

Binary Plankton Image Classification Using Random Subspace

Feng Zhao, Xiaoou Tang, and Feng Lin
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{fzhao0, xtang, flin0}@ie.cuhk.edu.hk

Scott Samson and Andrew Remsen
College of Marine Science
University of South Florida
Saint Petersburg, Florida, U. S. A.
{samson, aremsen}@marine.usf.edu

Abstract— In this paper, we implement a random subspace based algorithm to classify the plankton images detected in real time by the Shadowed Image Particle Profiling and Evaluation Recorder. The difficulty of such classification is compounded because the data sets are not only much noisier but the plankton are deformable, projection-variant, and often in partial occlusion. In addition, the images in our experiments are binary thus are lack of texture information. Using random sampling, we construct a set of stable classifiers to take full advantage of nearly all the discriminative information in the feature space of plankton images. The combination of multiple stable classifiers is better than a single classifier. We achieve over 93% classification accuracy on a collection of more than 3000 images, making it comparable with what a trained biologist can achieve by using conventional manual techniques.

I. INTRODUCTION

Plankton including phytoplankton and zooplankton form the base of the food chain in the ocean and are a fundamental component of marine ecosystem dynamics. The rapid mapping of plankton abundance together with taxonomic and size composition can help the oceanographic researchers understand how climate change and human activities affect marine ecosystems.

Earlier researchers investigated the temporal and spatial variability in plankton abundance and composition by manually counting the samples collected using traditional methods (e.g., towed nets, pumps, and Niskin bottles), which is laborious and time consuming. To improve sampling efficiency, some new instruments such as the Video Plankton Recorder (VPR) [1], the HOLOMAR underwater holographic camera system [2], and the Shadowed Image Particle Profiling and Evaluation Recorder (SIPPER) [3] have been developed to continuously sample magnified plankton images in the ocean.

The experimental data sets in this work come from the SIPPER system recently developed by University of South Florida. The SIPPER images differ from those used for most previous research in four aspects: 1) the images are much noisier, 2) the objects are deformable and often partially occluded, 3) the images are projection variant, i.e., the images are video records of 3D objects in arbitrary positions and orientations, and 4) the images in our experiments are binary thus are lack of texture information.

Fig. 1 shows some typical examples to illustrate the diversity of the SIPPER images. To deal with these difficulties, we combine the general features [4] (e.g., moment invariants [5], Fourier descriptors [6], and granulometric features [7]) with some specific features [8] (e.g., circular projections, boundary smoothness, and object density) to form a more complete description of the binary plankton patterns. To remove redundancy and reduce noise, we use the Principle Component Analysis (PCA) to compact the combined feature vectors, with the eigenvectors corresponding to small eigenvalues removed in the PCA subspace [4][8]. Since these eigenvectors may encode some useful information for recognition, their removal may introduce a loss of discriminative information.

To solve this problem, we propose an approach using random subspace [9]. The approach has been shown to be very effective for face recognition [11]. In the random subspace method, a number of low-dimensional subspaces are generated by randomly sampling from the original high-dimensional feature space. Finally, multiple classifiers constructed in the random subspaces are combined to make a powerful decision [10]. Using random sampling, the constructed classifiers are stable and multiple classifiers cover nearly the entire feature space without losing much discriminative information. Thus, good performance can be achieved. The experiments on seven classes of more than 3000 binary plankton images clearly demonstrate the efficiency and superiority of our algorithm.

II. FEATURE EXTRACTION

In order to form a more complete description of the binary plankton patterns, we combine the general features such as moment invariants, Fourier descriptors (FD and filled FD), and granulometries with some specific features such as circular projections (CMS, P_1 , P_2 , filled P_1 , and filled P_2), boundary smoothness, and object density. In this work, we add three types of structure elements (square, disk, and rhombus) of increasing sizes to compute the granulometric features since granulometries are relatively robust to noise, occlusion, and projection directions. All the extracted features are translation, scale, and rotation invariants. A brief description of them is shown in Table I. Refer to [4]-[8] for details.

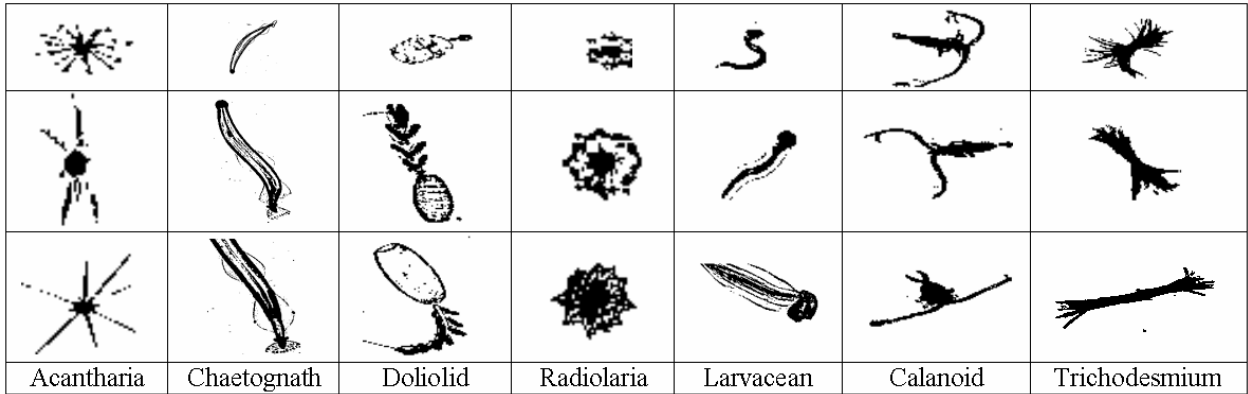


Fig. 1. Examples of seven plankton classes. (All images are rescaled for display purpose.)

III. RANDOM SAMPLING ON COMBINED FEATURE VECTOR

The combined feature vector has a high dimensionality. Normally the high-dimensional feature vector is projected to a low dimensional feature space using PCA to avoid classifier-overfitting problem due to the relatively small training set. In order to construct a stable classifier, the eigenvectors with small eigenvalues are usually removed in the PCA subspace. However, eigenvalue is not an indicator of the feature discriminability. Their removal may introduce a loss of useful discriminative information. To solve this problem, we randomly sample a small subset of features to reduce the discrepancy between the training set size and the feature vector length. Using such a random sampling method, we can construct multiple stable classifiers. We then combine these classifiers to construct a more powerful classifier that covers most of the feature space, so less discriminative information is lost.

As illustrated in Fig. 2, our random subspace based approach is designed as follows.

At the training stage,

1. Apply the first-stage PCA to all the long feature vectors ($LV_i, i=1, \dots, n_l$) such as P₁, P₂, filled P₁, filled P₂, CMS, FD, and filled FD to compute their eigenvectors U_i and eigenvalues λ_i , respectively. The eigenvectors with larger eigenvalues that preserve 98% of the total energy are preserved. The eigenvectors with very small eigenvalues mostly containing noise are removed.

TABLE I. FEATURE DESCRIPTION.

Features	Feature length
Boundary smoothness	1
Object density	1
Moment invariants	7
Granulometries	21
P ₁	50
P ₂	50
Filled P ₁	50
Filled P ₂	50
CMS	180
FD	180
Filled FD	180

2. For each training sample, project every long feature vector LV_i to the respective eigenvectors (PCA subspace) using $W_i = U_i^T (LV_i - m_i)$, where m_i is the mean of all the i_{th} long feature vectors.
3. Combine W_i ($i=1, \dots, n_l$) with the short feature vectors ($SV_j, j=1, \dots, n_s$) such as boundary smoothness, object density, moment invariants, and granulometries to form a new combined feature vector.
4. Normalize every component of the combined feature vector to the same scale.
5. Apply the second-stage PCA to the combined feature vectors of all the training samples. The first M largest eigenvectors $U_t = \{u_1, \dots, u_M\}$ that preserve 99% of the total energy are selected as candidates to construct the random subspaces.
6. Generate K random subspaces $\{S_n\}_{n=1}^K$. Each random subspace S_i is spanned by $N_0 + N_1$ dimensions. The first N_0 dimensions are fixed as the N_0 largest eigenvectors in U_t and the other N_1 dimensions are randomly selected from the remaining $M - N_0$ eigenvectors in U_t .
7. Construct K classifiers $\{C_n\}_{n=1}^K$ from the corresponding K random subspaces.

At the recognition stage,

1. For each testing image, project its long feature vectors to their respective PCA subspaces.
2. Combine the projected feature vectors with the short feature vectors to form a new combined feature vector.
3. Normalize every component of the combined feature vector to the same scale.
4. Project the normalized feature vector to the K random subspaces and fed them to the K classifiers in parallel.

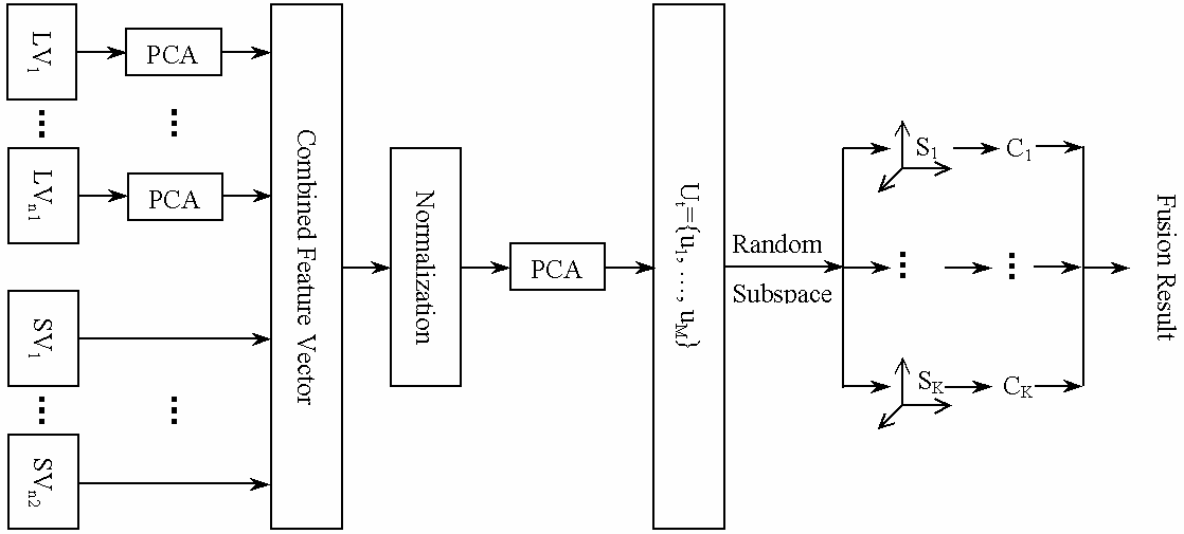


Fig. 2. Multi-feature multi-classifier recognition system using random sampling.

- Combine the outputs of the K classifiers using a fusion rule to make the final decision.

Compared with the traditional random subspace method that samples the original feature vector directly, our algorithm samples in the PCA subspace. The high dimension of the feature space is first greatly reduced without losing discriminative information. After doing PCA projection, the features on different eigenvectors are more independent. Better accuracy can be achieved if different random subspaces are more independent from one another.

Second, the random subspace is not completely random. The random subspace dimension is fixed as $N_0 + N_1$, which is determined by the training set to make a single classifier stable. The individual classifier constructed in each random subspace has a satisfactory accuracy since the first N_0 dimensions encode much information. In addition, the N_1 random dimensions cover most of the remaining small eigenvectors. Thus, our approach makes use of nearly all the discriminative information in the feature space.

Third, we use the first-stage PCA to compact every long feature vector by removing the redundant information. The noise within every long feature vector is removed as well. We then use the second-stage PCA to compact the combined feature vector by removing the correlative information among different types of features since they are all based on the same object shape. Furthermore, because the original feature vectors have different length and scale, the scale of the projected long feature vectors and the short feature vectors can be very different. One may overwhelm the other. We normalize all types of features in the combined feature vector to the same scale. Significant improvement can be achieved after normalization.

IV. EXPERIMENTS

We conduct experiments on seven classes of 3119 binary plankton images including 131 Acantharia, 172 Chaetognath,

450 Doliolid, 485 Radiolaria, 529 Larvacean, 563 Calanoid, and 789 Trichodesmium. All the images were manually classified by marine scientists. In our experiments, the Gaussian minimum error classifier is adopted to classify the plankton images and the tenfold cross-validation scheme is employed to evaluate the performance of our algorithm.

Fig. 3 shows the average accuracy of a single classifier constructed in the PCA subspace using different number of largest eigenvectors. We observe that a single classifier has the best accuracy of 91.05% using the largest 25 eigenvectors $\{u_1, \dots, u_{25}\}$, which seems to be suitable to construct a stable classifier for our data sets. In the following experiments, we choose 25 as the dimension of the random subspaces to construct a number of individual classifiers.

In each random subspace, we fix the first 15 dimensions as the 15 largest eigenvectors, and randomly select the other 10 dimensions from the remaining eigenvectors in U_i . As illustrated in Fig. 4, the individual classifiers constructed in the random subspaces are comparable to the single classifier using the largest 25 eigenvectors and have similar accuracies since much information is contained in the first 15 largest eigenvectors $\{u_1, \dots, u_{15}\}$. The results also indicate that the other 10 eigenvectors with large eigenvalues $\{u_{16}, \dots, u_{25}\}$ are not more discriminative than those smaller eigenvectors. In addition, these individual classifiers are complementary to one another. Better performance can be achieved when they are combined. The result of combining 20 individual classifiers using majority voting is shown in Fig. 4. It demonstrates that the random subspace method has a superior performance than the results in [8].

Table II shows the confusion matrix of our approach from a tenfold cross-validation experiment. The overall average accuracy is 93.27%. We have over 90% accuracy on most plankton except for Acantharia and Chaetognath. The reason is that we have fewer samples of the two classes.

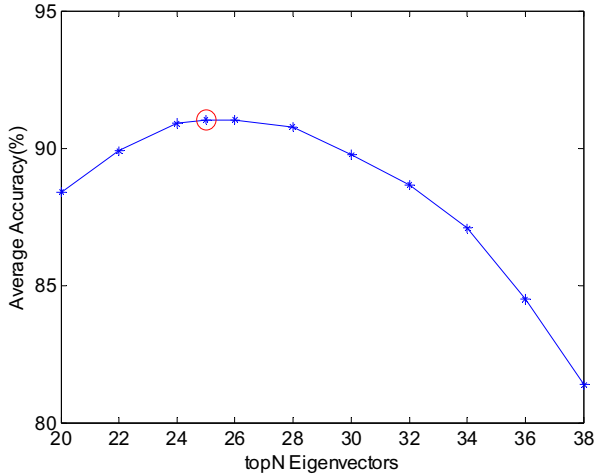


Fig. 3. Average accuracy of single classifier using different number of largest eigenvectors in the reduced PCA subspace. The red circle marks the best accuracy of 91.05%.

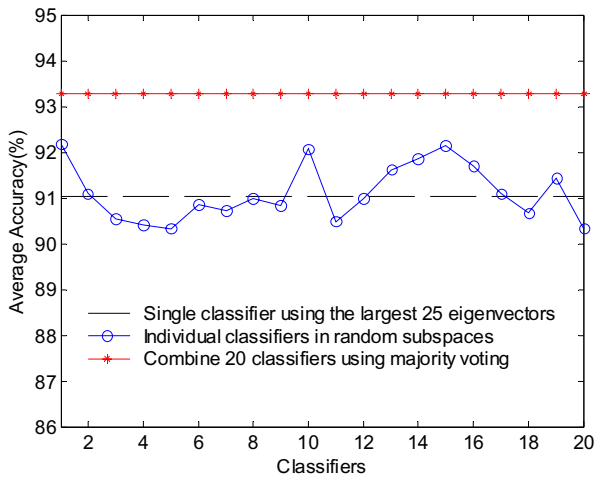


Fig. 4. Average accuracy of combining 20 individual classifiers constructed in the random subspaces using majority voting.

V. CONCLUSION

In this paper, we develop a random subspace approach for the integration of multiple features for binary plankton classification. Using random sampling, we construct a set of stable classifiers preserving nearly all the discriminative information in the high-dimensional feature space. The combination of multiple classifiers produces a better performance. Experimental results on a large data set show that the new algorithm can effectively classify plankton images with high accuracy acceptable for automatic plankton survey system.

TABLE II. CONFUSION MATRIX OF RANDOM SUBSPACE METHOD FROM A TENFOLD CROSS-VALIDATION ON 3119 SIPPER IMAGES. THE SEVEN PLANKTON TYPES ARE NUMBERED FROM 1 TO 7 IN TURN.

Class ID	1	2	3	4	5	6	7	Accuracy
1	115	0	4	2	0	10	0	87.79 %
2	0	147	15	0	8	0	2	85.47 %
3	0	5	416	19	8	0	2	92.44 %
4	16	0	16	448	0	0	5	93.37 %
5	1	3	25	0	491	1	8	92.82 %
6	4	0	0	0	3	544	12	96.63 %
7	0	1	9	10	10	11	748	94.80 %
Average Accuracy								93.27 %

ACKNOWLEDGMENT

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, and by the Office of Naval Research under grants N00014-02-1-0266 and N00014-96-1-5020.

REFERENCES

- [1] C.S. Davis, S.M. Gallager, N.S. Berman, L.R. Haury, and J.R. Strickler, "The Video Plankton Recorder (VPR): design and initial results," *Arch. Hydrobiol. Beith.*, vol. 36, pp. 67-81, 1992.
- [2] J. Watson, G. Craig, V. Chalvidan, J.P. Chambard, A. Diard, G.L. Foresti, B. Forre, S. Gentili, P.R. Hobson, R.S. Lampitt, P. Maine, J.T. Malmo, H. Nareid, G. Pieroni, S. Serpico, K. Tipping, and A. Trucco, "High resolution *in situ* holographic recording and analysis of marine organisms and particles (HOLOMAR)," *IEEE Int'l Conf. OCEAN'98*, pp. 1599-1604, 1998.
- [3] S. Samson, T. Hopkins, A. Remsen, L. Langebrake, T. Sutton, and J. Patten, "A system for high-resolution zooplankton imaging," *IEEE J. Ocean. Eng.*, vol. 26, no. 4, pp. 671-676, Oct. 2001.
- [4] X. Tang, W.K. Stewart, L. Vincent, H. Huang, M. Marra, S.M. Gallager, and C.S. Davis, "Automatic plankton image recognition," *Artificial Intelligence Rev.*, vol. 12, no. 1-3, pp. 177-199, 1998.
- [5] M.K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Information Theory*, vol. IT-8, pp. 179-187, 1962.
- [6] H. Kauppinen, T. Seppanen, and M. Pietikainen, "An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 201-207, Feb. 1995.
- [7] G. Matheron, *Random Sets and Integral Geometry*. New York: John Wiley and Sons, 1975.
- [8] X. Tang, F. Lin, S. Samson, and A. Remsen, "Binary image classification of plankton," *IEEE J. Ocean. Eng.*, in press.
- [9] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, Aug. 1998.
- [10] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, Mar. 1998.
- [11] X. Wang and X. Tang, "Random sampling LDA for face recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 259-265, 2004.