. As a consequence, in a MHT framework, a track is a hypothetical trajectory of the target that differs from the other hypotheses of the same target in the detections that it has been associated with.

Let's now see the main steps that characterize a MHT approach. The following implementation is inspired by the work presented in [1]. As every tracking system does, this algorithm iterates over the frames of a given data set and performs, on each frame, the operations that will follow.

The first step is the **association** between the existing track hypotheses and the detections coming from a specific frame of a data set. Given the *i-th* track, its next position at frame k, labeled as X_k^i, is assumed to be normally distributed:

$$X_k^i \sim \mathcal{N}(\hat{x}_k^i, \Sigma_k^i)$$

the mean $\hat{x}_k^i$ is the position predicted by the filter and the covariance Σ_k^i is the covariance of the filter, both at time step k. This assumption allows to identify the area in which the position of the track is expected to fall at time step k, which is called the "**gating area**" of the track. The **gating area** of the track is given by the points whose squared Mahalanobis distance from the track's predicted position $\hat{x}_k^i$ is smaller than a fixed threshold:

$$d^2 = (y - \hat{x}_k^i)^T (\Sigma_k^i)^{-1} (y - \hat{x}_k^i) \leq d_{th} \qquad (1)$$

The *gating area* is an ellipsoid which is delimited by the level curve of the density function of the random vector X_k^i. All the data points from frame k, called **detections**, that fall into this gating area are possible next positions of the *i-th* track and will thus be associated with it. If more than one detection falls into the gating area, then there will be new track hypotheses for the target, one for each detection. The best structure to represent the different track hypotheses referring to the same object is a tree structure: each node is a detection, one hypothesis corresponds to a path from the root node to one of the leaves, and every new detection associated with an existing hypothesis spawns a new branch of the tree. If a detection is not associated with any existing track, it starts its own tree, i.e., it is referred to a new target in the scenario.

The second phase is the **computation of the score**. To be able to choose the tracks that are most likely to represent real targets, it is necessary to assign a score to each of them. In this work, the score of the track is based on both motion and appearance features. Considering, for example, the *i-th* track hypothesis, its score at frame k can be designed as follows:

$$S^i(k) = \omega_{mot} S^i_{mot}(k) + \omega_{app} S^i_{app}(k)$$

where S_{mot} and S_{app} are denominated **motion score** and **appearance score**. They are linked to the evolution in position and appearance of the track. Their weights depend on the implementation. In this work both weights are set equal to one. Consider the *i-th* track and let k be the index of the current frame, then it is possible to show that the motion score can be obtained through a recursive formula:

$$S^i_{mot}(k) = \Delta S^i_{mot}(k) + S^i_{mot}(k-1)$$

$$\Delta S^i_{mot}(k) = \ln\left(\frac{V}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}\ln(|\Sigma_k^i|) - \frac{1}{2}d^2$$

with d^2 being the quantity in (1), V the image area and n the dimension of the position vector. Regarding the appearance score, it is important to mention that the appearance information that is usually provided in this framework is the bounding box of the object, which is a box that should represent the dimension of the object and its orientation in the space. As a consequence, the proposed appearance score corresponds to the **Intersection over Union** (IoU) between the bounding box of the track and the bounding box of the new associated detection, labeled as b_k. To maintain the recursive nature of the score, the bounding box of the track corresponds to the bounding box of the detection associated with the track in the previous frame centered in the position predicted for the track in the current frame, labeled as b_{k-1}:

$$S^i_{app}(k) = \Delta S^i_{app}(k) + S^i_{app}(k-1)$$

$$\Delta S^i_{app}(k) = \frac{Volume(b_{k-1}) \cap Volume(b_k)}{Volume(b_{k-1}) \cup Volume(b_k)}$$

Lastly, if a track has not been associated with any detection, its score is updated with a negative quantity that depends on the probability P_D that an object will be detected by the algorithm: $\Delta S^i(k) = \ln(1 - P_D)$.

The third step performed by the algorithm is the computation of the **best global hypothesis**. As a consequence of the previous step, each track has a score. This score is now used to determine the set of tracks that are more likely to represent real trajectories of the involved road users, called the *best global hypothesis*. It is important to say that two tracks are incompatible if they share at least one detection. Two incompatible tracks cannot coexist, so they cannot simultaneously be in the set of the best tracks. The problem to address is the following: find the set of compatible hypotheses with the highest sum of their scores. This problem can be reformulated as a **Maximum-Weight Clique Problem** (MWCP), which in this thesis is solved by exploiting the algorithm explained in [2], which gives an exact solution.

Now that the best global hypothesis is available, it is possible to choose which track hypotheses to retain and which to discard. The last step is thus the **pruning** one. The terminology derives from the fact that discarding a track corresponds, in the chosen representation, to pruning a branch of a tree. A pruning strategy is very important to keep the number of hypothesis limited and reduce the computational complexity of the algorithm. The pruning strategies adopted in this work are the ones suggested in [1]. The algorithm uses **N-scan pruning** to prune, i.e. discard, tracks. This means that if a tree doesn't have any of its track hypotheses in the best global hypothesis, then it is completely pruned; if one track belonging to the tree is in the best global hypothesis, then all the tracks that at time step $k - N$ differ from this track are discarded. The underlying assumption that an N-scan strategy has is that the ambiguities in the detection association step for frames 1 to $k - N$ can be resolved after looking ahead for a window of N frames. Moreover, if a tree grows too much, only N_{best} tracks are kept and the others are discarded based on their scores.

3. Detect and Split Merged Objects

One of the most common problems that a tracking system has to face is the presence of detections that refer to not only one target but two or more. This can happen whenever two road users are close to each other or one occludes the other, causing the sensors to detect them as one unique object. This issue causes tracks to be interrupted or merged into a single track. For this reason, a strategy to tackle this problem has been defined in this work alongside the MHT algorithm. Each track has a collection of detections that have been associated with it in the past frames. Each detection contains information on the position and size of the object, represented by a bounding box. The collection of the volumes of the bounding boxes of these detections forms a dataset from which a confidence interval for the volume of the underlying target can be obtained. In particular, a bootstrap strategy is used to compute these confidence intervals. This consists in sampling N times with replacement a new dataset from the dataset of volumes and in computing N times the sample mean on the new dataset. As a result, N estimates of the volume of the target are available. These estimates can be used to build the empirical distribution of the sample mean, from which an estimate of the quantiles of the distribution of the sample mean is obtained. These estimated quantiles provide a confidence interval for the volume of the target, which is called the **Bootstrap Confidence Interval**:

$$CI_\alpha(\theta) = [2\hat{\theta} - \hat{\theta}^*_{\alpha/2}, 2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}] \tag{2}$$

where θ is the volume of the target, $\hat{\theta}$ is the sample mean computed on the original dataset, $\hat{\theta}^*_\alpha$ is the estimated quantile and $1 - \alpha$ is the confidence level, which in this work will be 95%. It is worth mentioning that this confidence interval

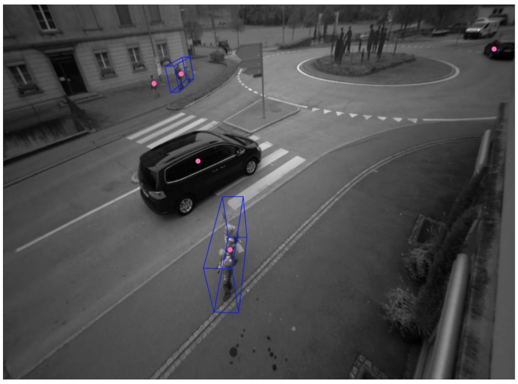

Figure 1: Frame number 10

has been built by means of the **Plug-in Principle**, which allows to substitute the quantiles of the unknown distribution of θ with the quantiles of the distribution of the estimator $\hat{\theta}$, which have been estimated through the empirical distribution.

4. Dataset and Implementation of the Algorithm

As mentioned before, the proposed algorithm will be tested on a provided dataset, which is a collection of information from 1 stereovision sensor, with detections corresponding to approximately 750 frames, or one minute of activity, of a roundabout in Kölliken, Switzerland. Data has been previously processed to provide points in 3D as centers of detected objects, and bounding boxes containing those objects. As a result, each row of the dataset contains information on the position and the bounding box related to a certain road user, together with the frame in which the road user appears. This data set contains the usual problems that a tracking system has to face: objects occlusion, presence of detections that are not referred to any road users but to stationary objects, presence of obstacles (road signs and statues in the middle of the roundabout) and detections that are missing. Figure 1 is an example of a frame in the data set, with some of the aforementioned problems.

The implementation of the algorithm follows the steps described in Section 2. The tracking system iterates over the frames extracting from the data set the detections belonging to each frame. Then, the association step is performed considering these detections and all the tracks that represent objects in the roundabout (called *Active* or *Pending*, in case they were not associated with any detection in the previous frames). In this phase, an association between a track and a detection is performed if and only if the detection falls in the gating area of the track and the Intersection over Union between the track and the detection is larger than zero. Moreover, if two tracks are associated with the same detection, the *bootstrap confidence intervals* of the volumes of the two tracks are computed to check if the detection represents two merged objects. If the volume of the associated detection is outside both the confidence intervals, then the detection is labeled as two merged objects. If this happens, the position of the track is updated either with the position predicted by the filter or with a weighted mean between the center of the fused bounding box and the predicted position. Once the association phase is complete, the score of the tracks is updated and the best global hypothesis is computed. Using the best global hypothesis, the tracks are pruned using the strategies mentioned in Section 2.

5. Results

Now let's analyze the results that the tracking system gives on the dataset described in the previous section. Given the fact that ground-truth trajectories are not available, it is not possible to resort to the usual framework to assess the quality of the tracker. As a consequence, the performance of the tracker will be analyzed qualitatively, i.e., by showing which type of problems the tracker is able to solve and which issues the tracker struggles the most with. The results that will be shown are related to the values of the parameters reported in the following table:

Parameters MHT				
d_{th}	P_D	V	N-scan	N_{best}
6	0.9	480000	5	100

Table 1: Values for parameters of MHT

The values are inspired by the implementation of MHT in [1].

First of all, the algorithm is designed to address the problem of the presence of merged objects. Looking at all cases of merged objects, it is possible to say that the algorithm is able to identify them and successfully solve them. The

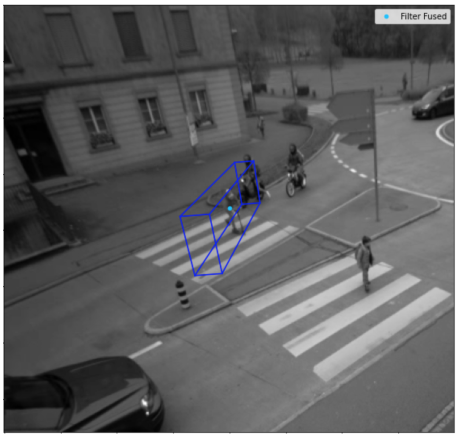

Figure 2: The scooter is crossing the street. The closest motorcyclist is merged with the pedestrian in a unique detection.

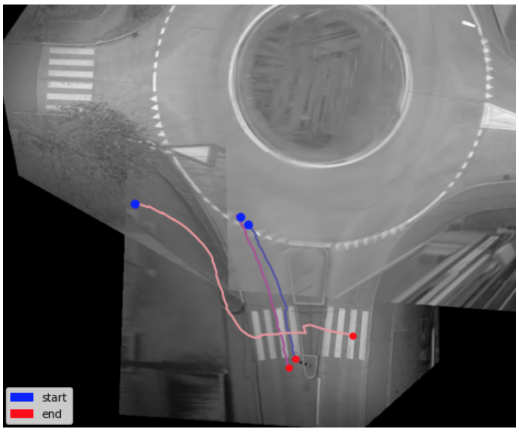

Figure 3: Tracks of the pedestrian and the two motocyclists. They are separated despite the presence of fused detections.

main problematic behaviors that can arise from this issue are the presence of interrupted tracks or the occurrence of an identity switch between tracks, i.e., a track that changes its target. Neither of those two problems arise thanks to the strategy that uses bootstrap confidence intervals. Figure 2 shows a frame of a scooter that is crossing the street. The proximity of the scooter with the other two road users (the motorcycles) causes the sensor in the roundabout to detect the scooter and one of the two motorcycles as unique objects, as the figure shows.

The detection of the scooter will be merged subsequently with the second motorcycle as well. Despite the lack of three distinct detections for the three road users for multiple frames, the algorithm manages to keep the three trajectories separated. As a result, in the output of the algorithm, i.e., the best global hypothesis after 750 frames, there are three distinct tracks for these three objects, an accomplishment that was not achievable with a single hypothesis tracker.

Figure 3 shows a bird's eye view of a part of the three distinct tracks.

This situation is not an isolated one. The tracker is able to recognize almost all cases of fused detections and keep the tracks of the involved targets separate.

It is also important to mention that after the fusion situation ends, the track of the scooter is associated with a detection that is not the closest one. In this case, using a Multiple Hypothesis approach helped the tracker consider not only the closest detection but also others as possible next position of the track and select the one that is most likely to be the one of the underlying target. This would not have been possible with a single hypothesis tracker. However, it is worth mentioning that only nine tracks in the best global hypothesis are associated with a detection that is not the closest one, and overall this happens in a total of 13 frames. These digits make it clear that in most of the cases in this scenario the best detection to be associated with a track is the closest one. As a consequence, the usage of Multiple Hypothesis Tracking is not strictly necessary in most of the cases, although it is very useful in these situations of fused detections if coupled with a proper strategy, as previously shown.

In addition to this, there are also some cases of detections that are labeled as fused even if they do not represent two fused objects. This usually happens when tracks related to stationary objects, and not road users, are then associated with vehicles or when there are two detections for the same vehicle for multiple frames. To solve this issue and improve the accuracy of identifying merged objects tighter controls on when to use the bootstrap confidence intervals should be implemented. However, overall, it is possible to say that the tracker performs successfully with real cases of merged objects.

A situation in which the tracker is not performing well is the case of occluded objects, which are also very frequent in the dataset. The statues in the middle of the roundabout and the road sign, visible in Figure 1, cause a lot of detections to be missing. In particular, in correspondence with the road sign all the tracks are interrupted. The occlusion does not last long, but the requirement of having the Intersection over Union bigger than zero in the association phase is the main reason why these tracks are interrupted. The constraint is necessary for the algorithm in cases in which the covariance of the filter is too large; thus, it is important to analyze this behavior to enhance the performances of the tracker.

To conclude the analysis of the results of the algorithm, it is important to mention that for 30 targets, the number of tracks that the algorithm produced is 50. Among these tracks, there are four cases of **Identity Switch**, that is, tracks that change their targets, eleven cases of **False Positive** tracks, that is, tracks that do not have a road user as a target and there are only two targets that do not have a corresponding track, mainly because there are no detections available for these targets for enough frames.

6. Additional Experiments

Among the additional experiments, it is worth mentioning that the usage of kinematic constraints, i.e., constraints on speed or acceleration, to limit the number of associations and avoid unfeasible values of speed and acceleration has been tested. The experiment led to the conclusion that these constraints limit too much the ability of the filter to capture the variability of the data, creating a tracker that interrupts multiple tracks due to the fact that the given data are noisy. In general, it is better to avoid this approach given the fact that detections are never precise and often noisy.

Moreover, the constant acceleration motion model has been tested in place of the constant speed one, without giving significantly different results. Table 2 shows a comparison between the two trackers with the two different motion models.

It is possible to say that on the given framework the two motion models produce basically the same output.

	CA model	CS model
False Positives	9	11
ID Switches	5	4
Non-tracked Objects	3	2

Table 2: Comparison of motion models

7. Conclusions

In this thesis, a Multiple Hypothesis Tracking approach coupled with a probabilistic approach to detect merged objects and keep their tracks separated has been proposed and tested on a given dataset. The work shows that a MHT approach is not strictly necessary in an urban scenario that is not as crowded as an only pedestrian one compared to a single hypothesis tracker. Nevertheless, it can be very useful in situations of merged objects if coupled with a proper strategy. The strategy proposed in this work exploits bootstrap confidence intervals for the volumes of the objects that have been merged to understand if there is a situation where two road users have been merged together. This algorithm is able to detect all the cases of merged objects and to solve them, i.e., to keep the tracks of the involved targets separated and not interrupt them.

Among the further improvements, it is worth mentioning that the accuracy of the algorithm in detecting fused objects can be enhanced using tighter constraints to guarantee that the confidence intervals are used only when two road users are possibly involved in a situation of merged detections. In addition to this, the performances of the tracker when objects are occluded needs to be improved.

References

[1] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4696–4704, 2015.

[2] Patric R.J. Östergård. A new algorithm for the maximum-weight clique problem. *Electronic Notes in Discrete Mathematics*, 3:153–156, 1999. 6th Twente Workshop on Graphs and Combinatorial Optimization.