ADAPTIVE MULTIRESOLUTION TECHNIQUES
FOR SUBCELLULAR PROTEIN LOCATION CLASSIFICATION

Gowri Srinivasa [1], *Thomas Merryman* [2], *Amina Chebira* [1], *Jelena Kovačević*[1,2] *and Alexia Mintos*[3]

[1] Dept. of Biomedical Engineering,
[2] Dept. of Electrical and Computer Engineering,
Carnegie Mellon University, Pittsburgh, PA.
[3] Dept. of Science and Mathematics,
University of Virgin Islands, St Thomas, VI.

Fig. 1. Typical images from the 2D HeLa collection.

ABSTRACT

We propose an adaptive multiresolution (MR) approach for classification of fluorescence microscopy images of subcellular protein locations, providing biologically relevant information. These images have highly localized features both in space and frequency which naturally leads us to MR tools. Moreover, as the goal of the classification system is to distinguish between various protein classes, we aim for features adapted to individual proteins. These two requirements further lead us to adaptive MR tools. We start with a simple classification system consisting of Haralick texture feature computation followed by a maximum-likelihood classifier, and demonstrate that, by adding an MR block in front, we are able to raise the average classification accuracy by roughly 10%. We conclude that selecting features in MR subspaces allows us to custom-build discriminative feature sets for fluorescence microscopy images of protein subcellular location images.

1. INTRODUCTION AND MOTIVATION

A study of cell function and behavior is important in understanding the development of certain diseases. As proteins are integral components of cell function, it is critical to understand their properties such as structure and localization. Advances in biochemistry and biocellular microscopy make it possible to collect large amounts of data sets depicting protein localization within the cell. However, processing of these data sets is still mostly manual, making the process inefficient and error-prone. This raises the need for automated classification of subcellular protein location images. Murphy et al. [1] pioneered automation for classification of subcellular protein location patterns and demonstrated that they can be distinguished with reproducible specificity and sensitivity.

In classifying subcellular protein locations, a host of methods should be considered. In particular, MR techniques provide adaptive space-frequency localization enabling a customized description of nonstationarities present in the signal. In particular, the wavelet packet transform, an MR tool, has been successfully applied in biometric fingerprint recognition [2]. While these adaptive flavors of MR techniques have been explored for pattern classification in various domains [3], to date, they have not been tried on subcellular location patterns. As our protein images are beset with nonstationarities, we hypothesize that designing MR features adapted to the signal rather than just the space or frequency features alone would yield improvement in classification accuracy. Testing this hypothesis is the goal of this paper.
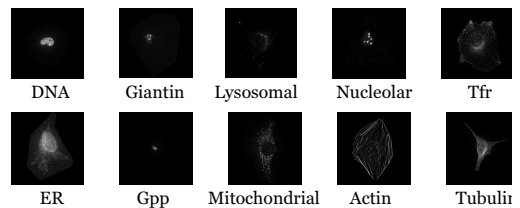
2. BACKGROUND

Location Proteomics. Automated classification in location proteomics (study of location for all proteins) developed by Murphy et al. [1], uses data sets depicting protein localization within the cell. These data sets are segmented into single-cell images, and after some preprocessing, features drawn from various categories—termed subcellular location features (SLF)—are computed. For example, in [4], Haralick texture features, which measure repetitive local patterns [5], have been used with some success in classification. Combining these texture features with a number of other feature types such as morphological features and Zernicke moments along with classical classifiers, lead to a classification accuracy of 86% [6]. On performing Stepwise-Discriminant-Analysis (SDA), to rank and select only the most discriminative features, a significant number of texture features were selected. To this were added the Gabor and Daubechies wavelet features to capture further information embedded in the images. Using a total of 174 features, followed by SDA and subsequent classification using an ensemble of classifiers, yielded an accuracy of 91.5% [7]. Our aim is to test whether adaptive MR features extract further information from those same data sets.

Formulation of a Signal Classification Problem. The design of an automated classifier can be formulated as designing a map from the *signal space* of protein localization images $\mathcal{X} \subset \mathbb{R}^{m \times m}$, to a *response space* $\mathcal{Y} \subseteq \{1, 2, ..., C\}$ of class labels. Thus, decision d is the map, $d \colon \mathcal{X} \mapsto \mathcal{Y}$ that associates an input image with a class label [3]. Since the signal space is overly redundant and direct manipulation of signals is prohibitive (typically $m = 512$), we must reduce dimensionality. Thus, we set up a feature space $\mathcal{F} \subset \mathbb{R}^k, k \leq m^2$, between the input space and the response space. The feature extractor θ is the map $\theta \colon \mathcal{X} \mapsto \mathcal{F}$, and the classifier ψ is the map $\psi \colon \mathcal{F} \mapsto \mathcal{Y}$. Further, the input space $\mathcal{X}$ is par-
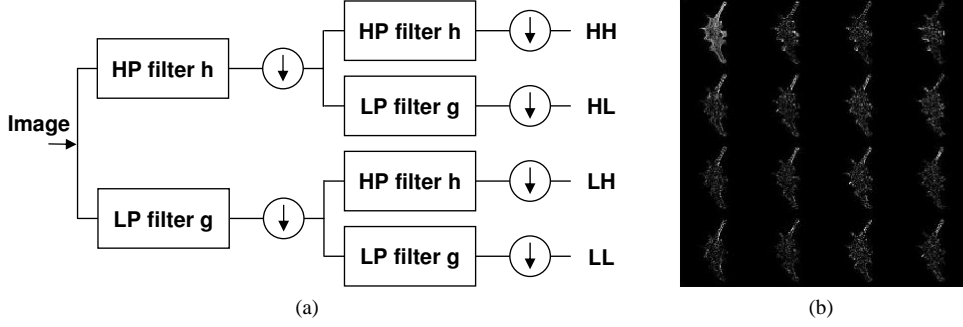
Fig. 2. (a) Two channel analysis filter bank for 2D signals. The filter h is a highpass filter and g is a lowpass filter. (b) A 2-level MR filter bank decomposition of Actin. The upper left image is the coarse representation. The other images are the detailed representations.

titioned into a *training set* $\mathcal{T} \subset \mathcal{X} \oplus \mathcal{Y} = \{x_{c,n}, y_c\}_{c=1,n=1}^{C,N}$ used to design d and a *testing set* $\mathcal{T}' = \{x'_{c,n'}, y'_c\}_{c=1,n'=1}^{C,N'}$, $\mathcal{T} \cap \mathcal{T}' \varnothing$, used to measure its performance. (N and N' refer to the number of images in each class of $\mathcal{T}$ and $\mathcal{T}'$ respectively.) The goal is to find a (θ, ψ) pair that maximizes the classification accuracy.

3. PROPOSED ALGORITHM

As we said, we aim to augment the features computed on the original image with any discriminative texture information in the sub-bands resulting from an adaptive MR decomposition. Those features are then modeled in order to measure their discriminative content (see Fig. 3). The goal is to ensure the augmented classifier does no worse than the classifier using the original image alone, and improve the average classification accuracy. A probabilistic measure of the discriminative power of a subband is used to weigh the decision obtained from a subband. We perform a convex combination of the decisions from the subbands to obtain a class label. The algorithm is split into two phases, the training phase and the testing phase, which we now describe (each step corresponds to a block in Fig. 3).

3.1. Training Phase

> *Input*: Training set $\mathcal{T}$.
> *Output*: A set G of Gaussian models for each class c and a weight vector ω.

We first choose the family of filters, that is, the bases h and g for the filter bank (see Fig. 2(a)), the number of levels L for the decomposition and the maximum number of clusters allowed per class, K, for the pdf modeling using the K-means algorithm.

Wavelet Tree: Conduct a full wavelet packet tree decomposition of L levels on the training images, that is, perform $\mathcal{W}: \mathcal{X} \mapsto \mathcal{V}$ where $\mathcal{W}$ is the projection of the image onto the transform domain and $\mathcal{V}$ is the set of subband representations of the training images. There are $S = (4^{L+1}-1)/3$ nonredundant subspaces corresponding to each input image, that is, $x_{c,n} = [\mathbf{t}_{c,n,1}, ..., \mathbf{t}_{c,n,S}]$, with each subband inheriting the class label y_c. ($t_{c,n,s}$ denotes the subband s of image n of class c.)

Feature Extraction: Compute F Haralick texture features on each subband, that is, $\mathcal{H}: \mathcal{V} \mapsto \mathcal{F}$; where $\mathcal{H}$ is the texture feature computation and $\mathcal{F}$ is the set of all features corresponding to the

various subbands of an image. An element of $\mathcal{F}$ is of the form $\mathbf{t}_{c,n,s,f}, 1 \leq f \leq F$.

K-Means and Gaussian Modeling: The training set is modeled as set of K_c F-variate Gaussian pdfs in each subspace as follows: for $s = 1, \ldots, S$, $i = 1, \ldots, N$, $c = 1, \ldots, C$, we perform the K-means clustering on $t_{c,n,s}$ for every $n \neq i$ to obtain at most K clusters in each of the C classes. We model each cluster, $\mathbf{G}_{c,i,s,k}$, as an F-variate Gaussian pdf with mean $\mu_{c,i,s,k}$ and variance $\sigma_{c,i,s,k}$. The probability of $\mathbf{t}_{c,i,s}$ belonging to a class j is computed as

$$
\begin{aligned}
P_j(\mathbf{t}_{c,i,s}) &= P(\mathbf{t}_{c,i,s} \in \text{class } j) \\
&= \frac{\sum_{k=1}^{K_j} l(\mathbf{t}_{c,i,s} \in \text{class } j, \text{cluster } k)}{\sum_{m=1}^{C} \sum_{k=1}^{K_m} l(\mathbf{t}_{c,i,s} \in \text{class } m, \text{cluster } k)},
\end{aligned}
$$

with

$$ l(t_{c,i,s} \in \text{class } m, \text{cluster } k) = $$
$$ \frac{1}{\sqrt{(2\pi)^F |\Sigma_{j,i,s,k}|}} \exp\{-\frac{1}{2}(\mathbf{t}_{c,i,s}-\mu_{j,i,s,k})^T \Sigma_{j,i,s,k}^{-1}(\mathbf{t}_{c,i,s}-\mu_{j,i,s,k})\}, $$

for $1 \leq j \leq C$ and $\Sigma_{j,i,s,k} = \text{diag}(\sigma_{j,i,s,k})$. Effectively, every element $t_{c,n,s}$ of the training set is now represented by a single probability vector $\mathbf{P}(t_{c,n,s})$ with $|(\mathbf{P}(t_{c,n,s})| = C$. The map from the *feature space* to the *probability space* is $\mathcal{M}: \mathcal{F} \mapsto \mathcal{P}$. Thus, our feature extractor $\theta = \mathcal{M} \circ \mathcal{H} \circ \mathcal{W}$. Note that the set G of Gaussian models is computed using all of the training images in $\mathcal{T}$.

Initialization of the Weight Vector: Define a decision vector $\mathbf{d} = [\sum_{n=1}^{N} d(\mathbf{t}_{1,n,s}), \ldots, \sum_{n=1}^{N} d(\mathbf{t}_{C,n,s})]$ for each subband of the training set, such that

$$
d(t_{c,n,s}) = \begin{cases} 1 & \text{if } \arg\max_c(\mathbf{P}(t_{c,n,s})) = y_c, \\ 0 & \text{otherwise.} \end{cases}
$$

We compute the weight for each subband as:

$$
\omega_s = \sum_{c=1}^{C} \sum_{n=1}^{N} d(t_{c,n,s})/(CN).
$$

We then form $\omega = [\omega_1, \ldots, \omega_S]$ and normalize so that $\omega = \omega/\sum_{s=1}^{S} \omega_s$, where ω_s gives the significance of the information contained in subband s.

Evolution of the Weight Vector: Compute the probability vector corresponding to each image as $\mathbf{P}(\mathbf{t}_{c,n}) = \sum_{s=1}^{S} \omega_s \mathbf{P}(t_{c,n,s})$. The overall decision the subbands reach is computed as

$$
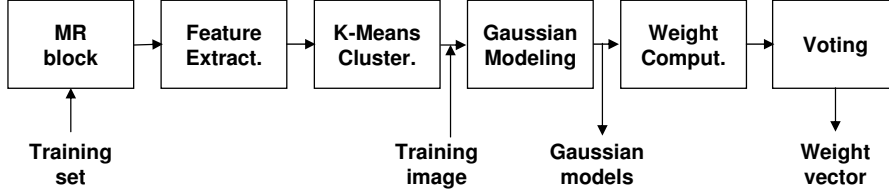d = \arg\max_c(\mathbf{P}(\mathbf{t}_{c,n})).
$$

Fig. 3. The training phase of our MR classifier.

If the class label d is correct, that is, $d = y_c$, the weights are unaltered; if not, then the weights are updated as

$$\omega_{s,new} = \begin{cases} (1 + \text{inc})\omega_{s,old} & \text{if subband } s \text{ contributes} \\ & \text{to a correct decision} \\ (1 - \text{inc})\omega_{s,old} & \text{otherwise,} \end{cases}$$

where $0 \leq \text{inc} < 1$. This step is repeated for η epochs over the entire training set. The weight vector ω is output at the end of this step. Thus, our classifier is the map $\psi \colon \mathcal{P} \mapsto \mathcal{Y}$ that selects the most likely class from a set of class membership likelihoods obtained as a convex combination of the decisions of the subbands.

3.2. Testing Phase

> **Input**: Testing set $\mathcal{T}'$, the set of Gaussian models G and the weight vector ω.
> **Output**: A set of decisions (class labels), one corresponding to each test image.

Preparation of Data: Decompose the data to L levels and compute Haralick texture features on each subband. Then, using the Gaussian model G, obtain the probability vector corresponding to each image $x_{n'}$, that is, perform feature extraction $\theta(\mathcal{X}')$ to obtain probability vectors $\mathbf{P}(\mathbf{t}'_{c,n',s})$.

Decision: Apply the weight vector ω to each of the image probability vectors $\mathbf{P}(\mathbf{t}'_{c,n'})$ and obtain the decision d that is a class label, using $\psi \colon \mathcal{P}' \mapsto \mathcal{Y}'$, where $d = \arg\max_c(\mathbf{P}(\mathbf{t}'_{c,n'}))$, $\mathbf{t}'_{c,n'}$ being the testing image n' from class c. A confusion matrix is computed from the decisions made on the set of testing images to evaluate the performance of the system.

3.3. Enhancing the System with PCA

Principal Component Analysis (PCA) is a method used to reduce the dimensionality of the feature vectors. It also allows us to express the features in a space spanned by an orthogonal set of vectors viz. eigenvectors. We use these two properties to our advantage. First, representing our feature vectors in their eigenspaces could help the clustering algorithm and certainly contribute to taking our models closer to our assumption of joint Gaussian vectors. Second, reducing the dimension of the vectors helps gain a deeper insight to the behavior of the data in the feature space. It also assists the classifier as there are fewer numbers characterizing each class. Based on the above considerations, we chose PCA over other techniques (such as SDA) for reducing the dimensionality of the feature vectors of our system.

We introduce the PCA block right after the feature computation step. We apply it on the individual feature vector from each

Average Classification Accuracy [%]		
	No PCA	PCA
No MR	69.0	81.8
MR	81.0	87.4

Table 2. Average accuracies across classes for $K = 10$. The MR results correspond to $\eta = 20$ iterations. Improvement in overall accuracy using a 2-level decomposition is 5.6% with PCA.

subband. To test its efficacy, we simply choose the same number of transformed features from each of the subbands. As PCA increases computational complexity, its use is contingent on a large-enough improvement in classification accuracy.

4. EXPERIMENTAL RESULTS

We used the 2D HeLa collection of cervical cancer cells with 50 single-cell images of size 512×512, in each class, courtesy of the MurphyLab [8]. There were 10 classes of subcellular location patterns obtained by marking the proteins Tubulin, Gpp130, Nucleolar, Giantin, Mitochondrial, ER, Lysosomal, Endosomal and Actin along with DNA. We partitioned our input set into a training set of 45 images and a testing set of 5 images in each class. We grew a full wavelet tree of $L = 2$ levels using the Haar basis. We performed a 10-fold cross validation and it is the average of these results that is reported here.

The main question we sought to answer was:

Does space-frequency adaptivity offered by the MR transform help?

That is, do subbands contain information useful for classification? The significant increase in classification accuracy (around 10%, see Table 1), demonstrates that subbands are indeed useful for classification.

Discussion. Having established that subbands indeed contain information useful for classification, how does the maximum number of clusters allowed influence the classification accuracy? In Fig. 4, we note that, irrespective of the value of K, we obtain a higher classification accuracy with decomposition than without. Further, the highest accuracy is obtained with $K = 5$. We are yet to interpret these results, that is, why $K = 5$ best describes our data set.

Another question is: how does η, the number of epochs (our stopping criterion) influence the accuracy? Again, Fig. 4 shows that the accuracies reach their maximum value fairly quickly and then oscillate in a finite and small range.

	Output of Classifier [%]										
	Tub	Gpp	Nuc	Gia	Mit	ER	L2	Tfr	Act	DNA	Accuracy
No MR	64	64	66	86	66	74	72	40	100	86	**71.8**
MR	74	84	98	90	68	80	86	48	100	94	**82.2**

Table 1. The accuracy per class, for $K = 5$. The MR results correspond to $\eta = 20$ iterations. Improvement in overall average accuracy using a 2-level decomposition is 10.4%.
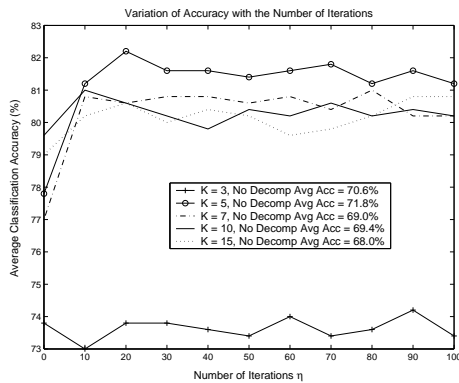

Fig. 4. Variation of the average classification accuracy with the number of iterations η used for weight training and different K. The legend reports the average accuracy for each K using the original images without decomposition.

As for PCA, we conclude that it is useful in improving the classification accuracy with or without MR (see Table 4). This increase may be explained as the advantage offered by expressing the features in an orthogonal representation, the eigenvectors, that brings our models closer to the underlying assumptions such as features' coordinate independence. As in the previous result, MR outperforms the nonMR.

We also tested the system with a set of morphological features. These features visually describe distinct aspects of the image as discerned by the human eye. The principal categories of morphological features are object (continuous portion of the fluorescent image), edge (boundary of the object), convex hull (a closed, convex contour of the object) and skeleton (a detailed grid of the fluorescent image) features. The results we obtained using 16 such features show an improvement of the classification accuracy from 53.6% without MR to 68.9% with MR (2-Level decomposition with $K = 10$ and $\eta = 20$), further establishing the contribution of MR.

A final comment: Since the K-means algorithm introduces some randomness in the modeling, we ran the experiments a number of times and found that the average classification accuracies are consistent and variation is small. We are also working on a more robust weight initialization procedure and a search routine for the stopping criterion to avoid fixing it apriori.

5. CONCLUSIONS AND FUTURE WORK

We motivated the use of an adaptive MR approach for subcellular protein classification. We obtained a substantial increase in the classification accuracy, suggesting that subbands do contain further information pertinent to classification and not included in the nondecomposed image alone. These results open up several venues for exploration. As our present system uses a very small number of features, the next step is combining different types of features, which may add to the discriminative strength of the algorithm. We may also explore different levels of decomposition and conduct a search on a library of bases for the filters used in the decomposition. We will also do that for the maximum number of clusters, to get a more adapted characterization on a specific data set. Finally, as the K-means clustering imposes a spherical topology which may not be natural to the data set, we will look into other nonparametric methods that do not impose such constraints.

6. REFERENCES

[1] R. F. Murphy, "Automated interpretation of protein subcellular location patterns–Implications for early cancer detection and assessment," *Ann. N.Y. Acad. Sci.*, pp. 1–8, March 2004.

[2] P. H. Hennings, J. Thornton, J. Kovačević, and B. V. K. Vijaya Kumar, "Wavelet packet correlation methods in biometrics," *Applied Optics, Special Issue on Biometric Recognition Systems*, vol. 44, pp. 637–646, February 2005.

[3] N. Saito and R. Coifman, "Local discriminant bases and their applications," *Math. Imaging Vision*, vol. 5, pp. 337–358, 1995.

[4] K. Huang, M. Velliste, and R. Murphy, "Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images," in *Proceedings of the SPIE*, vol. 4962, (San Jose, CA), pp. 307–318, 2003.

[5] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," in *IEEE Transactions on Systems, Man and Cybernetics*, 1979.

[6] R. Murphy, M. Velliste, and G. Porreca, "Robust classification of subcellular location patterns in fluorescence microscope images," in *Proceedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing*, pp. 67–76, September 2002.

[7] K. Huang and R. F. Murphy, "Boosting accuracy of automated classification of fluorescence microscope images for location proteomics," *BMC Bioinformatics*, 2004. In press.

[8] *http://murphylab.web.cmu.edu*. The Murphy Lab at Carnegie Mellon University, Pittsburgh, USA.