3D Tracking Swimming Fish School using a Master View Tracking First Strategy

Shuo Hong Wang*, Xiang Liu*†, Jingwen Zhao*, Ye Liu‡ and Yan Qiu Chen*
*School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai, China {sh_wang, xiangliu09, jingwenzhao13, chenyq}@fudan.edu.cn
†School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China
‡College of Automation, Nanjing University of Posts and Telecommunications, Jiangsu, China yeliu@njupt.edu.cn

***Abstract*—3D motion data of fish school is more valuable than 2D data for behavior and other researches. This paper proposes to use a master view tracking first strategy based on a novel master-slave camera setup. On this basis, fish are firstly tracked in master view in 2D after being extracted via an eye-focused Gaussian Mixture Model (E-GMM) detector. Then 3D trajectories are reconstructed by associating 2D tracking results in master view and detection results in slave views after fish in slave views are localized using an eye-focused Gabor (E-Gabor) detector. Experiments on data sets with different fish densities demonstrate that the proposed method outperforms two state-of-the-art methods in terms of 5 evaluation metrics.**

***Keywords*-3D tracking fish school; Master-slave camera setup; Master view tracking first strategy; Eye-focused fish detector; Cross-frame cross-view data association**

I. Introduction

The motion data of each individual in a fish school is of great value for collective behavior, ontogeny, hydrodynamics and bio-inspired robot design researches. Hence it has drawn significant attention of researchers from different areas. Video tracking is the most direct and effective way to obtain the accurate motion data of each individual [1].

Most fish behavior experiments use shallow water where fish swim in almost the same plane, therefore most fish tracking systems are based on 2D tracking. Although reducing the dimensionality by one did achieve good performance, the difference from the reality that fish swim in 3D space leads to lost accuracy and damaged integrity of researches on fish behavior. Therefore, the necessity is highlighted that fish behavior studies should be investigated in 3D space.

Particle image velocimetry (PIV) has been widely used to investigate the 3D motion [2], but it's not a direct way to obtain the motion data and the measurement accuracy is limited. Pereira *et al.* used mirrors to build stereo-vision systems [3]. Nimkerdphol *et al.* [4] tracked zebrafish in 3D using nonplanar 3D stereocameras in combination with perspective correction. Butail *et al.* [5] applied model based methods to track limited number of fish with body shape.

However, 3D tracking fish school remains a challenging task mainly due to the similar appearance among individuals, vast change of appearance in 2D images and frequent occlusions. Based on the observation that fish in top view images have more stable appearance, 2D tracking in top view is thus more robust than in the other views. Hence, this paper proposes a method using a novel master-slave camera setup where 3 synchronized and calibrated cameras are orthogonally placed, one master camera to capture the top view and the other two slave ones to capture the side view of the container. Fish are first tracked in master view in 2D and then the 3D trajectories are reconstructed by associating 2D tracking results in master view and detection results in slave views. The contributions of the paper are:

- Proposing a novel master view tracking first strategy using a top view camera and two side view ones.
- Proposing an eye-focused Gaussian Mixture Model (E-GMM) detector for fish head detection in master view and an eye-focused Gabor (E-Gabor) detector for slave views, both are capable of avoiding mutual occlusions.
- Proposing a novel cross-view data association method.

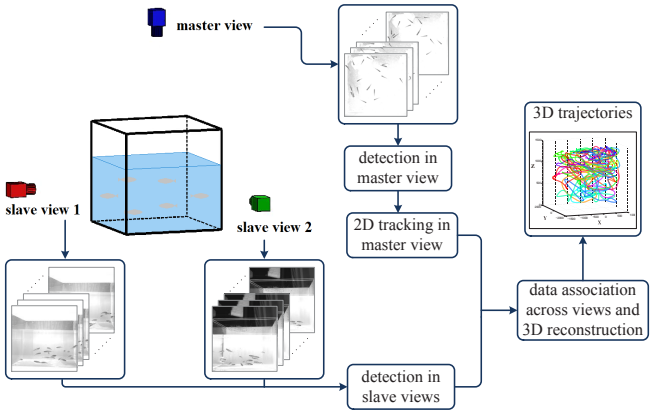

Figure 1: Procedure of the proposed method.

II. Proposed Method

A. Overview

Most existing 3D tracking systems locate the cameras at the side of the targets. However, the appearance of fish varies a lot during tracking in side view images. In top view images, fish can be viewed as composed of a rigid head and a belt-like body, which is more stable with less

*Corresponding author: Yan Qiu Chen. This work was supported by National Natural Science Foundation of China, Grant No.61175036.

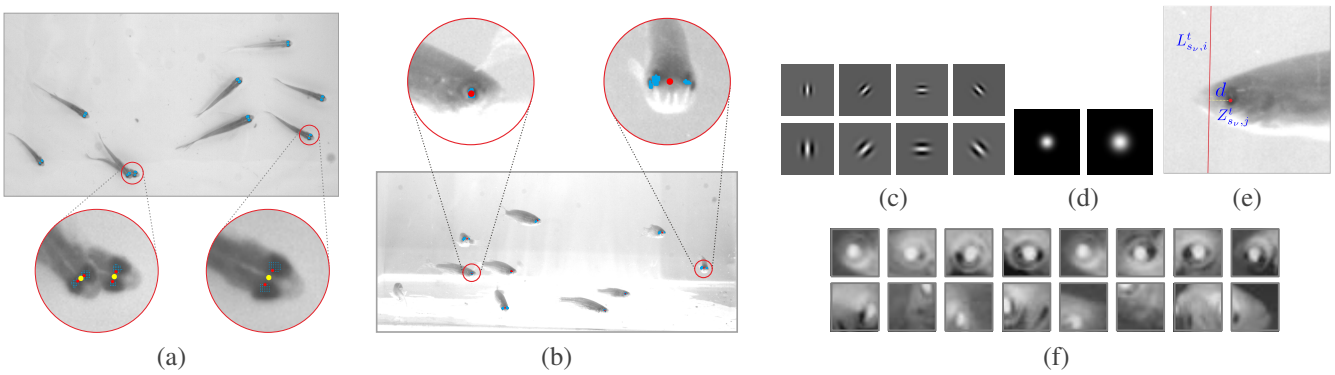

Figure 2: (a). Sample detection results of E-GMM detector in master view; (b). Sample detection results of E-Gabor detector in one slave view; (c). Illustration of the real part of the Gabor kernels with 2 different scales and 4 orientations; (d). Magnitude of the Gabor kernels corresponding to 2 scales used; (e) Illustration of $d(Z_j^{s_\nu,t}, L_i^{s_\nu,t})$ in Eq. (2); (f). Training samples of SVM. Top: Positive samples; Bottom: Negative samples.

appearance variance. 2D tracking in top view is thus more accurate and reliable. So we apply a master-slave setup, one master camera captures top view images, and the other two slave cameras capture side view images. Fish are first tracked in 2D in master view, then the resultant 2D trajectories in master view and detection results in two slave views are associated to reconstruct the 3D trajectories (*cf.* Fig. 1).

B. Fish detection

We propose different eye-focused detection methods for detecting fish in different views. In top view image, fish eyes appear as two dark blobs, and the appearance of eyes keeps stable during tracking, which inspires us to detect fish eyes in master view with Gaussian Mixture Model (GMM). It is efficient and robust even when partial occlusion occurs. In side view images, the shape and appearance of the fish body vary greatly, fish eye region which appears as concentric circles is the part of fish body with minimal change during the tracking period, motivating us to apply Gabor features to detect fish eyes in side views.

1) Fish detection in master view: Fish in master view are detected by an eye-focused Gaussian Mixture Model (E-GMM) detector. Assume the number of fish is n_{fish}, then there are $2 * n_{fish}$ eyes. For each frame, background is subtracted using temporal median filter [11]. Then Gauss filtering is performed, resulting in a filtered image I_0. In this way, uneven brightness of the fish body can be eliminated. Local maximum points in I_0 are then computed (blue dots in Fig. 2(a)), and GMM clustering [6] is performed on these local maximum points to localize fish eyes. The mean value of each Gaussian model corresponds to one fish eye (red points in Fig. 2(a)). Then the detected fish eyes are globally matched by Kuhn-Munkres algorithm [7] to find adjacent ones that belong to the same fish. The fish head position (x^i, y^i) is defined as the midpoint of two eyes (yellow dots in Fig. 2(a)). The observation vector of the detected fish i is defined by its head position: $Z^i = (x^i, y^i)^T$.

2) Fish detection in two slave views: Gabor filter can be used to extract oriented feature points with obvious characters and is robust to illumination, viewing direction and appearance changes [8]. And filters with varying sizes and orientations can be applied to detect and localize features at different scales. Therefore, fish eyes in slave views are detected by an eye-focused Gabor (E-Gabor) detector.

2D form Gabor filter takes the form of a plane wave restricted by a Gaussian envelope function [9]. To generate local descriptions of an image with different scales and orientations, different frequency levels (v) and orientations (u) are applied. Based on the observation that the eyes of different fish individuals in our experiments have little variance in size, to reduce the dimension of feature vector, we use only 2 different scales of Gabor mask, each with 4 different orientations (*i.e.*, $v \in \{0, 1\}$, $u \in \{0, 2, 4, 6\}$). Figs. 2(c), 2(d) show the real part of the Gabor kernels with 2 different scales and 4 orientations and the magnitude of Gabor kernels corresponding to the 2 scales used in the proposed method. The input of the Gabor filter set is 25×25 image patches, which can guarantee that the patch contains the whole eye region. The output response of each patch is normalized by subtracting mean value and dividing standard deviation, with dimension equals to 25×25×8=5000.

However, 5000 dimensions are too high to be directly used to train a classifier, hence Principal Component Analysis (PCA) is performed to reduce the feature vector to 40 dimensions, as according to our experiments the first 40 components preserve more than 95% of the information of the original feature vector. A trained SVM classifier is then applied to judge if an image patch is a fish eye or not. As a result, the image patches centered at several adjacent pixels at fish eye area will be judged as 'fish eye region' (blue points in Fig. 2(b), which are candidate fish eye points).

Afterwards, Max-Min Distance Clustering [10] is performed based on these blue points. Red points in Fig. 2(b) show the resultant fish head positions after clustering. The fish head position can be correctly localized when one or two of its eyes can be seen in the image. Remarkably, when one fish is missed in detection step in one slave view (*e.g.*, the left most fish in Fig. 2(b)), it can be detected in the other views. Hence it can still be successfully tracked.

To train the SVM classifier, we manually choose real fish eye image patches as positive samples and other image patches as negative samples (*cf.* Fig. 2(f)). And we calculate the mirror images of the positive samples to double the size of positive sample set. Altogether we selected 1200 positive samples and 8000 negative samples.

C. 2D tracking in master view

Considering these observations: (1) In top view images fish head can be viewed as undergoing a rigid transform during swimming; (2) The shape and appearance of fish head vary slightly compared to other parts of fish body; (3) The displacement of fish head between adjacent frames is small resulted from the high frame rate (100 fps), the motion of fish head can be accurately predicted by linear dynamic system model. Hence we apply Kalman filter to track fish in master view.

For multi-object tracking systems, association of the tracked targets and detection results is essential. The association should follow a one-to-one criterion which is defined as a global optimization problem. Assume $Z_j^{m,t}$ is the detected fish head points and $\tilde{X}_i^{m,t}$ is the predicted fish head point of each target in master view (superscript m indicates master view) respectively. The weight term $\omega(\tilde{X}_i^{m,t}, Z_j^{m,t})$ of target i being associated with detection j combines cues from head appearance coherency and motion continuity:

$$\begin{aligned}&\omega(\tilde{X}_i^{m,t}, Z_j^{m,t}) = \\ &\exp\{-[\alpha p_n(\tilde{X}_i^{m,t}, Z_j^{m,t}) + \beta p_d(\tilde{X}_i^{m,t}, Z_j^{m,t})]\} \\ &(i = 1, ..., n_{x,m},\ j = 1, ..., n_{z,m})\end{aligned} \tag{1}$$

where $p_n(\tilde{X}_i^{m,t}, Z_j^{m,t})$ is the normalized-cross-correlation (NCC) [11] between the corresponding head image patch of predicted head state $\tilde{X}_i^{m,t}$ and detection result $Z_j^{m,t}$ that reflects the similarity of the image patches. $p_d(\tilde{X}_i^{m,t}, Z_j^{m,t}) \propto \mathcal{N}(\|\tilde{X}_i^{m,t} - Z_j^{m,t}\|; 0, \Sigma)$ reflects the distance between predicted state of target $\tilde{X}_i^{m,t}$ and detection $Z_j^{m,t}$, which is inversely proportional to the Euclidean distance between them. $\alpha = 0.6$ and $\beta = 0.4$ are set empirically. The resultant 2D trajectories $\gamma = \{\gamma_i | i = 1, ..., n_\gamma\}$ in master view contain the coordinate of each object in each frame, denoted as $\gamma_i = \left\{X_i^{m,t} = (x_i^{m,t}, y_i^{m,t}) | t = st_{\gamma_i}, ..., ed_{\gamma_i}\right\}$, st_{γ_i} and ed_{γ_i} are the start and end frame of trajectory γ_i.

D. Cross-view association and 3D trajectory reconstruction

Cross-view data association aims to associate the 2D trajectories in master view and detection results in 2 slave views in order to reconstruct the 3D trajectories.

Hence if the tracked object can find its associated detection result in either of the two slave views, it will be regarded as successfully associated and the 3D coordinate will be retrieved by the coordinates in the master and one slave view together with intrinsic and extrinsic parameters of the two cameras. By applying this strategy, the influence of detection missing is reduced to the greatest extent.

Assume $Z^{s_\nu,t}$ is detection results (fish head coordinates) of slave view ν at moment t, $\nu = \{1, 2\}$, $Z^{s_\nu,t} = \left\{Z_i^{s_\nu,t} = (x_i^{s_\nu,t}, y_i^{s_\nu,t}) | i = 1, ..., n_{s_\nu}\right\}$. Cross-view association is also formulated as a global optimization problem and solved by Kuhn-Munkres algorithm [7]. The weight term $W(X_i^{m,t}, Z_j^{s_\nu,t})$ combines epipolar constraint and motion continuity cues to improve association accuracy.

- Epipolar constraint
  Assume $L_i^{s_\nu,t}$ is the corresponding epipolar line in slave view ν of $X_i^{m,t}$ in master view, $Z_j^{s_\nu,t}$ is one detected fish head point in slave view ν. The probability of detection $Z_j^{s_\nu,t}$ in slave view being associated with $X_i^{m,t}$ is inversely proportional to the Euclidean distance from $Z_j^{s_\nu,t}$ to $L_i^{s_\nu,t}$, formulated as Eq. (2) (illustrated in Fig. 2(e)).

$$p_e(X_i^{m,t}, Z_j^{s_\nu,t}) = \exp[-d(Z_j^{s_\nu,t}, L_i^{s_\nu,t})] \tag{2}$$

- Motion constraint
  Considering that the frame rate of the cameras is relatively high (100fps), the displacement of fish head in several consecutive frames is nearly uniform, we model the variation of fish in slave view with first order linear extrapolation. The state of the object j at current frame in slave view ν is predicted by linear extrapolation of the last two locations: $\tilde{X}_j^{s_\nu,t} = 2X_j^{s_\nu,t-1} - X_j^{s_\nu,t-2}$. Assume $X_{\kappa(i)}^{s_\nu,1:t-1}$ is the state sequence of the object $\kappa(i)$ in slave view ν that associated with target i in master view. The probability of detection $Z_j^{s_\nu,t}$ being associated with $\tilde{X}_{\kappa(i)}^{s_\nu,t}$ is inversely proportional to the distance between predicted state $\tilde{X}_{\kappa(i)}^{s_\nu,t}$ and detection $Z_j^{s_\nu,t}$, written as:

$$p_m(X_i^{m,t}, Z_j^{s_\nu,t}) = \exp[-d(Z_j^{s_\nu,t}, \tilde{X}_{\kappa(i)}^{s_\nu,t})] \tag{3}$$

Combining the epipolar constraint and motion continuity cues, the weight term $W\left(X_i^{m,t}, Z_j^{s_\nu,t}\right)$ is defined as:

$$W(X_i^{m,t}, Z_j^{s_\nu,t}) = \alpha p_e(X_i^{m,t}, Z_j^{s_\nu,t}) + \beta p_m(X_i^{m,t}, Z_j^{s_\nu,t}) \tag{4}$$

α and β are both set as 0.5 empirically. After associating the 2D tracking results in master view and detection results in two slave views, the associated coordinates of each object in at least two views (including master view) are obtained. The 3D coordinate of each object at each frame is then reconstructed by triangulation technique [12]. The resultant 3D trajectory set is written as $\Gamma = \{\Gamma_i \mid i = 1, ..., n_\Gamma\}$, $\Gamma_i = \{X_i^t = (x_i^t, y_i^t, z_i^t) \mid t = st_{\Gamma_i}, ..., ed_{\Gamma_i}\}$, st_{Γ_i}, ed_{Γ_i} are the start and end frame of trajectory Γ_i.

III. Experiments

Three calibrated and synchronized monochrome high speed cameras are orthogonally placed to capture 3 videos of zebrafish school containing 5, 10, 20 zebrafish respectively at 100fps. The result trajectories are plotted in Fig. 3.

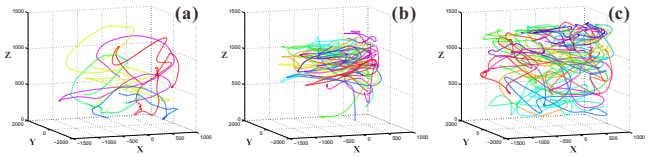

Figure 3: Result 3D trajectories of: (a). V1; (b). V2; (c). V3.

1) Performance of detection: We tested the performance of the proposed E-GMM and E-Gabor fish detector for master / slave views respectively on the 3 data sets. It can be concluded according to results in Table. I that the two detectors can both accurately and reliably detect fish individuals even when OP increases dramatically.

2) Performance of tracking: We adopt 5 widely used metrics [13] to evaluate the performance of the proposed and the other 2 state-of-the-art methods: (Liu *et al.*'s [11] and idTracker [14]). idTracker is a newly proposed 2D tracking method based on intensity distribution feature matching to preserve the identity of each individual during tracking. It can be extended to track fish in 3D space by associating the identical individual across views. Liu *et al.*'s method is a newly proposed 3D swarm tracking method. The evaluation results in Table. II show that idTracker is largely affected by the density of fish group, because frequent occlusions make 2D tracking in each view less robust, especially in side views. Liu *et al.*'s method applies particle filter technique, which may accept particles even if they are only observed in one view and thus lead to tracking errors. The proposed method is more robust than the other two methods thanks to its specific method to detect fish in master and slave views and the cross-view data association strategy.

IV. Conclusions

We have proposed in this paper a reliable 3D fish school tracking method based on a master-slave camera setup. We propose a master view tracking first strategy that first track fish in master view based on the detection results via E-GMM detector, then the 3D trajectories are reconstructed by associating 2D trajectories in master view and detection results by E-Gabor detector in slave views. Evaluation on data sets with different fish densities validates the effectiveness of the proposed method. It outperforms the other two state-of-the-art methods in terms of 5 evaluation metrics.

Table I: Evaluation of fish detection methods

Data Set	OP in master view (%)	OP in slave views (%)	DA in master view (%)	DA in slave views (%)
V1	0.42	6.67	99.24	98.94
V2	17.14	45.21	97.82	95.74
V3	24.17	63.33	96.13	90.76

OP: Occlusion Probability, DA: Detection Accuracy

Table II: Evaluation results of 3D tracking performance

Data Set	Method	P	R	F1	Frag	IDS
V1	Ours	**0.970**	**0.989**	**0.979**	**1.3**	**0.9**
	Liu *et al.*	0.966	0.973	0.969	2.8	1.3
	idTracker	0.884	0.951	0.916	6.1	1.8
V2	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	**0.913**	**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

The videos are available online: http://www.cv.fudan.edu.cn/fishtracking3d.htm

References

[1] J. Delcourt, M. Denoël, M. Ylieff *et al.*, "Video multitracking of fish behaviour: a synthesis and future perspectives," *Fish. Fish.*, vol. 14, no. 2, pp. 186–204, 2013.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] C. Fraley and A. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *Comput. J.*, vol. 41, no. 8, pp. 578–588, 1998.

[7] H. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Logist.*, vol. 2, no. 1-2, pp. 83–97, 1955.

[8] C. Liu, "Gabor-based kernel pca with fractional power polynomial models for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 572–81, 2004.

[9] M. Lades, J. C. Vorbrüggen, J. Buhmann *et al.*, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, no. 3, pp. 300–311, 1993.

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

[11] Y. Liu, H. Li, and Y. Q. Chen, "Automatic tracking of a large number of moving targets in 3d," in *ECCV*, 2012, pp. 456–463.

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

[13] 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.

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