# Skeleton-Based 3D Tracking of Multiple Fish From Two Orthogonal Views

Zhiming Qian[1](✉), Meiling Shi[2], Meijiao Wang[1], and Tianrui Cun[1]

[1] Chuxiong Normal University, Chuxiong 675000, China
qzhiming@126.com, {meijiaow,cuntianrui}@cxtc.edu.cn
[2] Qujing Normal University, Qujing 655011, China
mlshi@mail.ynu.edu.cn

**Abstract.** This paper proposes a skeleton-based method for tracking multiple fish in 3D space. First, skeleton analysis is performed to simplify object into feature point representation according to shape characteristics of fish. Next, based on the obtained feature points, object association and matching are achieved to acquire the motion trajectories of fish in 3D space. This process relies on top-view tracking that is supplemented by side-view detection. While fully exploiting the shape and motion information of fish, the proposed method is able to solve the problems of frequent occlusions that occur during the tracking process and has good tracking performance.

**Keywords:** 3D tracking · Fish tracking · Feature point
Stereo matching

## 1 Introduction

Animal behavior analysis has been a major research area in the natural sciences. As an important member of aquatic animals, fish occupy an essential position in the animal population. Fish ethology is of great significance to understanding animal behavior evolution and fisheries development, and thus is an active branch of study of animal behavior analysis [2,3,5].

Fish behavior is typically recorded as videos in the laboratory environment. Quantitative analysis of fish behavior using video images cannot be performed without acquiring the fish trajectory from the images. Traditionally, these trajectories would be obtained manually from each image frame. This method is inefficient and inaccurate. Due to developments in computer science, computer vision technology provides an effective approach for acquisition and quantitative representation of fish behavior.

Depending on the environment, video-based fish tracking can be classified into either 2D or 3D tracking. In 2D tracking, fish are confined to a container with shallow water. Thus, fish motion can be approximated as planar motion. Despite its ability to analyze fish behavior, this type of tracking can only be performed in 2D space and is unable to completely represent fish behavior. 3D

J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 25–36, 2017.
https://doi.org/10.1007/978-981-10-7299-4_3

tracking has drawn a lot of attention from academia because it is capable of providing fish trajectories that represent true fish behavior by simulating the motion pattern of fish in a real-world environment.

Many researchers had proposed some effective methods for 3D fish tracking. Zhu et al. [12] designed a catadioptric stereo-vision system for obtaining the motion trajectories of several fish by using a video camera and two planar mirrors. Nimkerdphol and Nakagawa [6] proposed a method of the 3D coordinates of zebrafish (*Danio rerio*) by using two regular digital video camcorders placed nonplanar to an object. Butail and Paley [1] proposed a 3D tracking system for tracking the full-body trajectories of schooling fish using multiple cameras. Maaswinkel et al. [4] presented an automatic video tracking system which can record the swimming trajectories of single and groups of zebrafish by using a mirror system. Voesenek et al. [10] presented a method that can track the fishs body position, curvature and direction in 3D space from multiple viewpoints.

Although existing tracking methods are able to obtain 3D motion trajectories of fish in certain environmental scenarios, there are some problems that are not completely solved, such as large appearance changes of objects in different views and frequent occlusions among different objects. Therefore, it is still a challenge to track multiple fish in 3D space. In this paper, a skeleton-based 3D tracking method is proposed to acquire motion trajectories of multiple fish from two orthogonal views. The contributions of this paper are as follows:

(1) Propose an effective strategy to reconstruct 3D trajectories of multiple fish by using a combination of top-view tracking and side-view detection.
(2) Propose a feature point model for representing objects in different views, which enhances detection performance and reduces tracking difficulty.
(3) Propose a novel stereo matching method that combines the epipolar constraint and motion consistency constraint in order to reduce matching ambiguity.

## 2 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 2D tracking results are acquired. Finally, 3D motion trajectories are reconstructed by matching top-view tracking with features points in the side view. Figure 1 shows the flow chart of the proposed tracking method.

### 2.1 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 2D region to the 1D curve. Next, the feature points which consist of the skeleton endpoints and motion direction are obtained to represent the object.

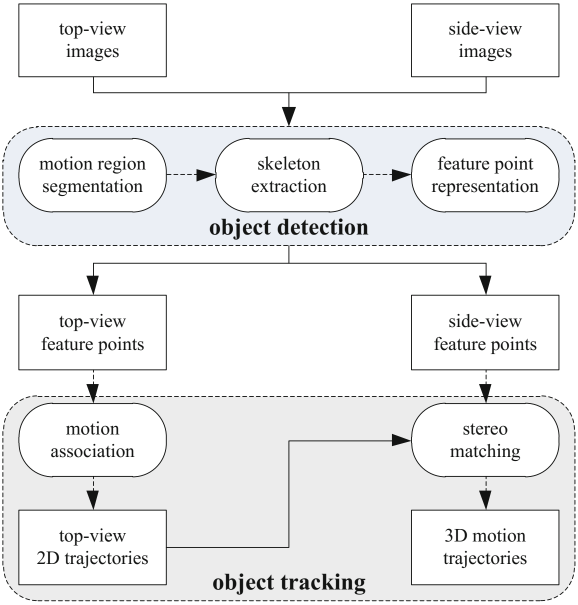


**Fig. 1.** Flow chart of the proposed method.

**Moving Region Segmentation.** In the laboratory environment, the video images usually consist of moving objects and nearly static background, and each 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.

$$R_t = \{(x,y) \in I| \quad |\underset{(I_1,\ldots,I_n)}{median}(x,y) - I_t(x,y)| > T_g\}\ . \tag{1}$$

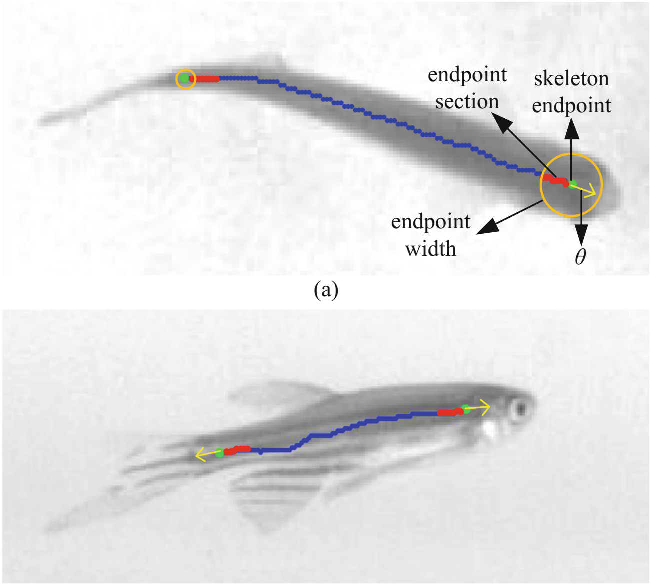
where $I_t(x,y)$ denotes the $t$-th frame image, and $median(x,y)$ represents the background image which consists of the median image of the first $n$ frames in each view. In order to minimize interference from the moving region's coarse edge during subsequent extraction of the skeleton, we first perform 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.

**Skeleton Extraction.** The skeleton of each moving region is extracted via the augmented fast marching method (AFMM) [9]. The AFMM is a fast and efficient skeleton extraction method. 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:

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

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, the skeleton represents more detail when the value of threshold $T_u$ is smaller. In this paper, the threshold is set to a large value for the purpose of extracting the main skeleton and eliminating interference from small branches in the skeleton.

**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.



(a)

(b)

**Fig. 2.** Feature points in different views. (a) Top view. (b) Side view.

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

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

The position ($p$) of skeleton endpoint $se$ and the direction ($\theta$) are combined into a feature point $F(p, \theta)$, where the feature point is used to represent the object, as shown in Fig. 2. 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 (Fig. 3), thereby greatly improving the ability of occlusion tracking.

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. 2(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.

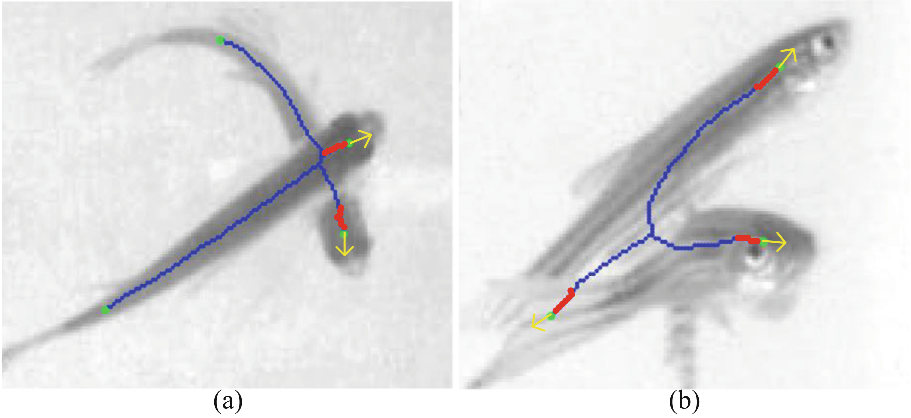


**Fig. 3.** Feature point representation when an occlusion exists between objects. (a) Top view. (b) Side view.

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.

### 2.2 Objects Tracking

**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 [8].

Assume $F_{i,t}^{top} = (p_{i,t}^{top}, \theta_{i,t}^{top})$ is the head feature point of an arbitrary object $i_t$ 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:

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

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. The association model can be defined as:

$$\begin{aligned} & z = \sum_{j_{t-1}=1}^{m} \sum_{i_t=1}^{n} cv(F_{j,t-1}^{top}, F_{i,t}^{top}) x_{j_{t-1}i_t} \\ & s.t. \begin{cases} \sum\limits_{j_{t-1}=1}^{m} x_{j_{t-1}i_t} = 1 \quad (i_t = 1, \ldots, n) \\ \sum\limits_{i_t=1}^{n} x_{j_{t-1}i_t} = 1 \quad (j_{t-1} = 1, \ldots, m) \\ x_{j_{t-1}i_t} = 1 \text{ or } 0 \quad (j_{t-1} = 1, \ldots, m;\ i_t = 1, \ldots, n) \end{cases} \end{aligned} . \tag{5}$$

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

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. Figure 4 shows the association process of feature points.

**Stereo Matching.** The purpose of stereo matching is to determine the third coordinate for each feature point in 2D 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.

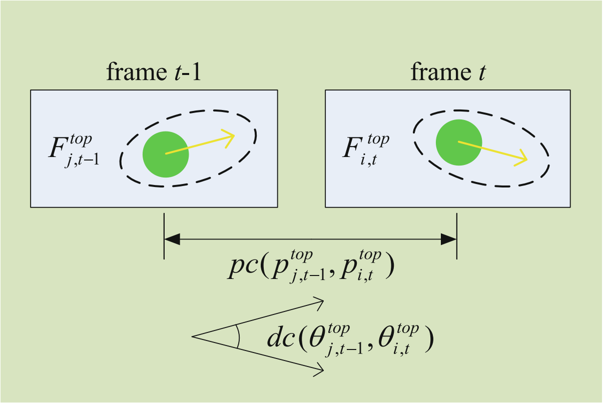

frame $t$-1
frame $t$
$F^{top}_{j,t-1}$
$F^{top}_{i,t}$
$pc(p^{top}_{j,t-1}, p^{top}_{i,t})$
$dc(\theta^{top}_{j,t-1}, \theta^{top}_{i,t})$

**Fig. 4.** Illustration of motion association in the top view.

(1) Epipolar constraint

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

$$ec(F^{top}_{i,t}, F^{side}_{q,t}) = \begin{cases} 1, & d(p^{side}_{q,t}, l^{side}_{i,t}) < T_e \\ 0, & otherwise \end{cases} . \tag{6}$$

where $T_e$ denotes the maximum matching distance under the epipolar constraint. If there is only one object under the epipolar constraint, the matching is performed successfully. If there is more than one object, we use motion consistency constraint to reduce matching ambiguity.

(2) Motion consistency constraint

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

$$mc(F^{top}_{i,t}) = \arg\min_q cv(F^{side}_{i,t-1}, F^{side}_{q,t}) \quad (q = 1, \ldots, k). \tag{7}$$

where $F^{side}_{i,t-1}$ denotes the matching of the feature point $F^{top}_{i,t-1}$ in frame $t-1$. This equation indicates that if the feature point of it achieves maximum motion consistency with that of the previous matching object in the side view, then matching is performed successfully. Figure 5 shows an example of stereo matching.

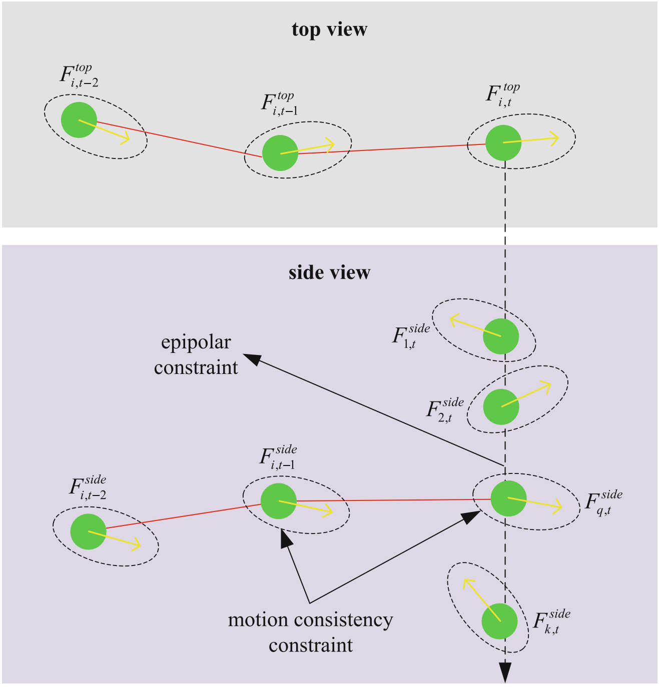


**Fig. 5.** Example of stereo matching. The dashed arrow indicates the epipolar line. An object in the top view can find $k$ candidates on corresponding epipolar line at frame $t$. The matching object is determined by the epipolar constraint and motion consistency constraint.

## 3 Experimental Results and Discussion

### 3.1 Data Sets

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 $20 \times 20 \times 20$ 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 of 90 fps and with a resolution of $2048 \times 2040$ pixels. According to the difference in the number of fish, we divide the test data into two sets: D1 for 10 fish and D2 for 15 fish. Each data set contains 1,000 consecutive frames.

### 3.2 Parameters Setting

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.

**Table 1.** Parameters setting in the test process.

| View | Detection | | Tracking | | | |
|---|---|---|---|---|---|---|
| | $T_g$ | $T_u$ | $\omega$ | $pc_{max}$ | $dc_{max}$ | $T_e$ |
| Top view | 30 | 50 | 0.6 | 40 | 180 | – |
| Side view | 40 | 70 | 0.6 | 40 | 180 | 20 |

### 3.3 Evaluation Metrics

We analyze the tracking results with the following four performance metrics.

Trajectory completeness (TC): Defined as the ratio of the number of tracked objects to the number of ground-truth objects. The higher the value of this metric, the more complete the tracked trajectories.

Trajectory fragmentations (TF): Defined as the ratio of the number of obtained trajectories to the number of actual trajectories. The larger the value of this metric, the more frequently the trajectories are broken.

Mostly tracked trajectories (MT): Defined as the number of tracked trajectories which are correctly tracked for more than 80% of ground-truth trajectories. The larger the value of this metric, the greater the number of effective tracked trajectories.

Identity switches (IDS): Defined as the number of object identity switches. The smaller the value of this metric, the stronger the ability of the method to handle occlusions.

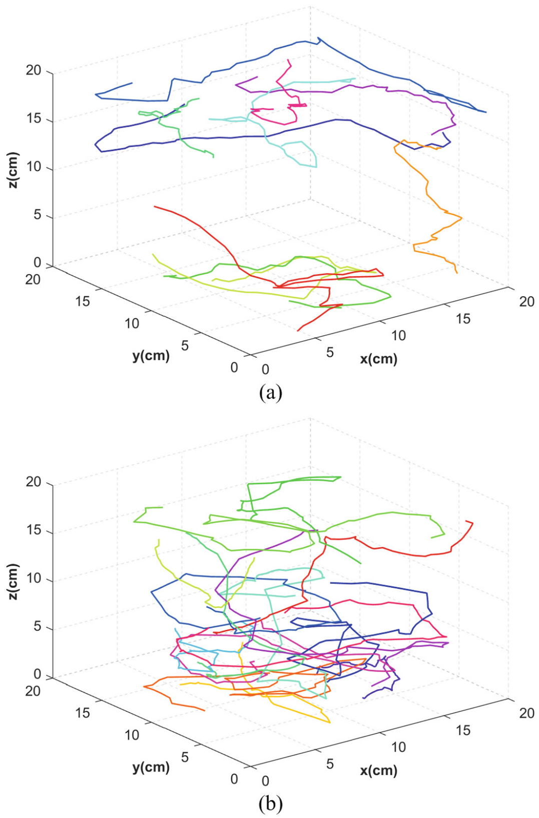
### 3.4 Results and Discussion

In order to evaluate the performance of the proposed method more effectively, we compare it with other two state-of-the-art methods, namely Wang et al.'s [11] and idTracker [7]. Experimental results are shown in Table 2.

From TC, it can be observed that the proposed method is able to produce very complete tracking results (over 94%) in the two tests and only 5.2% of the tracking information is lost at most. It can also be seen from TF and IDS that the proposed method can process occlusion effectively, because it is able to track the object accurately before and after occlusion occurs. There are at most twelve trajectory fragmentations and fourteen identity switches in the two tests, which means that track continuity and reliability are guaranteed. Finally, MT indicates that the proposed method can successfully obtain most of the objects trajectories. Only three trajectories are shorter than 80% of the lengths of ground-truth trajectories due to association and stereo matching errors. These results fully demonstrate feasibility and effectiveness of the proposed method. Acquired trajectories using the proposed method are shown in Fig. 6.

**Table 2.** Tracking performance on different test sets.

| Method | Test Set | TC | TF | MT | IDS |
|---|---|---|---|---|---|
| Wang *et al.* | D1 (10 fish) | 0.817 | 6.8 | 2 | 26 |
| | D2 (15 fish) | 0.764 | 11.6 | 3 | 54 |
| idTracker | D1 (10 fish) | 0.715 | 13.4 | 1 | 82 |
| | D2 (15 fish) | 0.536 | 28.3 | 0 | 142 |
| Proposed | D1 (10 fish) | **0.969** | **1.4** | **9** | **6** |
| | D2 (15 fish) | **0.948** | **1.8** | **13** | **14** |



(a)

(b)

**Fig. 6.** Tracking results on different groups. (a) 10 fish. (b) 15 fish.

Wang et al. first detected objects in the top view and side views by using the eye-focused Gaussian mixture model and the eye-focused Gabor model, respectively. Then, tracking results were obtained by associating top-view tracking and side-view detection. Their method performs well under good lighting conditions. However, in the experiments, detection performance of their method is poor because of normal lighting conditions. This leads to reduced tracking performance. Compared with the method proposed by Wang et al., the detection performance of the proposed method is more stable, and can adapt to a variety of lighting conditions. In addition, the proposed method uses both position and direction information to associate and match objects, which can improve tracking performance.

idTracker is one of the best methods for multi-fish tracking. It obtains motion trajectories by performing feature matching on all objects across image frames based on appearance analysis. In the experiments, we track objects in images from the top view and side view using idTracker at the same time. Then, tracking results from the two directions are fused to acquire trajectories in 3D space. Experimental results show that idTracker cannot track the object effectively due to object appearance variations. While the object is swimming in 3D space, its appearance in the image varies with the focal length of the camera. Especially in the side-view direction, the shooting distance causes variation in object appearance, and variation of the object in the swimming direction also leads to considerable changes in its appearance across neighboring frames. As we know, idTrackers tracking performance is heavily dependent on its remarkable appearance recognition ability. Instability of object appearance in the experiment thus makes it more challenging to match appearance features. And idTracker produces many mismatches due to this reason. The experiments also showed that, due to the complexity in the object's appearance, the motion-based method is superior to the appearance-based method for 3D tracking.

## 4 Conclusion

A skeleton-based multiple fish 3D tracking method is proposed in this paper. The contributions of this paper are as follows. First, a feature point model is constructed based on the shape characteristics of fish to represent objects in multi-view images. The constructed model enhances detection performance and reduces tracking difficulty. Second, motion consistency constraint is performed to eliminate uncertainty in stereo matching and improve the accuracy of object tracking. Third, the proposed method provides an effective strategy to address object occlusions by using top-view tracking, where object shapes change slightly, supplemented by side-view detection.

**Acknowledgments.** The research work presented in this paper was supported by National Natural Science Foundation of China (No. 61363023), and the Program for Innovative Research Team (in Science and Technology) in University of Yunnan Province.

## References

1. Butail, S., Paley, D.A.: 3d reconstruction of the fast-start swimming kinematics of densely schooling fish. J. R. Soc. Interface **9**(66), 77–88 (2012)
2. Delcourt, J., Denoël, M., Ylieff, M., Poncin, P.: Video multitracking of fish behaviour: a synthesis and future perspectives. Fish Fisheries **14**(2), 186–204 (2013)
3. Dell, A.I., Bender, J.A., Branson, K., Couzin, I.D., et al.: Automated image-based tracking and its application in ecology. Trends Ecol. Evol. **29**(7), 417–428 (2014)
4. Maaswinkel, H., Zhu, L., Weng, W.: Using an automated 3d-tracking system to record individual and shoals of adult zebrafish. J. Vis. Exp. **82**(82), e50681 (2013)
5. Martineau, P.R., Mourrain, P.: Tracking zebrafish larvae in group-status and perspectives. Methods **62**(3), 292–303 (2013)
6. Nimkerdphol, K., Nakagawa, M.: Effect of sodium hypochlorite on zebrafish swimming behavior estimated by fractal dimension analysis. J. Biosci. Bioeng. **105**(5), 486–492 (2008)
7. Pérez-Escudero, A., Vicente-Page, J., Hinz, R.C., Arganda, S., de Polavieja, G.G.: idtracker: tracking individuals in a group by automatic identification of unmarked animals. Nature Methods **11**(7), 743–748 (2014)
8. Qian, Z.M., Wang, S.H., Cheng, X.E., Chen, Y.Q.: An effective and robust method for tracking multiple fish in video image based on fish head detection. BMC Bioinform. **17**(1), 251–263 (2016)
9. Telea, A., Wijk, J.J.V.: An augmented fast marching method for computing skeletons and centerlines. In: Vissym Proceedings of the Symposium on Data Visualisation, pp. 251–259 (2002)
10. Voesenek, C.J., Pieters, R.P.M., Leeuwen, J.L.V.: Automated reconstruction of three-dimensional fish motion, forces, and torques. PloS One **11**(1), e0146682 (2016)
11. Wang, S.H., Liu, X., Liu, Y., Zhao, J., Chen, Y.Q.: 3d tracking swimming fish school using a master view tracking first strategy. In: IEEE International Conference on BIBM, pp. 516–519 (2016)
12. Zhu, L., Weng, W.: Catadioptric stereo-vision system for the real-time monitoring of 3d behavior in aquatic animals. Physiol. Behav. **91**(1), 106–119 (2007)