3D TRACKING SWIMMING FISH SCHOOL WITH LEARNED KINEMATIC MODEL USING LSTM NETWORK

Shuo Hong Wang[1], *Jingwen Zhao*[1], *Xiang Liu*[12], *Zhi-Ming Qian*[1], *Ye Liu*[3], *Yan Qiu Chen*[1*]

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China [2]Shanghai University of Engineering Science, China
[3]College of Automation, Nanjing University of Posts and Telecommunications, China

ABSTRACT

This paper proposes a reliable 3D fish tracking method using a novel master-slave camera setup. Instead of conventional dynamic models that rely on prior knowledge about target kinematics, the proposed method learns the kinematic model with a Long Short-Term Memory (LSTM) network. On this basis, the 3D state of fish at each moment is predicted by LSTM network. We propose to use an innovative master-view-tracking-first strategy. The fish are first tracked in the master view. Cross-view association is then established utilizing motion continuity and epipolar constraint cues. Experiments on data sets of different fish densities show that the proposed method is effective and outperforms the state-of-the-art methods.

***Index Terms*—** 3D tracking, fish school, master-view-tracking-first strategy, kinematic model, LSTM network

1. INTRODUCTION

Visual tracking is an effective, convenient, and economic way to acquire motion data of individuals in fish schools to study their behavior, which has attracted many researchers investigating the behavior of fish school, whose value is not limited to biological research but may also be helpful in areas *e.g.*, multi-agent robot design [1] and computer graphics [2].

Multi-object tracking using multiple synchronized and calibrated cameras is the most effective way to obtain accurate and complete motion data of fish school. Most existing fish tracking methods are limited in 2D space because many fish behavior experiments use shallow water, in which fish swim in almost the same plane. Software such as ANY-maze® and EthoVision® has been widely used by biologists [3, 4]. Qian *et al.* [5] applied Determinant of Hessian (DoH) to detect fish head and Kalman filter to track them. It can track dozens of fish under occlusions in 2D. Alfonso *et al.* [6] proposed an identity preserved 2D tracking method based on an intensity distribution feature. Cross-frame data association is accomplished by feature matching. However, as fish swim in 3D space, losing the depth information will affect the accuracy of motion data and integrity of behavior researches. Therefore, 3D tracking is definitely more essential and valuable. Techniques such as particle image velocimetry (PIV) can be used to investigate behavior in 3D [7], but are not direct ways to obtain the motion data and are largely limited by the space resolution. Several proposals used mirrors to construct a stereo-vision system [8, 9]. Nimkerdphol *et al.* [10] used stereo cameras and perspective correction techniques to obtain 3D coordinates of swimming zebrafish. Voesenek *et al.* [11] and Butail *et al.* [12] developed parameterized 3D fish models and the methods can obtain the full-body trajectory of each individual in a fish school with small quantity (<8 fish).

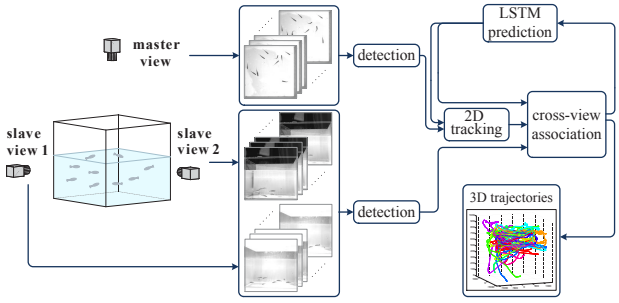

Fig. 1. Workflow of the proposed method.

In a word, 3D fish tracking remains a challenging task due to the vast change of appearance in images, similar appearance among individuals and frequent occlusions. We propose to learn the 3D kinematic model of fish individuals by a Long Short-Term Memory (LSTM) network. A novel master-view-tracking-first strategy is applied based on a master-slave camera setup. Different detection methods are proposed for master and slave views (*cf.* Sec 3.1). Fish are first tracked in master view (top view) guided by the LSTM prediction (*cf.* Sec 3.2). Then 2D tracking results in master view and detection results in slave views are associated to reconstruct the 3D trajectories based on motion continuity and epipolar constraint (*cf.* Sec 3.3). The workflow is shown in Fig. 1.

*The corresponding author gratefully acknowledges the financial support of National Natural Science Foundation of China (Grant No. 61175036 and No. 61602255), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No.16KJB520032).

2. KINEMATIC PATTERN MODELING

The goal of a tracking system is to recursively estimate the target's motion state X^t at each moment given the observation sequence $Z^{1:t}$. The dynamic model in a tracking system calculates the posterior density of the target's motion at each moment which is an important part of the system [13]. In conventional Bayesian tracking framework, the target's motion process satisfies the Markov property, such as first-order Markov chain rule. But in the fish tracking scenario, the state of fish depends on a motion process in several consecutive frames where first-order Markov assumption does not hold.

2.1. Learning kinematic model with LSTM network

Long Short-Term Memory (LSTM) network [14] has shown superior power in processing sequential data with varying lengths and learning long-term dependencies than standard recurrent neural network (RNN). Hence we model the fish's motion process by learning an LSTM network. The learned kinematic model which contains a single LSTM layer, (see Fig. 2(a)) can be implemented in a tracking method to guide the tracking process. The network input is the velocity sequence $V^{1:t}$; output is $h^{1:t}$, where h^t is the hypothetical velocity vector at time $t+1$, written as $\tilde{V}^{t+1}$. Velocity sequences are used instead of position sequences because the information obtained from spatio-temporal data has higher accuracy [15]. Then the predicted state at time $t+1$ is calculated as $\tilde{X}^{t+1} = X^t + \tilde{V}^{t+1}$.

2.2. LSTM training

The LSTM network for midline kinematic patterns modeling is trained offline before tracking. The fish are first tracked using conventional Kalman filter. The resultant midlines are then checked manually and the incorrect ones are removed. Then the velocity sequences are calculated based on the tracking results. The velocity sequences with different lengths are randomly selected from these long sequences as training samples for the LSTM network. We selected totally 50000 velocity sequences of different fish of 8$\sim$20 frames in length to be used as training sequences. The LSTM network is trained with Backpropagation Through Time (BPTT) under a matrix-based batch learning paradigm [16].

3. THE TRACKING METHOD

The appearance of fish is more stable and undergoes less appearance variance in top view images than side view ones, which consists of a rigid head part and a belt-like body part. Hence 2D tracking in top view can obtain higher accuracy, inspiring us to apply a master-slave camera setup: one camera capturing top view images is the master view and the other two cameras capturing side view images are the slave ones. The fish school is first tracked in master view in 2D guided by the 3D prediction of LSTM network, then the 2D trajectories in master view and detection results in the two slave views are associated to reconstruct the 3D trajectories.

3.1. Fish detection

The appearance of fish in master and slave view images varies. We therefore propose different methods for fish detection in different views.

3.1.1. Fish detection in master view

Based on the observation that in top view images, fish head appears as partial ellipses and the head pixels are darker than background ones. We detect fish head using a scale-space Determinant of Hessian (DoH) blob detection method [5]. After fish head detection in master view, the coordinates of fish head points locating at about the middle of the two fish eyes are obtained, as shown in Fig. 2(b).

3.1.2. Fish detection in slave view

In side view images, fish eye region appears as concentric circles, it is the part of fish with minimal change during the tracking period. Gabor filter is a powerful feature extraction strategy widely used in image texture analysis applications [17–19] that can extract orientated feature points with obvious characters, and it's robust under illumination, viewing direction and appearance changes [20]. The proposed eye-focused fish detector applies Gabor filters with varying sizes and orientations to localize fish eyes. The 2D Gabor filter is:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2}\exp(-\frac{\vec{k}^2 x^2}{2\sigma^2})[\exp(i\vec{k}\vec{x}) - \exp(-\frac{\sigma^2}{2})] \quad (1)$$

in which $\vec{k}$ determines the wavelength and orientation of the kernel, σ indicates the ratio of window width to wavelength. The filter needs to sample in both space and frequency domain, written as:

$$\vec{k} = k_v e^{i\phi_u} \quad (2)$$

where $k_v = \frac{k_{max}}{f^v}$, $\phi_u = \frac{u\pi}{8}$, $k_{max} = \frac{\pi}{2}$, $\sigma = 2\pi$. f is the spacing factor between kernels in frequency domain, set as $\sqrt{2}$. Totally 2 frequency levels ($v \in \{0, 1\}$) with 4 orientations ($u \in \{0, 2, 4, 6\}$) are applied to generate local descriptions of an image at different scales and orientations. The input of the Gabor filter set is image patches of size 25×25. The dimension of output feature is $25 \times 25 \times 8 = 5000$. However, the dimension is too high to be directly used as the input of a classifier. According to our experiment, the first 40 components preserve more than 95% of the information of the original feature vector. Hence Principal Component Analysis (PCA) is applied to reduce the feature vector to 40 dimensions. Then the dimension reduced features are fed into

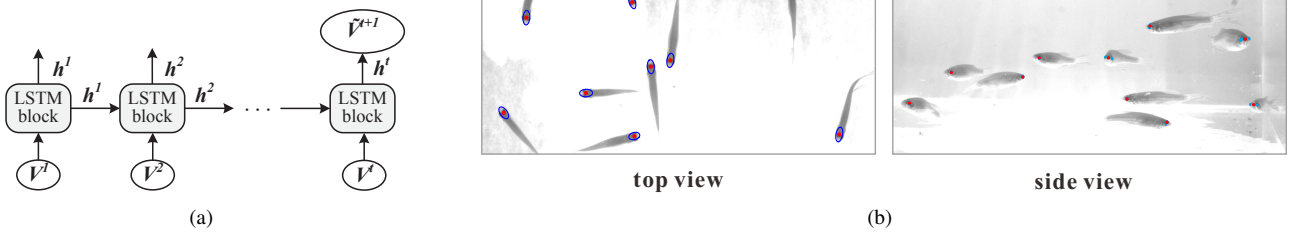

Fig. 2. (a). Illustration of the unfolded LSTM network in the proposed method; (b). Sample detection results of top and side views. In top view image, the blue ovals plot the fitted ellipses, the red points are the resultant fish head points. In side view image, blue points plot the candidate fish eye points, red points are the resultant fish head points after clustering.

a pre-trained SVM classifier to judge if it is a real fish eye or not, resulting in several adjacent candidate fish eye points for each fish eye. Afterwards, Max-Min Distance clustering [21] is performed based on these candidate points and center of each cluster corresponds to each fish head point, as shown in Fig. 2(b). The detected fish head points in the two slave views are $Z^{S_1,t}$ and $Z^{S_2,t}$ respectively.

Detection missing may occur when the back of one fish individual is facing the camera, however, in such case, the individual can be detected in the other slave view and the master view. Therefore, the missing object can still be tracked.

3.2. 2D preliminary tracking in master view

2D tracking in master view is the basis of the following cross-view association and 3D trajectory reconstruction steps. Inspired by the observation that in master view images the displacement of fish head is small resulted from the high frame rate (100fps) and the appearance of fish head varies slightly compared to other body parts, we apply Kalman filter [22] to track fish head points in 2D.

Cross-frame data association is one of the core steps of multi-object tracking, which we formulate as a global optimization problem. Totally two cues are applied to calculate the weight term $\omega(X_i^{m,t}, Z_j^{m,t})$, which measures the probability of target i being associated with target j in master view.

- Motion continuity

  Motion continuity cue measures the distance between the predicted 2D state in master view reprojected from the 3D state $\tilde{X}_i^t$ predicted by LSTM network and detection $Z_j^{m,t}$. The predicted 3D state functions as a constraint for 2D tracking in master view. Based on the output of LSTM network, the predicted velocity $\tilde{V}^t$ of each target is obtained. Then the predicted coordinates of head points are determined, written as $\tilde{X}^t = \{\tilde{X}_i^t = (\tilde{x}_i^t, \tilde{y}_i^t) | i = 1, ..., n\}$. Then the motion continuity term is calculated as:

$$p_m(X_i^{m,t}, Z_j^{m,t}) = \exp[-d(\mathcal{P}_m(\tilde{X}_i^t), Z_j^{m,t})] \quad (3)$$

  where $\mathcal{P}_m(\tilde{X}_i^t)$ is the predicted 2D state in master view reprojected from predicted 3D state $\tilde{X}_i^t$.

- Appearance coherency

  Appearance coherency term $p_a(X_i^{m,t}, Z_j^{m,t})$ calculated by Normalized Cross Correlation (NCC) [23] measures the similarity of fish head image patches determined by predicted 2D state $\tilde{X}_i^t$ and detection $Z_j^{m,t}$.

In summary, the weight term of cross-frame data association is defined as $\omega(X_i^{m,t}, Z_j^{m,t}) = \alpha p_m(X_i^{m,t}, Z_j^{m,t}) + \beta p_a(X_i^{m,t}, Z_j^{m,t})$. After cross-frame data association, for those targets that do not find association, their states are updated as the predicted one, written as $X_i^{m,t} = \mathcal{P}_m(\tilde{X}_i^t)$. With the guidance of LSTM prediction, the correct and complete 2D trajectory of each object is obtained. 2D tracking result at time t consists of the updated state of each object, denoted as $X^{m,t} = \{X_i^{m,t} = (x_i^{m,t}, y_i^{m,t})\}$.

3.3. Cross-view data association

The cross-view association step aims to associate the 2D tracking results in master view and detection results in two slave views. Then 3D trajectories can be reconstructed by the corresponding coordinates in at least 2 views.

- Motion continuity

  Denote $\mathcal{P}_{s_\nu}(\tilde{X}_i^t)$ as the point in slave view ν reprojected from predicted 3D state $\tilde{X}_i^t$ by LSTM network. The probability of detection $Z_j^{s_\nu,t}$ in slave view ν being associated with $\tilde{X}_i^t$ is inversely proportional to Euclidean distance between them, calculated as:

$$p_m(X_i^{m,t}, Z_j^{s_\nu,t}) = \exp[-d(\mathcal{P}_{s_\nu}(\tilde{X}_i^t), Z_j^{s_\nu,t})] \quad (4)$$

- Epipolar constraint

  Epipolar constraint determines the correspondence of objects in master and slave views. Assume $L_i^{s_\nu,t}$ is the epipolar line in slave view ν corresponding to $X_i^{m,t}$ in master view. The epipolar constraint term $p_e(X_i^{m,t}, Z_j^{s_\nu,t})$ is inversely proportional to the Euclidean distance from $Z_j^{s_\nu,t}$ to $L_i^{s_\nu,t}$, calculated as:

$$p_e(X_i^{m,t}, Z_j^{s_\nu,t}) = \exp[-d(L_i^{s_\nu,t}, Z_j^{s_\nu,t})] \quad (5)$$

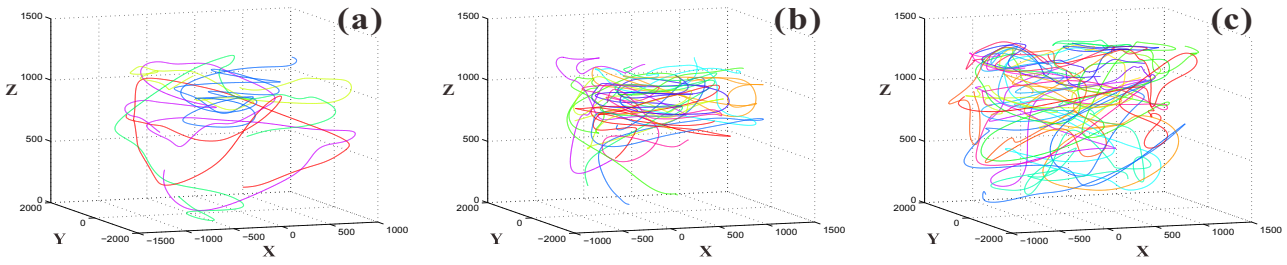

Fig. 3. Resultant 3D trajectories of: (a). V1; (b). V2; (c). V3.

Combing the two cues, the weight term is defined as $W\left(X_i^{m,t}, Z_j^{s_\nu,t}\right) = \alpha p_m\left(X_i^{m,t}, Z_j^{s_\nu,t}\right) + \beta p_e\left(X_i^{m,t}, Z_j^{s_\nu,t}\right)$. We set $\alpha = \beta = 0.5$, the 3D coordinate is then retrieved by triangulation technique [24].

4. EXPERIMENTS

4.1. Experiment setup

Three geometrically calibrated and temporally synchronized monochrome high speed cameras (100fps) were orthogonally placed to capture videos of the zebrafish school (2∼3 cm in length) swimming in a water tank. Each of the three video clips is 2000 frames in length. Descriptions of the videos are shown in Table. 1. The system is implemented with MATLAB$^{\mathrm{TM}}$ and the LSTM network which consists of one layer and 20 hidden units is adapted from [16].

4.2. Experiment results and discussions

We adopt 5 widely used metrics [25] to evaluate the performance of the proposed method and compare it with the other two state-of-the-art methods along with the proposed method without LSTM network. The results are shown in Fig. 3 and Table. 2. idTracker [6] is a recently proposed identity preserved 2D tracking method based on feature matching, which can be extended to 3D by associating the identical individual across views. Liu *et al.*'s method [26] is a 3D tracking method for large swarm of objects, which can be directly applied to track fish group. The results show that when group density increases that leads to high OF, the performance of idTracker drops significantly, because high OF makes feature matching in each view less robust, especially in side views, resulting in unreliable 3D trajectories. Liu *et al.*'s method obtains almost the same scores with the proposed method on V1 and V2, but shows fast decrease as OF increases a lot in V3 (>60% in slave views), because their method is based on particle filtering, the tracking system may accept particles even if they are only observed in one view, and the system lacks a well-designed data association strategy. The proposed method outperforms the others thanks to: 1) the specific detection methods for different views; 2) the learned kinematic model using LSTM network; 3) the master-view-tracking-first strategy; 4) the well-designed data association strategy.

Table 1. Description of the 3 data sets

	Group size	OF in top view	OF in side views
V1	5	0.40%	6.26%
V2	10	17.14%	45.21%
V3	20	24.17%	63.33%

OF indicates the occlusion frequency. The video clips are available online at: http://www.cv.fudan.edu.cn/fishtracking3d.htm

Table 2. Performance comparison with other methods

	Method	P	R	F1	Frag	IDS
V1	Ours	**0.977**	**0.992**	**0.984**	**0.9**	**0.7**
	Ours*	0.968	0.987	0.977	1.4	0.9
	Liu *et al.*	0.967	0.975	0.971	2.9	1.1
	idTracker	0.889	0.950	0.918	6.3	1.9
V2	Ours	**0.961**	**0.971**	**0.966**	**3.9**	**0.9**
	Ours*	0.953	0.967	0.960	4.8	1.1
	Liu *et al.*	0.942	0.958	0.950	6.3	3.8
	idTracker	0.833	0.907	0.868	36.9	7.3
V3	Ours	**0.920**	**0.925**	**0.922**	**5.7**	**1.9**
	Ours*	0.913	0.920	0.916	6.2	2.5
	Liu *et al.*	0.812	0.854	0.832	11.2	7.3
	idTracker	0.285	0.436	0.345	122.7	15.0

Ours* denotes the proposed method without LSTM. P, R, F1, Frag and IDS correspond to Precision, Recall, F1-score, Fragments and ID Switches respectively.

5. CONCLUSION

In this paper we propose to learn the kinematic model of fish by an LSTM network and implement it in a 3D fish tracking system to predict the targets' states at each moment. A novel master-view-tracking-first strategy is proposed on the basis of a master-slave camera setup. Fish are first tracked in master view guided by the LSTM prediction. Then cross-view association is established based on motion continuity and epipolar constraint cues. Experiments on data sets of different fish densities show that the proposed method outperforms the compared state-of-the-art methods.

6. REFERENCES

[1] J. Liu and H. Hu, "Biological inspiration: From carangiform fish to multi-joint robotic fish," *J. Bionic Eng.*, vol. 7, no. 1, pp. 35–48, 2010.

[2] X. Tu and D. Terzopoulos, "Artificial fishes: Physics, locomotion, perception, behavior," in *SIGGRAPH*. ACM, 1994, pp. 43–50.

[3] J. Delcourt, C. Becco, M. Y. Ylieff, et al., "Comparing the ethovision 2.3 system and a new computerized multitracking prototype system to measure the swimming behavior in fry fish," *Behav. Res. Methods.*, vol. 38, no. 4, pp. 704–710, 2006.

[4] D. B. Rosemberg, M. M. Braga, E. P. Rico, et al., "Behavioral effects of taurine pretreatment in zebrafish acutely exposed to ethanol," *Neuropharmacology*, vol. 63, no. 4, pp. 613–623, 2012.

[5] Z. M. Qian, X. E. Cheng, and Y. Q. Chen, "Automatically detect and track multiple fish swimming in shallow water with frequent occlusion," *PLoS One*, vol. 9, no. 9, pp. e106506, 09 2014.

[6] A. Pérez-Escudero, J. Vicente-Page, R.C. Hinz, et al., "idTracker: tracking individuals in a group by automatic identification of unmarked animals," *Nat. Methods*, vol. 11, no. 7, pp. 743–751, JUL 2014.

[7] J. Sakakibara, M. Nakagawa, and M. Yoshida, "Stereo-piv study of flow around a maneuvering fish," *Exp. Fluids*, vol. 36, no. 2, pp. 282–293, 2004.

[8] P. Pereira and R. F. Oliveira, "A simple method using a single video camera to determine the three-dimensional position of a fish," *Behavior Research Methods, Instruments, & Computers*, vol. 26, no. 4, pp. 443–446, 1994.

[9] L. Zhu and W. Weng, "Catadioptric stereo-vision system for the real-time monitoring of 3d behavior in aquatic animals," *Physiol. Behav.*, vol. 91, no. 1, pp. 106–19, 2007.

[10] K. Nimkerdphol and M. Nakagawa, "Effect of sodium hypochlorite on zebrafish swimming behavior estimated by fractal dimension analysis," *J. Biosci. Bioeng.*, vol. 105, no. 5, pp. 486–92, 2008.

[11] C. J. Voesenek, R. P. Pieters, and J. L. van Leeuwen, "Automated reconstruction of three-dimensional fish motion, forces, and torques," *PLoS One*, vol. 11, no. 1, 2016.

[12] S. Butail and D. A. Paley, "Three-dimensional reconstruction of the fast-start swimming kinematics of densely schooling fish," *J. R. Soc. Interface*, vol. 9, no. 66, pp. 77–88, 2012.

[13] L. Mihaylova, A. Y. Carmi, F. Septier, et al., "Overview of bayesian sequential monte carlo methods for group and extended object tracking," *Digit. Signal Prog.*, vol. 25, pp. 1–16, 2014.

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

[15] X. E. Cheng, S. H. Wang, and Y. Q. Chen, "Learning kinematic model of targets in videos from fixed cameras," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, July 2016, pp. 1–6.

[16] Q. Lyu and J. Zhu, "Revisit long short-term memory: An optimization perspective," 2015.

[17] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, vol. 2, no. 7, pp. 1160–9, 1985.

[18] A. Bodnarova, M. Bennamoun, and S. Latham, "Optimal gabor filters for textile flaw detection," *Pattern Recognit.*, vol. 35, no. 12, pp. 2973–2991, 2002.

[19] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–76, 2002.

[20] X. Tan, S. Chen, Z. H. Zhou, et al., "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.

[21] M. Friedman and A. Kandel, *Introduction to pattern recognition [electronic resource] : statistical, structural, neural, and fuzzy logic approaches*, World scientific, 1999.

[22] P. J. Hargrave, "A tutorial introduction to kalman filtering," in *Kalman Filters: Introduction, Applications and Future Developments, IEE Colloquium on. IET*, 1989, pp. 1–6.

[23] J.P. Lewis, "Fast template matching," *Vision Interface*, vol. 10, no. 1, pp. 120–123, May 1995.

[24] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.

[25] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[26] Y. Liu, S. H. Wang, and Y. Q. Chen, "Automatic 3d tracking system for large swarm of moving objects," *Pattern Recognit.*, vol. 52, no. C, pp. 384–396, 2015.