Multi-Person Tracking with a Multi-Hypothesis Approach for Ambiguous Assignments

Daniel Stadler

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
daniel.stadler@kit.edu

Abstract

Multi-person tracking is often solved with the tracking-by-detection paradigm, in that a distance measure is calculated for each possible track-detection assignment. Then, the sum of distances of all assignments has to be minimized, for which mostly the Hungarian method is used. Wheareas it is easy to design a distance measure that can clearly indicate the correct assignments in sequences with sparse person distributions, the distances of some assignments can be very similar in crowded scenes, where multiple persons share similar spatial positions and appearances. As a consequence, wrong assignments are inescapable, harming the tracking performance. In contrast of executing all assignments simultaneously, no matter if they are clear or ambiguous, this work treats ambiguous assignments with similar distances separately following a multi-hypothesis approach, updating the hypotheses until the assignment task is clear again. To determine which assignments are considered ambiguous, a method that compares the entries in the distance matrix of track-detection assignments is introduced. No further information next to the distance matrix is needed, which makes the proposed approach applicable to any tracking-by-detection based method. Experimental results show that the separate treatment of ambiguous assignments can improve the tracking performance in crowds and thus is a promising research directory.

1 Introduction

Multi-person tracking (MPT) is the task of detecting and identifying all persons in each frame of a video and is the basis for several applications ranging from action classification to crowd behavior analysis.

Most of the MPT approaches in literature follow the *tracking-by-detection* paradigm [2, 3, 7, 12, 14, 15, 16, 17, 18], which divides the problem into two subtasks: detection and association. In each iteration, the generated detections are assigned to the tracks from the previous time step on the basis of a distance measure, whereby mostly the Hungarian method [5] is applied for minimizing the overall costs. When designing the distance measure, different target information can be leveraged. Especially position information [2, 3] and visual cues [7, 15, 16, 18] are used. Some works additionally consider human poses [15, 17] or relation information w.r.t. other targets [7, 13, 18]. Designing such sophisticated distance measures aims to achieve a high degree of distinguishability between correct and incorrect assignments. However, there will still exist some situations, in that the assignment task is ambiguous, no matter how good the designed distance measure is. This holds especially true in crowded scenes, where multiple targets share similar positions and appearances. Furthermore, inaccurate and missing detections under heavy occlusion often prevent a clear association.

Because of the aforementioned reasons, a new association strategy is proposed, which treats *ambiguous* assignments separately with a multi-hypothesis approach, while solving the *clear* assignments with the Hungarian method as usual. To find ambiguous assignments between detections and tracks (and track hypotheses), a closer look at the distance matrix is taken. More precisely, the distances of all possible track-detection assignments are compared and if they differ by less than a similarity threshold, the involved tracks and detections are termed *similar*. If additionally, the numbers of tracks and detections that lead to similar assignments are different, i.e., either tracks or detections would remain unassigned, the assignments are considered ambiguous and multiple track hypotheses are built. These are updated in consecutive frames until the assignment task is clear again. With this strategy, ambiguous association decisions as often occurring under occlusion can be postponed und thus assignment errors prevented. As the determination of ambiguous assignments builds purely upon the distance matrix,

the proposed approach can be included in any tracking-by-detection method, independent from how the distances are calculated.

2 Find Ambiguous Assignments

In each time step t of a tracking-by-detection based approach, detections $\mathcal{D}^{(t)} = \{D_1^{(t)}, D_2^{(t)}, \dots, D_N^{(t)}\}$ are matched to tracks from the previous iteration $\mathcal{T}^{(t-1)} = \{T_1^{(t-1)}, T_2^{(t-1)}, \dots, T_M^{(t-1)}\}$. For each track $T_i \in \mathcal{T}^{(t-1)}$ and detection $D_j \in \mathcal{D}^{(t)}$, a distance $d(T_i, D_j)$ is calculated and saved at the respective position (i, j) in the distance matrix $\mathbf{D} \in \mathbb{R}^{M \times N}$. Hereby, various information can be used. For example, position, appearance, or motion cues are often considered. In contrast to the standard association, which applies the Hungarian method on the full distance matrix $\mathbf{D}$, it is argued that some detections and tracks might lead to *ambiguous* assignments, that should be treated separately. After that, the remaining *clear* assignments are handled by the Hungarian method.

In the following, a subset of possible assignments including track indices $\mathcal{I} \subset \{1, \dots M\}$ and detection indices $\mathcal{J} \subset \{1, \dots N\}$ is noted as tuple of sets $A = (\mathcal{I}, \mathcal{J}) = (A[1], A[2])$. For example, the possible assignments A of tracks T_1, T_3 and detections D_2, D_4 are noted as $A = (\{1, 3\}, \{2, 4\})$. The search for ambiguous assignments starts with *similar* assignments. Assignments are termed similar, if the respective entries in the distance matrix differ by less than a similarity threshold Δ. For example, the possible assignments of track T_1 to detection D_1 and of track T_1 to D_2 are similar, if $|\mathbf{D}[1, 1] - \mathbf{D}[1, 2]| < \Delta$ holds. If multiple detections *and* tracks are similar, the distance matrix $\mathbf{D}$ has to be scanned several times, iteratively searching for similar detections and tracks. The process for finding all similar assignments w.r.t. a specific detection (D_2) is shown for a toy example distance matrix in Figure 2.1. This procedure is done for each detection leading to a tentative set of similar assignments $\tilde{\mathcal{A}}^{\text{sim}} = \{A_j\}_{j=1\dots N}$. Then, the assignments that share track or detection indices are merged leading to the final set of similar assignments $\mathcal{A}^{\text{sim}}$. For instance, $\tilde{\mathcal{A}}^{\text{sim}} = \{(\{1, 2\}, \{3\}), (\{3\}, \{3, 4\})\}$ would turn to $\mathcal{A}^{\text{sim}} = \{(\{1, 2, 3\}, \{3, 4\})\}$ applying the merging operation. After determining $\mathcal{A}^{\text{sim}} = \{A_k\}_{k=1\dots K}$, it is decided for each element, whether the similar

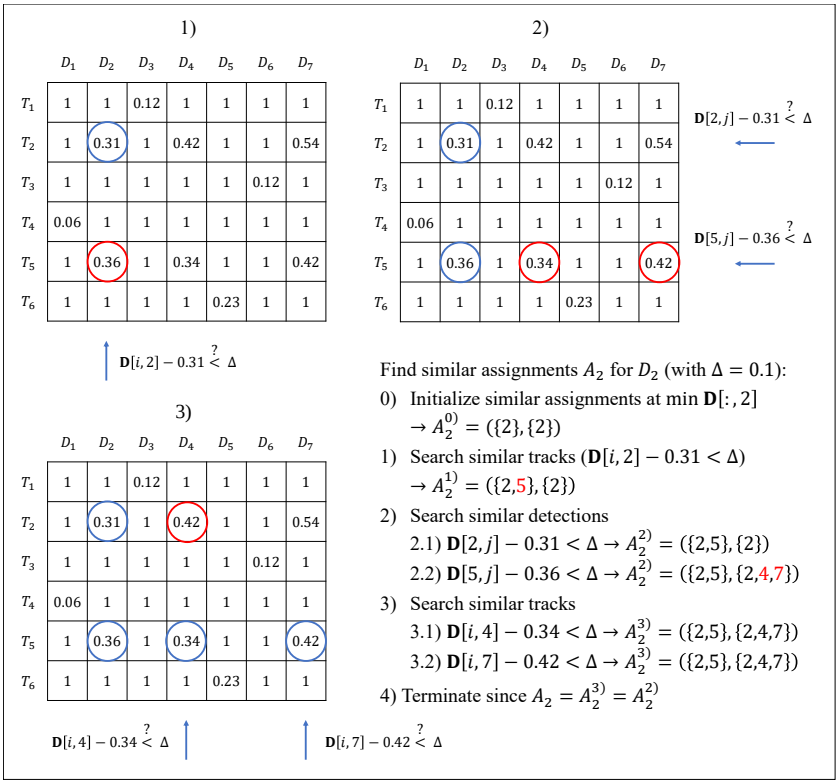


Figure 2.1: Process of finding similar assignments A_2 for detection D_2. **0)** Initialization of A_2 is done at the minimum distance of tracks w.r.t. D_2 which is $\mathbf{D}[2,2] = 0.31$ for track T_2 leading to $A_2^{0)} = (\{2\},\{2\})$. **1)** Similar tracks with lower distance difference to 0.31 than Δ are searched which yields T_5 (highlighted in red) and $A_2^{1)} = (\{2,5\},\{2\})$. **2)** Similar detections w.r.t. T_2 and T_5 are searched. D_4 and D_7 are added to the similar assignments $A_2^{2)} = (\{2,5\},\{2,4,7\})$. **3)** Again, similar tracks are searched, now for D_4 and D_7 yielding only T_2 which is already represented in $A_2^{2)}$. Thus $A_2^{3)}$ equals $A_2^{2)}$. **4)** Since A did not change in iteration 3), the algorithm stops, outputting $A_2 = A_2^{3)} = (\{2,5\},\{2,4,7\})$ as similar assignments for D_2.

assignments A_k are considered as ambiguous or not. It is argued that, if the number of tracks $n_T = |A_k[1]|$ and the number of detections $n_D = |A_k[2]|$ that

lead to similar assignments A_k is identical, the situation is not ambiguous, since each track and detection can be matched. This holds especially true for similar assignments with only one track and one detection ($n_T = n_D$). Those assignments are termed as *clear*, while similar assignments, for which the numbers of tracks and detections differ ($n_T \neq n_D$), i.e., there are missing detections or missing tracks, are termed *ambiguous*. Formally, the set of similar assignments $\mathcal{A}^{\text{sim}}$ is divided into a set of ambiguous assignments $\mathcal{A}^{\text{amb}}$ and a set of clear assignments $\mathcal{A}^{\text{clr}}$:

$$\mathcal{A}^{\text{amb}} = \{A | A \in \mathcal{A}^{\text{sim}} \wedge A[1] \neq A[2]\} \tag{2.1}$$

$$\mathcal{A}^{\text{clr}} = \{A | A \in \mathcal{A}^{\text{sim}} \wedge A[1] = A[2]\} \tag{2.2}$$

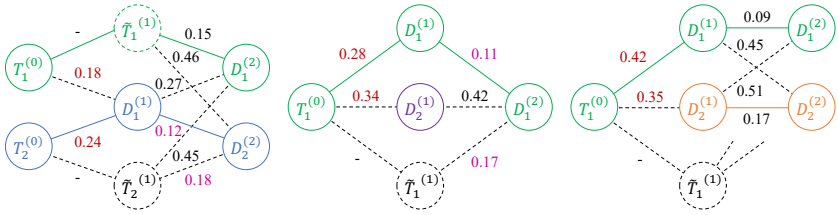
Note that $\mathcal{A}^{\text{sim}} = \mathcal{A}^{\text{amb}} \cup \mathcal{A}^{\text{clr}}$ holds. The track indices $\mathcal{I}^{\text{clr}}$ and detection indices $\mathcal{J}^{\text{clr}}$ of the clear assignments $\mathcal{A}^{\text{clr}}$ are used to generate a clear distance matrix $\mathbf{D}^{\text{clr}} = \mathbf{D}[\mathcal{I}^{\text{clr}}, \mathcal{J}^{\text{clr}}]$ on which the Hungarian method is applied. In contrast, the ambiguous assignments $\mathcal{A}^{\text{amb}}$ are treated separately with a multi-hypothesis tracking (MHT) approach that is described in the next section.

3 Solve Ambiguous Assignments with MHT

For each group of tracks and detections that lead to ambiguous assignments, multiple hypotheses are started and thus the assignment problem is postponed. In consecutive iterations, the set of hypotheses is updated until the assignment task is clear again. In the following, the procedure of building and updating the set of track hypotheses is exemplary described for an ambiguous situation with missing detection (FN). This situation is also depicted in Figure 3.1(a), however, note that the full figure has not to be understood at this point.

At time step $t = 1$, there is only one detection $D_1^{(1)}$ but there have been two tracks $T_1^{(0)}$ and $T_2^{(0)}$ in the previous iteration. Furthermore, the detection fits nearly equally well to both tracks ($0.24 - 0.18 < \Delta = 0.1$) so ambiguous assignments $A = (\{1, 2\}, \{1\}) \in \mathcal{A}^{\text{amb}}$ are present. Therefore, the two hypotheses $H_{11}^{(1)}$ and $H_{21}^{(1)}$ are built:

$$H_{11}^{(1)} = [T_1^{(0)}, D_1^{(1)}] \qquad H_{21}^{(1)} = [T_2^{(0)}, D_1^{(1)}] \tag{3.1}$$



(a) FN (b) FP (c) New detection

Figure 3.1: Illustration of the three types of ambiguous assignments ($\Delta = 0.1$) that can be handled by the proposed multi-hypothesis approach: (a) false negatives (FN), (b) false positives (FP), and (c) new detections which initialize new tracks. Detections and tracks are visualized with circles, whereby propagated tracks (without assigned detection) are indicated with dashed circles and a tilde. Hypotheses with the smallest distance that correspond to the resulting trajectories are marked in green, blue, and orange. Hypotheses with higher distances are drawn in dashed lines. Detections that are removed are depicted purple. Distances between detections and tracks are written near the edges of the graph. Distances that lead to ambiguous assignments are highlighted in red, whereas assignments with similar distances that can be clearly resolved are highlighted in pink. Note that there is no distance (-) for propagated tracks. Furthermore, time steps are specified in superscript and for situation (c), hypotheses with the propagated track are omitted in the second step for clarity.

In addition, two hypotheses $H_{10}^{(1)}$ and $H_{20}^{(1)}$ that are based on a Kalman filter prediction step (see details about the motion model in Section 4.2) are started:

$$H_{10}^{(1)} = [T_1^{(0)}, \tilde{T}_1^{(1)}] \qquad H_{20}^{(1)} = [T_2^{(0)}, \tilde{T}_2^{(1)}] \tag{3.2}$$

Note that propagated tracks are indicated with a tilde. The overall set of track hypotheses $\mathcal{H}^{(1)} = \{H_{10}^{(1)}, H_{11}^{(1)}, H_{20}^{(1)}, H_{21}^{(1)}\}$ is saved for the next time step. At $t = 2$, there are two detections $D_1^{(2)}$ and $D_2^{(2)}$ that update the set of hypotheses to $\mathcal{H}^{(2)} = \{H_{101}^{(2)}, H_{102}^{(2)}, H_{111}^{(2)}, H_{112}^{(2)}, H_{201}^{(2)}, H_{202}^{(2)}, H_{211}^{(2)}, H_{212}^{(2)}\}$ with:

$$\begin{aligned}
H_{101}^{(2)} &= [T_1^{(0)}, \tilde{T}_1^{(1)}, D_1^{(2)}] & H_{102}^{(2)} &= [T_1^{(0)}, \tilde{T}_1^{(1)}, D_2^{(2)}] \\
H_{111}^{(2)} &= [T_1^{(0)}, D_1^{(1)}, D_1^{(2)}] & H_{112}^{(2)} &= [T_1^{(0)}, D_1^{(1)}, D_2^{(2)}] \\
H_{201}^{(2)} &= [T_2^{(0)}, \tilde{T}_2^{(1)}, D_1^{(2)}] & H_{202}^{(2)} &= [T_2^{(0)}, \tilde{T}_2^{(1)}, D_2^{(2)}] \\
H_{211}^{(2)} &= [T_2^{(0)}, D_1^{(1)}, D_1^{(2)}] & H_{212}^{(2)} &= [T_2^{(0)}, D_1^{(1)}, D_2^{(2)}]
\end{aligned} \tag{3.3}$$

For each hypothesis $H \in \mathcal{H}^{(2)}$, a distance $d(H)$ has to be determined in order to find out whether the assignment problem is still ambiguous or clear again. The distance of a hypothesis is set to the distance between the last two entries of the hypothesis. For instance, the distance of $H_{111}^{(2)} = [T_1^{(0)}, D_1^{(1)}, D_1^{(2)}]$ is $d(H_{111}^{(2)}) = d(D_1^{(1)}, D_1^{(2)}) = 0.27$. With the hypothesis distances and the information about track and detection numbers n_T and n_D, respectively, within a set of hypotheses, one of the following two requirements has to be fulfilled that the assignment problem is considered clear again:

1. Track number and detection number are identical: $n_T = n_D$.
2. There are more detections than tracks ($n_D > n_T$) in each iteration *and* the number of detections is the same for two successive time steps: $n_D^{(t)} = n_D^{(t-1)}$.

The first item corresponds to cases (a) and (b) and the second item applies for case (c) in Figure 3.1. The three types of ambiguous assignments (false negatives, false positives, new detections) are discussed in the following subsections.

The attentive reader may have noticed that not all distances of the eight hypotheses in $\mathcal{H}^{(2)}$ are depicted in Figure 3.1(a). As the assignment of a detection to a track changes its motion prediction, the propagated track position differs for two tracks that would be assigned the same detection. Thus, for example, $d(H_{111}^{(2)}) \neq d(H_{211}^{(2)})$ holds, which is not considered in Figure 3.1(a) for clarity. Another side note is that the overall number of hypotheses is limited to maintain a low computational complexity when multiple time steps are involved. More precisely, the $h_{\max}$ hypotheses with the lowest distances are kept in each iteration.

3.1 Missing Detections

Missing detections (FN) appear frequently in crowded scenes, where the detector cannot recognize all targets due to occlusion. At the same time, the assignment of the available detections can be ambiguous due to inaccuracies of the detection boxes or the propagated track boxes. This is also the case for $D_1^{(1)}$ in Figure 3.1(a) which can be assigned either to $T_1^{(0)}$ ($d = 0.18$) or to $T_2^{(0)}$ ($d = 0.24$) as already

discussed. In the next iteration, however, the assignment becomes clear as the distance $d(D_1^{(1)}, D_2^{(2)}) = 0.12$ is significantly lower (in terms of $\Delta = 0.1$) than $d(D_1^{(1)}, D_1^{(2)}) = 0.27$. Therefore, the assignment task is clear again and the ambiguous situation can be resolved. Here, three things should be noted. First, with the multi-hypothesis approach, an identity switch is prevented, since detection $D_1^{(1)}$ would be erroneously assigned to $T_1^{(1)}$ in the standard association. Second, the missing detection for track T_1 is bridged by track propagation with the motion model. Third, the two hypotheses $H_{202}^{(2)}$ and $H_{212}^{(2)}$ are similar (pink) but not ambiguous, as hypotheses with propagated track boxes are not considered competing to hypotheses of the same track with a detection assigned.

3.2 Duplicate Detections

In crowded scenes, it can also happen that the detector produces duplicate detections (FP), struggling to recognize the precise boundaries of the targets. In the simplest case, two detections $D_1^{(1)}$ and $D_2^{(1)}$ fit nearly equally well to a track $T_1^{(0)}$ as in Figure 3.1(b). In the standard association, that starts new tracks with unassigned detections, an additional duplicate track would be initialized that could introduce further tracking errors. In contrast, the multi-hypothesis approach postpones the decision, which detection should be assigned to the track, until the situation is clear again. Then, the unassigned duplicate detection (purple) can be identified and removed. Note that it would also be possible that the propagated track box $\tilde{T}_1^{(1)}$ is the best match to $D_1^{(2)}$, namely if both detection boxes $D_1^{(1)}$ and $D_2^{(1)}$ are inaccurate due to occlusion. In that case, the propagated track $\tilde{T}_1^{(1)}$ would be used and both detections removed.

3.3 New Detections

In the previous subsection, a situation with $n_D > n_T$ has been treated, where the additional detections have been considered as FP. However, this is not the case, when a new target is about to show for the first time, still partly occluded by a nearby target. Fortunately, the proposed multi-hypothesis approach can identify such situations implicitly as shown in Figure 3.1(c). Whereas the situation is similar to case (b) in the first time step, two detections are again

present in the second iteration. Furthermore, the assignment task is clear at $t = 2$. Therefore, it is likely that the additional detections belong to a separate target, also because FP mostly occur only in single frames. As a consequence, the hypothesis with the smallest distance of 0.09 (green) is taken for track $T_1^{(0)}$, whereas $D_2^{(1)}$ and $D_2^{(2)}$ start a new track $T_{\text{new}}^{(2)} = [D_2^{(1)}, D_2^{(2)}]$ (orange). Note that an identity switch is prevented with the multi-hypothesis approach that postpones the assignment decision until the situation is clear again.

4 Experiments

4.1 Dataset and Evaluation Metrics

The MOT17 dataset [8] is used in the experiments, since it is one of the most popular benchmarks for evaluating multi-target tracking performance. It comprises a train and a test split with 7 videos each. Since the annotations of the test split are not publicly available, the train split is divided into two halves, enabling the evaluation of the tracker in combination with a fine-tuned detection model, which is necessary for achieving good results. More precisely, the images of the first part of each sequence are taken for fine-tuning the detector, while the tracker is evaluated on the second parts of the sequences.

The tracking performance is measured in IDF1 [10], that emphasizes on identity preservation abilities, and MOTA [1], which focuses more on detection quality. Furthermore, the components of MOTA are reported, i.e., number of false negatives (FN), false positives (FP), and identity switches (IDSW).

4.2 Implementation Details

As detection model for the tracking-by-detection based approach, a Faster R-CNN [9] with FPN [6] and ResNet-50 [4] as backbone is used. The model is first pre-trained on the CrowdHuman dataset [11] with a batch size of 16 and an initial learning rate of 0.01 for 30 epochs, which is lowered by factor 10 after epochs 24 and 27. Then, the model is fine-tuned on the first half of MOT17 with an initial learning rate of 0.001 with the same schedule. For the tracking process,

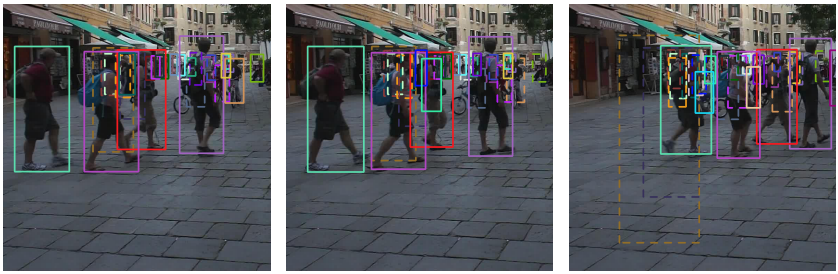
the generated detections are filtered with a minimum score threshold of 0.9 and an Intersection over Union (IoU) threshold of 0.5 is applied in the non-maximum suppression (NMS) step. The distance measure between a track T and detection D is calculated as $d(T, D) = 1 - \mathrm{IoU}(T, D)$. Both in the standard association and the proposed multi-hypothesis approach for ambiguous assignments, a maximum distance of $d_{\max} = 0.8$, which corresponds to a minimum IoU of 0.2, is enforced for matching tracks and detections. Tracks are propagated with a Kalman filter as motion model, whereby the implementation of [16] is used. Inactive tracks and track hypotheses are maintained for a maximum number of 40 iterations without assigned detection before termination or deletion. At re-activation, a linear interpolation is performed to close the gap of missed detections. For finding ambiguous assignments in the distance matrix $\mathbf{D}$, a similarity threshold $\Delta = 0.1$ is applied. The number of track hypotheses h is limited to $h_{\max} = 10$ to keep a low computational complexity of the approach.

4.3 Results

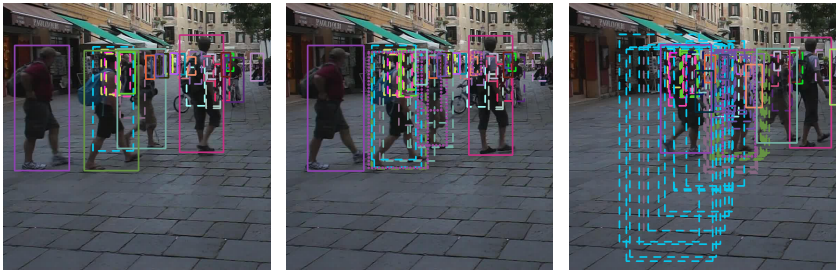
To get a feeling, in which situations the proposed multi-hypothesis approach for ambiguous assignments is superior to the standard association, in that all assignments are made with the Hungarian method, several experiments with different settings are run. The results are summarized in Table 4.1.

MH ($n_D < n_T$)	MH ($n_D > n_T$)	IDF1	MOTA	FN	FP	IDSW
✗	✗	76.1	73.2	24798	18036	600
✓	✗	**77.8**	72.6	**24264**	19593	597
✗	✓	76.9	**73.6**	25935	**16218**	**591**
✓	✓	76.8	73.1	25563	17403	615

Table 4.1: Tracking results of the proposed multi-hypothesis approach for ambiguous assignments in comparison with the standard association (first row). MH ($n_D < n_T$) corresponds to applying the multi-hypothesis approach only in situations with missing detections, while in the MH ($n_D > n_T$) variant, the approach is only applied in situations with more detections than tracks. The last row shows results, where for all ambiguous assignments, multiple track hypotheses are built.



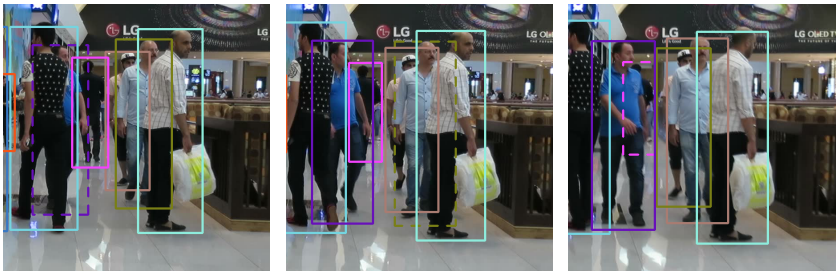
(a) Baseline



(b) Proposed approach

Figure 4.1: Failure case of the proposed multi-hypothesis approach. Active tracks are drawn in solid lines, inactive tracks or track hypotheses in dashed lines. (a) In the standard association, some incorrect inactive tracks are propagated over the image, which is not optimal but inevitable if a re-activation of tracks after occlusion should be possible. (b) In the multi-hypothesis approach, the incorrect inactive tracks mistakenly lead to ambiguous assignments. The involved ambiguous detections (marked as purple dotted boxes) increase the set of hypotheses for the incorrect inactive tracks. As a consequence, many incorrect hypotheses emerge that can introduce tracking errors.

One can see that the multi-hypothesis approach for ambiguous assignments with missing detections ($n_D < n_T$) improves IDF1 by 1.7 points, however, MOTA is reduced by 0.6 points. Qualitatively, it is observed that the approach is vulnerable to some cases with incorrect inactive tracks as shown in Figure 4.1. Incorrect inactive tracks always pose a risk for introducing tracking errors. The multi-hypothesis approach increases this risk, when multiple hypotheses for



(a) Baseline

(b) Proposed approach

Figure 4.2: Positive example of the proposed multi-hypothesis approach. (a) In the standard association, the ID of the man with light blue shirt switches with the ID of the occluded man with cap. The cause of the IDSW is the ambiguous detection indicated with a purple dotted box in the lower middle frame. (b) Instead of directly assigning this ambiguous detection, multiple hypotheses (4 in total, for both tracks one with track propagation and one with assigned detection) are built. Later, the assignment task is clear again. Postponing the association decision prevents the IDSW.

such an incorrect inactive track are built. In future experiments, this problem should be further investigated. It might be better to consider only active tracks for building hypotheses. One situation, where the multi-hypothesis approach successfully solves an ambiguous assignment problem can be found in Figure 4.2. Whereas in the baseline association, the ambiguous detection is assigned to the wrong track leading to an identity switch, all targets can be successfully tracked when the assignment decision is postponed with the help of multiple hypotheses

until the situation is clear again. Note that the example in Figure 4.2, where two tracks are competing for one detection, is exactly as depicted in Figure 3.1(a).

Having again a look at Table 4.1, it can be seen that the multi-hypotheses approach for situations with more detections than tracks ($n_D > n_T$) both enhances IDF1 and MOTA by 0.8 and 0.4 points, respectively. As expected, the number of FP can be greatly reduced, since many duplicate detections appear only in single frames and thus are removed with the multi-hypothesis approach as in Figure 3.1(b). Furthermore, the number of IDSW is slightly decreased.

Unfortunately, applying the multi-hypothesis approach for all types of ambiguous assignments (last line in Table 4.1), the results cannot be further improved. Whereas the decrease in MOTA is expected as the MH ($n_D > n_T$) variant also lowers MOTA, the reduction of IDF1 is surprising.

In future works, a deeper analysis has to be made to better understand the negative impact of incorrect inactive tracks in the multi-hypothesis approach. Furthermore, more ablative experiments could be run (on tracking parameters) to get a deeper understanding of the proposed approach and find out other possible error sources. Nevertheless, it has been shown, that separately treating ambiguous assignments is a promising idea, for which the development of other strategies, next to the proposed multi-hypothesis approach, should be explored.

5 Conclusion

In this report, a novel association technique for multi-target tracking is proposed that treats ambiguous assignments separately with a multi-hypothesis approach. The ambiguous assignments are determined purely based on the distance matrix of tracks and detections and thus, the proposed method can be applied in any tracking-by-detection based approach. The track hypotheses allow to postpone the association decision for ambiguous assignments until the situation is clear again, which can improve the association accuracy in ambiguous situations. Besides showing the superiority of the proposed approach in comparison to the standard association in some scenarios, also some weaknesses are identified and suggestions for possible improvements in future works are made.

References

[1] Keni Bernardin and Rainer Stiefelhagen. “Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics”. In: *EURASIP Journal on Image and Video Processing* 2008 (2008).

[2] Alex Bewley et al. “Simple Online and Realtime Tracking”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3464–3468.

[3] 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)*. 2017.

[4] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[5] Harold W. Kuhn. “The Hungarian Method for the Assignment Problem”. In: *Naval Research Logistics Quarterly* 2.1–2 (1955), pp. 83–97.

[6] Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 936–944.

[7] Qiankun Liu et al. “GSM: Graph Similarity Model for Multi-Object Tracking”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. 2020, pp. 530–536.

[8] Anton Milan et al. “MOT16: A Benchmark for Multi-Object Tracking”. In: *arXiv:1603.00831* (2016).

[9] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149.

[10] Ergys Ristani et al. “Performance Measures and a Data Set for Multi-target, Multi-camera Tracking”. In: *Computer Vision – ECCV 2016 Workshops*. Vol. 9914. Lecture Notes in Computer Science. 2016, pp. 17–35.

[11] Shuai Shao et al. “CrowdHuman: A Benchmark for Detecting Human in a Crowd”. In: *arXiv:1805.00123* (2018).

[12] Andreas Specker et al. “An Occlusion-Aware Multi-Target Multi-Camera Tracking System”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2021, pp. 4173–4182.

[13] Daniel Stadler and Jürgen Beyerer. “Improving Multiple Pedestrian Tracking by Track Management and Occlusion Handling”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 10958–10967.

[14] Daniel Stadler, Lars Wilko Sommer, and Jürgen Beyerer. “PAS Tracker: Position-, Appearance- and Size-Aware Multi-object Tracking in Drone Videos”. In: *Computer Vision – ECCV 2020 Workshops*. Vol. 12538. Lecture Notes in Computer Science. 2020, pp. 604–620.

[15] Siyu Tang et al. “Multiple People Tracking by Lifted Multicut and Person Re-identification”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3701–3710.

[16] 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)*. 2017, pp. 3645–3649.

[17] Bin Xiao, Haiping Wu, and Yichen Wei. “Simple Baselines for Human Pose Estimation and Tracking”. In: *Computer Vision – ECCV 2018*. Vol. 11210. Lecture Notes in Computer Science. 2018, pp. 472–487.

[18] Jiarui Xu et al. “Spatial-Temporal Relation Networks for Multi-Object Tracking”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 3987–3997.