IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 4, 2019, accepted September 23, 2019, date of publication October 4, 2019, date of current version October 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945606

3-D Video Tracking of Multiple Fish in a Water Tank

XIAOQING LIU[1], YINGYING YUE[2], MEILING SHI[3], AND ZHI-MING QIAN[1]
[1]School of Information Science and Technology, Chuxiong Normal University, Chuxiong 675000, China
[2]School of Mathematics and Information Technology, Yuxi Normal University, Yuxi 653100, China
[3]School of Physics and Electronic Engineering, Qujing Normal University, Qujing 655000, China

Corresponding author: Zhi-Ming Qian (qzhiming@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61363023, and in part by the Joint Fund Project for Basic Research in Local Undergraduate Universities of Yunnan Province under Grant 2017FH001-063.

ABSTRACT Study on fish behavior is essential to fishery development and water environment protection. Quantitative analysis of fish behavior is impossible without information on fish trajectories. Although the computer vision techniques have provided an effective approach to the collection of fish trajectories, it is still challenging to track fish groups accurately and robustly due to fish appearance variations and frequent occlusions. In this paper, a skeleton-based method for 3-D tracking of fish group is proposed. First, skeleton analysis is performed to simplify the top- and side-view objects into representation of feature points. Next, the feature points under the top view are associated to obtain the top-view trajectories of objects. Finally, the epipolar constraint and the motion continuity constraint are used to match the top-view trajectories with the side-view feature points, thereby obtaining the 3-D trajectories of objects. Experimental results demonstrate the ability of the proposed method to track fish group more effectively than other state-of-the-art methods.

INDEX TERMS Fish tracking, 3-D tracking, skeleton analysis, motion trajectory.

I. INTRODUCTION

Animal behavior analysis is a hot research topic in natural science. As a key member of aquatic animal family, fish bears an important status in the animal world. Fish behavior analysis is important to promote the development of sport dynamics and collective ethology. It is of great significance to fishery and water environment protection as well. The most intuitive and effective method to study fish behaviors and extract their pattern is to first obtain their motion trajectories. This is done by sampling and quantifying the motion data of each individual fish, and then by analyzing the intra-track and inter-track relationships. Fishes live across a wide region. Data of their motion in the rivers or lakes can be collected using sensor-based or acoustics-based methods [1], [2]. These two methods can acquire real-world data of fish behavior from the living environment. However, deploying these methods requires purchasing expensive hardware in advance. Besides, equipment installation and configuration are very complex. Therefore, only a few institutions conduct research in this way. To facilitate research, the fishes are usually housed in the laboratory. Under the laboratory environment, the camera is used to take photo of the fish. The most effective method for extracting fish trajectories from captured images is to track the target fish in the image individually by using computer vision technique [3], [4]. Based on the processes used, the approach of computer vision-based fish tracking can be classified into 2-D and 3-D [5].

For 2-D tracking, the objects are placed into a container with shallow water. A camera is used to track each individual fish in 2-D plane. In the case of 3-D tracking, several cameras are used to capture the motion of fish from different views to estimate their motion trajectories in 3-D space. The trajectories obtained through 3-D tracking are valuable and useful to research, as they closely resemble to those of real-world behavior of fish.

Unlike the common scenario of 3-D tracking, camera imaging is prone to be affected by unsettled water surface due to fish swimming in the water. In the traditional method for 3-D tracking based on binocular vision, the parallax between two cameras is very little. Due to this, stereo matching can be performed using appearance similarity of objects under two views. But in underwater environment, 3-D reconstruction is inaccurate if the fish trajectory is captured in this way. In order to alleviate problems arising from water surface refraction, the most effective way of capturing image is to vertically align the camera's optical axis to water surface. But the parallax of cameras is large in such a way that the

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

object appearance differs considerably under different views. As a result, appearance similarity is no longer useful for stereo matching, and subsequently it becomes more difficult to model the objects.

To solve these problems, a tracking method based on three views had been proposed by us in a previous research [6]. Then, it was improved at the CCCV2017 [7] (denoted as SK-3D). Especially, the top view plays a main role in tracking, while the detection results of the side views are not used in tracking. Instead, they are used for stereo matching with the tracking result of the top view. On the basis of SK-3D, this paper offers extensions in the following aspects:

(1) Extend and supplement Sec.I Introduction and Sec.II Related work, and elaborate on the parameters of the proposed method.

(2) Modify and improve the top-view tracking and stereo matching in the proposed method and further enhance the tracking performance of the method.

(3) Enrich the experimental data and compare and analyze the tracking performance of the proposed method with more contrast methods.

II. RELATED WORK

Many researchers had proposed some effective methods for 3-D fish tracking. Hughes and Kelly [8] first began to study for tracking fish in 3-D space, and designed a method to track fish swimming movements by using multiple video cameras. This method uses a point model to represent the object, which simplifies the tracking process. However, the calculation of the object centerline is complicated, occlusion processing is not considered, and the association accuracy is not high. Goodwin *et al.* [9] presented an Eulerian–Lagrangian–agent method which can handle dynamics at multiple scales simultaneously for tracking 3-D movement patterns of individual fish. Their method can perform 3-D tracking of fish in real environments, but the spatial transformation process is complicated and the hardware requirements are high. Zhu and Weng [10] designed a catadioptric stereo-vision system for obtaining the motion trajectories of several fish by using a video camera and two planar mirrors. This system can accurately reconstruct the 3-D trajectory of the fish, but the hardware installation and the debugging process are cumbersome (two mirrors are required), the ability of tracking occluded objects is weak, and only a small number of objects can be tracked. Nimkerdphol and Nakagawa [11] proposed a method to compute the 3-D coordinates of zebrafish by using nonplanar 3-D stereo cameras in combination with 3-D coordinate computation with perspective correction. This method can continuously track a single fish for a long time but does not have the capability of multi-object tracking, which leads to certain limitations. Wu *et al.* [12] presented a multi-object multi-camera framework for tracking large numbers of flying objects by using a greedy randomized adaptive search procedure. This framework can track a large number of objects simultaneously and is universally applicable. However, only the position information of the object is considered. For this reason, the tracking effect is poor in case of complicated motion or occlusion. Wu [13] introduced a tracking system to obtain the 3-D kinematic parameters in fish swimming by simultaneously taking images from the ventral view and the lateral view with two cameras. This system can obtain the motion parameters of a fish swimming, which is beneficial to behavior analysis. However, the hardware setting is complicated and only a single object can be tracked, which limits the application scenarios. Butail and Paley [14] designed a 3-D tracking framework to reconstruct the motion trajectories of densely schooling fish using 2-D silhouettes from multiple camera views. This framework can track multiple fish simultaneously and can obtain multiple motion parameters of the fish, but the calculation process is complicated, and the ability of tracking of occluded objects is not strong. Maaswinkel *et al.* [15] presented an automatic video tracking system to acquire the swimming trajectories of single and groups of zebrafish by using a mirror system and a calibration procedure. This system can accurately locate the object in the 3-D space, and it provides high accuracy regarding the tracking result. However, only a small number of objects can be tracked, and a lot of parameters are required for the installation and the debugging. Pérez-Escudero *et al.* [16] proposed a multiple fish tracking method that extracts a characteristic fingerprint from each object in a video recording. Their method uses simple texture features to recognize the objects, and has high tracking accuracy, but its calculation speed is slow. Pautsina *et al.* [17] designed an infrared reflection (IREF) system for indoor 3-D tracking of fish based on the effect of strong absorption of near infrared (NIR) range light by water. This system is not affected by illumination conditions; the hardware installation and the debugging are conveniently designed. However, the tracking performance is greatly affected by occlusion interference, and the tracking accuracy is not high. Voesenek *et al.* [18] presented a morphology-based method to track a fish in 3-D space by reconstructing its position, orientation and body curvature from multiple cameras. Their method can obtain multiple motion parameters of the object, which is beneficial to behavior analysis. However, it has high requirements regarding the image resolution, and only a single object can be tracked, which limits the application scenarios. Saberioon and Cisar [19] presented a tracking system which used the structured light (SL) emission sensor to monitor multiple fish activities in 3-D space. This system uses a depth camera to track the object. The hardware installation and the debugging are simple, and multiple objects can be tracked. However, the tracking performance is susceptible to appearance similarity and occlusion. Wang *et al.* [20] proposed a 3-D tracking method of multiple fish based on a master-slave camera setup from three views. Their method uses two-view fusion to achieve 3-D tracking, which is currently a better tracking mode, but the tracking accuracy is not high because the motion direction of the object is not considered.

Unlike the usual multi-object tracking scenario, the fish is characterized by a large number of individuals, a high

level of similarity between individual appearances as well as frequent occurrence of occlusions. These factors make it a big challenge to track the fish. When it comes to 3-D tracking, uncertainty in stereo matching is another problem that needs to be addressed. Although the above methods can obtain the fish school's 3-D motion trajectories, these problems are not solved very effectively. Accurate and robust 3-D tracking of fish remains a real challenge.

In order to enhance the performance of multiple fish 3-D tracking and reduce the tracking difficulties brought about by occlusion and stereo matching, this paper proposes a skeleton based multiple fish 3-D tracking method. This method first adopts skeleton analysis to simplify the object into feature points that include head position and moving direction for description; then it sets up the metric function of motion continuity for the feature points of the consecutive frames in the top view and carries out local optimization association according to this function to obtain the top-view motion trajectories of the objects; at last, it brings the top-view tracking results into stereo matching with the feature points in the side view by means of the epipolar constraint and motion continuity constraint, obtaining the motion trajectories of the objects in 3-D space. The contribution of this paper lies in the following aspects:

(1) The accuracy of the data association of occluded objects in the top view is improved by combining the information of the side view in the data association of the top view.

(2) The continuity and integrity of the tracking results is enhanced by using the trajectory connection to process the top view trajectories.

(3) The accuracy of the trajectory is improved by using the main skeleton points to achieve stereo matching of occluded objects in the side view.

III. THE PROPOSED METHOD

Two synchronization cameras are used to record videos of objects motion in the rectangular container from the top and side view vertical to the water surface. First, the main skeletons for moving regions are extracted and the feature points that are most expressed in the skeletons are obtained. Next, feature points in neighboring frames are associated in the top view and 2-D tracking results are acquired. Finally, 3-D motion trajectories are reconstructed by matching top-view tracking with features points in the side view. Fig. 1 shows an overview of the proposed tracking method.

A. OBJECT DETECTION

Fish assumes a strip structure in two view images. Inspired by this observation, a skeleton analysis is performed to simplify objects from the 2-D region to the 1-D curve. Next, the feature points which consist of the skeleton endpoints and motion direction are obtained to represent the object.

1) MOVING REGION SEGMENTATION

In the laboratory environment, the video images usually consist of moving objects and nearly static background, and each

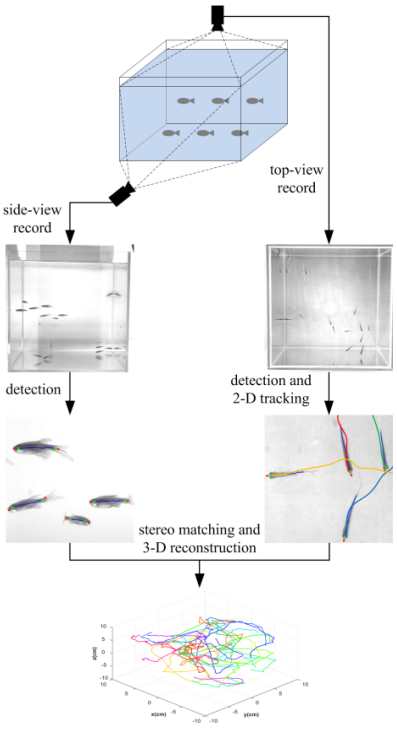

FIGURE 1. An overview of the proposed method.

object stays only for a short time in an area. Thus, moving regions can be segmented from the difference between the current frame and a background image.

$$R_t = \{(x, y) \in I \mid |median(x, y) - I_t(x, y)| > T_g\} \quad (1)$$

where R_t stands for the obtained moving regions, $I_t(x, y)$ for the t-th frame image, T_g for the segmentation threshold, and median(x, y) for the background image which consists of the median image of the first N frames in each view. In order to minimize interference from the moving region's coarse edge during subsequent extraction of the skeleton, we first perform a morphologic operation on the moving region, fill in the holes within the region, and delete small interfering regions. The moving region is then smoothed using median filter.

2) SKELETON EXTRACTION

Many skeleton extraction algorithms exist at present, which are mostly used for object matching and recognition [21], [22]. However, most of such algorithms cannot provide support for the tracking process due to considerable change of the skeleton and frequent occlusion of the objects. Considering its ability to enormously reduce the difficulty in object analysis, the skeleton is particularly used in this paper to simplify the object's appearance and structure. Moreover, the large population of the objects underscores the need to guarantee tracking speed by maximizing skeleton extraction efficiency. Therefore, the augmented fast marching method (AFMM) [23] based on the level set is chosen to extract

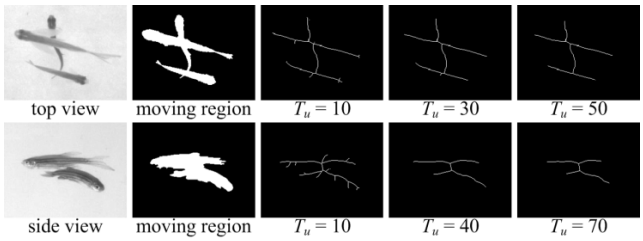

FIGURE 2. The obtained skeleton results of the moving region with different T_u.

skeleton from the motion region. In addition to efficiently extracting skeleton of certain regions in the image, it can self-define skeleton structure and effectively eliminate interferences from small branches like burrs without compromising skeleton integrity. First, an arrival time U is set for each point at the edge of the region, and then the value of U for the entire region is obtained by iteration of the fast marching method. Based on the distribution of U, the skeleton points can be defined as:

$$s = \{(i,j)|\max(|u_x|,|u_y|) > T_u\}$$
$$u_x = U(i+1,j) - U(i,j), \quad u_y = U(i,j+1) - U(i,j) \tag{2}$$

This equation shows that, for a given point(i, j), we regard this point as the skeleton point when the larger of u_x and u_y is larger than threshold T_u, where u_x and u_y denote the difference of U between this point and its neighbor in the x and y directions. In the skeleton extraction process, T_u refers to the capacity of skeleton to describe the object structure. Smaller T_u and more skeleton branches mean the stronger capacity to describe the details of object; larger T_u and less skeleton branches reveal the greater capacity to describe the major structure of object. Fig. 2 shows the skeleton extraction results of the object under different T_u. It can be seen from the figure that the increasing T_u is accompanied by less small branch structures, such as the burrs in skeletons. In object detection, a larger T_u is set for the decrease in skeleton analysis difficulty and increase in detection performance.

3) FEATURE POINT REPRESENTATION

The main skeleton of fish usually has two endpoints only, one at the head and the other at the tail. Based on this observation, we simplify the object's structure from the skeleton curve to the feature point.

Assume the skeleton endpoint is se, where an endpoint segment $es = \{(x_i, y_i)|i = 1, \ldots, K\}$ consists of K skeleton points closest to se. The direction of the endpoint segment can be calculated based on the least squares method.

$$\theta = \arctan\left(\frac{\sum x_i \sum y_i - K\sum x_i y_i}{(\sum x_i)^2 - K\sum x_i^2}\right) \tag{3}$$

The position p of skeleton endpoint se and the direction θ are combined into a feature point $F(p, \theta)$, where the feature point is used to represent the object, as shown in Fig. 3. This representation has the following advantages: (1) Less data: only one point is necessary to effectively represent the object, thus improving the tracking efficiency, (2) Direction information: reduces the difficulty of stereo matching and effectively improves the accuracy of data association, and (3) Occlusion handling: even if an occlusion exists between objects, the occluded objects can still be represented effectively (see Fig. 4), thereby greatly improving the ability of occlusion tracking.

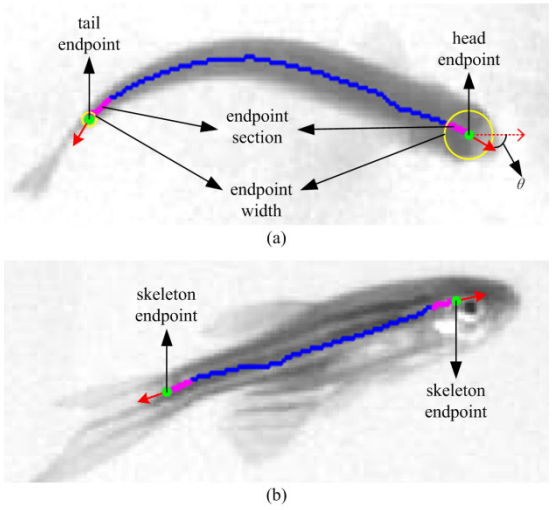

FIGURE 3. Feature points in different views. (a) Top view. (b) Side view.

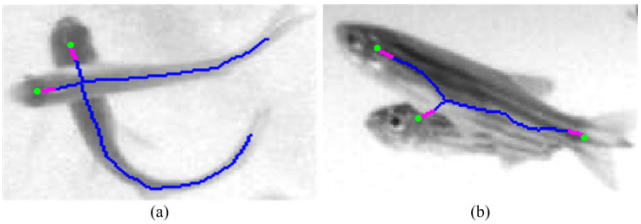

FIGURE 4. Feature point in occlusion scene. (a) Top view. (b) Side view.

The fish body in the top view becomes thinner from head to tail. Based on the characteristic, we draw a circle with each feature point in the top view as the center, and the shortest distance between the center and the edge is the radius r. The fish body width located at this point can be approximated as $2r$. Let w_h and w_t denote the width of the feature points in the head and tail of the fish body, respectively. Based on the shape of the fish body in the top view, it is observed that $w_h > w_t$, as shown in Fig. 3(a). Therefore, the tail feature point can be eliminated by comparing the width. The position and direction of the object is represented by the head feature point.

The shape variation of fish in the side view is complicated and it is challenging to recognize the head feature point from all feature points. Hence, all feature points are first used to represent the object, and an effort is made during the stereo matching process to determine the head feature point.

B. OBJECT TRACKING

1) TOP-VIEW TRACKING

The object's motion in the top view has the following characteristics: (1) The shape changes are relatively small compared to the side view, and (2) The motion states of the objects has good consistency between neighboring frames, which can be expressed as: the changes of position and direction are small for the same object, and the changes are large for different objects. According to these cues, we construct an association cost function for the objects based on the method proposed in [24].

Assume $F_{i,t}^{top} = (p_{i,t}^{top}, \theta_{i,t}^{top})$ is the head feature point of an arbitrary object i in the top view in frame t, $p_{i,t}^{top}$ and $\theta_{i,t}^{top}$ denote the position and direction of the feature point, respectively. The association cost function can be defined as:

$$cv(F_{j,t-1}^{top}, F_{i,t}^{top}) = \omega_1 \left(\frac{pc(p_{j,t-1}^{top}, p_{i,t}^{top})}{pc_{\max}} \right) + (1-\omega_1) \left(\frac{dc(\theta_{j,t-1}^{top}, \theta_{i,t}^{top})}{dc_{\max}} \right) \quad (4)$$

where $pc_{\max}$ and $dc_{\max}$ denote the maximum motion distance and the maximum deflection angle of the object in neighboring frames, respectively. $pc(p_{j,t-1}^{top}, p_{i,t}^{top})$ and $dc(\theta_{j,t-1}^{top}, \theta_{i,t}^{top})$ represent the position change and direction change between $F_{j,t-1}^{top}$ and $F_{i,t}^{top}$, respectively. ω_1 and 1-ω_1 stand for the weight of change of position and direction in metric function respectively, a setting that allows the dynamic adjustment of the importance of position and direction in association to bring itself in line with the motion state of objects in different views.

Suppose m objects in frame $t-1$ are to be associated with n objects in frame t. A matrix of $m \times n$ can be set up according to Equation (4) to express the function value of motion continuity between objects in two frames. Data association means the process of matching the objects with the most similar motion continuity iteratively searched in the matrix. The association model can be defined as:

$$z = \min \sum_{j_{t-1}=1}^{m} \sum_{i_t=1}^{n} cv(F_{j,t-1}^{top}, F_{i,t}^{top}) x_{j_{t-1} i_t}$$

$$s.t. \begin{cases} \sum_{j_{t-1}=1}^{m} x_{j_{t-1} i_t} = 1 & (i_t = 1, \ldots, n) \\ \sum_{i_t=1}^{n} x_{j_{t-1} i_t} = 1 & (j_{t-1} = 1, \ldots, m) \\ x_{j_{t-1} i_t} = 1 \text{ or } 0 & (j_{t-1} = 1, \ldots, m;\ i_t = 1, \ldots, n) \end{cases} \quad (5)$$

where $x_{j_{t-1} i_t} = 1$ means the feature point $F_{j,t-1}^{top}$ is associated with the feature point $F_{i,t}^{top}$; $x_{j_{t-1} i_t} = 0$ means the feature point $F_{j,t-1}^{top}$ is not associated with the feature point $F_{i,t}^{top}$.

The greedy algorithm can be used to solve this equation and to obtain the local optimal association of all objects. To improve performance, if the inter-object distance is larger than $pc_{\max}$, the association is abandoned. Fig. 5 shows the association process of feature points. If the state of the object $F_{i,t}^{top}$ in frame t coincides with that in frame t-1 in the association process, $F_{i,t}^{top}$ is not detected due to head occlusion, and the object is labeled as $O_{i,t}^{top}$.

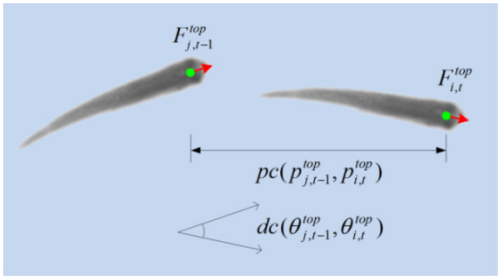

FIGURE 5. An illustration of the association cost function in the top view.

2) STEREO MATCHING

The purpose of stereo matching is to determine the third coordinate for each feature point in 2-D trajectories of top-view tracking. In many binocular vision-based stereo matching methods, the number of matched objects is typically reduced first via the epipolar constraint. Next, the object is confirmed based on appearance similarity. However, this strategy is not feasible in the proposed method because the large angle between cameras causes the shape of the object to vary greatly between any two views, making it impossible to match objects based on appearance similarity. In order to address this problem, we match feature points from the top view to side view subject to the epipolar constraint and motion consistency constraint.

a: EPIPOLAR CONSTRAINT

Assume $F_{q,t}^{side} = (p_{q,t}^{side}, \theta_{q,t}^{side})$ is the feature point of an arbitrary object q in the side view in frame t, and $l_{i,t}^{side}$ is the corresponding epipolar line of $F_{i,t}^{top}$ in the side view. The association probability is determined by the Euclidean distance from $p_{q,t}^{side}$ to $l_{i,t}^{side}$, and the result of the epipolar constraint can be expressed as:

$$ec(F_{i,t}^{top}, F_{q,t}^{side}) = \frac{d(p_{q,t}^{side}, l_{i,t}^{side})}{T_e} \quad (6)$$

where $d(p_{q,t}^{side}, l_{i,t}^{side})$ represents the Euclidean distance from $p_{q,t}^{side}$ to $l_{i,t}^{side}$, and T_e denotes the maximum matching distance under the epipolar constraint (see Fig. 6). If $ec(F_{i,t}^{top}, F_{q,t}^{side})$ is less than 1, the feature point q subjects to the epipolar constraint. If there is only one object under the epipolar constraint, stereo matching is performed successfully. The frame for which the epipolar constraint will suffice to complete the stereo matching is selected as the starting frame. If the stereo matching fails, the starting frame is processed using manual calibration. If there is more than one object, we use motion consistency constraint to reduce matching ambiguity.

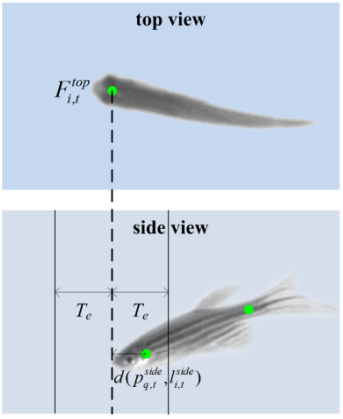

FIGURE 6. An Example of the epipolar constraint. The dotted line indicates the epipolar line.

b: MOTION CONSISTENCY CONSTRAINT

If there are k feature points of possible matching $(F_{1,t}^{side}, \ldots, F_{k,t}^{side})$ for $F_{i,t}^{top}$ in the side view subjected to the epipolar constraint. According to Equation (4), motion consistency constraint can be expressed as:

$$mc(F_{i,t}^{top}, F_{q,t}^{side}) = cv(F_{i,t-1}^{side}, F_{q,t}^{side}) \quad (q = 1, \ldots, k) \tag{7}$$

where $F_{i,t-1}^{side}$ denotes the matching of the feature point $F_{i,t-1}^{top}$ in frame t-1.

The result of stereo matching can be expressed as:

$$sm(F_{i,t}^{top}) = \arg\min_{q} \Big[\omega_2 ec(F_{i,t}^{top}, F_{q,t}^{side}) + (1-\omega_2) mc(F_{i,t}^{top}, F_{q,t}^{side})\Big] \tag{8}$$

where ω_2 and 1-ω_2 stand for the weight of the epipolar constraint and motion consistency constraint in stereo matching respectively. This equation indicates that if the feature point of it has the best matching value under the epipolar constraint and motion consistency constraint, then matching is performed successfully. Fig. 7 shows an example of stereo matching. If the moving direction of the object in the side view is perpendicular to the camera, there is only one feature point detected with the loss of direction information. Under such circumstances, the object could only be subject to stereo matching with top-view trajectory through position information, which may reduce the matching accuracy. As the time for the moving direction of object to stay perpendicular with camera is very short, the tracking performance is thus less affected.

3) OCCLUSION PROCESSING

For data association, the tracking performance is most significantly influenced by missed detections. Thus, reducing the missed detection rate is the most effective way to improve the tracking performance. In the single view, the missed detection of occluded objects is inevitable, independent from the detection method used. In order to solve this problem, the information of the side view is integrated to reduce the

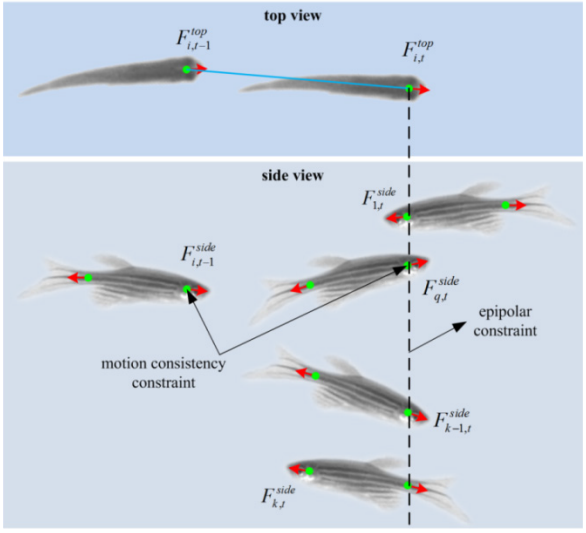

FIGURE 7. An Example of stereo matching. An object in the top view can find k candidates on corresponding epipolar line at frame t. The matching object is determined by the epipolar constraint and motion consistency constraint.

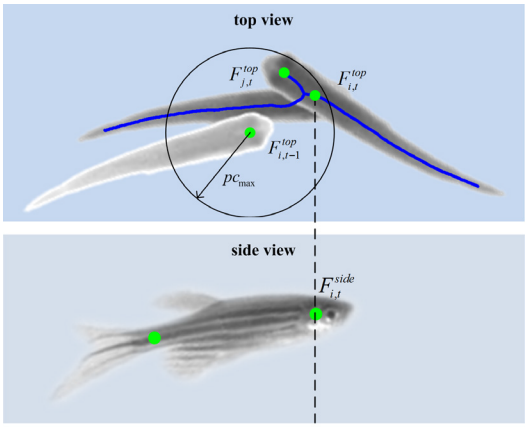

FIGURE 8. An example of stereo matching for the occluded object in the top view.

missed detection rate of the top view. Assuming object i in frame t is labeled as $O_{i,t}^{top}$ in the top view, the epipolar constraint is used to find the point for which the skeleton point in the region $pc_{\max}$ best matches with the skeleton endpoint in the side view.

$$oa(O_{i,t}^{top}) = \arg\min_{s} ec(s, F_{i,t}^{side}) \tag{9}$$

where s represents the skeleton point within the circle with $F_{i,t-1}^{top}$ as its center and $pc_{\max}$ as the radius in the top view (see Fig. 8).

If the objects are occluded under the side view, the top-view trajectories may fail in stereo matching. Either of the two methods can be adopted in this context: under the epipolar constraint, match the major skeleton of the occluded objects with the top-view objects; or, postpone the matching process until the occluded objects appear again. To improve the accuracy of the tracking trajectory, the former method is

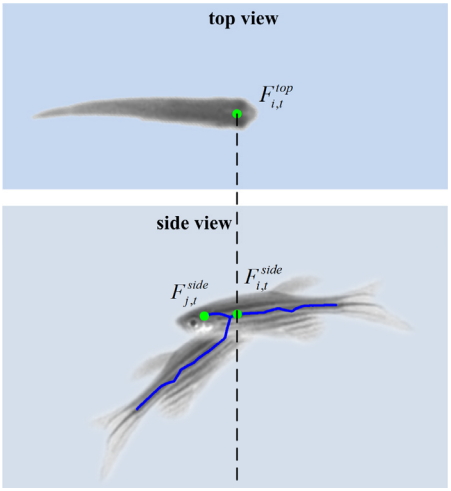

FIGURE 9. An example of stereo matching for the occluded object in the side view.

employed to complete the stereo matching of the occluded objects. In particular, when an object is occluded in the side view, the matching result on the main skeleton in the side view based on the top view trajectory is used as the detection result of the occluded object (see Fig. 9). Therefore, the problem of locating occluded objects in the side view can be solved, and the accuracy of the trajectory can be improved.

Although we have handled the occluded objects as much as possible, there are still a few occluded objects being omitted during the tracking process. Such omission may cause trajectory break. In order to guarantee trajectory continuity, the broken trajectories are connected to form complete trajectories. Note that trajectory connection is dependent on spatial-temporal relationships between trajectories. Consider trajectory T_i with an ending time $T_i(et)$ and ending position $T_i(ep)$; and trajectory T_j with a starting time $T_j(st)$ and starting position $T_j(sp)$. The following constraint is established:

$$tl(T_i, T_j) = \begin{cases} 1, & |T_i(ep) - T_j(sp)| < (pc_{\max} * f_o) \\ \& & T_i(et) < T_j(st) \\ 0, & otherwise \end{cases} \quad (10)$$

where f_o denotes the number of occluded frames between two trajectories. According to the above equation, the two trajectories satisfying the spatial and temporal constraints at the same time should be connected (see Fig. 10). If one trajectory has several qualified candidates to connect with, the candidate with the best continuity is selected based on Equation (4) for connection.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASETS

In order to evaluate the proposed method, we choose zebrafish as the tracking object. The length of zebrafish is 1-3 cm. They were placed in a 20×20×20 cm container with a water depth of about 18 cm. Two Flare 4M180-CL synchronized cameras were installed in the top-view and side-view directions. Video images were recorded at a rate

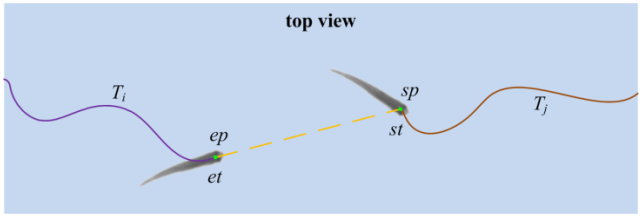

FIGURE 10. An example of trajectory connection. If two trajectories meet both the spatial and temporal constraints, they will be connected as a continuous trajectory.

of 90 fps and with a resolution of 2048×2040 pixels. Given the different object quantity, two groups video data (10 fish and 15 fish) are shot with each video length accounting to 5 minutes. 2000 consecutive frames with active objects in each of the video have been chosen as the testing datasets and named as D1 for 10 fish and D2 for 15 fish respectively.

B. PARAMETERS SETTING

Eight parameters need to be set for tracking, which are T_g, N, T_u, ω_1, ω_2, $pc_{\max}$, $dc_{\max}$ and T_e respectively. T_g mainly influences the integrity of object segmentation. Smaller T_g brings about more integral segmentation results and is more sensible to such interference as noise and water wave. As we simply need to obtain the main skeleton of the object and do not have high requirement for the integrity of segmentation results, it is reasonable to set a relatively large T_g. N affects the quality of the background image, it can be set in accordance with the motion state of objects in the datasets. ω_1 mainly reflects the importance of the change of position and direction in data association. With smaller object quantity, the change in position plays a dominant role in association. The increasing object quantity and constantly strengthening occlusion frequency heighten the role of change in direction in association. Similar to parameter ω_1, ω_2 mainly reflects the importance of the epipolar constraint and motion continuity constraint in stereo matching. The increase in object quantity is accompanied by the greater uncertainty of stereo matching under the epipolar constraint and larger weight of motion continuity constraint in stereo matching. T_e shows the error range of the epipolar constraint and allows dynamic adjustment according to the calibration error of the camera. $pc_{\max}$ and $dc_{\max}$ can be set in accordance with the motion state of object in adjacent frames of video sequence.

In order to set the detection and tracking parameters, we select 300 frames from each view as parameter samples. The tracking results obtained under different parameters are compared with the ground-truth generated by visual examination to determine the optimal setting. The final setting of parameters is shown in Table 1.

C. EVALUATION METRICS

The detection performance is evaluated using the following two metrics:

$$\begin{aligned} precision &= TP/(TP + FP) \\ recall &= TP/(TP + FN) \end{aligned} \quad (11)$$

TABLE 1. Parameters setting in the test process.

Test Set	View	Detection		Tracking					
		T_g	N	T_u	ω_1	ω_2	$pc_{\max}$	$dc_{\max}$	T_e
D1(10 fish)	Top view	35	500	50	0.6	0.5	40	180	-
	Side view	45	500	70	0.6	0.5	40	180	60
D2(15 fish)	Top view	30	800	55	0.5	0.4	50	180	-
	Side view	40	800	75	0.5	0.4	50	180	60

TABLE 2. Detection performance on the testing datasets.

Test Set	View	precision	recall
D1 (10 fish)	Top view	98.3%	86.8%
	Side view	96.5%	81.2%
D2 (15 fish)	Top view	97.9%	82.4%
	Side view	95.3%	74.3%

where *TP*, *FP*, *FN* denote the number of true positives, false positives and false negatives, respectively.

The tracking results are analyzed with the several widely used performance metrics [25].

(1) Mostly Tracked Trajectories (MT): The number of ground-truth trajectories which are successfully tracked for more than 80%.

(2) Mostly Lost Trajectories (ML): The number of ground-truth trajectories which are successfully tracked for less than 20%.

(3) Fragments (Frag): The number of times that a ground-truth trajectory is interrupted by the tracking trajectories.

(4) Identity Switches (IDS): The number of times that a tracking trajectory changes its matched ground-truth identity.

Lager value of MT indicates better tracking performance; smaller values of ML, Frag and IDS indicate better tracking performance.

D. RESULTS AND DISCUSSION

The detection results are shown in Table 2. It can be seen from this table that for the two groups of tests, the precision value of the proposed method is more than 95%. This indicates that the proposed detection method has higher accuracy. The recall value gradually decreases for an increasing number of objects. This is due to the fact that the frequency of head occlusions increases with an increase in the object density, which results in a gradual increase of the number of missed detection objects. In addition, the detection performance in the top view is higher than in the side view since fish appearance is more stable in the top view. As a result, the main skeleton is easier to extract in the top view. Also, since the area occupied by the fish body is larger in the side view than in the top view, motion occlusions are more likely to occur in the side view.

In order to evaluate the tracking performance of the proposed method more effectively, we compare it with other four methods, namely Wu *et al.* [12], idTracker [16], Wang *et al.* [20] and SK-3D [7]. The method of Wu et al. is a general framework for 3-D tracking. Many object tracking methods use Wu et al.'s method as a benchmark when conducting a performance analysis. idTracker is currently the best method to achieve 2-D tracking of fish based on artificial feature. The method of Wang et al. is one of the advanced methods for 3-D tracking of multiple fish at present. Experimental results are shown in Table 3.

TABLE 3. Tracking performance on our datasets.

Method	Test Set	MT	ML	Frag	IDS
Wu *et al.* [12]	D1 (10 fish)	2	5	12.1	7.6
	D2 (15 fish)	1	8	14.5	8.8
idTracker [16]	D1 (10 fish)	0	9	24.3	15.2
	D2 (15 fish)	0	15	29.1	17.3
Wang *et al.* [20]	D1 (10 fish)	4	1	6.4	4.2
	D2 (15 fish)	4	3	7.8	5.1
SK-3D [7]	D1 (10 fish)	6	1	2.3	1.4
	D2 (15 fish)	8	4	3.4	2.1
Proposed	D1 (10 fish)	**9**	**0**	**1.1**	**0.7**
	D2 (15 fish)	**12**	**1**	**1.7**	**1.2**

From ML, it can be observed that in the two groups of test, only one invalid trajectory is obtained by the proposed method, which demonstrates its great performance in object tracking. MT suggests that the proposed method is able to obtain the trajectories of most objects, but the mean value decreases with the number of objects. The reason for such decrease is the increased frequency of object occlusion that causes trajectory error. Frag and IDS show the impressive ability of the proposed method to track occluded objects and ensure trajectory integrity by accurately finding matches of most occluded objects. The tracking results are shown in Fig. 11.

The method proposed by Wu et al. can track many objects which fly very fast. In their method, each object is simplified into a point. The multi-view information is fused for data association and stereo matching. However, the appearance of objects varies radically under different views and the objects move very randomly. Representing the object with a single point makes it more difficult to perform stereo matching and occluded tracking. As a result, the algorithm's tracking performance is affected. Furthermore, the moving direction of objects is not estimated. Location information is thus insufficient for accurate data association.

idTracker begins with obtaining fish fingerprints by analyzing object appearances. Then, it matches these fingerprints in different frames and determines their trajectories. In our experiment, the objects are tracked simultaneously from the top- and side-view directions. The tracking results of the two views of idtracker are analyzed. In the starting frame, the relationship between the object in the top view and the object in the side view is established using manual calibration. During the tracking process, the tracking result of the top view provides (x, y) coordinates, while that of the side view provides z coordinates. In this way the 3-D position of the

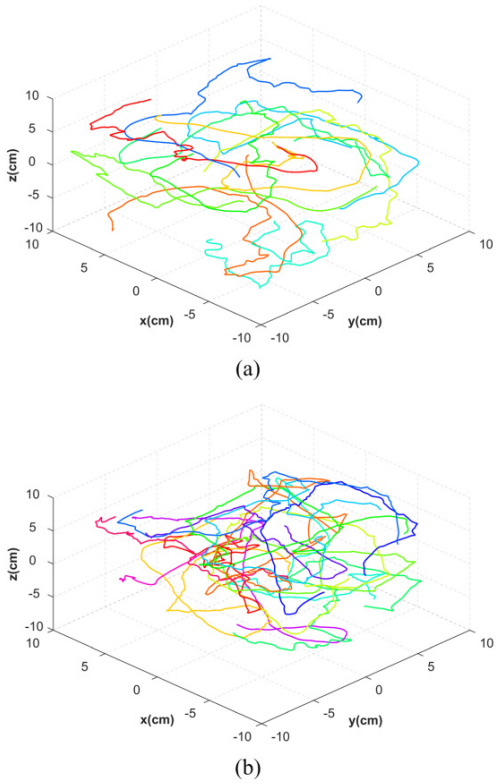

FIGURE 11. The obtained trajectories on different groups. (a) 10 fish. (b) 15 fish.

object is obtained. In the trajectory analysis, an error in the 2-D trajectory of a certain view, is bound to result in errors in the corresponding 3-D trajectory as well. Experimental results indicate that idTracker does not perform effectively, due to appearance variation of objects in the 3-D space. It is known that, variation in locations and moving directions of the object may cause variation in appearance of the object. This is especially true along the side-view direction. Appearance variation makes it harder to match fingerprints. Due to these reasons, idTracker leads to many mismatches. It is also observed from the experiment that, considering variability of the object's appearance, kinetics-based tracking is superior to appearance analysis-based tracking in the 3-D space.

Wang et al. determined the location of fisheye under the top and side views using mixed Gaussian model and Gabor model respectively. Next, the 3-D motion trajectories of the objects are obtained by associating top-view tracking results with the trajectories of two side views. In their method, multi-view information was fully exploited to avoid the problem of object omission that may arise from single-view tracking. In other words, the objects that are not detected under certain view will assuredly be detected under other views. It can enhance detection performance and thus improve the accuracy of stereo matching, an advantage that is directly related to shooting conditions and training samples. In the testing data, the low distinction degree of the eye area characteristics of the object determines the poor detection performance of their method. In addition, without using the motion direction of

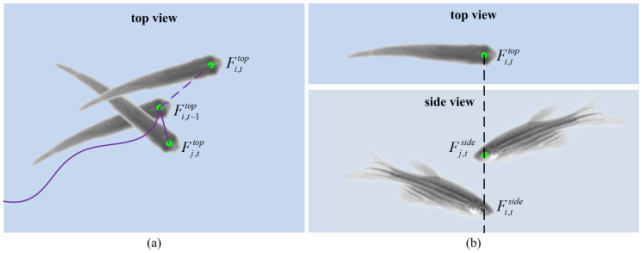

FIGURE 12. Tracking errors of Wang et al. (a) Association error caused by not considering moving direction. (b) Stereo matching error caused by misdetection.

the object during tracking, their method reduces the accuracy of association and stereo matching (see Fig. 12). Finally, this method involves capturing and analyzing objects under three views, resulting in complicated equipment installation, configuration and data association. Compared to the scheme by Wang et al., the proposed method is able to detect objects more robustly. By exploiting their location and direction, the proposed method addresses frequent occlusions of objects more effectively. Moreover, 3-D tracking is enabled in the proposed method by tracking objects under two views that greatly simplifies the tracking process.

Compared with SK-3D, the improved one employs the top view and the side view to locate the occluded objects in a complementary way, which better solves the most difficult problem of motion occlusion in multi-object tracking and further improves the tracking performance. In the process of object motion, the likelihood that occlusion occurs in both views is small due to the mechanism of collision avoidance between the objects. Therefore, it is feasible to combine the information of two views to solve the occlusion problem. In addition, since the trajectory connection is used to process the trajectory fragments, the tracking result of the proposed method has better continuity.

For better analysis of the performance of the proposed method, we choose the datasets of Wang *et al.* [20] for test. For the datasets the zebrafish is placed in a container of $15\times15\times15$ cm and a water depth of 10 cm. The image resolution is 2048×2040 pixels, and the frame rate is 90 fps. According to the number of fish, two groups, namely D3 (5 fish) and D4 (10 fish), are selected from the dataset as test data. Each group contains 1000 frames. Compared with our datasets, Wang et al.'s datasets have higher occlusion frequency and greater tracking difficulty. The test results are presented in Table 4, from which, it can be seen that the proposed method has achieved the best tracking performance in the two groups of test. In contrast, Wu et al. and idTracker are still low-performing in the datasets, and Wang et al. has its tracking performance enhanced to some extent. The reason for such results is that the distinction degree of the eye area characteristics of the object in the datasets is higher than that of ours, thus resulting in the certain improvement of the detection performance of theirs. However, the failure to take into account the moving direction of the object undermines its accuracy in data association and stereo matching.

TABLE 4. Tracking performance on Wang et al. 's datasets.

Method	Test Set	MT	ML	Frag	IDS
Wu *et al.* [12]	D3 (5 fish)	3	0	4.2	2.4
	D4 (10 fish)	4	2	7.5	3.6
idTracker [16]	D3 (5 fish)	1	2	6.1	2.8
	D4 (10 fish)	0	5	13.4	7.6
Wang *et al.* [20]	D3 (5 fish)	5	0	1.3	0.9
	D4 (10 fish)	8	1	4.8	1.1
SK-3D [7]	D3 (5 fish)	4	1	1.6	1.2
	D4 (10 fish)	6	3	6.4	2.3
Proposed	D3 (5 fish)	**5**	**0**	**0.9**	**0.3**
	D4 (10 fish)	**8**	**0**	**1.4**	**0.9**

Further analysis of the experimental results reveal that the tracking errors of the proposed method concentrate more in top-view tracking than in stereo matching. In other words, any error found in top-view trajectory will directly affect the 3-D tracking results after stereo matching. With correct top-view trajectory but erroneous side-view detection results, only the 3-D trajectory in the current frame or in a few frames will be influenced. Therefore, the top-view tracking results have greater impact on the tracking performance. When the density of objects is low, the proposed method can track objects effectively along the top-view direction. Object occlusions happen more frequently when the number of objects increases. Location and moving direction alone are no longer sufficient to guarantee accurate matching of occluded objects. As a result, more matching errors occur and the method's tracking performance deteriorates. How to benefit top-view tracking from more side-view information is an important topic that will be studied in the future.

V. CONCLUSION

A skeleton-based multiple fish 3-D tracking method is proposed in this paper. Experimental results show that the proposed method, with sound tracking performance, can acquire the 3-D motion trajectories of multiple fish. Since the proposed detection method is based on the asymmetric strip structure of the object, it has certain limitations. For example, the method may not provide correct results, if the fish body width is larger than or equal to body length, or fish with complex fin and fishtail shape. Therefore, we shift our research focus towards establishing a universal detection method that is using more appearance features. In addition, with the development of computer vision technology, deep learning has shown a strong ability of automatic feature extraction, and provides a new solution for multi-object tracking [26], [27]. Next, further improving the tracking performance through machine learning is another focus of our research.

REFERENCES

[1] N. C. Makris, P. Ratilal, D. T. Symonds, S. Jagannathan, S. Lee, and R. W. Nero, "Fish population and behavior revealed by instantaneous continental shelf-scale imaging," *Science*, vol. 311, no. 5761, pp. 660–663, 2006.
[2] N. C. Makris, P. Ratilal, S. Jagannathan, Z. Gong, M. Andrews, M. Bertsatos, I. Bertsatos, O. R. Godø, R. W. Nero, and J. M. Jech, "Critical population density triggers rapid formation of vast oceanic fish shoals," *Science*, vol. 323, no. 5922, pp. 1734–1737, 2009.
[3] A. I. Dell, J. A. Bender, K. Branson, I. D. Couzin, G. G. de Polavieja, L. P. J. J. Noldus, A. Pérez-Escudero, P. Perona, A. D. Straw, M. Wikelski, and U. Brose, "Automated image-based tracking and its application in ecology," *Trends Ecol. Evol.*, vol. 29, no. 7, pp. 417–428, 2014.
[4] P. R. Martineau and P. Mourrain, "Tracking zebrafish larvae in group–status and perspectives," *Methods*, vol. 62, no. 3, pp. 292–303, 2013.
[5] J. Delcourt, M. Denoël, M. Ylieff, and P. Poncin, "Video multitracking of fish behaviour: A synthesis and future perspectives," *Fish Fisheries*, vol. 14, no. 2, pp. 186–204, Jun. 2013.
[6] Z.-M. Qian and Y. Q. Chen, "Feature point based 3D tracking of multiple fish from multi-view images," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0180254.
[7] Z. Qian, M. Shi, M. Wang, and T. Cun, "Skeleton-based 3D tracking of multiple fish from two orthogonal views," in *Proc. CCF Chin. Conf. Comput. Vis.*, 2017, pp. 25–36.
[8] N. F. Hughes and L. H. Kelly, "New techniques for 3-D video tracking of fish swimming movements in still or flowing water," *Can. J. Fisheries Aquatic Sci.*, vol. 53, no. 11, pp. 2473–2483, 1996.
[9] R. A. Goodwin, J. M. Nestler, J. J. Anderson, L. J. Weber, and D. P. Loucks, "Forecasting 3-D fish movement behavior using a Eulerian–Lagrangian–agent method (ELAM)," *Ecol. Modell.*, vol. 192, nos. 1–2, pp. 197–223, 2006.
[10] 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. 106–119, 2007.
[11] 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. 486–492, 2008.
[12] Z. Wu, N. I. Hristov, T. L. Hedrick, T. H. Kunz, and M. Betke, "Tracking a large number of objects from multiple views," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1546–1553.
[13] G. Wu, "Measuring the three-dimensional kinematics of a free-swimming koi carp by video tracking method," *J. Bionic Eng.*, vol. 7, no. 1, pp. 49–55, 2010.
[14] S. Butail and D. A. Paley, "Three-dimensional reconstruction of the fast-start swimming kinematics of densely schooling fish," *J. Roy. Soc. Interface*, vol. 9, no. 66, pp. 77–88, 2011.
[15] H. Maaswinkel, L. Zhu, and W. Weng, "Using an automated 3D-tracking system to record individual and shoals of adult zebrafish," *J. Visualized Exp.*, vol. 82, Dec. 2013, Art. no. e50681.
[16] A. Pérez–Escudero, J. Vicente-Page, R. C. Hinz, S. Arganda, and G. G. De Polavieja, "idTracker: Tracking individuals in a group by automatic identification of unmarked animals," *Nature Methods*, vol. 11, no. 7, pp. 743–748, 2014.
[17] A. Pautsina, P. Císař, D. Štys, B. F. Terjesen, and Å. M. O. Espmark, "Infrared reflection system for indoor 3D tracking of fish," *Aquacultural Eng.*, vol. 69, pp. 7–17, Nov. 2015.
[18] C. J. Voesenek, R. P. M. Pieters, and J. L. van Leeuwen, "Automated reconstruction of three-dimensional fish motion, forces, and torques," *PLoS ONE*, vol. 11, no. 1, 2016, Art. no. e0146682.
[19] M. M. Saberioon and P. Cisar, "Automated multiple fish tracking in three-dimension using a structured light sensor," *Comput. Electron. Agricult.*, vol. 121, pp. 215–221, Feb. 2016.
[20] S. H. Wang, X. Liu, J. Zhao, Y. Liu, and Y. Q. Chen, "3D tracking swimming fish school using a master view tracking first strategy," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 516–519.
[21] O. K.-C. Au, C.-L. Tai, H.-K. Chu, D. Cohen-Or, and T.-Y. Lee, "Skeleton extraction by mesh contraction," in *ACM Trans. Graph.*, vol. 27, no. 3, Aug. 2008, Art. no. 44.
[22] X. Bai, L. J. Latecki, and W.-Y. Liu, "Skeleton pruning by contour partitioning with discrete curve evolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 449–462, Mar. 2007.
[23] A. Telea and J. J. van Wijk, "An augmented fast marching method for computing skeletons and centerlines," in *Proc. VISSYM*, May 2002, pp. 251–258.
[24] Z.-M. Qian, S. H. Wang, X. E. Cheng, and Y. Q. Chen, "An effective and robust method for tracking multiple fish in video image based on fish head detection," *BMC Bioinf.*, vol. 17, no. 1, 2016, Art. no. 251.

[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, Nov. 2007.
[26] F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. H. Heras, and G. G. De Polavieja, "idtracker.ai: Tracking all individuals in small or large collectives of unmarked animals," *Nature Methods*, vol. 16, no. 2, pp. 179–182, 2019.
[27] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

XIAOQING LIU received the M.S. degree in computer science and technology from the Kunming University of Science and Technology, Kunming, China, in 2010. She is currently an Associate Professor with Chuxiong Normal University, Chuxiong, China. Her research interests include computer vision and signal processing.

YINGYING YUE received the M.S. degree in computer system architecture from Yunnan University, Kunming, China, in 2010. She is currently a Lecturer with Yuxi Normal University, Yuxi, China. Her research interests include computer vision and machine learning.

MEILING SHI received the Ph.D. degree in computer software and theory from Yunnan University, Kunming, China, in 2011. She is currently an Associate Professor with Qujing Normal University, Qujing, China. Her research interests include computer vision and pattern recognition.

ZHI-MING QIAN received the Ph.D. degree in computer application technology from Fudan University, Shanghai, China, in 2018. He is currently a Professor with Chuxiong Normal University, Chuxiong, China. His research interests include computer vision and pattern recognition.