Multiple Fish Tracking via Viterbi Data Association for Low-Frame-Rate Underwater Camera Systems†

Meng-Che Chuang, Jenq-Neng Hwang
Department of Electrical Engineering, Box 352500
University of Washington
Seattle, WA 98105, USA
{mengche, hwang}@uw.edu

Kresimir Williams, Richard Towler
Alaska Fisheries Science Center
National Oceanic and Atmospheric Association
Seattle, WA 98115, USA
{kresimir.williams, rick.towler}@noaa.gov

***Abstract*—Non-extractive fish abundance estimation with the aid of visual analysis has drawn increasing attention. Low frame rate and variable illumination in the underwater environment, however, makes conventional tracking methods unreliable. In this paper, a robust multiple fish tracking system for low-frame-rate underwater stereo cameras is proposed. With the result of fish segmentation, a computationally efficient block-matching method is applied to perform successful stereo matching. A multiple-feature matching cost function is utilized to give a simple but effective metric for finding the temporal match of each target. Built upon reliable stereo matching, a multiple-target tracking algorithm via the Viterbi data association is developed to overcome the poor motion continuity of targets. Experimental results show that an accurate underwater live fish tracking result with stereo cameras is achieved.**

I. INTRODUCTION

For the conservation and management of commercially important fish populations in fisheries science, fish abundance estimation is required [1]. To improve the quality of abundance survey, we developed the Cam-trawl [2], a self-contained stereo-camera system fit to the aft end of a trawl. The collected image sequences are analyzed using automatic image processing techniques for object segmentation, tracking, classification and modeling. Specifically, object tracking provides a mean of fish counting and allows for more accurate length estimation by averaging several measurements of the same fish.

There are, however, several challenges for underwater image/video analyses, including the variable lighting conditions and the ubiquitous noise from non-fish objects. Related works on underwater fish video analysis using conventional multiple-target tracking algorithms were proposed [3,4]. Other methods including an approach based on linear programming were also proposed [5]. However, due to the low frame capturing rate and low contrast of intensity, targets move abruptly from one frame to another and enter/exit the field of view (FOV) frequently (4.3 frames of target lifespan in average). Fig. 1 shows this scenario, which makes conventional tracking methods infeasible for this task.

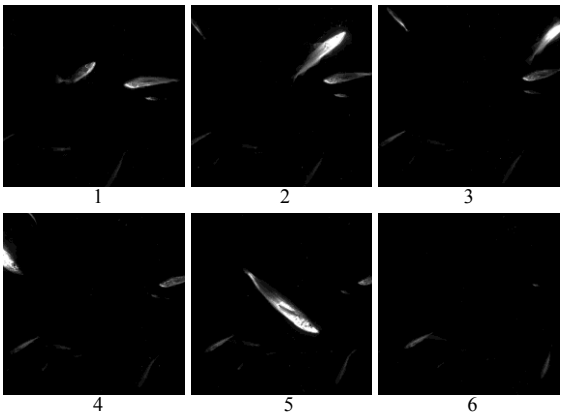

Figure 1. Abrupt movement and frequent entrance/exit of underwater fish captured at 5 frames per second.

These issues motivate us to resort to a stereo video trellis-based dynamic programming solution. In this paper, a novel multiple fish tracking algorithm for low frame rate stereo cameras is proposed, as shown in Fig. 2. Using the result of fish segmentation in [6] (see Section II), a block-matching stereo matching approach by object-height blocks is effectively used to identify the matching fish pairs (see Section III). Furthermore, by treating each stereo matching fish pair as one observation and a multiple-feature matching cost as a dissimilarity metric (see Section IV), a multiple-target version of Viterbi data association [7, 8] (see Section V) is proposed to find the optimal fish moving path in the observation trellis and thus determine the best tracking trajectory.

II. PREVIOUS WORK ON FISH SEGMENTATION

A novel trawl-based underwater camera system named Cam-trawl [2] was constructed to provide critical additional information for fisheries survey and reduce impact of surveying on fish population. Cam-trawl consists of a pair of

† This project was made possible through funds from NMFS' Advanced Sampling Technology Working Group, NOAA.

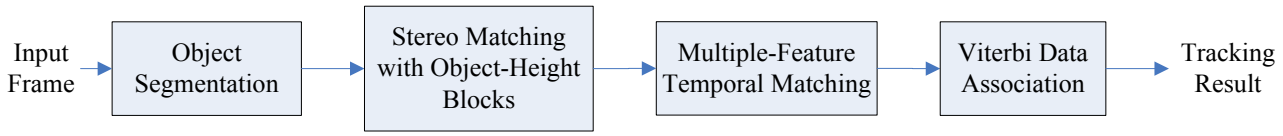

Figure 2. Overview of the proposed system.

high-resolution cameras, a series of LED strobes, a computer microcontroller, sensors and battery power supply. The cameras are capable of capturing 4 megapixel images at low frame rate (at most 10 fps) due to the limited bus bandwidth.

For this underwater camera systems, an automatic fish segmentation algorithm has been proposed [6]. We adopt a histogram backprojection method, with double local thresholding, to ensure a reliable segmentation on the boundary of fish. Results are further validated by area and variance criteria to reject non-fish objects. Simulations show that a 78% recall on segmentation and 10% of mean of absolute error in fish length measurement is achieved.

III. Stereo Matching by Object-Height Blocks

One major drawback of conventional dense stereo matching techniques is the intensive computational power consumption, making them impractical for a real-time imaging system. With knowledge of the target locations available in the segmentation stage, a fast stereo matching approach is proposed to match targets successfully while reducing much of computational redundancy.

Given a segmented object in the left image, its bounding box is divided to 4 equal sub-blocks in horizontal direction, as shown in Fig. 3. These sub-blocks are referred to as *object-height blocks*. The best match in the right image for the object-height block is then determined by a simple block-matching algorithm based on minimum sum of absolute difference (SAD) criterion. Note that only those object-height blocks within the bounding box of the object are taken as candidates, i.e., there are four candidates in total for each block. This results in great saving of computations without losing the accuracy of matching.

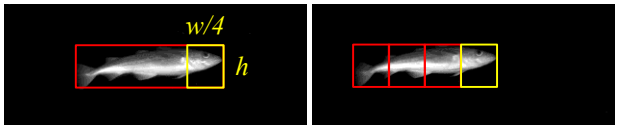

Figure 3. Object-height blocks on the rectified left and right image.

IV. Multiple-Feature Temporal Matching

A. Object Features

The ubiquitous noise and abrupt movement of targets are the major difficulties for tracking in a low-frame-rate underwater video. An object matching approach is therefore proposed for associating observations and targets. Various useful features are considered for measuring the dissimilarity between objects. Given an object O_t^j in frame t and an object O_{t-1}^i in frame $(t-1)$, four cues are exploited as follows.

1) Vicinity cue: The Euclidean distance is given by $d = \left\| \mathbf{x}_t^j - \hat{\mathbf{x}}_t^i \right\|$, where $\mathbf{x}_t^j$ and $\hat{\mathbf{x}}_t^i$ denote the center point coordinates of the observation j and the prediction of target i at frame t, respectively. Details about target prediction obtained by a motion projection scheme is described in Section V-*B*.

2) Area cue: The difference of area, after normalized by stereo triangulation, between the observation and the target is supposed to be small. The object area, denoted as $A(\cdot)$, is calculated by the classic connected components algorithm [9].

3) Direction of motion: Assume that O_t^j and O_{t-1}^i are matched, we define the motion vector as $\mathbf{v}_t^{i,j} = \mathbf{x}_t^j - \mathbf{x}_{t-1}^i$. The direction of motion is then represented by the angle between $\mathbf{v}_t^{i,j}$ and a predefined reference vector $\mathbf{v}_{ref}$, given by

$$\theta(\mathbf{v}_t^{i,j}, \mathbf{v}_{ref}) = \cos^{-1} \frac{\mathbf{v}_t^{i,j} \cdot \mathbf{v}_{ref}}{\left\| \mathbf{v}_t^{i,j} \right\| \left\| \mathbf{v}_{ref} \right\|} . \quad (1)$$

In the experiments, the reference vector is chosen according to the motion trend of fish schools, due to the movement of Cam-trawl, as $\mathbf{v}_{ref} = (-1, 0)$.

4) Histogram distance: In addition to geometric features, pixel value also plays an important role. To exploit the dissimilarity of grayscale intensity distribution between two objects, the earth mover's distance [10] is computed as the distance metric between 16-bin histograms for its desirable properties of allowing partial matches.

B. Matching Cost Function

Combining all the cues above, a likelihood function for object temporal matching is given by

$$L = \exp\left(-\frac{\left\| \mathbf{x}_t^j - \hat{\mathbf{x}}_t^i \right\|}{\sigma_v^2}\right) \cdot \exp\left(-\frac{(A(O_t^j) - A(O_{t-1}^i))^2}{\sigma_a^2}\right) \cdot \exp\left(-\frac{\theta(\mathbf{v}_t^{i,j}, \mathbf{v}_{ref})}{\sigma_m^2}\right) \cdot \exp\left(-\frac{(EMD(O_t^j, O_{t-1}^i))^2}{\sigma_h^2}\right) , \quad (2)$$

and the "matching cost" is defined as

$$C = -\ln L . \quad (3)$$

This cost is assigned to the edge between O_{t-1}^i and O_t^j in the Viterbi trellis as discussed in the next section. The σ values in (2) are determined empirically as $\sigma_v = 2.5$, $\sigma_a = 2.3$, $\sigma_m = 2.5$ and $\sigma_h = 0.67$ in the experiments.

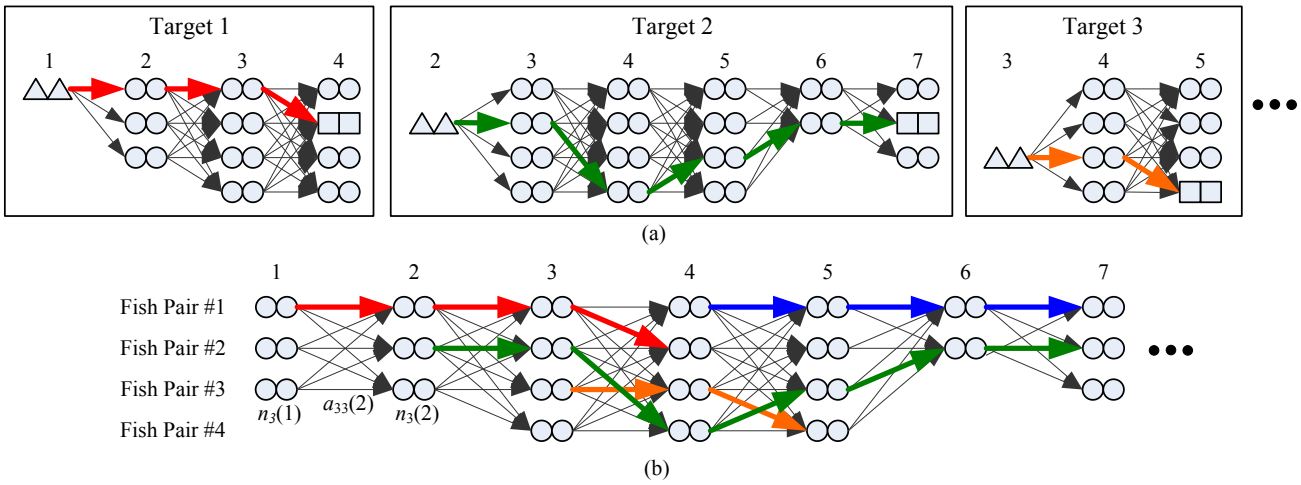

Figure 4. Multiple-target Viterbi data association. (a) Each target maintains a separate trellis during its lifespan with its own starting node (triangles) and ending node (squares). The optimal path in each trellis is labeled by colored arrows. (b) An overall trellis showing several paths from targets in (a).

V. Viterbi Data Association

In the proposed system, the stereo information is utilized by regarding a pair of stereo-matched object, i.e., the same object in the left and right image, as one observation for tracking. The matching cost of temporal matching is then given by the sum of the costs from two images, namely, $C_{i,j}(t) = C^L_{i,j}(t) + C^R_{i,j}(t)$. To overcome the difficulties in tracking fish due to their frequent entrance/exit, a multiple-target Viterbi data association algorithm is introduced.

A. Basic idea

The Viterbi data association [7, 8] is performed based on a trellis of the observations in each frame. A trellis is a type of directed graph where nodes are partitioned into ordered subsets $N(t) = \{n_j(t) \mid j = 1,2,...,|N(t)|\}$ for $t = 1,2,...,T$, and edges $a_{ij}(t)$ lie between any pair of node in adjacent subsets $\{n_i(t-1), n_j(t)\}$. Nodes in a subset represent objects in one frame, and each edge is assigned a matching cost $C_{i,j}(t)$. The total cost of a path P may be written as

$$C(P) = \sum_{t=2}^{T} C_{i,j}(t) \text{, where } \{n_i(t-1), n_j(t)\} \in P. \quad (4)$$

The Viterbi algorithm [11] is applied here to find the minimum-cost path during single target tracking. When a new observation is detected, a new node is initialized with zero cost and a null predecessor. For each iteration, the matching cost for every node $n_j(t), j = 1,2,...,|N(t)|$ is given by (3). Then the predecessor and cost are assigned to node $n_j(t)$:

$$\pi_j(t) = \arg \max_{1 \le i \le |N(t-1)|} [C_i(t-1) + C_{i,j}(t)], \quad (5)$$

$$C_j(t) = C_{\pi_j(t)}(t-1) + C_{\pi_j(t),j}(t). \quad (6)$$

Once the target leaves FOV, i.e., the final stage of trellis is reached, a backtracking step is performed. Starting from the minimum-cost node in the final stage, the optimal sequence is recovered by traversing backward all the way to the first stage according to the predecessors stored at each stage.

B. Multiple-Target Case

Inspired by [8], a multiple-target Viterbi data association algorithm is proposed for our case of low-frame-rate fish tracking. Since the starting frame may differ among targets, the predecessor and minimum cost at each node may also differ. Therefore, we create a separate trellis for every target to track. Data association by (5) and (6) is then performed separately for each target with all observations, as shown in Fig. 4. Observations that are not associated with any target corresponds to new targets or false alarms. Note that occlusion is inherently handled by our method since paths in different trellises can pass through the same nodes.

In each frame, a motion projection mechanism is utilized to estimate and update the position of the existing target. Given the position $\mathbf{x}^k_{t-1}$ and velocity $\mathbf{v}^k_{t-1}$ of the k-th tracked target from frame $(t-1)$, the predicted position at current frame is given by $\hat{\mathbf{x}}^k_t = \mathbf{x}^k_{t-1} + \mathbf{v}^k_{t-1}$. After the data association, the observation node with the minimum cost is chosen to update the position and velocity by

$$\mathbf{x}^k_t = \mathbf{x}^{j^*}_t, \quad (7)$$

$$\mathbf{v}^k_t = \alpha \mathbf{v}^{j^*}_t + (1-\alpha)\mathbf{v}^k_{t-1}, \quad (8)$$

where $\mathbf{x}^{j^*}_t$ and $\mathbf{v}^{j^*}_t$ denote the position and velocity of the minimum-cost observation, respectively. The value of α is determined empirically as $\alpha = 0.3$ in the experiments.

The abrupt movement and short lifespan of a fish target make the criteria of track creation and deletion less reliable. For this reason, a track is restricted to end only when its predicted position is within 100 pixels of the boundary. If a track is lost before approaching the frame boundary, the predicted position is used as the actual position and the velocity remains. A new prediction is then made for the next frame.

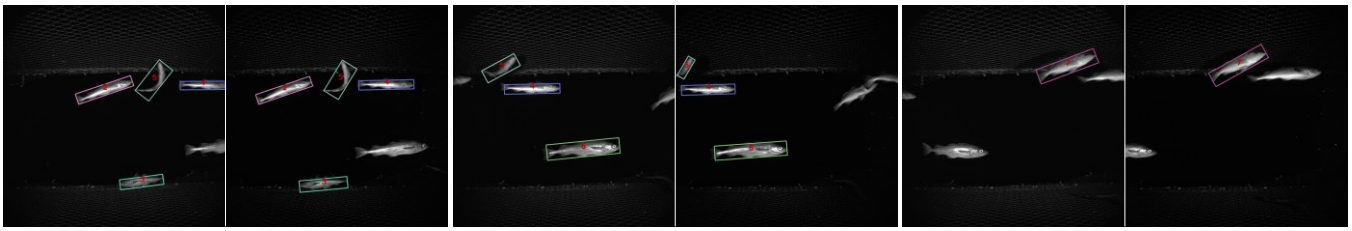

Figure 5. Tracking multiple fish in an underwater stereo video clip. Targets are labeled by numbers and oriented bounding boxes with different colors.

Rather than using a batch scheme as mentioned in [7], backtracking is performed and the optimal sequence of observations is recovered once there is a target leaving the FOV. The proposed system is thus able to not only perform online tracking but also exempt from potential failures due to the gaps between batches.

VI. Experimental Result

Before the object segmentation stage, a background subtraction is performed to eliminate the trawl web behind the fish based on a combination of morphological operations. The stereo frames are calibrated and rectified via camera parameters obtained offline prior to stereo matching.

The proposed system is used to track multiple fish targets simultaneously in several sample video clips. All these video clips are grayscale and recorded underwater by the stereo cameras on Cam-trawl. The frame size is 2048×2048 pixels, and the frame rate is 5 frames per second. The whole process is fully automatic and requires no manual intervention. Tracked fish are labeled with numbers and bounding boxes with different colors in order to make them differentiable.

Some tracking statistics of the proposed system comparing with other data association methods are listed in this section. Table I shows that the proposed system improves the accuracy of underwater fish detection by utilizing temporal information. In Table II, tracking success rate is defined as the ratio of correctly labeled targets to correctly detected targets. One can see that the proposed system, i.e., matching cost plus Viterbi data association (MC+VDA) outperforms other data association methods. The conventional nearest neighbor (NN) suffers from poor motion continuity and short lifespan of targets, and thus tracks fish in a low success rate. Note that the proposed matching cost (MC) is also tested alone to show the effectiveness of matching objects based on various type of feature. Fig. 5 shows a representative experimental result in 3 consecutive frames with tracked fish labeled on both sides of stereo image, which clearly demonstrates the robustness of the proposed system.

VII. Conclusion

A novel multiple fish tracking system based on Viterbi data association for low-frame-rate underwater stereo cameras is proposed. By exploiting various appearance features, the matching cost function gives an effective metric to find a temporal match in the noisy underwater environment. The multiple-target Viterbi data association takes advantage of dynamic programming to overcomes the difficulties of abrupt target motion and frequent entrance/exit. Experimental result shows that the proposed system gives a success rate at 87.60% in terms of fish tracking for low-frame-rate and low-contrast underwater stereo videos.

TABLE I. Precision and Recall of The Proposed System

Num. of Targets	Detection Precision	Detection Recall
62	0.9831	0.9355

TABLE II. Tracking Success Rate Among Different Data Association Methods

Clip	NN	MC	MC+VDA
1	0.3750	0.4688	0.9375
2	0.4167	0.5833	0.8333
3	0.2857	0.5000	0.8571
Avg.	0.3591	0.5412	**0.8760**

References

[1] D. G. Hankin and G. H. Reeves, "Estimating total fish abundance and total habitat area in small streams based on visual estimation methods," *Can. J. Fish. Aquat. Sci., NRC Research Press*, vol. 45, no. 5, pp. 834-844, 1988.

[2] K. Williams, R. Towler, C. Wilson, "Cam-trawl: a combination trawl and stereo-camera system," *Sea Technol.*, vol. 51, no. 12, Dec. 2010.

[3] D. Walther, D.R. Edgington and C. Koch, "Detection and tracking of objects in underwater video," *Proc. of Computer Vision and Pattern Recognition, IEEE Intl. Conf. on* (CVPR '04), Jun. 2004.

[4] S. Butail and D. A. Paley, "3D reconstruction of fish schooling kinematics from underwater video," *Proc. of Robotics and Automation, IEEE Intl. Conf. on* (ICRA '10), May 2010.

[5] H. Jiang, S. Fels and J. J. Little, "A linear programming approach for multiple object tracking," *Proc. of Computer Visioni and Pattern Recognition, IEEE Intl. Conf. on* (CVPR '07), Jun. 2007.

[6] M.-C. Chuang, J.-N. Hwang, K. Williams and R. Towler, "Automatic fish segmentation via double local thresholding for trawl-based underwater camera systems," *Proc. of Image Processing, IEEE Intl. Conf. on* (ICIP '11), pp.3145-3148, Sep. 2011.

[7] G. W. Pulford, "Multi-target Viterbi data association," *Proc. of Information Fusion, IEEE Intl. Conf. on*, pp. 1-8, Jul. 2006.

[8] A. Azim and O. Aycard, "Multiple pedestrian tracking using Viterbi data association," *Proc. of Intelligent Vehicles Symposium (IV), 2010 IEEE*, pp.706-711, Jun. 2010.

[9] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*. Reading, MA: Addison-Wesley, 1992, pp. 28–48.

[10] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Intl. J. of Computer Vision,* vol. 40, no. 2, pp. 99-121, 2000.

[11] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. of the IEEE*, vol. 61, no. 3, pp. 268-278, Mar. 1973.