TRACKING THE 3D POSITION AND ORIENTATION OF FLYING SWARMS WITH LEARNED KINEMATIC PATTERN USING LSTM NETWORK

Anonymous ICME submission

ABSTRACT

Accurately and reliably tracking the 3D position and orientation of individuals in large flying swarms is valuable not only for scientific researches but also practical applications. However, large quantity, frequent occlusions, similar appearance, tiny body size and abrupt motion make it remain an open problem. The 3D flying swarm tracking method proposed in this paper tracks both position and orientation of each individual in the swarm using the particle filter framework. Particles are scattered more pertinently by the dynamic model based on the learned kinematic pattern of a single target with a Long Short-Term Memory (LSTM) network. In addition, the observation model combines the Weighted Occupancy Ratio (WOR) and Temporal Appearance Coherency (TAC) cues in each view to improve the accuracy and robustness of the reconstructed body orientation. Experiments on both simulation and real-world data sets demonstrate the effectiveness and superiority of the proposed method.

***Index Terms*—** 3D multi-object tracking, flying swarms, LSTM network, kinematic pattern modeling, body orientation

1. INTRODUCTION

The underlying mechanism of the collective behavior patterns exhibited by large flying swarms such as insect swarms, bird flocks, *etc.* have attracted the attention of researchers from many research areas [1, 2]. Video tracking is the most effective way to obtain motion data of flying swarms [3]. Both the position and orientation sequences of the targets provide important information for behavior studies [4]. 3D coordinate and orientation of each target at each frame can be reconstructed by associating observations in two or more geometrically calibrated and temporally synchronized cameras.

The challenges of multi-view tracking 3D position and orientation of flying swarms mainly result from large swarm quantity, frequent occlusions, similar appearance, tiny body size and abrupt motion. Different strategies have been proposed to cope with these difficulties. The most intuitive one is to first detect objects in each view at each frame, then establish cross-view, cross-frame association and reconstruct the 3D trajectories based on the association results. Wu *et al.* [5] first tracked targets in 2D in each view, then matched these 2D tracklets by linear assignment. Trajectories will break up if in one frame the target cannot find its associated detection and even a relinking step cannot solve fragmentation to a satisfactory extent. Wu *et al.* [6] relaxed the one-to-one matching constraint, so that the fragmentation of trajectories can be reduced. But all of these methods highly depend on detection performance. Another strategy is to first find cross-view association by feature matching to reconstruct 3D observations, then establish cross-frame association based on the 3D observations. Ardekani *et al.* [7] applied this strategy to track several fruit flies. The key point of this strategy is to distinguish each detected object in 2D and reconstruct 3D observations correctly. Therefore, it's difficult to track a large number of objects with small body size and similar appearance.

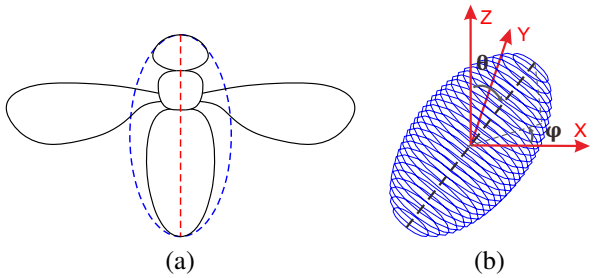


Fig. 1. Body model of a flying target.

To make up for the deficiencies of the two strategies described above, researchers proposed to track targets by projection and pruning. Hypothetic targets in 3D are generated based on newly detected objects. Then the posterior distribution of each target's state at each moment is estimated by incorporating observation from each camera view. False alarms can be eliminated after several frames when the observation model is no longer satisfied. Liu *et al.* [8] applied this strategy to obtain 3D trajectories of flying swarms using particle filtering technique. Cheng *et al.* [9] used the Long Short-Term Memory (LSTM) network to learn the probability of an input velocity sequence of an object. The probability is combined with the observation model to weight each particle under the particle filter framework. However, the dynamic model of their tracking system is a simple linear extrapolation model, so large number of particles are needed to approach the true state if the object undergoes an abrupt motion, which lowers the system's operating efficiency or even leads to loss of targets. In [10], Cheng *et al.* proposed a 3D-POT algorith-

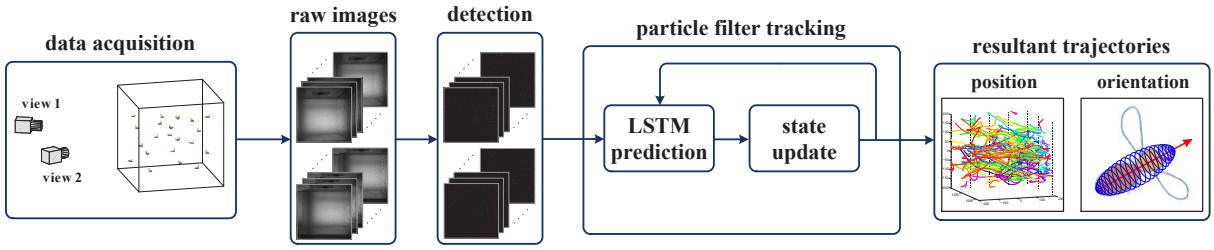


Fig. 2. Workflow of the proposed 3D tracking method.

m to track both position and orientation of each target. The position of the targets at each frame is tracked by the Rao-Blackwellized particle filter and the posterior probability of orientation is estimated by the Kalman filter. The accuracy of the reconstructed orientation depends on the performance of ellipse fitting, which will be unreliable when the silhouette in the 2D image is similar to a circle, or when the wings of the object are not removed neatly. Furthermore, the estimated orientation will delay when the target takes a sharp turn.

In summary, the limitations of existing methods lie in two aspects: (1) Most existing methods model the flying target as a single point. As a result, the orientation of the target, which also provides important motion data, is missing; (2) The performance of most methods relies on the detection accuracy. When tracking insect swarms such as fruit flies and midges, where each target occupies only tens of pixels in the captured images, these methods will face difficulties obtaining reliable trajectories. This paper proposes a 3D tracking method based on the particle filter framework which can reliably obtain the position as well as orientation of each target even when the target is very small in size or performs fast maneuvers. The contributions of the proposed method are:

- The kinematic pattern of a single object is learned by an LSTM network and used as the dynamic model. Thus the particles are scattered more pertinently, closer to the location of the true state. In this way, a lower number of particles is needed to obtain a satisfying tracking performance.
- The target's orientation is included in the state vector and estimated using particle filter framework. In this way the accuracy of reconstructed orientation is improved and not affected by the ellipse fitting errors in detection stage.
- Proposing to improve tracking accuracy by combining the Weighted Occupancy Ratio (WOR) and the Temporal Appearance Coherency (TAC) to build the observation model.

2. KINEMATIC PATTERN LEARNING

The dynamic model in a tracking system determines the evolution of the motion state of each target. In particle filter framework, dynamic model is used to transfer the particles to the predicted state [11]. Conventional tracking methods apply models such as random walk [12] or first-order linear extrapolation [9] as the dynamic model, which do not confirm to the motion pattern of maneuvering targets such as insect swarms. Considering that the trajectories of targets in real 3D space is physically meaningful, the prior states could provide more information for the dynamic model to predict the new state more accurately than 2D trajectories on single view images, the state of the target in current frame can be predicted using the motion state of it in several prior consecutive frames.

Long Short-Term Memory (LSTM) network [13] is a special kind of recurrent neural network (RNN) which replaces the conventional neurons in hidden layers with memory blocks. Applications of LSTM networks in many temporal processing tasks demonstrate the superior power of it than standard RNN owing to its ability of learning long-term dependencies. And LSTM networks have been applied by several proposals to predict the state of targets [9, 14, 15]. We thus model the kinematic pattern of a flying target directly in 3D space with an LSTM network and the learned kinematic pattern is employed in the proposed tracking method as the dynamic model.

2.1. The body model

Based on the observation that a flying object rarely bends its body, the body part is bilaterally symmetrical and nearly rigid [10], and 3D tracking methods usually focus on obtaining the motion data of the body of each target, thus wing motion is not taken into consideration in the proposed method. The target body is thereby approximated as an ellipsoid determined by the mean axis (*cf.* Figure 1(a)). The state vector of each target consists of the coordinate of the ellipsoid center (x, y, z), the orientation (θ, φ) and the body length l:

$$\mathbf{X}^t = \{x, y, z, \theta, \varphi, l\} \tag{1}$$

where θ is the polar angle and φ is the azimuthal angel, as illustrated in Figure 1(b). In our experiments, we applied the proposed method to track insect swarms with similar body length. So we set l as a constant.

2.2. The LSTM network

The unfolded LSTM network in the proposed method is illustrated in Figure 3. The network input $L^{1:\tau}$ is the sequence of

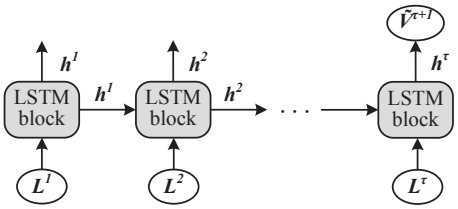


Fig. 3. Illustration of the unfolded LSTM network in the proposed method.

velocity and change rate of the orientation from time $t-\tau$ to time $t-1$ of the target, as defined in Equ. (2). Velocity sequence is used instead of positional sequence, as spatio-temporal data has higher accuracy [15]. The network output $\mathbf{h}^\tau$ is the hypothetical velocity at time t and the variation of orientation between time $t-1$ and t. More details of the LSTM network are described in the supplementary material.

$$\mathbf{L}^{t-\tau+\kappa-1} = \left[\mathbf{v}_x^\kappa, \mathbf{v}_y^\kappa, \mathbf{v}_z^\kappa, \delta_\theta^\kappa, \delta_\varphi^\kappa, \right]^T, \ (\kappa = 1, ..., \tau) \quad (2)$$

2.3. LSTM Training

To train the LSTM network, large numbers of ground truth velocity sequences are needed, we thus captured 400 video clips with only one flying target in the arena, so that mismatches in cross-view, cross-frame association won't occur. The resultant 3D trajectories including orientations are then projected onto each camera view to check and correct tracking errors manually. In total, 40000 velocity sequences are randomly selected from the raw trajectories and used as the training set. The LSTM network is trained with Backpropagation Through Time (BPTT) under a matrix-based batch learning paradigm [16].

3. THE TRACKING METHOD

The proposed tracking method performs detection and tracking stages sequentially until the whole video has been processed, as the workflow shown in Figure 2. In the detection stage, the image blob of each target is detected and extracted from the background-subtracted image. The tracking stage is based on the particle filter framework. The kinematic pattern of a single target learned with an LSTM network is implemented as the dynamic model. The state of each target, which includes position and orientation, is estimated by weighted particles. In this way, the target's orientation can be accurately reconstructed even under abrupt motion and the accuracy is not affected by detection errors.

3.1. Detection

Although the captured images of each view are cluttered with swarms of targets, the background of the images remains stable over a short period of time. The background of each view ν at time t is thus calculated by subtracting the mean image of raw images from time $t - w/2$ to $t + w/2$, written as:

$$I'^t_\nu = |I^t_\nu - mean(I^{t-w/2}_\nu, ..., I^{t+w/2}_\nu)| \quad (3)$$

w is chosen according to the average velocity of the objects and frame rate, and $w = 9$ in our experiment. The blob of each target is then segmented based on image I'. The barycenter of each blob is $\mathbf{p}^t_{\nu,i} = (x^t_{\nu,i}, y^t_{\nu,i})$.

3.2. Tracking via particle filter framework

As the dynamic system of flying swarms is highly non-linear and the posterior density is usually non-Gaussian, the particle filtering framework is applied to approximate the posterior of the target's state with a set of N weighted particles, denoted as $\{(\mathbf{X}^t_i, w^t_i)\}_{i=1,...,N}$. Each particle drawn from the previous state using importance sampling and shifted by the dynamic model is weighted by the likelihood of the observation at time t given the particle state, denoted as $w^t_i \propto p(\mathbf{Z}^t \mid \tilde{\mathbf{X}}^t_i)$, which is called the observation model. The expectation of the target's state $\hat{X}^t$ is then calculated as:

$$\hat{\mathbf{X}}^t = E(\mathbf{X}^t \mid \mathbf{Z}^{1:t}) = \sum_{i=1}^{N} w^t_i \mathbf{X}^t_i \quad (4)$$

3.2.1. Dynamic model

The dynamic model in a tracking method predicts the target's state at each frame. In the proposed method, an LSTM network is applied to learn the kinematic pattern of the flying object and the learned kinematic pattern is used as the dynamic model. The output of the LSTM network is the hypothetical velocity and the change rate of orientation. Thus the predicted state of each target is:

$$\tilde{\mathbf{X}}^t = \mathbf{X}^{t-1} + \mathbf{h}^\tau \quad (5)$$

As state prediction by the LSTM network is more accurate than conventional dynamic models, the number of particles needed to approach the true state of the object is reduced. Thus, the computational efficiency of the system is improved.

3.2.2. Observation model

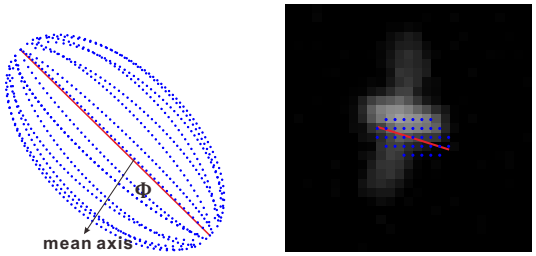
The observation model $p(\mathbf{Z}^t \mid \mathbf{X}^t)$ measures the likelihood of each particle. As the observations are the detection results in 2D, the 3D state is first projected onto each 2D image plane and the observation model combines the likelihood calculated in each view. For each camera view ν, two complementary cues namely the Weighted Occupancy Ratio (WOR) and the Temporal Appearance Coherency (TAC) are utilized.

- Weighted Occupancy Ratio (WOR)

As introduced in Section 2.1, the 3D shape of the target is modeled as an ellipsoid based on the fact that the target's body is similar to an ellipsoid in 3D space. We then project the ellipsoid onto each camera view ν and calculate the WOR as:

$$p_{wor,\nu}(\mathbf{Z}^t \mid \mathbf{X}_i^t) = {I'}_{\nu}^{t} \odot [\mathcal{P}_\nu \mathscr{E}] \tag{6}$$

where $\mathcal{P}_\nu$ is the projection matrix of view ν, $\mathscr{E}$ is the ellipsoid reconstructed by each particle, and $\odot$ is the pixel-wise multiplication. Figure 4 is an illustration of Equ. (6). The 2D pixels are weighted by ${I'}_{\nu}^{t}$, so that the pixels with lighter values, which are closer to the center of the body, are given larger weights. WOR calculated in this way is more accurate than the occupancy ratio with the body approximated as a sphere without the weighting strategy [10].



(a) (b)

Fig. 4. (a). The corresponding ellipsoid $\mathscr{E}$ of a particle in 3D. Φ is the mean axis of $\mathscr{E}$; (b). Illustration of $p_{wor}(\mathbf{Z}^t \mid \mathbf{X}_i^t)$ in Equ. (6).

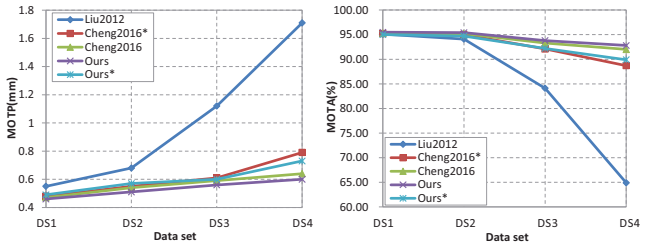
- Temporal Appearance Coherency (TAC)

  Different from other flying objects, insect swarms are very small in size, each target only takes up tens of pixels, resulting in loss of visual details in the captured images. Features such as color, texture and key points that commonly used in state-of-the-art tracking systems are not applicable for insect swarm tracking, because they are not robust while there is very little visual information on such few pixels. Normalized Cross Correlation (NCC) can reflect the subtle variation between two images at pixel level [17]. Thus we found it effective to measure the image similarity of 2D blobs in consecutive frames by NCC. Assume that the 2D blob in view ν corresponding to the target at time $t-1$ is $\mathcal{B}_\nu^{t-1}$, and the blob in view ν corresponding to particle i at time t is $\mathcal{B}_{\nu,i}^{t}$, then the TAC of particle i is calculated as:

$$p_{ap,\nu}(\mathbf{Z}^t \mid \mathbf{X}_i^t) = \mathcal{NCC}(\mathcal{B}_\nu^{t-1}, \mathcal{B}_{\nu,i}^{t}) \tag{7}$$

  The two blobs are rectified to the up-right position based on the orientation of the projected mean axis Φ, and are resized to the same size prior to NCC calculation.

Finally, the observation model is formulated by combining the WOR and TAC in each camera view ν:



(a) (b)

Fig. 5. Evaluation results using CLEAR MOT metrics on simulation data sets.

$$p(\mathbf{Z}^t \mid \mathbf{X}_i^t) = \prod_{\nu=1}^{n_\nu} p_{wor,\nu}(\mathbf{Z}^t \mid \mathbf{X}_i^t) p_{ap,\nu}(\mathbf{Z}^t \mid \mathbf{X}_i^t) \tag{8}$$

4. EXPERIMENTS

Images of flying swarms captured by multiple cameras were applied to test the performance of the proposed method and compare it to three state-of-the-art methods: Liu *et al.* [8], Cheng *et al.* [9] and Cheng *et al.* [10] (as introduced in Section 1 and denoted as Liu2012, Cheng2016 and Cheng2016* respectively) and our method with reduced number of particles (originally, there are 500 particles in our method Ours and 300 particles in the particle-reduced method Ours*). However, the groundtruth trajectories for real-world data are difficult to obtain due to the large number of tiny targets and ambiguities of cross-view association. Hence it is more feasible to use simulation data to quantitatively evaluate the performance of the methods. We therefore applied both simulation data and real world data for performance evaluation.

4.1. Evaluation on simulation data

Boids model [18] was adopted to generate the 3D trajectories of a flying swarm. The polar angle of each target was set as: $\theta = \pi/4 + \delta, \delta \sim \mathcal{N}(0, \Sigma)$, in which δ is a small Gaussian white noise. The shapes of the targets were created using a generative model [10]. Four simulation data sets DS1∼DS4 were created with 20, 80, 150, and 250 targets respectively, based on the two-camera stereo system simulated using "Machine Vision Toolbox" in MATLAB [19]. The widely used CLEAR MOT metrics [20] and one self-defined metric named Orientation Accuracy (OA) were applied to evaluate these tracking methods on simulation data. More detailed descriptions of the metrics are illustrated in the supplementary material. The evaluation results are shown in Figure 5 and 6.

4.2. Evaluation on real-world data

We captured a swarm of fruit flies (*Drosophila melanogaster*) which were housed in a cubic transparent acrylic box of size

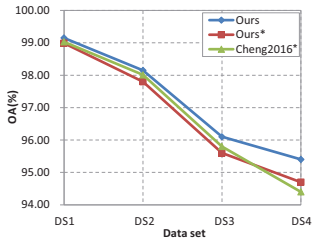
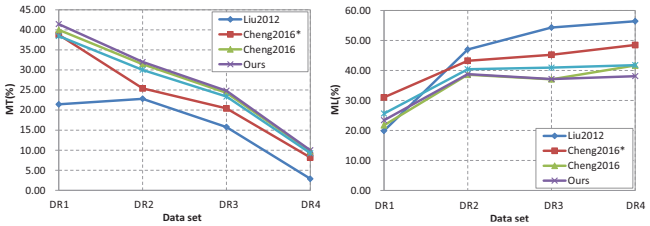


Fig. 6. Evaluation of orientation accuracy on simulation data sets.



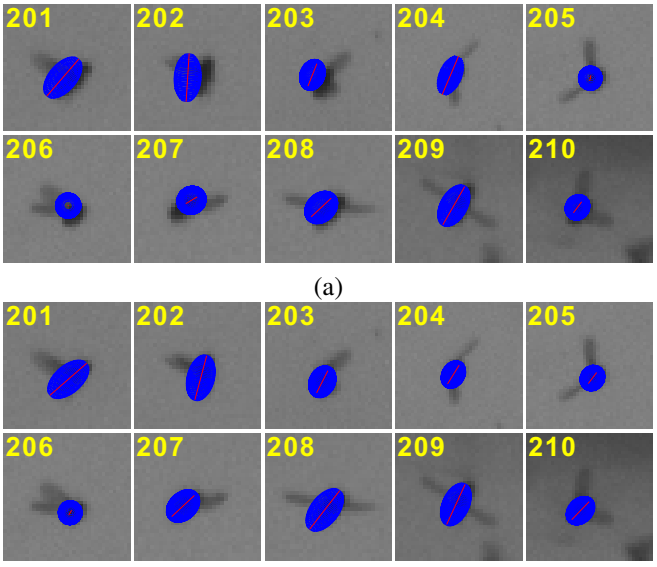
(a) (b)

Fig. 8. Evaluation results on real-world data sets.

$40*40*40\,cm$ using two geometrically calibrated and temporally synchronized high speed monochrome CMOS cameras (IO Industries Canada, Flare 4M 180-CL 2048*2040 pixels at 100 fps). The arena was back-lit by two planar lights. Four video clips (named DR1∼DR4) with 500 frames in length were captured and the group quantities were 50, 150, 300, and 600 flies, respectively. The resultant 3D trajectories are shown in Figure 7. Due to the impracticality of obtaining the groundtruth data from real-world videos, the two metrics that applicable here are Mostly tracked (MT) and Mostly lost (ML) [21], which measure the percentage of trajectories that are tracked more than 80% and less than 20% of the acquisition time respectively. The two metrics reflect the ability of the tracking method to obtain complete trajectories. The results are shown in Figure 8.

4.3. Discussion

According to the results in Figure 5 and 8, Liu2012 obtained satisfactory performance on data set DS1, DS2, DR1 and DR2, but the performance dropped dramatically on DS3, DS4, DR3 and DR4, where the group density is high, because their method may accept a particle even if it is observed in only one view. Some of these false targets cannot be successfully removed in the following frames, which also leads to trajectory fragmentation. Cheng2016* performed slightly inferior to the results presented in their paper, especially the orientation accuracy (*cf.* Figure 6), because their method requires a constraint of eccentricity of the fitted ellipse, so that usually more than two cameras are required to guarantee orientation accuracy. As the estimation of the orientation is



(a)

(b)

Fig. 9. Orientation accuracy comparison (projected onto one camera view) on an abrupt turn case. (a). Result of Cheng2016* [10]; (b). Result of the proposed method.

based on Kalman filter, the accuracy of the estimated orientation is largely dependent on ellipse fitting in the detection phase. When the target takes an abrupt turn, the estimated orientation will delay, as is the case in Figure 9. Cheng2016 applied an LSTM network to learn the probability of an input velocity sequence and the probability is combined with the observation model to weight the particles. The performance is improved compared to Cheng2016*, but slightly inferior to the proposed method when using the same number of particles. Moreover, Cheng2016* can only track positions of each target. In the proposed method, we do not perform segmentation and ellipse-fitting in the detection phase (except the first several frames for initialization), so that the accuracy of the reconstructed orientation will not be affected by the detection accuracy. We also reduced the number of particles, which resulted in nearly the same metric scores as Cheng2016. It can thus be concluded that the dynamic model in the proposed method, which applied an LSTM network for state prediction, can scatter particles more pertinently (*i.e.*, closer to the true state). Accordingly, the number of particles needed is reduced and in this way the efficiency of the system is improved.

5. CONCLUSION

We have proposed in this paper a 3D tracking method capable of tracking both position and orientation of each target in a flying swarm using the particle filter framework. The kinematic pattern of a single target is learned with an LSTM network and functions as the dynamic model in the proposed method. Thus the particles are scattered more pertinently.

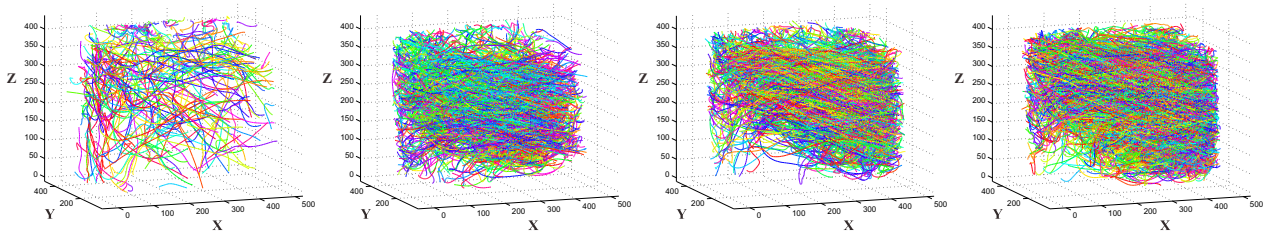


Fig. 7. Resultant 3D trajectories of: (a). DR1; (b). DR2; (c). DR3; (d). DR4.

The observation model which combines the Weighted Occupancy Ratio (WOR) and Temporal Appearance Coherency (TAC) in each camera view contributes to the high accuracy of orientation reconstruction. Experiments were conducted on both simulation and real-world data sets. The evaluation results demonstrate the superior performance of the proposed method compared to three other state-of-the-art methods.

6. REFERENCES

[1] H. Dankert, L. Wang, E. D. Hoopfer, et al., "Automated monitoring and analysis of social behavior in drosophila," *Nat. Methods*, vol. 6, no. 4, pp. 297–303, 2009.

[2] M. Nagy, Z. Ákos, D. Biro, et al., "Hierarchical group dynamics in pigeon flocks," *Nature*, vol. 464, no. 7290, pp. 890–893, 2010.

[3] A. I. Dell, J. A. Bender, K. Branson, et al., "Automated image-based tracking and its application in ecology," *Trends Ecol. Evol.*, vol. 29, no. 7, pp. 417–428, 2014.

[4] P. T. Weir and M. H. Dickinson, "Flying drosophila orient to sky polarization," *Curr. Biol.*, vol. 22, no. 1, pp. 21–27, 2012.

[5] H. S. Wu, Q. Zhao, D. Zou, et al., "Automated 3d trajectory measuring of large numbers of moving particles," *Opt. Express*, vol. 19, no. 8, pp. 7646–7663, 2011.

[6] Z. Wu, N. I. Hristov, T. L. Hedrick, et al., "Tracking a large number of objects from multiple views," in *ICCV*. IEEE, 2009, pp. 1546–1553.

[7] R. Ardekani, A. Biyani, J. E. Dalton, et al., "Three-dimensional tracking and behaviour monitoring of multiple fruit flies," *J. R. Soc. Interface*, vol. 10, no. 78, 2012.

[8] Y. Liu, H. Li, and Y. Q. Chen, "Automatic tracking of a large number of moving targets in 3d," in *ECCV*. Springer, 2012, pp. 730–742.

[9] X. E. Cheng, S. H. Wang, and Y. Q. Chen, "3d tracking targets via kinematic model weighted particle filter," in *ICME*. IEEE, 2016, pp. 1–6.

[10] X. E. Cheng, S. H. Wang, and Y. Q. Chen, "Estimating orientation in tracking individuals of flying swarms," in *ICASSP*. IEEE, 2016, pp. 1496–1500.

[11] M. S. Arulampalam, S. Maskell, N. Gordon, et al., "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[12] D. A. Ross, J. Lim, R. S. Lin, et al., "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] A. Alahi, K. Goel, V. Ramanathan, et al., "Social lstm: Human trajectory prediction in crowded spaces," in *CVPR*, June 2016.

[15] S. H. Wang, X. E. Cheng, and Y. Q. Chen, "Tracking undulatory body motion of multiple fish based on midline dynamics modeling," in *ICME*. IEEE, 2016, pp. 1–6.

[16] Q. Lyu and J. Zhu, "Revisit long short-term memory: An optimization perspective," in *Advances in neural information processing systems workshop on deep Learning and representation Learning*, 2014.

[17] J. P. Lewis, "Fast template matching," in *Vision interface*, 1995, vol. 95, pp. 15–19.

[18] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *Computer Graphics*, vol. 21, no. 4, pp. 25–34, 1987.

[19] P. Corke, *Robotics, vision and control: fundamental algorithms in MATLAB*, vol. 73, Springer, 2011.

[20] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP J. Image Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[21] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *CVPR*. IEEE, 2009, pp. 2953–2960.