A Multiview, Multimodal Fusion Framework for Classifying Small Marine Animals with an Opto-Acoustic Imaging System

Paul L. D. Roberts and Jules S. Jaffe
Marine Physical Laboratory
Scripps Institution of Oceanography
9500 Gilman Dr. La Jolla, CA 92093-0238
{paulr,jules}@mpl.ucsd.edu

Mohan M. Trivedi
Electrical and Computer Engineering Department
University of California, San Diego
9500 Gilman Dr. La Jolla, CA 29093-0434
mtrivedi@ucsd.edu

Abstract

A multiview, multimodal fusion algorithm for classifying marine plankton is described and its performance is evaluated on laboratory data from live animals. The algorithm uses support vector machines with softmax outputs to classify either acoustical or optical features. Outputs from these single-view classifiers are then combined together using a feedback network with confidence weighting. For each view or modality, the initial classification and classifications from all other views and modalities are confidence-weighted and combined to render a final, improved classification. Simple features are computed from acoustic and video data with an aim at noise robustness. The algorithm is tested on acoustic and video data collected in the laboratory from live, untethered copepods and mysids (two dominant crustacean zooplankton). It is shown that the algorithm is able to yield significant ($>$ 50%) reductions in error by combining views together. In addition, it is shown that the algorithm is able boost performance by giving more weight to views or modalities that are more discriminant than others, without any a priori *knowledge of which views are more discriminant.*

1. Introduction

Remote classification of marine animals is an important goal in oceanography. During the last three decades, many systems have been developed that make use of optical sensors for plankton classification [1, 2, 3, 4, 5] and marine animal tracking [6], or acoustic sensors for fish classification [7, 8, 9]. Advantages of these indirect methods over direct sampling methods include: reduced sensitivity to animal avoidance, incorporation into automated processing algorithms, larger survey volumes, and faster survey rates. However, these benefits come at the cost of needing more sophisticated methods to infer animal taxa from data. Optical systems typically sample small volumes (although large-sample-volume optical systems also exist, for example [10]) with high resolution and record images of plankton appearance from which many features can be extracted and used to discriminate on the taxa level. In contrast, acoustic systems typically sample larger volumes at much lower resolution and only a small set of features related to echo intensity are used for classification. In many cases, optical and acoustic methods are combined together as these two modalities offer complementary information [11, 12].

Multimodal, multiperspective (multiview) fusion is a key component in modern vision-based human activity analysis systems. It has been used to enhance speech recognition by combing audio and visual cues [13], improve person tracking using color and thermal images [14], and perform driver activity analysis using multi-camera, thermal and color imagery [15]. In underwater imaging using acoustic sensors, multiview systems have been shown to improve estimates of animal size [16] as well as improve animal classification [17]. One of the key features of multiview, multimodal systems is that they collect complementary information that enhances the system's understanding of the different aspects of the subject of interest. This can be used to improve performance for subjects who's appearance is sensitive to orientation, and add robustness to changing environmental conditions.

Here, a bimodal, multiview opto-acoustic system that uses eight acoustic receivers and two video cameras is examined for classification of two dominant types of crustacean zooplankton: copepods and mysids. A fusion algorithm and feature extraction method is developed to yield a system that can combine views together and improve performance independently of the total number of views and modalities.

978-1-4244-5498-3/09/$25.00 ©2009 IEEE

2. Multiview and Multimodal Classification and Fusion Algorithms

The classification algorithm is an extension of multiview methods developed recently in underwater target classification [18, 19] that employs support vector machines (SVMs) for feature-level classification and uses confidence weighting to combine multiple views together.

2.1. Single-View Classification

A single-view feature vector $\mathbf{y}$ is classified using an SVM. The trained SVM has the form [20]

$$f(\mathbf{y}) = \sum_{n=1}^{N} w_n \Phi\left(\mathbf{y}_n, \mathbf{y}\right) + b, \tag{1}$$

where the weights $\mathbf{w}$ are learned during training and the kernel function $\Phi(\mathbf{y}_n, \mathbf{y})$ is Gaussian. For a C-class classification problem, C SVMs are trained to separate each class from the other $C-1$ classes. Let the output for the c^{th} classifier be $f_c(\mathbf{y})$. An estimate for the posterior class probability over all SVMs is computed as

$$P(c|\mathbf{y}) = \frac{\exp\left[f_c(\mathbf{y})\right]}{\sum_{c=1}^{C} \exp\left[f_c(\mathbf{y})\right]}. \tag{2}$$

This function is equivalent to the softmax activation function commonly used in neural networks. Note that at this stage, all information is contained in $P(c|\mathbf{y})$.

2.2. Multiview Fusion

Let the probabilities computed from (2) for the j^{th} view be

$$\mathbf{p}_j = \left(P(c=1|\mathbf{y}_j), ..., P(c=C|\mathbf{y}_j)\right)^T. \tag{3}$$

Probability vectors for each view are then combined together using an extension of the collaborative agent framework used in [19]. For the j^{th} agent, all other agents send their initial probability vectors which are then combined to yield a joint probability vector

$$P_{i\neq j}(c=k|\mathbf{Y}_{i\neq j}) = \prod_{i\neq j}^{M} \mathbf{p}_i[k], \tag{4}$$

where $\mathbf{Y}_{i\neq j}$ is the matrix of feature vectors for the $M-1$ views excluding the feature vector from the j^{th} view.

To quantify the degree to which the probability is spread between classes, the side-lobe ratio [21] is used. The ratio is defined as

$$C_{SL} = \frac{P(c_1|\mathbf{Y}) - P(c_2|\mathbf{Y})}{P(c_1|\mathbf{Y})}, \tag{5}$$

where

$$P(c_1|\mathbf{Y}) \geq P(c_2|\mathbf{Y}) \geq ... \geq P(c_C|\mathbf{Y}). \tag{6}$$

When most of the probability is given to c_1, C_{SL} is close to 1, when the probabilities are roughly equation between classes, C_{SL} is close to 0. To combine the individual predictions with the joint probabilities, the side-lobe ratio is computed for each and normalized to sum to unity. This gives the weights

$$w_j = \frac{C_{SL}^{j}}{C_{SL}^{j} + C_{SL}^{i\neq j}}, \tag{7}$$

and

$$w_{i\neq j} = \frac{C_{SL}^{i\neq j}}{C_{SL}^{j} + C_{SL}^{i\neq j}}. \tag{8}$$

The final prediction output from the j^{th} agent is then given as a vector of posterior probabilities for each class label $\mathbf{P}_j^f$, where

$$P_j^f[k] = P_j(c=k|\mathbf{y}_j)w_j + P_{i\neq j}(c=k|\mathbf{Y}_{i\neq j})w_{i\neq j}. \tag{9}$$

Using these final predictions from each agent, the algorithm again computes the joint posterior over all M agents as

$$P^*(c=k|\mathbf{Y}) = \prod_{j=1}^{M} P_j^f[k]. \tag{10}$$

The class is then selected as the one with the highest, final posterior probability

$$c^* = \underset{c}{\text{argmax}}\ P^*(c|\mathbf{Y}). \tag{11}$$

The key aspect of this algorithm is that all feature extraction and classification is performed on single-view data, and therefore does not depend on the manner in which views are collected, or even the modality used to collect data from each view. Therefore, the fusion component of the algorithm can be applied to new data sets without retraining provided that at least one view from each modality is available for training.

3. Experimental Analysis

3.1. Data Collection

Data were collected using a laboratory multiview scattering apparatus (Fig. 1) [22]. The system consisted of eight acoustic receivers and two video cameras that were all focused on a single location in the tank. The field of view was roughly 500 mL. Live copepods and mysids were pumped through the FOV while the system recorded synchronized acoustic and video data at a rate of 10 Hz. Details of the laboratory system are given in [22]. Transmit signals were linear frequency-modulated (LFM) chirps ranging from 1.5 to 2.5 MHz. These signals were windowed with a cosine-squared envelope yielding an effective bandwidth of 500 kHz.

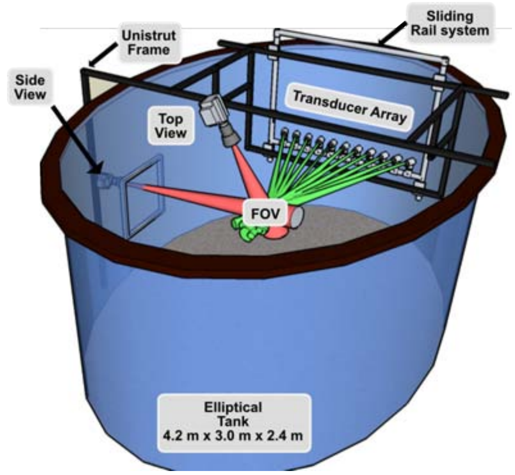


Figure 1: Drawing of the scattering apparatus showing the transducer array, sliding rail system for moving the array in and out of the water, unistrut frame, top- and side-view cameras, and elliptical tank.

Examples of zooplankton data are shown in Fig. 2. It can be seen that there are strong difference between the image of the copepod and mysid. Acoustic data show significant differences as well, however there is clearly less discrimination power in acoustic data than video data for these particular animals.

3.2. Acoustic Features

Acoustic features were computed from the echo envelope on each receiver. The echo envelope was estimated by matched filtering the raw pressure signal [23], and using a Hilbert transform [24] to extract the envelope. Examples of echo envelopes for copepods and mysids are shown in Fig. 2c.

Copepods are typically smaller and more spherical in body shape than mysids. The peak echo amplitude and echo duration are related to animal scattering cross-section and animal thickness, respectively. Let the peak of the echo on the j^{th} receiver be $\mathcal{P}_j$ and the time that the peak occurs be $T_{\mathcal{P}_j}$. The echo duration was then defined as $W_j = T^+_{\mathcal{P}_j} - T^-_{\mathcal{P}_j}$, where $T^+_{\mathcal{P}_j}$ is the time when the echo drops to 10% of the peak value, and $T^-_{\mathcal{P}_j}$ is the time when the echo rises to 10% of the peak value. The acoustic feature vector is then

$$\mathbf{y}_j = (W_j, \mathcal{P}_j)^T, \tag{12}$$

As can be seen in Figs. 3a and 3b, both the echo duration, and echo peak are significantly different between copepods vs. mysids.

3.3. Optical Features

There has been significant research in classifying zooplankton from high-resolution optical images [3, 5, 4]. Image features such as binary region properties [25], moment invariants and granularity [26], and biological-inspired shape semantics [27] have been used in the past. In these cases many image features are computed from a single animal (typically requiring high-resolution images) and used to differentiate between taxa. Here, an alternative approach is taken that is aimed at systems with very low resolution where small scale features of animals are not discernible.

An image of the animal was segmented from background using a unimodal Gaussian model [28]. Let the segmented image of the animal be $I(x_1, x_2)$. The shape of the animal was quantified by the major and minor axis lengths of $I(x_1, x_2)$. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the vectors of coordinates for which $I(x_1, x_2)$ is non zero. Then, let R define the sample covariance matrix between $\mathbf{x}_1$ and $\mathbf{x}_2$. The axis lengths were estimated using the eigenvalue decomposition [29] of R, $\text{Diag}(\lambda_1, \lambda_2) = Q^T R Q$, and optical features were selected to be the two eigenvalues

$$\mathbf{y} = (\lambda_1, \lambda_2)^T. \tag{13}$$

Example feature distributions for acoustic and video data show that both modalities offer some separation between classes (3). However, It can be seen that optical features offer much better separation between classes than acoustic features, and there is significant variability between the discrimination power of different acoustic views.

4. Classification Results

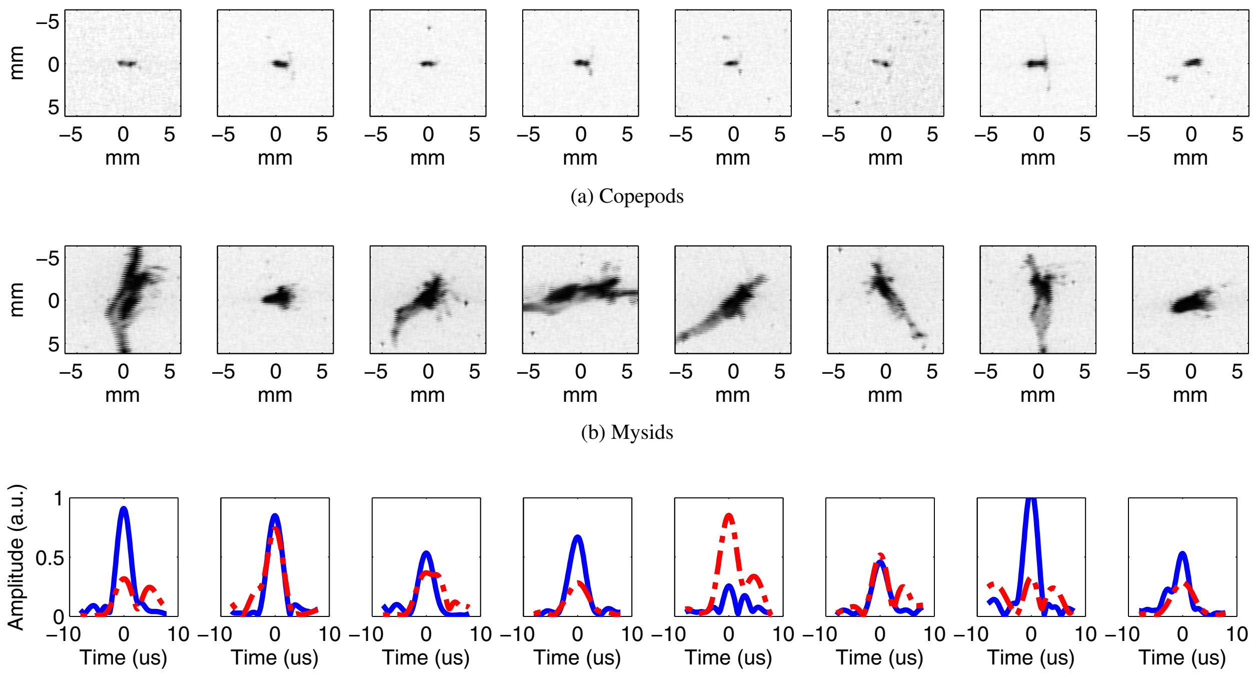
Classifier performance was evaluated as the average probability of error computed using 5-fold cross-validation. For each fold, 160 training samples and 40 testing samples were used. Standard errors of the cross-validation estimate were computed using the adjusted variance estimate [30] of the form

$$\hat{\sigma}^2_{CV} = \left(\frac{1}{F} + \frac{p}{1-p}\right)\text{Var}[x], \tag{14}$$

where $p = \frac{1}{F+1}$, F is the number of folds in the cross-validation, and $Var[x]$ is the sample variance. For the purpose of comparing the performance with different sets of views, let A_j, and V_j denote an acoustic and optical view as defined above. The set A_{1-4} then means that acoustic views 1 through 4 were used, and likewise, $V_{1,2}$ means that video views 1 and 2 were used.

4.1. Performance for Single-View Classification

Performance was found to varying significantly as a function of both acoustical and optical views (Fig. 4a).

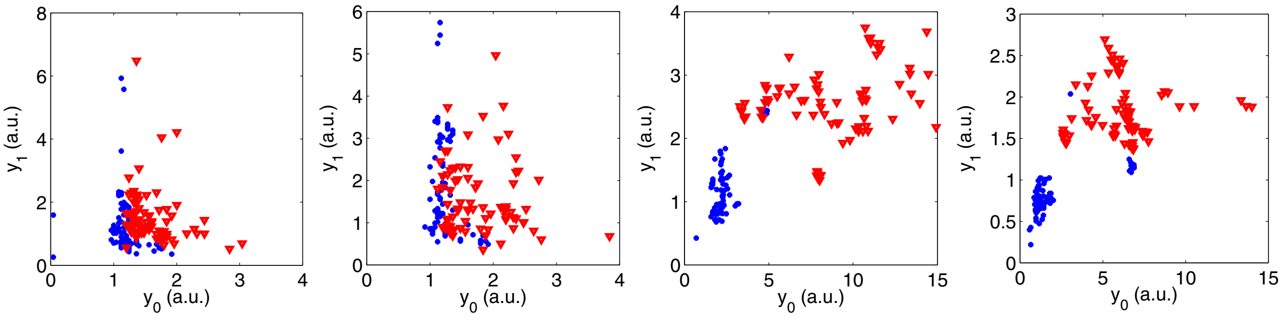


(a) Copepods

(b) Mysids

(c) Acoustic Echo Envelope

Figure 2: Example video and acoustic data from the laboratory system. Video frames for copepods (a) and mysids (b) are shown in the top two rows. Acoustic data for copepods (solid) and mysids (dashed) are shown in the bottom row (c).



(a) Acoustic View 1

(b) Acoustic View 3

(c) Top view camera

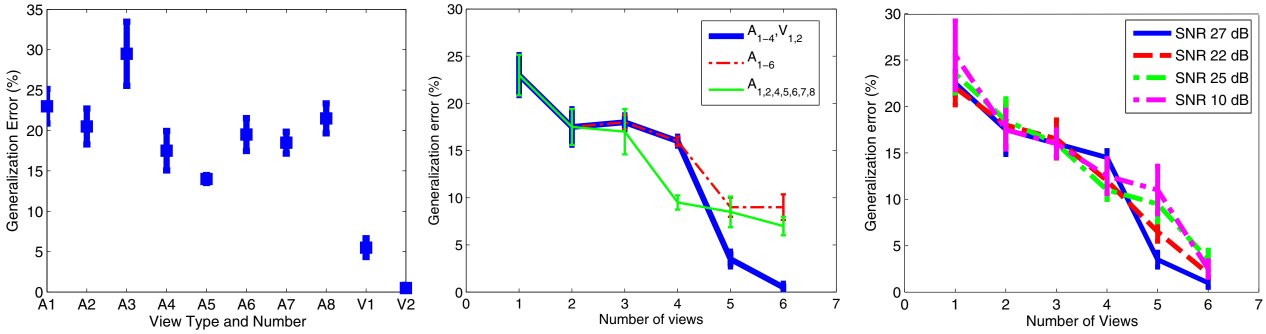
(d) Side view camera

Figure 3: Example feature distributions for the first (a) and third (b) acoustic views and the top- (c) and side-view (d) cameras. Triangles denote mysid features and circles denote copepod features. Note that the smaller size of copepods vs. mysid can be seen in both acoustic and optical features.

However, it can be seen clearly that the performance using optical views is dramatically better than with acoustic views. This results from the strong difference in the appearance of copepods vs. mysids in video data (Fig. 2). Optical features effectively capture these differences (Figs. 3d and 3c).

4.2. Fusion Performance vs. Number of Views

Two different cases were considered: (1) fusing both acoustic and optical features and (2) fusing a subset of acoustic features (Fig. 4b). It can be seen that adding views significantly reduces classification error in all cases. However, it is clear that the features computed from video data are much more discriminant that acoustic features. It is im-



(a) Single-View Classification

(b) Multiview Classification

(c) Performance for Varied SNR

Figure 4: (a) Algorithm performance for single-view classification. The type of view is listed on the horizontal axis where A_j denotes the j^{th} acoustic view, and V_j denotes the j^{th} optical view. (b) Algorithm performance vs. the number of views. The thick, solid curve shows the case where four acoustic views and two optical views were combined in the order: $A_{1-4}, V_{1,2}$. The thinner dashed curves show different combinations of acoustic views (A_{1-6} and $A_{1,2,4,5,7,8}$). The optical views offer significantly better discrimination for this problem. (c) Algorithm performance using A_{1-4} and $V_{1,2}$ as a function of peak Signal to Noise Ratio at the output of the matched filter.

portant to note that the fusion algorithm does not know *a priori* that the optical features are more discriminant. It discovers this by computing the confidence of the classifications produced using video data alone and then puts more weight on these predictions and ignores predictions from the other acoustic views. This demonstrates a key advantage of the fusion algorithm.

4.3. Performance vs. Signal to Noise Ratio

One of the challenging aspects of remotely classifying plankton is receiving enough signal to detect the presence of the animal and estimate features at ranges greater than 1 meter. In this study, acoustic and optical features were selected based on their robustness to low SNR data. Classifier performance is plotted as a function of SNR ranging from the default SNR in the lab of 27 dB down to 10 dB (Fig. 4c). It can be seen that the performance is relatively unaffected by reductions in SNR of nearly 20 dB. However, note that as the SNR of video data is reduced, the performance decreases more than for an equivalent reduction in SNR for acoustic data.

5. Conclusions

The fusion algorithm is general, offers dramatic improvements in accuracy without *a priori* knowledge of which modalities are most accurate, and has demonstrated good performance for zooplankton classification considered here, and also fish classification [31]. Once single-view features have been classified, the fusion process is applicable to any modality and any number of views. This is a key advantage in cases where the number of views (and the type of view) are unknown during testing. Only one data set per modality is required during training. In cases where *a priori* information about the fidelity of each view as available, the final fusion step [equation (10)] could be modified to weight each view based on its expected accuracy. One possible limitation of the algorithm is its dependence on confidence weighting for fusion. The fusion algorithm will not work well in cases where the sidelobe ratio is not dependent on the accuracy of the of the classifier output. This case is likely to occur when single-view error rates are nearly as bad as random guessing.

Although the results presented here are promising, the limited amount of data restricts the evaluation of the algorithm to a scenario in which video data is sufficient to yield nearly perfect classification of these animals. An important future direction is the deployment of the multiview system in different environments in which it can be evaluated on more diverse plankton groups and more challenging conditions. A multiview field system is currently under development, and will be tested later this year.

The multimodal nature of the fusion algorithm makes it applicable to a wide range of problems in computer vision such as multimodal scene understanding and human activity analysis [13]. A future direction of research will be to apply the algorithm in detecting and classifying events in rooms outfitted with multiple audio and video sensors, and to multimodal vehicle data streams.

References

[1] P. Wiebe and M. Benfield, “From the hensen net twoard four-dimensional biological oceanography,” *Prog. Oceanogr.*, vol. 1, pp. 130–138, Jan 2003.

[2] M. Blaschko, G. Holness, M. Mattar, D. Lisin, P. Utgoff, A. Hanson, H. Schultz, and E. Riseman, “Automatic in situ identification of plankton,” in *IEEE Workshops on Application of Computer Vision.*, vol. 1, Jan. 2005, pp. 79–86.

[3] Q. Hu and C. Davis, “Automatic plankton image recognition with co-occurrence matrices and support vector machine,” *Mar. Eco.-Prog. Ser.*, vol. 295, pp. 21–31, June 2005.

[4] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, “Active learning to recognize multiple types of plankton,” *J. Mach. Learn. Res.*, vol. 6, pp. 589–613, Apr. 2005.

[5] J. L. Bell and R. R. Hopcroft, “Assessment of zooimage as a tool for the classification of zooplankton,” *J. Plankton Res.*, vol. 30, pp. 1351–1367, Dec. 2008.

[6] D. Edgington, D. Cline, D. Davis, I. Kerkez, and J. Mariette, “Detecting, tracking and classifying animals in underwater video,” in *OCEANS 2006*, Sep. 2006, pp. 1–5.

[7] A. G. Cabreira, M. Tripode, and A. Madirolas, “Artificial neural networks for fish-species identification,” *ICES J. Mar. Sci.*, vol. 66, pp. 1119–1129, July 2009.

[8] E. Rogers, G. Fleischer, P. Simpson, and G. Denny, “Broadband fish identification of laurentian great lakes fishes,” *IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, pp. 1430–1434, Sept. 2004.

[9] L. Martin, T. Stanton, P. Wiebe, and J. Lynch, “Acoustic classification of zooplankton,” *ICES J. Mar. Sci.*, vol. 53, pp. 217–224, 1996.

[10] R. K. Cowen and C. M. Guigand, “In situ ichthyoplankton imaging system (isiis): system design and preliminary results,” *Limnol. Oceanogr.-Meth.*, vol. 6, pp. 126–132, Feb. 2008.

[11] J. Jaffe, M. Ohman, and A. Derobertis, “Oasis in the sea: measurement of the acoustic reflectivity of zooplankton with concurrent optical imaging,” *Deep-Sea Res. Pt. II*, vol. 45, pp. 1239–1253, 1998.

[12] E. A. Broughton and R. G. Lough, “A direct comparison of mocness and video plankton recorder zooplankton abundance estimates: Possible applications for augmenting net sampling with video systems.” *Deep-Sea Res. Pt. II*, vol. 53, pp. 2789–2807, 2006.

[13] S. Shivappa, M. M. Trivedi, and B. D. Rao, “Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms,” in *Computer Vision and Pattern Recognition Workshop*, June 2009, pp. 107–114.

[14] S. Krotosky and M. Trivedi, “Person surveillance using visual and infrared imagery,” *IEEE T. Circ. Syst. Vid.*, vol. 18, pp. 1096–1105, Aug. 2008.

[15] S. Y. Cheng, S. Park, and M. M. Trivedi, “Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis,” *Comput. Vis. Image. Und.*, vol. 106, pp. 245 – 257, May 2007.

[16] J. S. Jaffe, “Using multiple-angle scattered sound to size fish swim bladders,” *ICES J. Mar. Sci.*, vol. 63, pp. 1397–1404, 2006.

[17] P. L. D. Roberts and J. S. Jaffe, “Multiple angle acoustic classification of zooplankton,” *J. Acoust. Soc. Am.*, vol. 121, pp. 2060–2070, Apr. 2007.

[18] M. Azimi-Sadjadi, D. Yao, Q. Huang, and G. Dobeck, “Underwater target classification using wavelet packets and neural networks,” *IEEE T. Neural. Networ.*, vol. 11, pp. 784–794, May 2000.

[19] J. Cartmill, N. Wachowski, and M. R. Azimi-Sadjadi, “Buried underwater object classification using a collaborative multiaspect classifier,” *IEEE J. Ocean. Eng.*, vol. 34, pp. 32–44, Jan. 2009.

[20] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Min. Knowl. Disc.*, vol. 2, pp. 121–167, 1998.

[21] B. T. Morris and M. M. Trivedi, “Learning, modeling, and classification of vehicle track patterns from live video,” *IEEE T. Intell. Transp.*, vol. 9, pp. 425–437, Sept. 2008.

[22] P. L. D. Roberts and J. S. Jaffe, “Classification of live, untethered zooplankton from observations of multiple-angle acoustic scatter,” *J. Acoust. Soc. Am.*, vol. 124, pp. 796–802, Aug. 2008.

[23] D. Chu and T. Stanton, “Application of pulse compression techniques to broadband acoustic scattering by live individual zooplankton,” *J. Acoust. Soc. Am.*, vol. 104, pp. 39–55, July 1998.

[24] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Prentice-Hall, 1998.

[25] S. Iwamoto, D. M. C. Jr., and M. M. Trivedi, “Reflics: Realtime flow imaging and classification system,” *Mach. Vison Appl.*, vol. 13, pp. 1–13, 2001.

[26] X. Tang, F. Lin, S. Samson, and A. Remsen, “Binary plankton image classification,” *IEEE J. Ocean. Eng.*, vol. 31, pp. 728–735, July 2006.

[27] H. Zhou, C. Wang, and R. Wang, “Biologically-inspired identification of plankton based on hierarchical shape semantics modeling,” in *Int. Conf. Bioinformatics and Biomedical Engineering*, May 2008, pp. 2000–2003.

[28] M. Piccardi, “Background subtraction techniques: a review,” in *Int. Conf. on Systems, Man and Cybernetics*, vol. 4, Oct. 2004, pp. 3099–3104.

[29] T. K. Moon and W. C. Stirling, *Mathematical methods and algorithms for signal processing*. Prentice Hall, 2000.

[30] M. Markatou, H. Tian, S. Biswas, and G. Hripcsak, “Analysis of variance of cross-validation estimators of the generalization error,” *J. Mach. Learn. Res.*, vol. 6, pp. 1127–1168, July 2005.

[31] P. L. D. Roberts, “Multi-view, broadband, acoustic classification of marine animals,” Ph.D. dissertation, University of California, San Diego, May 2009.