Stereo and Motion Based 3D High Density Object Tracking

Junli Tao, Benjamin Risse, and Xiaoyi Jiang

University of Auckland, Computer Science
University of Münster, Computer Science, Neurobiology
jtao076@aucklanduni.ac.nz,
b.risse@wwu.de,
xjiang@uni-muenster.de

Abstract. In order to understand the behavior of adult *Drosophila melanogaster* (fruit flies), vision-based 3D trajectory reconstruction methods are adopted. To improve the statistical strength of subsequent analysis, high-throughput measurements are necessary. However, ambiguities in both stereo matching and temporal tracking appear more frequently in high density situations, aggravating the complexity of the 3D tracking situation. In this paper we propose a high density object tracking algorithm. Instead of approximating trajectories for all frames in a direct manner, in ambiguous situations, tracking is terminated to generate robust tracklets based on the modified tracking-by-matching method. The terminated tracklets are linked to ongoing (unterminated) tracklets with minimum linking cost in an on-line fashion. Furthermore, we introduce a set of new evaluation metrics to analyze the tracking results. These metrics are used to analyse the effect of detection noise and compare our tracking algorithm with two state-of-the-art 3D tracking methods based on simulated data with hundreds of flies. The results indicate that our proposed algorithm outperforms both, the tracking-by-matching algorithm and a global correspondence selection approach.

Keywords: Drosophila melanogaster, fruit flies, 3D tracking, tracklets, stereo matching, Kalman filter, evaluation metrics.

1 Introduction

For almost all animals, the ability to move is pivotal for finding food, mating partners or escaping from dangerous situations. During evolution, an increasingly complex nervous system allowed sophisticated locomotion control. Therefore, vision based locomotion analysis of various organisms is an important subject in neurobiological research [13].

Drosophila melanogaster (i.e. fruit fly) is one of the most popular model organisms to study the nervous system. It is a holometabolous insect. In the larval stage, movement is restricted to two dimensions and behavioral experiments are well established using 2D tracking [12,15]. In the adult stage several 2D behavioral experiments are done by cutting the wings [3] or using an arena with flat

R. Klette, M. Rivera, and S. Satoh (Eds.): PSIVT 2013, LNCS 8333, pp. 136–148, 2014.

ceiling [17] to restrict the locomotion to two dimensions. However these manipulations lead to unnatural behavior [5].

Reconstructing 3D trajectories of hundreds of objects with similar appearance is challenging. On the one hand, more than one camera is necessary to determine the 3D positions with cross view correspondences. On the other hand, those 3D positions are associated over time to generate trajectories. Thus, two subprocesses are involved to obtain 3D trajectories: stereo matching between different views and temporal tracking over time. Both leading to the so-called general multi-index assignment problem, which is non-deterministically polynomial-time hard ($\mathcal{NP}$-hard) [2].

For a small number of simultaneously tracked objects (about 10 objects), epipolar constraint is sufficient for stereo matching [16]. However, in high density scenes (e.g. hundreds of objects) the ambiguity of both tasks increased significantly: If there are multiple objects close to an epipolar line in view 2 corresponding to a single object in view 1, stereo matching is ambiguous (Figure 1 (left,mid)). Temporal tracking is more ambiguous if there are more than one possible successors within the search region of a tracked object (Figure 1 (right)). Furthermore, the number of occlusion increases in high density scenes which affects both, cross views and temporal associations.

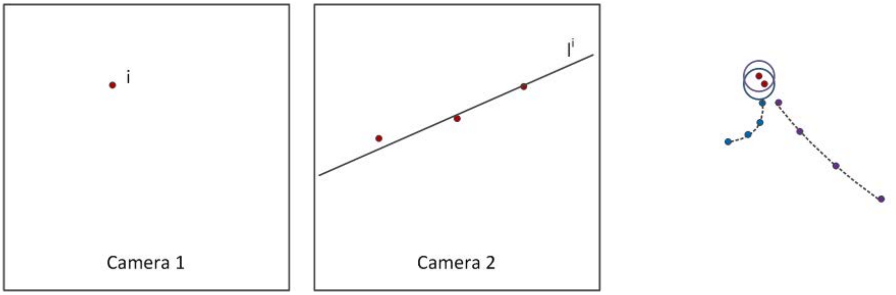

Fig. 1. Stereo and temporal ambiguities. The projection i of a object in Camera 1 (left) may have multiple candidates, as multiple projections are located on or close to its epipolar line l^i in Camera 2 (mid). Right: Blue and purple empty circles denote the search region for two tracklets shown with blue and purple solid circles respectively. Projections in current frame are represented by red solid circles.

1.1 Related Work

Methods for estimating 3D trajectories are typically based on stereo matching and temporal tracking; consequently they require two or more cameras. In [4,18], 2D trajectories are calculated in the image plane, and then matched between cameras to reconstruct 3D trajectories. Alternatively, in [11,6,14], stereo matching is used to reconstruct 3D coordinates, then followed by tracking 3D points to obtain 3D trajectories. However, these methods are vulnerable to either stereo or tracking ambiguities. In [9], several more frames are considered together to deal with stereo matching ambiguities. A modified Hungarian method is proposed to handle stereo matching for sequences containing up to seven objects [1]. Beside utilizing several cameras, a single-camera setup in combination with two mirrors is used in [7] to track up to 10 flies in a comparatively small test tube. In [6,14], more than two cameras are used for the 3D tracking task. [6] adopts three cameras for reconstructing a 3-dimensional hull for a fly, and then tracks the hull using the extended Kalman filter (EKF). In a similar approach [14] employs up to eleven cameras for realtime trajectory estimation.

[20] handles stereo matching and tracking simultaneously by minimizing a cost function related to the epipolar constraint, kinetic coherency, and observation matches. However the domain of the cost function increases exponentially for both, the number of objects and the number of frames. In [16], a third camera is employed to verify stereo pairs in the other two views. Stereo matching and temporal tracking are conducted alternatively to further reduce the ambiguities, but generating fragmented or incorrectly merged trajectories.

Multiple pedestrian tracking is a different task but shares certain similarities with fruit fly tracking. Both aim to reconstruct trajectories of multiple objects. In high density scenes, occlusions happen more frequently, which increases the difficulty to obtain good results. Pedestrian tracking adopts appearance, motion and temporal cues to deal with occlusions [19,8]. Appearance is considered as the most important cue to avoid identity switch [8]. However, fruit flies share similar appearance so that appearance is not a useful cue for reducing identity switches in temporal tracking. As the size of the object is small, occlusion time is relatively short. Thus, motion and temporal cues are reasonable selections for fruit fly temporal tracking.

1.2 Our Approach

In this paper, we propose a robust 3D tracking algorithm for high density object trajectory reconstruction. Trinocular stereovision is adopted to reduce stereo ambiguity in binocular stereovision by utilizing the projection consistency [16]. The tracking algorithm is an extended version of the Tracking-by-Matching (TbM) algorithm, which uses the epipolar constraint for stereo matching and the Kalman filter for temporal tracking [16]. In the conventional TbM approach trajectories were extended as long as there are valid successors available. If no valid successors are found, tracks are terminated and reinitialized if the ambiguity is solved in at least two of the three views [16]. Unfortunately this leads to fragmented or incorrectly merged trajectories and prevents the preservation of fly identities.

In the proposed approach only unambiguous situations are used to generate robust *tracklets* (i.e. trajectory segments). Tracklets which can not be associated to a unique successor are terminated and subsequently linked to appropriate temporally and spatially ongoing tracklets in an on-line fashion. We use both, motion and location context for tracklet association. The 3D trajectories are reconstructed with paired 2D trajectories.

A set of new evaluation metrics is proposed to quantify the tracking performance. The effect of detection noise is analyzed with these new metrics.

The proposed algorithm outperforms the state-of-the-art algorithms [16,20] in sequences with up to 200 simulated objects so that ground truth is available.

2 Notation

Given three time-synchronized sequences recorded from calibrated cameras Camera 1, Camera 2 and Camera 3. Then $\mathbf{I}_t^i$ represents an image obtained from Camera i ($i = 1, 2, 3$), at time t. In each $\mathbf{I}_t^i$, the detected fruit flies (i.e. detections) in each view at each frame are denoted by $D_t^i = \{d_{n^i,t}^i\} = \{(u_{n^i,t}^i, v_{n^i,t}^i)\}$ for $n^i = 1, \ldots, N_t^i$, where $(u_{n^i,t}^i, v_{n^i,t}^i)$ is the centroid of a blob (i.e. projection of a fly) in view i.

Stereo pairs are generated by matching blobs $d_{n^i,t}^i$ from D_t^1, D_t^2 and D_t^3 between the views. These matches can be associated over time to generate trajectories $S = \{s_{t_s:t_e}^k\}, k = 1, \ldots, K$, where K denotes the number of trajectories and t_s, t_e denote the start and end time of a trajectory. A trajectory $s_{t_s:t_e}^k = \{s_{t_s}^k, \ldots, s_{t_e}^k\}$ consists of a set of states s_t. A state is defined by

$$s_t = ((d_{n^1,t}^1, d_{n^2,t}^2, d_{n^3,t}^3), (\mathbf{v}_{n^1,t}^1, \mathbf{v}_{n^2,t}^2, \mathbf{v}_{n^2,t}^3)) \tag{1}$$

containing the projection term $(d_{n^1,t}^1, d_{n^2,t}^2, d_{n^3,t}^3)$ and the corresponding velocity term $(\mathbf{v}_{n^1,t}^1, \mathbf{v}_{n^2,t}^2, \mathbf{v}_{n^2,t}^3)$ belonging to a object in three views, where $(\mathbf{v}_{n^i,t}^i = (\nu_{u,n^i,t}^i, \nu_{v,n^i,t}^i))$. The 3D trajectories $\mathbf{T} = \{T_{t_s:t_e}^k\}$, where $T_t^k = (x, y, z)$ is the 3D location, are obtained from S by triangulating stereo pairs from triplets.

3 Proposed Algorithm

The proposed causal approach aims to handle 3D trajectory reconstruction in high density scene, see Figure 2. Because of the high occlusion rate, trajectories are frequently fragmented (missing valid triplet) or merged (multiple trajectories sharing one triplet). Thus, our approach proposes to generate robust tracklets by modifying the TbM method [16], and then associate tracklets using 2D motion and location context information.

3.1 Robust Tracklet Generation

The TbM method proposed in [16] is adopted to generate robust tracklets with more tracklet termination constraints. All candidate stereo pairs are verified based on epipolar line between any two views. Thus, the resultant triplets all satisfy the projection consistency. For the first three time-synchronized frames, triplets are found by exhausted search cross the three views, see Figure 2. Then, valid triplets are employed to initialize a triplet Kalman filter tracker. One Kalman filter is applied to one element of a triplet respectively. With the estimated velocities $(\mathbf{v}_{n^1,t}^1, \mathbf{v}_{n^2,t}^2, \mathbf{v}_{n^2,t}^3)$ from the filters, predicted locations in the image planes are optained for frames $\mathbf{I}_{t+1}^i$ by utilizing a constant velocity model.

Subsequently, search regions around the predicted locations are used to compare the detections with the predictions. Ambiguities (e.g. more than one detection is located in the search region) are addressed by verification, correction and fetching as described in [16].

In order to obtain robust tracklets, triplet tracking is terminated if one of the following termination conditions is satisfied:

- no valid triplet found within the search region;
- missing detections in more than one view;
- more than one tracklets associated with one valid triplet.

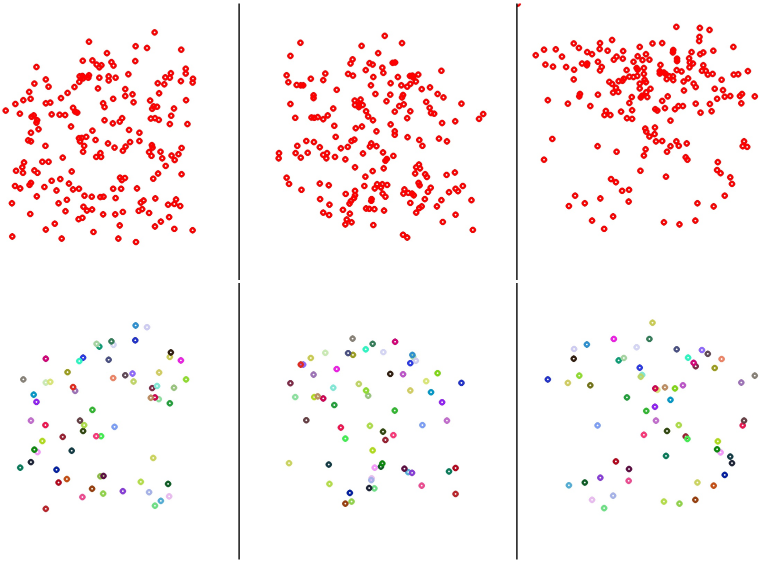

Fig. 2. Detections (up) and valid triplets (down) in three time-synchronized frames with 200 objects. Up: one red circle denotes a detection (object). Down: corresponding triplets are given by the same color across the views. Comparing detections and valid triplets indicates that invalid triplets are discarded.

3.2 Causal Tracklet Association

At time t, the terminated traklets are denoted by $S^{-} = \{s^{l}_{t_s:t_e}\}$ and the unterminated tracklets, namely ongoing tracklets, are denoted by $S^{+} = \{s^{k}_{t_s:t}\}$. In a first step, unambiguous detections are associated to the ongoing tracklets S^{+}. The tracklets satisfying at least one of the termination conditions are terminated and assigned to the terminated tracklets S^{-}. Afterwards, terminated tracklets S^{-} are linked with ongoing tracklets S^{+} *online* using the current frame instead of the whole seauence as proposed in [18,19]. As mentioned above, only motion and temporal cues can be used for linkage because of the similar appearance of fruit flies. $s^l_{t_s:t_e}$ and $s^k_{t_s:t}$ is linkable if the number of missing frames is within the range:

$$0 < t - t_e < \tau \tag{2}$$

where τ is the maximum gap between the linkable tracklet pairs and is set according to the frame rate and the occlusion time.

In order to calculate the association cost between a ongoing tracklet s^k and a terminated tracklet s^l, the tail triplet locations of the terminated tracklet $(d^{1,l}_{n_i,t_e}, d^{2,l}_{n_2,t_e}, d^{3,l}_{n_3,t_e})$ are extended to $p^{i,l}_{fwd}$, defined by:

$$p^{i,l}_{fwd} = (u^{i,l}_{n_i,t_e} + \nu^{i,l}_{u,n_i,t_e} \times \mathit{\Delta} t, \quad v^{i,l}_{n_i,t_e} + \nu^{i,l}_{v,n_i,t_e} \times \mathit{\Delta} t) \tag{3}$$

$$\mathit{\Delta} t = t^k_s - t^l_e \tag{4}$$

where $\mathit{\Delta} t$ is the time difference between s^l and s^k. The head triplet locations of the ongoing tracklet $(d^{1,k}_{n_i,t_s}, d^{2,k}_{n_2,t_s}, d^{3,k}_{n_3,t_s})$ are extended to $p^{i,k}_{bwd}$, defined by:

$$p^{i,k}_{bwd} = (u^{i,k}_{n_i,t_s} - \nu^{i,k}_{u,n_i,t_s} \times \mathit{\Delta} t, \quad v^{i,k}_{n_i,t_s} - \nu^{i,k}_{v,n_i,t_s} \times \mathit{\Delta} t) \tag{5}$$

The linear motion extended head and tail locations are compared to the real head and tail locations to produce the motion based linking cost L_m:

$$L_m = \frac{1}{3}\sum_{i=1}^{3}(w_1 \mathrm{dist}(d^{i,k}_{n^i,t_s}, p^{i,l}_{fwd}) + w_2 \mathrm{dist}(d^{i,l}_{n^i,t_e}, p^{i,k}_{bwd})) \tag{6}$$

$$w_1 + w_2 = 1 \tag{7}$$

where w_1, w_2 are the weight value for considering the forward and backward triplet location differences respectively.

This motion based linking costs L_m are used to identify candidates within the terminated tracklets S^- to be linked to ongoing tracklets in S^+. If there is only one close terminated tracklet $s^{l_1} \in S^-$ satisfying

$$L^*_m = \min_{s^* \in S^+} \frac{1}{3}\sum_{i=1}^{3}(w_1 \mathrm{dist}(d^{i,*}_{n^i,t_s}, p^{i,l_1}_{fwd}) + w_2 \mathrm{dist}(d^{i,l_1}_{n^i,t_e}, p^{i,*}_{bwd})) < \tau_1 \tag{8}$$

s^{l_1} is matched to the ongoing tracklet s^* and s^* is removed from S^+ for this time step (Figure 3 left). τ_1 is selected based on the frame rate and the maximal flight speed.

If two terminated tracklets $s^{l_1}, s^{l_2} \in S^-$ are temporally and spatially close, equation 9 and 10 are satisfied:

$$|t^{l_1}_e - t^{l_2}_e| < \tau_2 \tag{9}$$

$$\frac{1}{3}\sum_{i=1}^{3} \mathrm{dist}(d^{i,l_1}_{n^i,t_e}, d^{i,l_2}_{n^i,t_e}) < \tau_3 \tag{10}$$

Again, τ_2 and τ_3 are constant thresholds based on the frame rate and the maximal flight speed. The respective other tracklet, e.g. s^{l_2}, is considered as context when

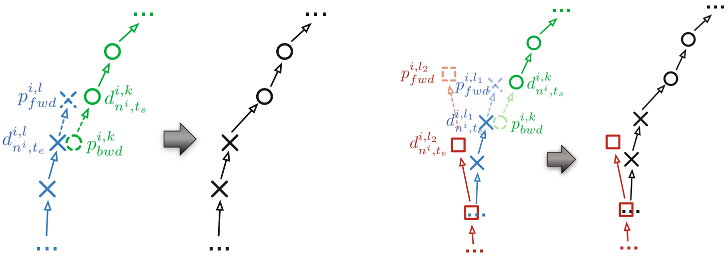

Fig. 3. Illustration of the motion cost based (left) and the context cost based (right) associations. The terminated tracklets are highlighted in blue and red and the ongoing tracklets are highlighted in green.

linking s^{l_1}. The context term for the linking costs L_c of s^{l_1} is defined similar to the motion based linking costs by:

$$L_c = \frac{1}{3}\sum_{i=1}^{3}(w_1 \mathrm{dist}(d^{i,k}_{n^i,t_s}, p^{i,l_2}_{fwd}) + w_2 \mathrm{dist}(d^{i,l_2}_{n^i,t_e}, p^{i,k}_{bwd})) \tag{11}$$

The ongoing tracklet s^* is matched to s^{l_1}, if equation (12) is satisfied:

$$L^*_c = \min_{s^* \in S^+} \frac{1}{3}\sum_{i=1}^{3}(w_1 \mathrm{dist}(d^{i,*}_{n^i,t_s}, p^{i,l_2}_{fwd}) + w_2 \mathrm{dist}(d^{i,l_2}_{n^i,t_e}, p^{i,*}_{bwd})) > L^*_m \tag{12}$$

given the motion based linking costs L^*_m (equation (8); see Figure 3 right).

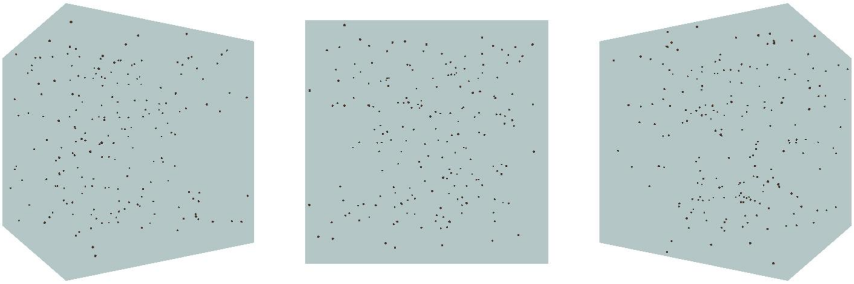

Fig. 4. Frame triplet generated by the simulator

4 Simulator and Evaluation Metrics

4.1 Ground Truth

Quantitative evaluation of fruit fly trajectories is very difficult in real-world high-density scenes. In order to measure the performance of the proposed method, a simulator used in [16] is adopted to generate test sequence. It generates both, synthetic images from three cameras including all camera matrices and 2D/3D ground truth. Detections are generated by adopting background subtraction to extract the blobs from the synthetic sequences. The center of mass of each segmented blob is used as $d^i_{n^i,t}$. As a result, occlusions and image noise is within the 2D detections D^i_t similar to real world conditions (Figure 4).

4.2 Evaluation Metrics

In [16] and [20] E_{ca} and RAE are used to measure the tracking algorithms performance. The number of inaccurate 3D locations and wrong associations are divided by the number of frames (E_{ca}) or by the number of all objects in all frames (RAE). Both offer a general measure for performance. We propose several metrics to evaluate the tracked trajectories in more details. *MT*, *Acc*, and *IDS* are proposed to measure the stereo accuracy:

- *MT* specifies the number of missed triplets, i.e. the ground truth triplets which are not matched to any valid detected triplets.
- *Acc* specifies the number of inaccurate 3D locations. If the Euclidean distance between the reconstructed 3D locations from detected triplets and 3D ground truth locations is between 5 and $10cm$, Acc is incremented.
- *IDS* specifies the number of identity switches. If the Euclidean distance between the reconstructed 3D locations from detected triplets and 3D ground truth locations is lager than $10cm$, it is counted as a wrong match.

In addition, *Occ* is used to measure detected occlusions:

- *Occ* specifies the number of detected occlusions. If one detection is matched to multiple triplets, *Occ* is incremented.

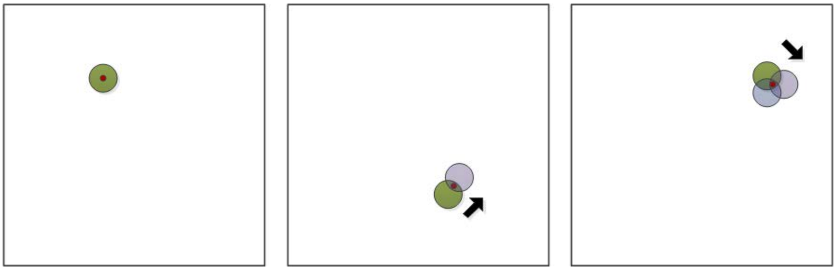

Fig. 5. Occlusions lead to shifted centers. The detected centers (small red circles) of an object (big green circles) in Camera 2 and Camera 3 are shifted due to occlusions. Arrows denote the shifting directions.

To evaluate the quality of the trajectories, *Frag* is proposed to measure the fragmentation of the trajectories, and *Complete Tracks*, *Partial Tracks*, and *Lost Tracks* are employed to measure the completeness of the tracked paths:

- *Frag* specifies the number of fragments. It counts the number of fragments for all trajectories in a sequence.
- *Complete Tracks* specifies the number of completed trajectories. If both, 95% of the trajectory is tracked and 95% 3D locations are accurate, this trajectory is counted as a complete track.
- *Lost Tracks* specifies the number of lost trajectories. If more then 50% of the trajectory is lost it is counted as a lost track.
- *Partial Tracks* specifies the number of partially tracked trajectories, which are neither *Complete Tracks*, nor *Lost Tracks*.

5 Experiments

Sequences with 10 to 200 objects are used to test the algorithm. All cameras cover the whole tracking chamber and are located with a distance of 80cm to the chambers center. Rotations around the y-axis are 0^o, 120^o and -120^o for camera 1, 2 and 3 respectively. The scene is captured with a resolution of 800×800 pixels and a frame rate of 150fps. The chamber is set to be $20 \times 20 \times 20cm^3$. The flies are represented with a radius of $2mm$. The maximum speed is set to $0.8m/s$ [10]. A screen shot from the resultant synthetic images is given in Figure 4. Based on the maximum speed and experiment experience, the linking parameters are set to be $\tau = 7, \tau_1 = 10, \tau_2 = 3, \tau_3 = 20$.

5.1 Detection-Based Ground Truth

Due to occlusions, nearby targets and noise, detections in the images do not match to the 2D ground truth. As a consequence, triangulated pairs of detections do not match to the 3D ground truth. Therefore, detection-based 3D ground truth is generated: For all detections for time t (i.e. D_t^1, D_t^2, D_t^3), triplets are determined by employing stereo matching and projection consistency [16]. These triplets are compared to the 2D ground truth, selecting only detections with an average Euclidean distance below a certain threshold. The detection-based ground truth was then calculated by triangulating the remaining detections.

The aberration between the detection-based ground truth and the 3D ground truth from the simulator is given in Table 1 (*MT*, *Acc*, *IDS* and *Occ* are given in % by dividing the measures by the number of frames). Obviously, the more objects need to be tracked, the more missed triplets, inaccurate locations, identity switches and detected occlusions can be measured, which decreases the overall tracking performance. In fact, *MT*, *Acc* and *IDS* increase with the number of occlusions, since the blob centers are shifted in case of overlapping silhouettes (Figure 5): Shifted blob centers lead to inaccurate stereo pairs which can not be corrected by projection consistency [16], some triplets are missed in the detection-based ground truth (compare to *MT*). If matching the shifted centers is still successful (Section 3), triangulation leads to wrong 3D locations (compare to *Acc*). Other shifted blob centers are, however, erroneously assigned to the real 3D ground truth locations because they are located in a certain range of tolerance, leading to identity switches (compare to *IDS*).

Table 1. Detection-based ground truth measure results

		General(1000 frames)					High-Density (150 frames)		
# objects		10	20	30	40	50	100	150	200
MT	Abs.	128	521	681	1106	1436	147	489	921
	%	0.13	0.52	0.68	1.11	1.44	0.98	3.26	6.14
Acc	Abs.	0	5	11	48	40	6	15	25
	%	0.0	0.01	0.01	0.05	0.04	0.04	0.1	0.17
IDS	Abs.	129	522	682	1107	1437	148	490	922
	%	0.13	0.52	0.68	1.11	1.44	0.99	3.27	6.15
Occ	Abs.	21	119	289	846	1248	889	1850	3423
	%	0.02	0.12	0.29	0.85	1.25	5.93	12.33	22.82

5.2 Comparison

The results of proposed algorithm are compared with detection-based ground truth (see Table 2 and 3). The numbers shown in brackets are the corresponding results from Table 1. Obviously, our proposed algorithm outperforms detection-based ground truth measurements by reducing *IDS* and *MT* while slightly increasing *Acc*. *IDS* is reduced significantly due to temporal tracking information. Similar to the detection-based ground truth *MT* and *Acc* increase with the number of objects. Trajectories are more fragmented in high density scenes (e.g. 53 fragmentations occur for 200 objects, whereas no fragmentations are measured for 10 objects). Almost all objects are tracked for all frames (compare to *Complete Tracks*). Only for very high object densities this measure is decreased (Table 3).

Furthermore, the proposed algorithm is compared to the algorithms proposed in [20] and [16], namely Global-Correspondence Selection (GCS) and Tracking-by-Matching (TbM). The results measured with E_{ca} are shown in Table 4. Based on E_{ca} both, the TbM method and the proposed method outperform the GCS approach. Most accurate trajectories are generated by the tracklet-based method.

Table 2. Comparison between the our algorithm and the detection-based ground truth using the proposed metrics for 10-50 objects

(# objects, # frames)	(10,1000)	(20,1000)	(30,1000)	(40,1000)	(50,1000)
MT	0(128)	16(521)	23(681)	54(1106)	67(1436)
Acc	0(0)	12(5)	32(11)	111(48)	107(40)
IDS	0(129)	0(522)	0(682)	0(1107)	0(1437)
Frag	0	1	3	7	16
Complete Tracks	10	20	30	40	49
Partial Tracks	0	0	0	0	1
Lost Tracks	0	0	0	0	0

Table 3. Comparison between the our algorithm and the detection-based ground truth using the proposed metrics for 100-200 objects

(# objects, # frames)	(100,150)	(150,150)	(200,150)
MT	29 (147)	473 (489)	745 (921)
Acc	21 (6)	76 (15)	132 (25)
IDS	0 (148)	0 (490)	36 (922)
Frag	4	26	53
Complete Tracks	99	135	172
Partial Tracks	1	13	26
Lost Tracks	0	2	2

Table 4. Comparison between the our algorithm and the TbM and GCS approach using the E_{ca} quality measure

# objects	10	20	30	40	50	100	150	200
TbM	0.009	0.052	0.33	0.326	0.544	1.16	3.353	6.727
GCS	0.085	1.657	0.791	0.317	2.517	n/a	n/a	n/a
Proposed	0.00	0.034	0.064	0.163	0.206	0.38	1.98	4.63

A further comparison between the TbM and proposed method is given for very high-density scenes (Table 5). The results of proposed algorithm is shown in brackets. The proposed method reduces fragmentation with the proposed linking strategy. More complete trajectories are obtained as the termination and association process terminates ambiguous tracklet tracking and retrieves the corresponding tracklets later on. With the termination and association, less identity switches occur. As the corrected tracklets from *IDS* might contain inaccurate 3D locations as a result of shifted centers, *Acc* is slightly increased.

Table 5. Comparison between our algorithm and the TbM approach in very high-density situations using the proposed metrics. The numbers shown in brackets are the corresponding results from Table 3.

(# objects, # frames)	(100,150)	(150,150)	(200,150)
MT	394 (29)	1093 (473)	1880 (745)
Acc	17 (21)	68 (76)	108 (132)
IDS	73 (0)	248 (0)	382 (36)
Frag	14 (4)	42 (26)	75 (53)
Complete Tracks	87 (99)	122 (135)	138 (172)
Partial Tracks	13 (1)	24 (13)	58 (26)
Lost Tracks	0 (0)	4 (2)	4 (2)

6 Conclusions

In this paper, we proposed a tracklet-based probabilistic 3D tracking algorithm for high-density situations. This algorithm is compared to two state-of-the-art algorithms, by utilizing a set of new evaluation metrics to analyze the tracking results in details. It has been shown that tracklet-based probabilistic tracking outperforms both, a Global Correspondence Selection algorithm and a conventional probabilistic tracking algorithm.

Furthermore, the proposed metrics offer information to perceive more about the tracking results by evaluating detection-based ground truth. These metrics in combination with synthetic data offer help to set up real-world experimental settings because of the possibility to quantify trade-offs between the number of fruit flies to be observed, the frame rate used for recording, the image resolution or other crucial parameters.

Thus, our future work will mainly focus on two mutually influencing challenges: On the one hand, we will design and improve our real-world setups by quantifying above mentioned trade-offs using simulated data. On the other hand, we will adjust our tracking algorithm to facilitate high-throughput behavioral experiments for freely flying fruit flies. For example, a more precise velocity model could be used for an overall better prediction performance.

References

1. Ardekani, R., Biyani, A., Dalton, J.E., Saltz, J.B., Arbeitman, M.N., Tower, J., Nuzhdin, S., Tavaré, S.: Three-dimensional Tracking and Behaviour Monitoring of Multiple Fruit Flies. J. Royal Society Interface 10, 1–13 (2013)
2. Burkard, R.E., Dell'Amico, M., Martello, S.: Assignment Problems, 1st edn. Society of Industrial Mathematics (2009)
3. Colomb, J., Reiter, L., Blaszkiewicz, J., Wessnitzer, J.: Open Source Tracking and Analysis of Adult Drosophila Locomotion in Buridan's Paradigm with and without Visual Targets. PloS One 7(8), e41642 (2012)
4. Du, H., Zou, D., Chen, Y.Q.: Relative Epipolar Motion of Tracked Features for Correspondence in Binocular Stereo. In: 11th Internitional Conference Computer Vision, pp. 1–8. IEEE Press, Rio de Janeiro (2007)
5. Fry, S.N., Bichsel, M., Müller, P., Robert, D.: Tracking of Flying Insects Using Pan-tilt Cameras. J. Neurosci Methods 101(1), 59–67 (2000)
6. Grover, D., Tower, J., Tavaré, S.: O fly, Where Art Thou? J. Royal Society 5, 1181–1191 (2008)
7. Kohlhoff, K.J., Jahn, T.R., Lomas, D.A., Dobson, C.M., Crowther, D.C., Vendruscolo, C.M.: The iFly tracking system for an automated locomotor and behavioural analysis of Drosophila melanogaster. Integrative Biology (2011)
8. Kuo, C., Nevatia, R.: How Does Person Identity Recognition Help Multi-person Tracking? In: Computer Vision and Pattern Recognition, pp. 1217–1224. IEEE Press, Colorado Springs (2011)
9. Liu, Y., Li, H., Chen, Y.Q.: Automatic Tracking of a Large Number of Moving Targets in 3D. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 730–742. Springer, Heidelberg (2012)

10. Marden, J.H., Wolf, K.E.: Areal performance of Drosophila melanogaster from populations selected for upwind flight ability. Journal of Experimental Biology 200, 2747–2755 (1997)
11. Pereira, F., Stuer, H., Graff, E.C., Gharib, M.: Two-frame 3d Particle Tracking. J. Measurement Science Technology 17, 1–8 (2006)
12. Risse, B., Thomas, S., Otto, N., Löpmeier, T., Valkov, D., Jiang, X., Klämbt, C.: FIM, a Novel FTIR-Based Imaging Method for High Throughput Locomotion Analysis. PloS One 8(1), e53963 (2013)
13. Sokolowski, M.A.: Drosophila: Genetics Meets Behavior. J. Nature Reviews Genetics 2(11), 879–890 (2001)
14. Straw, A.D., Branson, K., Neumann, T.R., Dickinson, M.H.: Multi-camera Realtime Three-dimensional Tracking of Multiple Flying Aanimals. J. Royal Society 8, 395–409 (2011)
15. Tao, J., Klette, R.: Tracking of 2D or 3D Irregular Movement by a Family of Unscented Kalman Filters. J. Information Communication Convergence Engineering 1, 307–314 (2012)
16. Tao, J., Risse, B., Jiang, X., Klette, R.: 3D Trajectory Estimation of Simulated Fruit Flies. In: 27th Image Vision Computing New Zealand, pp. 31–36. ACM, Dunedin (2012)
17. Valente, D., Golani, I., Mitra, P.P.: Analysis of the Trajectory of Drosophila Melanogaster in a Circular Open Field Arena. PloS One 2(10), e1083 (2007)
18. Wu, H.S., Zhao, Q., Zou, D., Chen, Y.Q.: Acquiring 3d Motion Trajectories of Large Numbers of Swarming Animals. In: 12th Internitional Conference Computer Vision Workshop, pp. 593–600. IEEE Press, Kyoto (2009)
19. Yang, B., Nevatia, R.: An Online Learned CRF Model for Multi-target Tracking. In: Computer Vision and Pattern Recognition, pp. 2034–2041. IEEE Press, Oregon (2012)
20. Zou, D., Zhao, Q., Wu, H.S., Chen, Y.Q.: Reconstructing 3d Motion Trajectories of Particle Swarms by Global Correspondence Selection. In: 12th Internitional Conference Computer Vision, pp. 1578–1585. IEEE Press, Kyoto (2009)