AGGREGATED SEGMENTATION OF FISH FROM CONVEYOR BELT VIDEOS†

Meng-Che Chuang, Jenq-Neng Hwang

Department of Electrical Engineering
University of Washington
Box 352500, Seattle, WA 98195, USA
{mengche, hwang}@uw.edu

Craig S. Rose

Alaska Fisheries Science Center
National Oceanic & Atmospheric Administration
Seattle, WA 98115, USA
craig.rose@noaa.gov

ABSTRACT

Automation of fishery survey through the aid of visual analysis has received increasing attention. In this paper, a novel algorithm for the aggregated segmentation of fish images taken from conveyor belt videos is proposed. The watershed algorithm driven by an automatic marker generation scheme successfully separates clustered fish images without damaging their boundaries. A target selection based on appearance classification then rejects non-fish objects. By applying histogram backprojection and kernel density estimation, an innovative algorithm for combining object masks of one tracked fish from multiple frames into a refined single one is also proposed. Experimental results show that accurate fish segmentation from conveyor belt videos is achieved.

Index Terms— aggregated segmentation, conveyor belt, fish/non-fish classification, kernel density estimation, soft segmentation

1. INTRODUCTION

Counting and isolation of fish after capture are normally carried out manually on fishing vessels or at onshore processing facilities [1]. This conventional process is extremely laborious and limits the efficiency of fish catch processing for either commercial, regulatory, or research purposes. Replacing the current manual system with an automated fish monitoring system requires video analysis techniques that include object segmentation and tracking. The successful development of algorithms to support this automated system could significantly speed up the sampling of fish during processing.

While there are clear advantages to using an automated monitoring system, there still exist practical challenges to implementing such a system that are particular in aspects of processing conveyor belt videos. The surface texture of moving belts, for example, degrades the effectiveness of conventional object segmentation methods. Fish targets that overlap partially or that have slight displacement while passing along on the belt are also difficult to isolate. Moreover, various types of irrelevant objects such as debris and other marine animals often appear on the belt simultaneously with fish. Storbeck *et al.* [2] used a laser source aimed oblique to the belt plane to identify the size and shape of passing fish targets. However, this approach requires additional equipment, making it less applicable to the general scenario which uses only a single camera.

In this paper, a novel algorithm for aggregated segmentation of fish images captured from conveyor belt videos is proposed. The contributions of this work include: 1) a fully systematic marker-driven watershed segmentation algorithm to separate clustered fish that overlap with one another; 2) by exploiting useful features from segmented targets, an automated target selection algorithm rejects unwanted objects; 3) inspired by [3], an innovative algorithm of segmentation aggregated from multiple frames into a refined single result is proposed by applying kernel density estimation to its alpha maps generated from histogram backprojection.

This paper is organized as follows: Section 2 briefly gives an overview of the proposed system. Section 3 describes the foreground extraction. Section 4 introduces the target selection, followed by the proposed algorithm of aggregated segmentations from multiple frames in Section 5. Section 6 describes the experimental results, followed by the conclusion in Section 7.

2. SYSTEM OVERVIEW

A flow chart of our proposed aggregated segmentation system is shown in Fig. 1. The first step is foreground extraction. Next, the resultant binary mask is effectively refined by histogram backprojection, where the alpha map (opacity of the foreground) is generated. Target selection extracts useful appearance features, which are used to reject non-fish objects.

Extracted fish targets are then handled by the multiple-target tracking stage. If there are no targets exiting the field

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

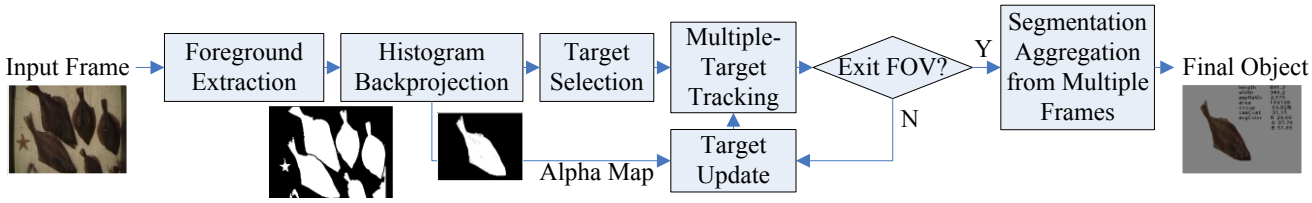

Figure 1. Overview of the proposed aggregated segmentation system.

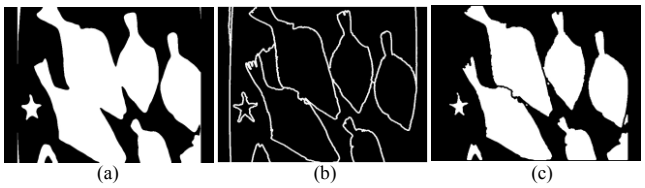

Figure 2. Separating clustered fish in foreground extraction. (a) Object mask by Otsu thresholding. (b) Edge map generated by watershed algorithm. (c) Subtracted object mask.

of view (FOV), tracks are updated with the current observations to be used in the next frame; otherwise, segmentation aggregation collects alpha maps of the leaving target from every frame it has appeared and combines them into a single one, which generates the final segmentation.

3. FOREGROUND EXTRACTION

To isolate the image of fish passing along the conveyor belt, the rough silhouettes of all foreground objects must first be detected. Prior to this process, Gaussian smoothing is performed to remove noise. An adaptive thresholding technique using Otsu's method [4] is then employed to generate an initial binary mask of foreground objects. Note that this is only a preliminary segmentation of fish body, which needs to be carefully refined by the following steps.

Fish often are found clustered on a conveyor belt; thus their bodies touch or even overlap each other. This presents a large blob which encompasses several fish from the thresholding result. To detect and separate fish bodies in a robust way, the watershed segmentation with an automatic marker generation scheme [5] is utilized to extract the boundary around each fish. Distance transform is applied to the initial mask. The distance map is then thresholded and the resulting connected components serve as the "markers" for objects. The watershed algorithm controlled by these markers is performed to the initial mask and produces the edge map of the input image. Finally, by subtracting the watershed segmentation from the initial mask, a refined binary segmentation result is generated. An illustrative example of separating clustered fish is shown in Fig. 2.

4. TARGET SELECTION

Objects are located roughly during the foreground extraction stage by an initial binary object mask. However, false alarms still exist, including irrelevant objects and small fragments which represent the noise created from segmentation. These are rejected by our target selection algorithm, which consists of size thresholding and appearance classification.

4.1. Size Thresholding

The classic connected components algorithm [6] is applied to determine each isolated foreground region in the object mask. Specifically, for each pixel (x, y) within the k-th segmented object O_k, its corresponding pixel on the object mask is revised by

$$B(x,y) = \begin{cases} 1 & \text{if } \theta_A^L \le A(O_k) \\ 0 & \text{otherwise} \end{cases}, \ (x,y) \in O_k, \qquad (1)$$

where $A(\cdot)$ gives the area of an object, and θ_A^L denotes the lower bound for the area to preserve.

The aspect ratio of foreground object O_k is defined as $R(O_k) = \max(w_k^{OB}/h_k^{OB}, h_k^{OB}/w_k^{OB})$, where w_k^{OB} and h_k^{OB} denotes respectively the width and height of the k-th object's oriented bounding box. Given the aspect ratio, the foreground mask is thresholded by

$$B(x,y) = \begin{cases} 1 & \text{if } \theta_R^L \le R(O_k) \le \theta_R^U \\ 0 & \text{otherwise} \end{cases}, \ (x,y) \in O_k, \qquad (2)$$

where (θ_R^L, θ_R^U) denotes the lower and upper bound for the aspect ratio. In the experiments, the threshold for area θ_A^L is determined empirically as 0.05 times the total pixels in one frame and the bounds for aspect ratio as $(\theta_R^L, \theta_R^U) = (2, 7)$.

4.2. Appearance Classification

Many marine animals have similar size and aspect ratios to fish. This makes it difficult to distinguish them using the above mentioned size thresholding method. To overcome this, more sophisticated features are exploited and fed to a multi-class classifier in order to give a more reliable approach to target selection.

1) Occupancy Rate: Fish bodies tend to be elliptical while lying flat on a conveyor belt. In other words, a foreground object fits its oriented bounding ellipses better if

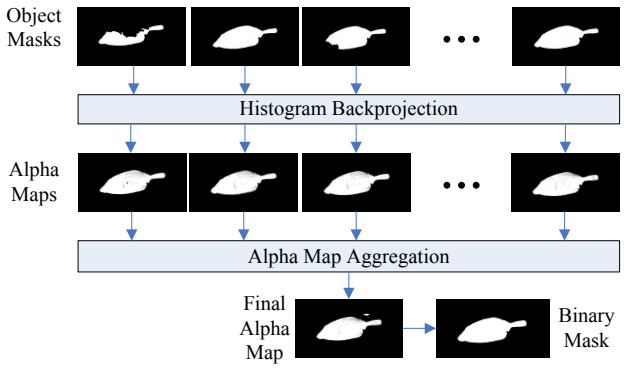

Figure 3. Segmentation aggregation procedure from multiple frames.

it is a fish. The occupancy rate within the oriented bounding ellipse is given by $OR(O_k) = A(O_k)/\pi ab$, where a and b denote the major and minor axis of the oriented bounding ellipse separately.

2) Shape: A reliable shape descriptor has to be invariant to translation, rotation and scaling. The curvature scale space (CSS) representation [7] of the object contour is thus employed. A model set consisting of 3 fish that belong to different species is created. For every extracted object, its shape dissimilarity values with models are calculated by a matching algorithm also proposed by [7], which compares two sets of descending-ordered local maxima of CSS images.

3) Chromaticity: In addition to geometric information, the pixel values within object region are also taken into account. The average of pixel values within an object is calculated and converted to normalized-RGB color space as $(r,g) = (R/(R+G+B), G/(R+G+B))$. In this way, the factor of illumination is removed and only the chromaticity is taken into account.

These appearance signatures form the feature vector. A classification and regression tree (CART) [8] is trained to differentiate fish and non-fish objects and used as the second part of target selection.

5. AGGREGATING SEGMENTATIONS FROM MULTIPLE FRAMES

In many cases, the segmentation of a video object succeeds in some frames but fails in other frames. For this reason, we present the notion of combining multiple segmentations of one target from every video frame in which it has appeared and generate an aggregated segmentation. The procedure is shown in Fig. 3. Specifically, the linear Kalman filter [9] is applied to track the fish object that is preserved by target selection. According to this, alpha maps generated by histogram backprojection over frames are collected for the same target. Once a target exits, the set of its object masks are aggregated to one final mask by alpha map aggregation.

5.1. Histogram Backprojection

The alpha value of a pixel is defined as the opacity of the foreground. Specifically, a video frame is assumed to be a linear combination of a foreground image F and a background image B as $I_p = \alpha_p F_p + (1-\alpha_p)B_p$, $0 \le \alpha_p \le 1$, where I_p denotes the color at pixel p.

For an extracted object, the acquisition of alpha map and refinement of segmentation can be carried out concurrently by using the histogram backprojection procedure [10]. First, a swollen object mask is generated from the original one by morphological dilation using a disk of radius 3 pixels as the structuring element. The two masks are used to generate two color histograms, $H_{ori}(c)$ and $H_{dil}(c)$, where c denotes the a color vector. Here, we use three-dimensional RGB color histograms with 16 bins in each channel. A ratio histogram $H_R(c) = \min(H_{ori}(c)/H_{dil}(c), 1), \forall c$ is then calculated and backprojected to the image plane. This procedure not only computes the alpha map of foreground but also provides a successful approach to refining segmentation on the boundary by applying a threshold to the backprojection of ratio histogram.

5.2. Alpha Map Aggregation

Having collected the alpha maps from every frame containing the target, an effective aggregation method is used to further improve the accuracy of segmentation. The concept of kernel density estimation [11] is adopted by viewing the alpha values of one pixel from different frames as samples from an underlying probability density function. The estimator of such a density is given by

$$\hat{f}_h(\alpha) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\alpha - \alpha_i}{h}\right), \qquad (3)$$

where $K(\cdot)$ is the kernel, h is the bandwidth of kernel. The Gaussian function is chosen for the kernel, and the bandwidth (i.e., the variance of Gaussian function) is determined empirically as $h = 0.15$ in the experiments.

The aggregated alpha map is calculated by setting each pixel to the maximum of $\hat{f}_h(\alpha)$. This is achieved by the mean-shift algorithm [12] as

$$m(\alpha) = \frac{\sum_{i=1}^{n} \alpha_i k'\left(\left\|\frac{\alpha-\alpha_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} k'\left(\left\|\frac{\alpha-\alpha_i}{h}\right\|^2\right)} - \alpha, \qquad (4)$$

where $k(\cdot)$ is the kernel profile. To avoid a local maximum, majority voting is used to estimate the initial value of α. A threshold is then applied to get the binary aggregated object mask. The threshold value is determined empirically as $\theta_\alpha = 0.3$ in our experiments.

(a) (b)

Figure 4. Segmentation of fish on the conveyor belt (a) using a high-definition camcorder and (b) using a GoPro wide-angle camera.

TABLE I
PRECISION AND RECALL OF FISH/NON-FISH CLASSIFICATION

Class	Number	Precision	Recall
Non-fish	86	0.943	0.965
Fish	73	0.958	0.932

TABLE II
AVERAGE OF NORMALIZED OVERLAPPING AREA FOR 14 TARGETS

BinMV	AlphaAvg	AlphaKDE
0.9278	0.9444	**0.9648**

6. EXPERIMENTAL RESULT

The proposed system is evaluated by processing several video clips of moving conveyor belts. These clips were captured by using a digital high-definition camcorder positioned directly over the belt. The resolution is 1440×1080 pixels, and the frame rate is 30 frames per second. In addition to the aggregated segmentation, the optimal object mask from all frames of one target is also provided as the output. The optimality of a mask is assessed by its distance from frame boundaries and the normalized overlapping area

$$Overlap(O_k, T_k) = \frac{A(O_k \cap T_k)}{A(T_k)}. \quad (5)$$

where O_k denotes the object mask from one frame and T_k denotes the aggregated mask.

For target selection, the fish/non-fish classifier is trained by 159 manually-labeled samples through a 10-fold cross-validation procedure. Table I presents the precision and recall of classification. For aggregated segmentation, the performance is measured by (5) between the computed mask and the hand-labeled ground truth. Table II presents a comparison of the average of normalized overlapping area among several segmentation methods. The proposed algorithm, i.e., using alpha maps and kernel density estimation (AlphaKDE) outperforms the methods using binary masks and majority vote (BinMV) and using alpha maps and averaging (AlphaAvg). Note that the accuracy of AlphaAvg is also higher than BinMV, showing the advantage of soft decision by using alpha maps.

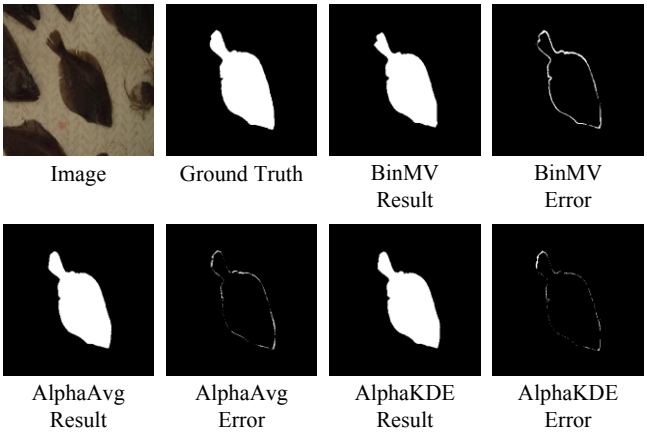

Figure 5. Comparison of target segmentation results and absolute error maps using different algorithms.

The proposed system is also applied to a video clip captured by a GoPro HERO3 camera. Because of the heavy distortion introduced by its wide-angle lens, camera calibration is required in prior to processing. Some representative results of segmentation for fish on the conveyor belt, with systematically derived useful parameters of the target added as the meta-data, from both kinds of videos are exhibited in Fig. 4. A comparison of the results and absolute errors among different segmentation methods are also shown in Fig. 5.

7. CONCLUSION

A novel framework of aggregated segmentation based on object tracking for conveyor belt videos is proposed. Irrelevant objects are successfully detected and rejected by exploiting various kinds of features, especially the use of CSS representation for shape. By target tracking, temporal information is utilized by combining segmentation result from several frames. Histogram backprojection generates a soft segmentation result, and kernel density estimation provides an effective way to aggregate them into a well-refined segmentation. Experimental result shows that the proposed system produces an accuracy at 96.48% in terms of correct pixels in the object mask.

7. REFERENCES

[1] D.J. White, C. Svellingen and N.J.C. Strachan, "Automated measurement of species and length of fish by computer vision," *Fisheries Research, Elsevier*, vol. 80, no. 2, pp. 203-210, Feb. 2006.

[2] F. Storbeck and B. Daan, "Fish species recognition using computer vision and a neural network", *Fisheries Research*, *Elsevier*, vol. 51, no. 1, pp.11-15, Jan. 2001.

[3] A. Alush and J. Goldberger, "Ensemble segmentation using efficient integer linear programming," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 34, no.10, pp.1966-1977, Oct. 2012.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *Sys., Man, Cyber., IEEE Trans. on.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[5] J. Cheng and J. C. Rajapakse, "Segmentation of clustered nuclei with shape markers and marking function," *Biomedical Engineering, IEEE Trans. on*, vol. 56, no. 3, pp.741-748, Mar. 2009.

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

[7] S. Abbasi, F. Mokhtarian and J. Kittler, "Curvature scale space image in shape similarity retrieval," *Multimedia systems*, *Springer*, vol. 7, no. 6, pp.467-476, Jun. 1999.

[8] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall (Wadsworth, Inc.): New York, 1984.

[9] G. Welch and G. Bishop, "An introduction to the Kalman filter," *Technical Report, University of North Carolina at Chapel Hill*, 1995.

[10] 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.

[11] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley-Interscience, 1992.

[12] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," Pattern Ana*lysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603-619, May 2002.