A Multi-Camera System for Underwater Real-Time 3D Fish Detection and Tracking

Rodrigo Verschae, Hiroaki Kawashima, and Shohei Nobuhara
Graduate Schoof of Informatics, Kyoto University Yoshida-Honmachi, 606-8501 Sakyo-ku Kyoto, Japan
Emails: rodrigo@verschae.org, hiroaki@vision.kuee.kyoto-u.ac.jp, nob@vision.kuee.kyoto-u.ac.jp

***Abstract*—Analyzing fish and fish schools behavior can help in studying fish-fish interaction, analyzing characteristics of fish species, studying prey avoidance maneuvers of fish schools, etc. Such analysis requires the estimation of each fish's 3D location, 3D pose, and 3D shape over time. Moreover if we are interested in studying the interaction of fish by injecting visual / acoustic stimuli artificially according to their motions, this information is required in real-time. In this context, our goal is to track in 3D each fish location and pose, accurately, in real-time, and in the future we foresee the use of underwater vehicles with multiple cameras for 3D fish school behavior analysis. As a step in this direction, in the current paper we propose the use of a calibrated multi-camera system, where each camera captures images through a flat surface, and the cameras observe a common region from different point of views (through one or more flat surfaces). The proposed system allows to detect and track in 3D each fish location in real-time, while taking into account light refraction though flat surfaces. We test the proposed approach using a fish tank with flat surfaces and present validation results and obtained processing times.**

I. Introduction

Acquiring accurate information of fish and fish school collective behavior, analyzing their behavior, and in particular estimating each fish's pose, shape and trajectory [1][2][3][4] can help engineers and fish biologists in *i*) studying fish behavior when swimming in regions of interest [1], *ii*) studying fish-fish interaction [2], *iii*) analyzing characteristics of fish species (e.g swimming ability) [3], *iv*) studying prey avoidance maneuvers of fish schools [4], among many others. Such analysis requires the estimation of 3D information of each fish: 3D location, 3D pose, and 3D shape needs to be estimated over time [5]. Moreover if we are interested in studying the *interaction* of fish by injecting visual / acoustic stimuli artificially according to their motions, this information is required in real-time.

When doing such analysis there are various additional challenges, including: *i*) a fish is a deformable object, *ii*) a fish can make sudden changes in direction & speed, and *iii*) a fish can be seen from multiple views and under occlusions.

While fish detection and tracking has been studied [4][6][7][8] most systems are designed in scenarios: *i*) with a small number of fish, *ii*) running off-line, or *iii*) where fish move in shallow waters (i.e. each fish moves in a 2D plane). Moreover, light refraction in glass and water is usually not taken into account.

In this context, our long term goal is to track in 3D each fish location and pose, accurately, in real-time and in large fish schools, and we foresee the use of underwater vehicles with multiple cameras for 3D fish school behavior analysis. This could be achieved using a single underwater vehicle with multiple cameras or multiple vehicles with a single camera.

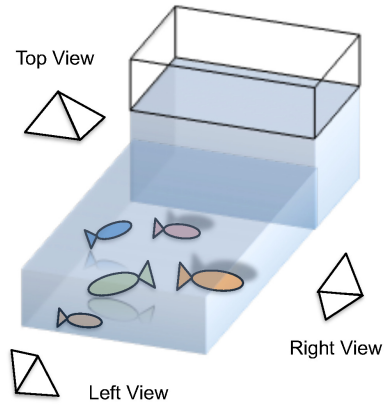

Fig. 1: Cameras and fish tank. There are three cameras facing three flat tank surfaces. The tank is designed for the top tank surface to touch the water. In addition to light refraction, virtual (mirror) fish and shadows occur.

We address this by using a calibrated multi-camera system, where each camera captures images through a flat surface (e.g. air, glass, water), and the cameras observe a common region from different point of views (through one or more flat surfaces). This corresponds to the case of a single vehicle with multiple cameras and approximates the case of multiple vehicles having a single camera (where the relative position of the vehicles is known). The presented approach can be also of interest for fish behavior analysis in aquariums. We consider the scenario in Figure 1, where one or more fishes move in a custom designed water tank with flat surfaces.

The scenario we consider (multiple camera system with a group of fish moving in a water tank) introduces additional challenges, namely: *i*) multiple virtual fish (due to reflections in the tank surface), *ii*) fish shadows, *iii*) reflections of objects outside the tank may be visible, etc.

The problems of occlusion and 3D modeling are challenging, but using views can help, and for this, considering light refractions is critical, especially for an accurate 3D estimation. Thus a key part of our work is taking into account light refraction through flat surfaces in an efficient manner.

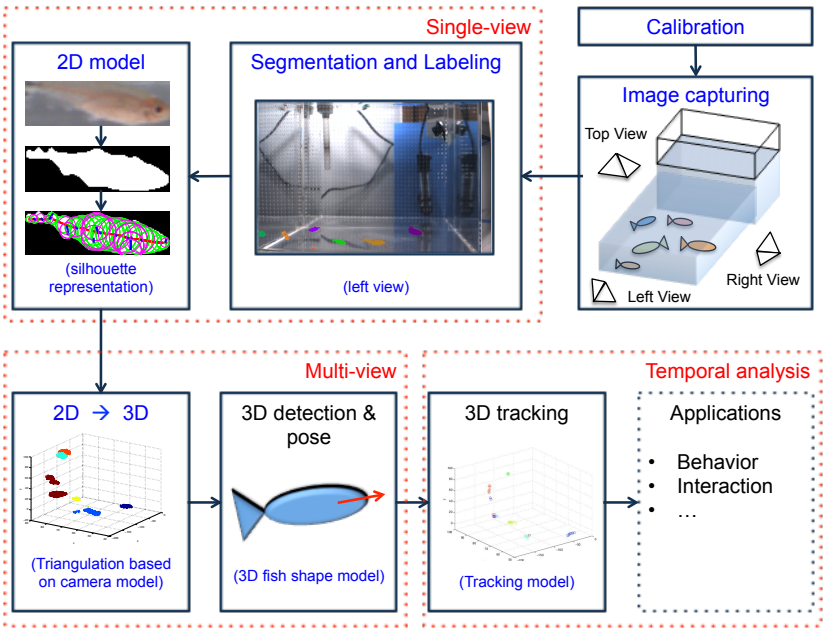

Fig. 2: System diagram

Fig. 3: Labeled blobs: left, top and right views.

II. System Overview

We have organized the system in four major modules as shown in Figure. 2: *i*) image capture, *ii*) single-view 2D analysis, *iii*) multi-view 3D analysis, and *iv*) temporal analysis, with an additional off-line calibration step.

The overall system works as follows. Images are captured from multiple views (3 views in the experiments), and for each of these views we obtain 2D fish silhouettes where more than one fish may be present. Then, using a calibrated camera model, the 2D silhouettes are combined to obtain 3D regions corresponding to one or more fishes. Finally, the 3D location and pose of each fish can be estimated and tracked. For efficiency, the system is implemented using parallel computing, with each single-view processing running in parallel in a multi-core CPU.

We give an overview of the system, with a focus on the single-view 2D processing. We present an overview of the multi-view 3D processing, in particular on 2D to 3D processing. Finally, the temporal analysis is briefly discussed.

A. Setup

As test scenario we consider a few Rummy nose tetra fishes (3 - 9 fishes), each about 3 cm length, swimming freely within a space of: 30 cm (width), 25 cm (depth), and 20-22 cm (height). The top surface is slightly sloped to reduce bubbles. The images are captured from three views using cameras from top, front and right views.

B. Image acquisition & calibration

The images are acquired using a multi-camera system and three Flea3 FL3-U3-13E4C Point grey USB cameras are used. We use a 3D multi-camera model that takes into light refraction efficiently [9]. The camera model considers light refraction through flat surfaces, thus information from the multiple views can be combined in the multi-view 3D processing. The camera model calibration is done off-line and has intrinsic and extrinsic steps.

C. Single-view 2D analysis

This module takes a single image, segments foreground objects, labels them, and generates a set of 2D silhouettes, with each silhouette corresponding to one or more fish (the representation is designed to manage occlusions in the 2D to 3D analysis). It has 3 main submodules:

1) Foreground segmentation: is obtained using a simple background subtraction method, with each background pixel

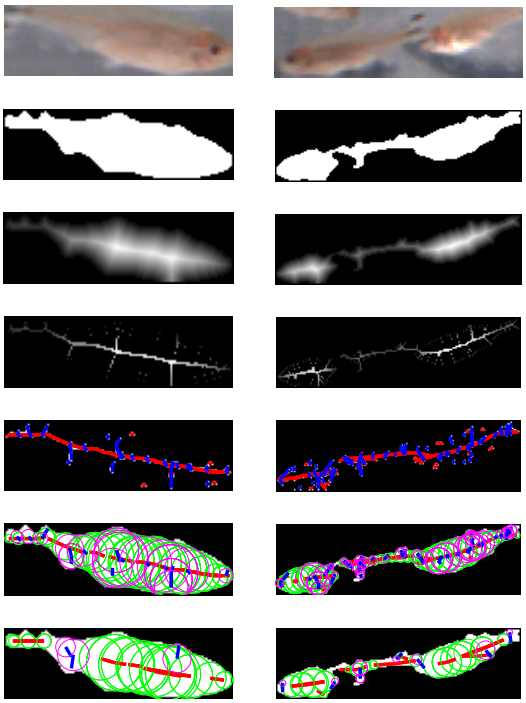

Fig. 4: 2D silhouette representation. *From top-to-bottom*: (i) cropped image region; (ii) silhouette, (iii) distance transform, (iv) local maxima, (v) skeleton, (vi) 2D representation (segment end-points/circles), (vii) 2D representation (selected end-points/circles). *Left*: One-fish; Region area: 36×117; blob area: 2189, skeleton area: 240. Number of circles: 109 (vi), 46 (vii). *Right*: Two-fishes; Region area: 47×164; blob area: 2150; skeleton area: 461, number of circles: 176 (vi), 84 (vii). *Note*: the end-points/circles density can be reduced, e.g. by keeping only longer segments (as done in the last row).

modeled using the median color. An opening operator, implemented using the integral image [10] for efficiency, is used to eliminate false detections. This method is good enough under controlled illumination, but in complex scenarios, advanced methods such as Robust PCA could be used.

2) Connected-component labels: are obtained using an efficient run-base algorithm [11]. See Figure 3 for an example. Each obtained labeled object (each silhouette) corresponds to one or more fishes.

3) A 2D silhouette model: is used to represent a labeled object. The 2D silhouette model consists of a skeleton and a set of covering circles (see examples in Figure 4 bottom row), where line segments approximate the skeleton, and circles (centered at the end-points of each segment) cover the silhouette. This representation is robust under occlusions, and thus is used later in the 3D analysis.

The steps of the process to obtain this representation is illustrated in Figure 4, and consists of:

i) Input region.
ii) Segmented silhouette.
iii) Distance Transform (DT): it gives, for every pixel in the silhouette, the minimum distance to the silhouette edge. We obtain it using the efficient linear algorithm in [12].
iv) Silhouette skeleton: DT local maxima (from Laplacian).
v) Segment-based skeleton approximation: method inspired in the real-time line segment detector Edlines [13] [14].
vi) Silhouette representation: We associate a circle to the end-points of each segment. The radius of each circle corresponds to the distance given by the DT at that point.
vii) The number of segments can be selected, e.g. by keeping only long segments.

This allows to efficiently obtain a *compact* representation of each silhouette, which is later used in the 2D-to-3D matching.

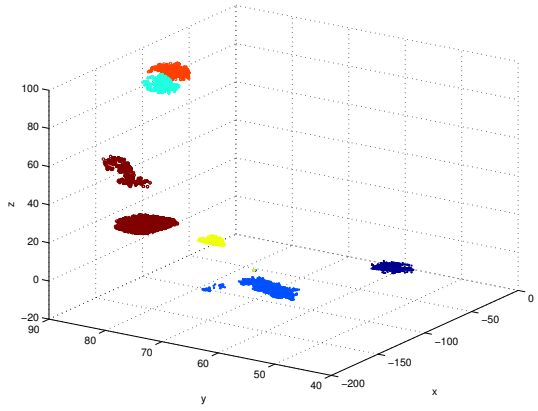

Fig. 5: 3D matched points.

D. Multi-view 3D analysis

This module integrates the 2D segmented silhouettes from the multiple views, generating 3D fish candidates (location & pose) using triangulation. Two key parts of this module are the efficient forward-projection (3D to 2D), and the fitting of the 3D fish shape model.

Multiview triangulation & matching. To efficiently integrate information from multiple cameras, we triangulate the silhouettes from the multiple views. This is achieved using the *pixel-wise varifocal camera model* in [9], model that allows efficient forward-projection for cameras in front of flat surfaces (i.e. considers light refraction in the flat surface and in water). The information from multiple 2D views is efficiently integrated to obtain a 3D representation, as follows:

- We first match the silhouettes' mass center from pairs views in 3D (for all pairs of views) using triangulation. To handle possible occlusions, we assume a silhouette may match more than one silhouette.
- We then refine the matched silhouettes using the detailed silhouette model (line segments+covering circles) following a process similar to the one above, but taking into account the all three views and matching all covering circles for all pairs of matched silhouettes.

3D pose model & estimation: Modeling the fish 3D shape is important and it is a problem that has been studied in the past. Most models assume a midline and a deformable shape around that midline or a ellipsoid-based representation (see e.g

Fig. 6: Re-projection on the three views.

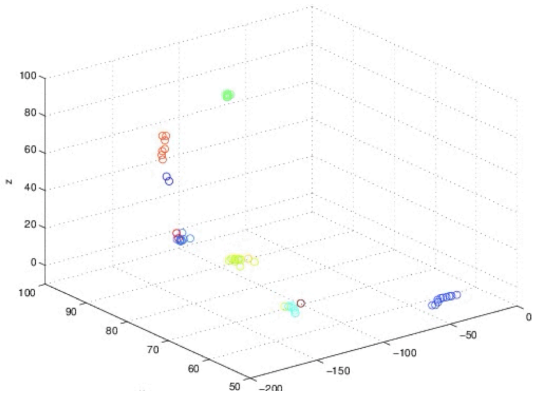

Fig. 7: Fish center tracks.

TABLE I: Processing time of main modules. Average over 50 frames (1024x800 pixels); Intel I5-4690 @3.50GHz (4 cores), 32 GB RAM. Ubuntu 15.4, gcc 4.9.2, C/C++.

Number of fishes	*3*			*6*		
	Processing time [ms]					
2D BG/FG	11.83			12.3		
2D Labeling	9.02			12.05		
2D DT & Skeleton	0.39			0.80		
Total single view	**21.24**			**25.15**		
Triangulation (center)	9.78			17.27		
Triangulation (skeleton)	44.9	163	2543	73.27	183.6	2345
Total multi-view	**54.7**	**173**	**2553**	**90.5**	**201**	**2362**
Total [msec]	**76**	**194**	**2564**	**116**	**226**	**2387**
Total (fps)	**13.2**	**5.15**	**0.39**	**8.62**	**4.42**	**0.41**
Circle cover number	*15*	*45*	*135*	*15*	*45*	*135*

[15] [16] [5] [7]). We match the triangulated data (see Figures 5 and 6), which later can be matched to a 3D shape model that can be efficiently integrated in the tracking module.

E. Temporal analysis

The temporal analysis module integrates 3D detection information overtime, and it can track each fish position, pose/shape in 3D. In addition, the obtained tracking information can be used to analyze fish interaction and behavior. In the present paper we use simple temporal association, as the 3D fish interaction and behavior analysis is out of scope of this paper.

III. Results

We present results on the described scenario (three cameras facing three flat tank surfaces), to validate our approach. This scenario is closer to the case of an aquarium than to underwater vehicles, but it allows to study the key issues of the problem (fish detection and matching) under refractions, and to evaluate the efficiency of the proposed methods.

3D matching & re-projection. Figure 5 presents the result of matching the silhouettes, while Figure 6 presents results of the forward-projection of matched points. Note that each fish and each virtual fish (i.e. mirror) is detected (in the 3D map & re-projections), and occlusions are correctly handled.

Processing time. Table I presents the processing time of the implemented modules for two experiments considering three and six fishes. Each image of each view has a 1024x800 pixel resolution. In addition to the number of fishes, we evaluated the effect of using a more detailed fish model (the number of circle covers in the triangulation: 15, 45 and 135). We can observed that the single view processing runs at about 40-45 fps (the three views run in parallel), while the 2D information triangulation runs at 55-103 fps, with the triangulation time being longer for larger number of fishes and for larger number of segments, as expected. As it can be observed, increasing the number of fishes increases the processing time, but when the 2D silhouette model is not too detailed (circle cover number less that 45), the processing time is about 5 fps or less. Processing time could be reduced by reducing image resolution.

IV. Conclusion

To study fish behavior (including the injection of artificial visual / acoustic stimuli), we require the accurate and efficient estimation of 3D models of fish trajectories and shape in water. As a first step in this direction, we have presented a real-time 3D fish detection and tracking system. The system uses multiple cameras capturing images though flat surfaces, which makes it an appropriate design for underwater vehicles as well as aquariums. The method is based on matching silhouettes from multiple views, and it uses an efficient model for efficient forward-projection for light for cameras in front of flat housings. Detection results in a water tank are presented for validation. In addition, a detailed analysis of the processing time of the proposed system is presented, showing that the system is fast enough for small and middle size fish swarms.

There are three key issues that still need to be addressed. The first one corresponds to improving the foreground segmentation robustness under illumination changes & reflections of external objects. The second point is to achieve robustness

under occlusions, in particular for large fish schools, while having real-time processing. Finally, we are interested in modeling fish-fish interaction, individual and group fish behavior, and fish behavior response to artificial stimuli.

Acknowledgment

This work was supported by JSPS KAKENHI grants 26240023 and 26540084. We thank Yanghong Zhong, a former graduate student, for the multiview fish data used in this paper.

REFERENCES

[1] E. F. Morais, M. F. M. Campos, F. L. C. Padua, and R. L. Carceroni, "Particle filter-based predictive tracking for robust fish counting," in *XVIII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'05)*, Oct 2005, pp. 367–374.

[2] D. T. Swain, I. D. Couzin, and N. E. Leonard, "Coordinated speed oscillations in schooling killifish enrich social communication," *Journal of Nonlinear Science*, vol. 25, no. 5, pp. 1077–1109, 2015.

[3] C. Spampinato, D. Giordano, R. Di Salvo, Y.-H. J. Chen-Burger, R. B. Fisher, and G. Nadarajan, "Automatic fish classification for underwater species behavior understanding," in *ACM workshop on Analysis & Retrieval of Tracked Events & Motion in Imagery Streams*, 2010, pp. 45–50.

[4] D. T. Swain, I. D. Couzin, and N. E. Leonard, "Real-time feedback-controlled robotic fish for behavioral experiments with fish schools," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 150–163, 2012.

[5] 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, pp. 1–17, 01 2016.

[6] Z.-M. Qian, X. E. Cheng, and Y. Q. Chen, "Automatically detect and track multiple fish swimming in shallow water with frequent occlusion," *PLoS ONE*, vol. 9, no. 9, pp. 1–12, 09 2014.

[7] S. Butail and D. A. Paley, "Three-dimensional reconstruction of the fast-start swimming kinematics of densely schooling fish," *Journal of The Royal Society Interface*, vol. 9, no. 66, pp. 77–88, 2011.

[8] ——, "3d reconstruction of fish schooling kinematics from underwater video," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2010, pp. 2438–2443.

[9] R. Kawahara, S. Nobuhara, and T. Matsuyama, "A pixel-wise varifocal camera model for efficient forward projection and linear extrinsic calibration of underwater cameras with flat housings," in *ICCV 2013 Workshop on Underwater Vision.*, Dec 2013, pp. 819–824.

[10] G. Facciolo, N. Limare, and E. Meinhardt-Llopis, "Integral Images for Block Matching," *Image Processing On Line*, vol. 4, pp. 344–369, 2014.

[11] L. He, Y. Chao, K. Suzuki, and K. Wu, "Fast connected-component labeling," *Pattern Recognition*, vol. 42, no. 9, pp. 1977 – 1987, 2009.

[12] A. Meijster, J. B. T. M. Roerdink, and W. H. Hesselink, *Mathematical Morphology and its Applications to Image and Signal Processing*. Boston, MA: Springer US, 2000, ch. A General Algorithm for Computing Distance Transforms in Linear Time, pp. 331–340.

[13] C. Akinlar and C. Topal, "Edlines: A real-time line segment detector with a false detection control," *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1633 – 1642, 2011.

[14] C. Topal, C. Akinlar, and Y. Genc, "Edge drawing: A heuristic approach to robust real-time edge detection," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, Aug 2010, pp. 2424–2427.

[15] N. F. Hughes and L. H. Kelly, "New techniques for 3-d video tracking of fish swimming movements in still or flowing water," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 53, no. 11, pp. 2473–2483, 1996.

[16] M. A. MacIver and M. E. Nelson, "Body modeling and model-based tracking for neuroethology," *Journal of neuroscience methods*, vol. 95, no. 2, pp. 133–143, 2000.