Multiple Features Based Two-stage Hybrid Classifier Ensembles for Subcellular Phenotype Images Classification

Bailing Zhang bailing.zhang@xjtlu.edu.cn
Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University,
Suzhou,215123, P.R.Chin

Tuan D. Pham t.pham@adfa.edu.au
School of Engineering and Information Technology,
The University of New South Wales,
Canberra, ACT 2600, Australia.

Abstract

Subcellular localization is a key functional characteristic of proteins. As an interesting ``bio-image informatics'' application, an automatic, reliable and efficient prediction system for protein subcellular localization can be used for establishing knowledge of the spatial distribution of proteins within living cells and permits to screen systems for drug discovery or for early diagnosis of a disease. In this paper, we propose a two-stage multiple classifier system to improve classification reliability by introducing rejection option. The system is built as a cascade of two classifier ensembles. The first ensemble consists of set of binary SVMs which generalizes to learn a general classification rule and the second ensemble, which also include three distinct classifiers, focus on the exceptions rejected by the rule. A new way to induce diversity for the classifier ensembles is proposed by designing classifiers that are based on descriptions of different feature patterns. In addition to the Subcellular Location Features (SLF) generally adopted in earlier researches, three well-known texture feature descriptions have been applied to cell phenotype images, which are the local binary patterns (LBP), Gabor filtering and Gray Level Co-occurrence Matrix (GLCM). The different texture feature sets can provide sufficient diversity among base classifiers, which is known as a necessary condition for improvement in ensemble performance. Using the public benchmark 2D HeLa cell images, a high classification accuracy 96% is obtained with rejection rate 21% from the proposed system by taking advantages of the complementary strengths of feature construction and majority-voting based classifiers' decision fusions.

Keywords: subcellular phenotype images classification; hybrid classifier; local binary patterns; Gabor filtering; Gray level co-occurrence matrix; support vector machine; multiple layer perceptron; random forest

1. INTRODUCTION

Eukaryotic cells have a number of subcompartments termed organelles, each of which contains a unique localization of proteins and hence different biochemical properties. Determining a protein's location within a cell is critical to understanding its function and to build models that capture and simulate cell behaviors. It has been shown that mislocalization of proteins correlates with several diseases that range from metabolic disorders to cancer [1], thus knowledge of the location of all proteins will be essential for early diagnosis of disease and/or monitoring of therapeutic effectiveness of drugs. Given that mammalian cells are believed to express tens of thousands of proteins, a comprehensive analysis of protein locations requires the development of an automated massive analysis method. If such analyses can be converted into high throughput ``location proteomics'' assays, the resulting information would help us to understand the functions, properties and distribution of proteins in cells, and how a protein changes its characteristics in response to drugs, diseases and various stages of the cell cycle.

The most widely used method for determining protein subcellular location is fluorescence microscopy, which combines fluorescence detection with high-powered digital microscopy. Advances in fluorescent probe chemistry, protein chemistry, and imaging techniques have made fluorescence microscopy a valuable method for determining protein subcellular locations [2,3]. Over the past decade, there has been much progress in the classification of subcellular protein location patterns from fluorescence microscope images. The pioneering contributions to this problem should be attributed to Murphy and his colleagues [4-8]. Machine learning methods such as artificial neural networks and Support Vector Machine (SVM) have been utilized for the predictive task of protein localization in conjunction with various feature extraction methods from fluorescence microscopy images. Most of the proposed approaches employed feature set which consist of different combinations of morphological, edge, texture, geometric, moment and wavelet features. For example, [5] used images of ten different subcellular patterns to train a neural network classifier, which has been shown to correctly recognize an average of 83% of the patterns.

In previous studies of subcellular phenotype images classification, classification accuracy was the only pursuit, aiming to produce a classifier with the smallest error rate possible. In many applications, however, reject option for classifiers by allowing for an extra decision expressing doubt is important. For instance, in early diagnosis of disease or monitoring of therapeutic effectiveness of drugs, it is more important to be able to reject an example of subcellular phenotype image when there is no sufficiently high degree of accuracy, since the consequences of misclassification are severe and scientific expertise is required to exert control over the accuracy of the classifier thus making reliable determination. Therefore, we are motivated to investigate the option of classification scheme with rejection paradigm to meet the desirable functionality of automated subcellular phenotype images classification whereby the system generates decisions with confidence larger than some prescribed threshold and transfers the decision on cases with lower confidence to a human expert. For the 2D HeLa images [5,6], evidence from many published works and our own extensive experiments confirmed that no single method of classification could achieve high classification accuracy for all localizations. It has become a consensus in machine learning community that an integrative approach by combining multiple learning systems often offer higher and more robust classification accuracy than a single learning system [19]. The so-called ensemble system that combines the outputs of several diverse classifiers or experts has been broadly applied and proven an efficient approach to improve the performance of recognition systems. The intuition is that the diversity in the classifiers allows different decision boundaries to be generated, which can be implemented by using different learning algorithms corresponding to different errors or by using different representations of the same input to make different features apparent and provide supplementary information.

As a typical multi-class classification issue, subcellular phenotype images classification involves two interweaved parts: feature representation and classification. Many of the off-the-shelf standard classifiers such as multiple layer perceptron can be directly applied together with different possible feature sets which are potentially useful for separating different classes of subcellular phenotype [32]. However, a multi-class subcellular phenotype images dataset is often featured with large intra-class variations and inter-class similarities, which poses serious problems for simultaneous multi-class separation using the standard classifiers. On the other hand, it is almost impossible to find a feature set that is universally informative for separating all classes simultaneously. A better alternative solution to the problem, therefore, is to train different classifiers on distinct feature sets to fit the different characteristics. In our study, three kind of texture feature representations were considered, together with the Subcellular Location Features (SLF) [5,7]. The three texture feature expressions are the local binary patterns (LBP) [12], Gabor filtering [17] and Gray Level Co-occurrence Matrix (GLCM) [18] . The LBP operator has been proved a powerful means of texture description, which is relatively invariant with respect to changes in illumination and image rotation, and computationally simple [13, 14]. Gabor filter is another widely adopted operator for texture properties description and has been shown to be very efficient in many applications [17]. The Gray Level Co-occurrence Matrix (GLCM) method is characterized by its capability of extracting second order statistical texture features when considering the spatial relationship of pixels and has been proved to be a promising method in many image analysis tasks. These kinds of texture features alone might, however, have limited power in describing the complex features from microscopy images related to the subcellular protein location patterns. This again strengthens our avocations to propose a two-stage classifiers system which cater for a design-based method to fuse the features from LBP, Gabor filter, GLCM and SLF in order to obtain an improved classification performance.

Our work follows the hybrid classification paradigm, which combines classifiers to yield more accurate recognition rates when different classifiers contributes partially with different features. Unlike relative works that combine different base classifiers (trained with same samples) for image recognition systems, we use an effective approach to utilize complementary texture information and provide sufficient diversity among base classifiers of ensemble. With the 2D HeLa images, a sample can be either classified or rejected. The objective of reject option is to improve classification reliability and leave the control of classification accuracy to human expert. Comparing with some earlier cascading classifier paradigms, our proposed system is composed of different classifiers each specializes with different set of features. In our implementation, one-vs-all SVMs are employed in the first stage to obtain high accuracy for easier inputs and reject a subset of class assignments which is harder or ambiguous. A second stage classifier ensemble consists of three different kind of multi-class classifiers working in parallel (random forest, neural networks and support vector machines) and the final decision is based on the majority voting for the final combination.

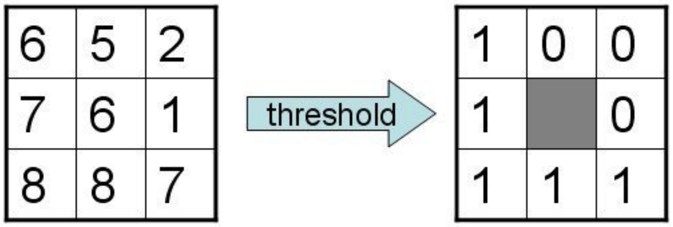
The paper is organized as follows. In Section 2, we introduce feature descriptions, including three texture descriptors LBP, Gabor filter and GLCM, together with the Subcellular Location Features (SLF). In Section 3, we elaborate the details of the proposed two-stage hybrid classification system. Experiments using the 2D HeLa images are provided in Section 4 and conclusion is outlined in Section 5.

2. FEATURE DESCRIPTIONS FOR CELL PHENOTYPE IMAGES

In order to automated analyse and classify microscopic cellular images, some kind of features have to be extracted to express the statistical characteristics in the image. And given two sets of sub-cellular localization images under differing experimental conditions, an efficient image feature can be used to evaluate if there is a statistically significant difference, even to the extent that visually indistinguishable images of distinct localizations may be differentiated [4] .The feature sets proposed in the literature include, for instance, morphological data of binary image structures, Zernike moments and edge information [5,6]. Use of a single technique for the extraction of diverse features in an image usually exhibits limited information description. Features extracted using different techniques can be combined in an attempt to enhance their description capability.

2.1 Local Binary Pattern

Local Binary Pattern (LBP) operator was introduced as a texture descriptor for summarizing local gray-level structure [12]. LBP labels pixels of an image by taking a local neighborhood around each pixel into account, thresholding the pixels of the neighborhood at the value of the central pixel and then using the resulting binary-valued image patch as a local image descriptor. In another word, the operator assigns a binary code of 0 and 1 to each neighbor of the neighborhoods. The binary code of each pixel in the case of 3x3 neighborhoods would be a binary code of 8 bits and by a single scan through the image for each pixel the LBP codes of the entire image can be calculated. Figure 1 shows an example of an LPB operator utilizing 3x3 neighborhoods.



6	5	2
7	6	1
8	8	7

1	0	0
1		0
1	1	1

Binary code = **11110001**
LBP = 1 + 16 +32 + 64 + 128 = **241**

Figure 1. Illustration of the basic LBP operator.

Formally, the LBP operator takes the form

$$\mathrm{LBP}(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n$$

where in this case n runs over the 8 neighbors of the central pixel c, i_c and i_n are the gray-level values at c and n, and s(u) is 1 if u ≥0 and 0 otherwise.

An useful extension to the original LBP operator is the so-called uniform patterns [12]. An LBP is ``uniform'' if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For example, 11100001 is a uniform pattern, whereas 11110101 is a non-uniform pattern. The uniform LBP describes those structures which contain at most two bitwise (0 to 1 or 1 to 0) transitions. Uniformity is an important concept in the LBP methodology, representing important structural features such as edges, spots and corners. Ojala et al. [12] observed that although only 58 of the 256 8-bit patterns are uniform, nearly 90 percent of all observed image neighbourhoods are uniform. We use the notation $\mathrm{LBP}^{u}_{P,R}$ for the uniform LBP operator. $\mathrm{LBP}^{u}_{P,R}$ means using the LBP operator in a neighborhood of P sampling points on a circle of radius R. The superscript u stands for using uniform patterns and labeling all remaining patterns with a single label. The number of labels for a neighbourhood of 8 pixels is 256 for standard LBP and 59 for $\mathrm{LBP}^{u}_{8,1}$.

A common practice to apply the LBP coding over an image is by using the histogram of the labels, where a 256-bin histogram represents the texture description of the image and each bin can be regarded as a micro-pattern. Local primitives which are coded by these bins include different types of curved edges, spots, flat areas, etc. The distribution of these patterns represents the whole structure of the texture. The number of patterns in an LBP histogram can be reduced by only using uniform patterns without losing much information. There are totally 58 different uniform patterns at 8-bit LBP representation and the remaining patterns can be assigned in one non-uniform binary number, thus representing the texture structure with a 59-bin histogram.

LBP scheme has been extensively applied in face recognition, face detection and facial expression recognition with excellent success, outperforming the state-of-the-art methods [13].

The methodology can be directly extended to microscopy image representations as outlined in the following. First, a microscopy image is divided into M small no-overlapping rectangular blocks R_0, R_1, …, R_M. On each block, the histogram of local binary patterns is calculated. The procedure can be illustrated by Figure 2. The LBP histograms extracted from each block are then concatenated into a single, spatially enhanced feature histogram defined as:

$$H_{ij} = \sum_{x,y} I(f_l(x,y) = i) \qquad i = 0, \ldots, L-1, \; j = 0, \ldots, M-1$$

where L is the number of different labels produced by the LBP operator and I(A) is 1 if A is true and 0 otherwise. The extracted feature histogram describes the local texture and global shape of microscopy images.

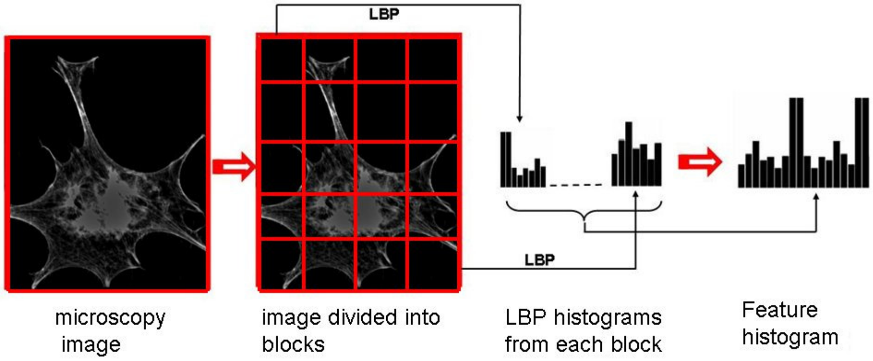


FIGURE 2: Feature extraction diagram for image recognition with local binary patterns.

LBP has been proved being a good texture descriptor with high extra-class variance and low intra-class variance. Recently, a number of variants of LBP have been proposed [15]. In [16], a completed modeling of the local binary pattern operator is proposed and an associated completed LBP (CLBP) scheme is developed for texture classification. In this scheme, a local region is represented by its center pixel and a local difference sign-magnitude transform. And the center pixels represent the image gray level and they are converted into a binary code by global thresholding. For many applications like face recognition, CLBP can offer better performance.

2.2 Gabor Based Texture Features

Gabor filters [17] have been used extensively to extract texture features for different image processing tasks. Image representation using Gabor filter responses minimises the joint space-frequency uncertainty. The filters are orientation- and scale-tunable edge and line detectors. Statistics of these local features in a region relate to the underlying texture information. The convolution kernel of Gabor filter is a product of a Gaussian and a cosine function, which can be characterized by a preferred orientation and a preferred spatial frequency:

$$g_{\lambda,\theta,\varphi}(x,y) = \exp(-\frac{(x'^2 + \gamma y'^2}{2\sigma^2}) \cos(2\pi \frac{x'}{\lambda} + \varphi)$$

where

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta$$

The standard deviation σ determines the effective size of the Gaussian signal. The eccentricity of the convolution kernel g is determined by the parameter λ, called the spatial aspect ratio. λ determines the frequency (wavelength) of the cosine. θ determines the direction of the cosine function and finally, φ is the phase offset.

There exists several useful properties with Gabor functions which are important for texture analysis. Gabor function optimally concentrate both in space and space-frequency domain by the smallest time-bandwidth product of the Gaussian function. Due to the ability to tune a Gabor filter to specific spatial frequency and orientation, and achieve both localization in the spatial and the spatial-frequency domains, textures can be encoded into multiple channels each having narrow spatial frequency and orientation. The local information regarding the texture elements is described by the orientations and frequencies of the sinusoidal grating and the global properties are captured by the Gaussian envelope of the Gabor function. Hence the local and global properties of the texture regions can be simultaneously represented by making use of the Gabor filters.

Typically, an image is filtered with a set of Gabor filters of different preferred orientations and spatial frequencies that cover appropriately the spatial frequency domain, and the features obtained form a feature vector that is further used for classification. Given an image I(x,y), its Gabor wavelet transform is defined as

$$W_{mn}(x,y)=\int I(x_1,y_1)g^{*}_{mn}(x-x_1,y-y_1)dx_1dy_1$$

where * indicates the complex conjugate. With assumption of spatially homogeneous local texture regions, the mean μ_{mn} and standard deviation σ_{mn} of the magnitude of transform coefficients can be used to represent the regions [17]. A feature vector f (texture representation) is thus created using μ_{mn} and σ_{mn} as the feature components.

2.3 Gray Level Co-occurrence Matrices

Gray level co-occurrence matrix (GLCM) proposed by Haralick [18] is another common texture analysis method which estimates image properties related to second-order statistics. GLCM matrix is defined over an image to be the distribution of co-occurring values at a given offset. Mathematically, a co-occurrence matrix C is defined over an nxm image I, parameterized by an offset

$$C_{\Delta x,\Delta y}(i,j)=\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(p,q)=i \text{ and } I(p+\Delta\ x,q+\Delta\ y)=j\\ 0, & \text{otherwise}\end{cases}$$

Note that the (△x, △y) parameterization makes the co-occurrence matrix sensitive to rotation. An offset vector can be chosen such that a rotation of the image not equal to 180 degrees will result in a different co-occurrence distribution for the same (rotated) image.

In order to estimate the similarity between different GLCM matrices, Haralick proposed 14 statistical features extracted from them [18]. To reduce the computational complexity, only some of these features will be selected. The 4 most relevant features that are widely used in literature include: (1)Energy, which is a measure of textural uniformity of an image and reaches its highest value when gray level distribution has either a constant or a periodic form; (2) Entropy, which measures the disorder of an image and achieves its largest value when all elements in C matrix are equal; (3) Contrast, which is a difference moment of the C and measures the amount of local variations in an image; (4) Inverse difference moment (IDM) that measures image homogeneity.

2.4 Subcellular Location Features (SLF)

Murphy group has developed and published several sets of informative features, termed Subcellular Location Features (SLFs), that describe protein subcellular location patterns in 2D fluorescence microscope images [5-7]. There are three major subsets of features. The first set is 49 Zernike moment features through order 12, which are calculated from the moments of each image relative to the Zernike polynomials, an orthogonal basis set defined on the unit circle. The second set is 13 Haralick texture features [18], which is related to intuitive descriptions of image texture, such as coarseness, complexity and isotropy. The third set of 22 features was derived from morphological and geometric analysis that correspond better to the terms used by biologists, including the number of objects, the ratio of the size of the largest object to the smallest object, the average distance of an object from the center of fluorescence, and the fraction of above-threshold pixels along an edge et al. Each cell in the dataset is thus represented by a SLF feature vector x of length d = 84. Though SLF includes a much simplified Haralick texture features, we still applied GLCM analysis in a general scenario by specifying the different distance between the pixel of interest and its neighbor and including more statistical measurements as introduced in last subsection.

3. TWO-STAGE HYBRID CLASSIFICATION ENSEMBLES

After feature extraction, a statistical model needs to be learned from data that accurately associates image features with predefined phenotype classes. Some supervised learning algorithms such neural networks, k-nearest neighbor algorithm and SVM [5-8] have been applied to solve this problem. In pattern recognition systems, it has been proven that ensemble of classifiers have the potential to improve classification performance. How to combine multiple classifiers has been studied for decades, with a number of successful methods proposed in the literature [19]. The most popular method for creating an ensemble classifier is to build multiple parallel classifiers, and then to combine their outputs according certain decision fusion strategy. Alternatively, serial architecture can be adopted with different classifiers arranged in cascade and the output of each classifier is the input to the classifier of the next stage of the cascade.

Our approach is based on a hybrid topology that combine parallel and serial schemes. The idea is motivated by a human category learning theory rule-plus-exception model (RULEX) proposed in [20]. According to RULEX, people learn to classify objects by forming simple logical rules and remembering occasional exceptions to those rules. In machine learning, many off-the-shelf methods like support vector machine (SVM) and multi-layer perceptron (MLP) are able to approximate the Bayes optimal discriminant function, which is equivalent to discover the knowledge or patterns hidden in the dataset. Such a knowledge can be represented in terms of a set of rules underlying most of the training examples [22]. A rule consists of an antecedent (a set of attribute values) and a consequent (class):

$$\text{IF} < attrib = value > \text{AND} \cdots \text{AND} < attrib = value >$$
$$\text{THEN} < class > .$$

It is not realistic to expect such a rule to explain all of the data. The examples which are failed to be explained should be considered as exceptions and processed with a rejection option separately. For many real-world applications, such a rejection option is important to satisfy the classification constraints and many multi-stage classifier architectures have been proposed to automatically treating the rejects [23, 25 , 26] .

Extending from the previous works, we proposed a two-stage hybrid classifier ensemble in which a second classifier ensemble is concatenated to the first ensemble. At all stages, a pattern can be either classified or rejected. Rejected patterns are fed into the next stage. The overall system can be illustrated in Figure 3, which shows that second stage need only operate on the surviving inputs from the previous stage.

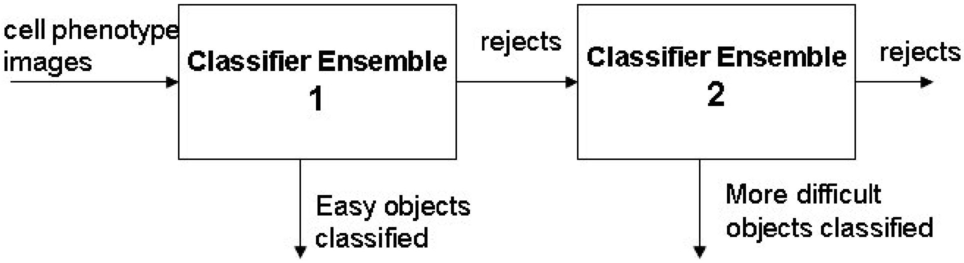


FIGURE 3: Illustration of the overall system which is a cascade of classifier ensembles. Samples rejected at first stage are passed on to second stage during classification.

The major issue for designing the above hybrid classification system is to decide when a pattern is covered by the rule and should be learned by the first classifier ensemble and when it is an exception and should be learned by the second classifier ensemble. The reject option has been formalized in the context of statistical pattern recognition, under the minimum risk theory [31, 23]. It consists in withholding the automatic classification of a pattern, if the decision is considered not sufficiently reliable. Intuitively, objects should be rejected when the confidence in their classification is too low. The standard approach to rejection in classification is to estimate the class posteriors, and to reject the most unreliable objects, that is, the objects that have the lowest class posterior probabilities [24, 23] . As the posteriors sum to 1, there will be complete ambiguity if all posteriors are equal to 1/d with d classes and complete certainty when one posterior is equal to 1 and all others equal to 0.

To simplify the design of the first stage ensemble with appropriate posteriors estimation, we can decompose the multi-label classification problems with k classes into k independent two-class problems, each one consisting in deciding whether an object should be assigned or not to the corresponding class. This is the idea of the *one-versus-all* approach to divide the classes into two groups each time, with one group consisting of a single class and the other group consisting of samples in all the other classes. In other words, a set of k independent binary classifiers are constructed for k classes where the ith classifier is trained to separate samples belonging to class i from all others. Then the multiclass classification is carried out according to the maximal output of the binary classifiers. Though there are many candidates to implement such a scheme, we choose to apply SVMs due to their ability to map features into arbitrarily complex spatial dimensions to find the optimal margin of separation. To estimate class posteriors from SVM's outputs, a mapping can be implemented using the following sigmoid function [28]:

$$P(y=+1|\mathrm{x}) = \frac{1}{1+\exp(\mathrm{a}\rho(\mathrm{x})+\mathrm{b})}$$

where the class labels are denoted as y = +1, -1, while a and b are constant terms to be defined on the basis of sample data. Such a method provides estimates of the posterior probabilities that are monotonic functions of the output ρ(x) of a SVM. This implies that Chow's rule applied to such estimates is equivalent to the rejection rule obtained by directly applying a reject threshold on the absolute value of the outputρ(x) [27].

In our scheme, M binary SVM classifiers are constructed for M different image features. The ith SVM output function P_i is trained taking the examples from ith class as positive and the examples from all other classes as negative. In another word, each binary SVM classifier in the ensemble was trained to act as a class label detector, outputting a positive response if its label is present and a negative response otherwise [21]. So, for example, a binary SVM trained as a ``Nuclei detector'' would classify between cell phenotypes which are Nuclei and not Nuclei. For a new example x, the corresponding SVM assigns it to the class with the largest value of P_i following

$$Class = \arg\max \; P_i, \quad i = 1, \ldots, n$$

where Pi is the signed confidence measure of the ith SVM classifier. The maximum confidence rule with $P(Y_i = 1)$ is used as the confidence measure.

We assume that k classifier ensemble or experts are deployed in the first stage, and that for each input sample, each expert produces a unique decision regarding the identity of the sample. This identity could be one of the allowable classes, or a rejection when no such identity is considered possible. In the event that the decision can contain multiple choices, the top choice would be selected [29]. In combining the decisions of the k experts, the sample is assigned the class for which there is a consensus or when at least t of the experts are agreed on the identity, where

$$t = \begin{cases} \frac{k}{2} + 1 & \text{if } k \text{ is even} \\ \frac{k+1}{2} & \text{if } k \text{ is odd} \end{cases}$$

Otherwise, the sample is rejected. Since there can be more than two classes, the combined decision is correct when a majority of the experts are correct, but wrong when a majority of the decisions are wrong and they agree. A rejection is considered neither correct nor wrong, so it is equivalent to a neutral position or an abstention. Figure 2 further explains the process chart of the stage 1 classifier ensemble.

It is worthy to emphasize that different representations of same set of images were considered for different ``expert", which allow a single expert to take decision about class memberships and thus have different probable decisions. This presents a way to use fusion to have more authenticated decisions by considering many representations of set of patterns.

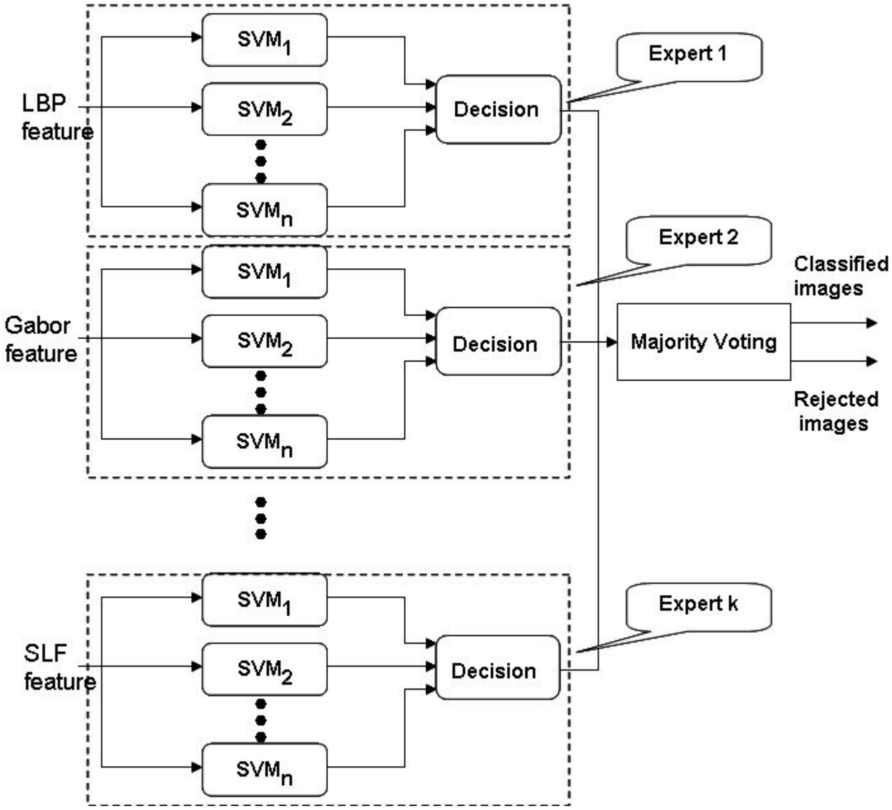


FIGURE 4: Process chart of the stage 1 classifier ensemble, which consist of a set of binary SVMs with high rejection rate.

The set of rejected patterns found by the first stage classifier ensemble will be handled by next stage ensemble, which is a multiple classifier combination with the aim of overcoming the limitations of individual classifiers. In our design, diversity is achieved by choosing classifiers differing in feature representation, architecture and learning algorithm in order to bring complementary classification behavior. In stage 2, the multi-class classification is handled directly by three individual classifiers, including neural network (NN), support vector machine (SVM), and Random Forest classifier [34], which are simultaneously trained with stage 1 ensemble. The three classifiers are of different types: NN classifier is weight-based, SVM classifier is distance or margin based, and Random Forest is rule based. Using different types of classifiers as the constituent classifiers in classifier fusion is one of our design strategies in obtaining necessary diversity, thus achieving improved performance.

The neural network classifier is a 2-layer feed-forward network. It has one hidden layer with a few hidden neurons and has 10 output nodes, each representing a class label. The activation functions for hidden and output nodes are logistic sigmoid function and linear function, respectively. Support Vector Machines (SVM) is a developed learning system originated from the statistical learning theory [30]. One distinction between SVM and many other learning systems is that its decision surface is an optimal hyperplane in a high dimensional feature space. The optimal hyperplane is defined as the one with the maximal margin of separation between positive and negative examples. Designing SVM classifiers includes selecting the proper kernel function and the corresponding kernel parameters and choosing proper C value.

The histogram intersection, $k_{HI}(h_a, h_b) = \sum_{i=1}^{n} \min(h_a(i), h_b(i)),$ is often used as a measurement of similarity between histograms h_a and h_b, and because it is positive definite, it can be used as a kernel for discriminative classification using SVMs. Recently, intersection kernel SVMs have been shown to be successful for detection and recognition [33].

Traditional decision tree classifiers are presented in a binary tree structure constructed by repeatedly splitting the data subsets into two descendant subsets. Each terminal subset is assigned a class label and the resulting partition of the dataset corresponds to the classifier. A random forest (RF) classifier [34] consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The RF algorithm combines ``bagging'' idea to construct a collection of decision trees with controlled variations. There are a number of advantages of RF classifiers, including: (1). it can efficiently handle high dimensional data; (2) it can simultaneously estimates the importance of variables in determining classification; (3). It maintains accuracy when a large proportion of the data are missing.

The last step of the second ensemble is to combine the above base models to give final decision. There are different types of voting systems, the frequently used ones are simple voting and weighted voting [29]. Simple voting, also called majority voting and select all majority (SAM) , considers each component classifier as an equally weighted vote. The classifier that has the largest amount of votes is chosen as the final classification scheme. In weighted voting schemes, each vote receives a weight, which is usually proportional to the estimated generalization performance of the corresponding component classifier. Weighted voting schemes usually give better performance than simple voting. In our study, however, we only experimented with the simple voting.

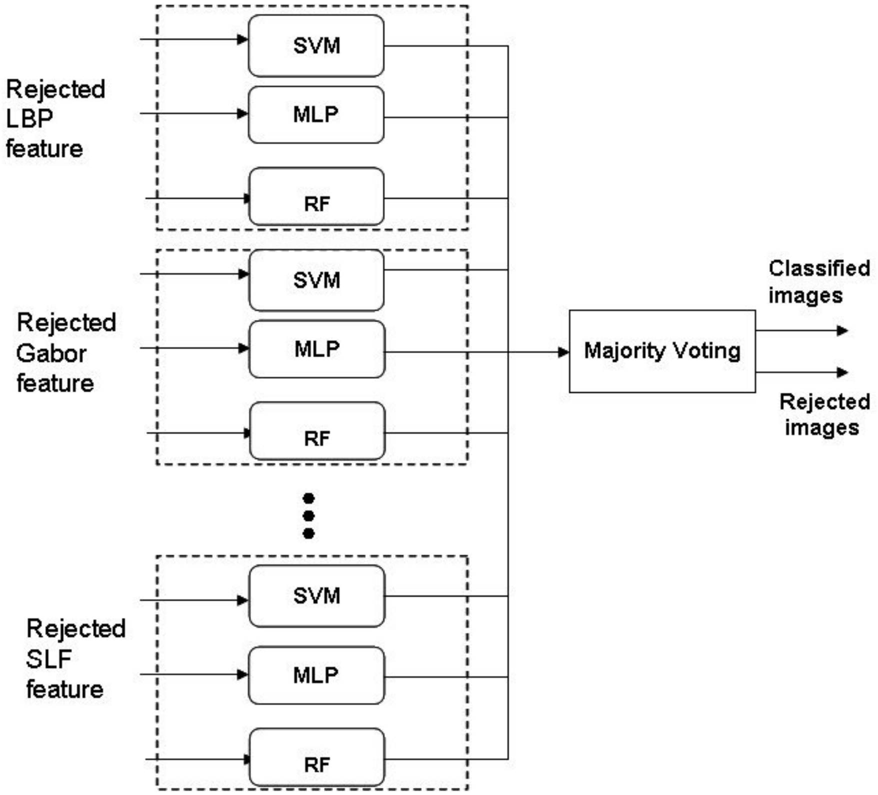


FIGURE 5: Illustration of the stage 2 classifier ensemble which consist of a set of binary SVMs with high rejection rate.

4. EXPERIMENTS

The dataset used for evaluating the system is the 2D HeLa dataset, a collections of HeLa cell immunofluorescence images containing 10 distinct subcellular location patterns [5,6]. The subcellular location patterns in these collections include endoplasmic reticulum (ER), the Golgi complex, lysosomes, mitochondria, nucleoli, actin microfilaments, endosomes, microtubules, and nuclear DNA. The 2D HeLa image dataset is composed of 862 single-cell images, each with size 382x512. Sample images for each class are illustrated in Figure 6. The 2D HeLa image datasets have been used as benchmark for automatically identifying sub-cellular organelles [9-11]. A good verifiable performance for 2D HeLa image classification is currently 91.5% [8], by including a set of multi-resolution features. The best published accuracy 97.5% was recently reported in [9], for which we could not confirm from our own experiments.

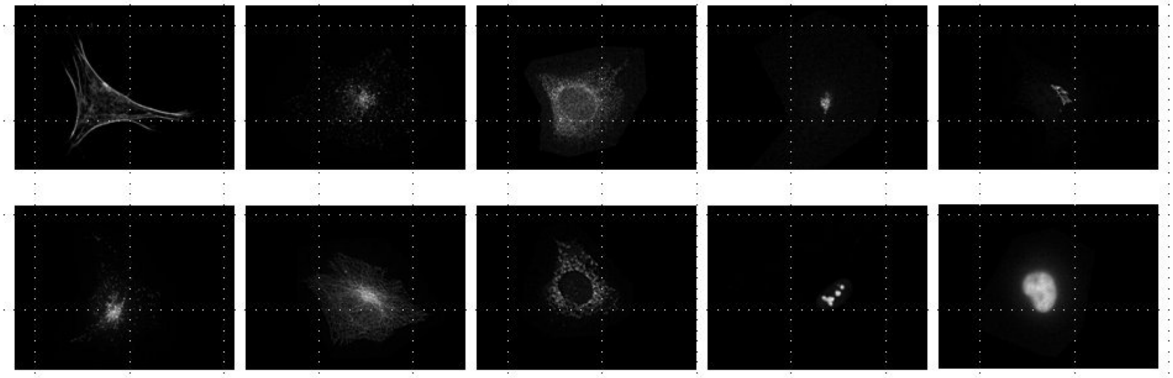


FIGURE 6: Sample 2D HeLa images

As elaborated in Section 2, we are interested in those numerical features that are generally applied in computer vision to describe the pattern in the images. Regarding the LBP feature, a 59-label $LBP^{u}_{8,1}$ operator was used as most of the texture information is contained in the uniform patterns. Specifically, $LBP^{u}_{8,1}$ operator is applied to non-overlapping image subregions to form a concatenated histogram. The performance of LBP representation is not sensitive to the subregion divisions, which do not need to be of the same size or cover the whole image. It is also quite robust with respect to the selection of parameters when looking for the optimal window size. Changes in the parameters may cause big differences in the length of the feature vector, but the overall performance is not necessarily affected significantly. Therefore, in all the experiments we fixed a subregion (window) size 95x102 for the HeLa images, yielding LBP feature vector with length 4x5x59 =1180 . As a comparison, we also applied a newly published variant of LBP operator, called Complete LBP (CLBP for short) [16]. A problem of CLBP is its much higher dimension, which is 2400 with a much larger subregion (size 125 x128) and parameters radius=3 and neighborhood =8.

The Gabor feature vector contains pairs for all the scales and orientations of the wavelets. From a number of experiments we found that a filter bank with six orientations and four scales gave the best classification performance for the classifiers used, which means 24x2 component features will be extracted for a given image patch. Therefore, the figuration is applied to 6x8 non-overlapping image subregions each with the size 60x64, yielding overall feature vector with length 4x5x48=960 for each image. For GLCM feature case, 16 gray co-occurrence matrices were created for each image with an offset that specifies four orientations 0, π/4, π/2 and 3π/4 and 4 distances (1,2,3 and 4 pixels) for each direction. Then for each normalized co-occurrence matrix P(i,j), 12 different type of statistic measurements were estimated, including correlation, variance, contrast, energy, difference variance, entropy, and homogeneity, as described in Section 2. Thus the dimension of GLCM feature is 16x12 = 192. To normalize for the differences in range, each of the LBP, CLBP, Gabor and GLCM feature components is scaled to have a mean of zero and a standard deviation of one across the dataset.

As first set of experiment, we compared the classification performance from the three base classifiers, *i.e.*, random forest, SVM and three-layer perceptron (MLP) neural network, for each of the features (LBP, CLBP, Gabor, GLCM and SLF). The experiment settings for all the classifiers are summarized as follows. For MLP, we experimented with a three-layer network. Specifically, the number of inputs is the same as the number of features, one hidden layer with 20 units and a single linear unit representing the class label. The network is trained using the Conjugate Gradient learning algorithm for 500 epochs. To prevent saturation, the target values are scaled to 0.9 for positive cases and to 0.1 for negative cases.

The popular library for support vector machines LIBSVM (*www.csie.ntu.edu.tw/~cjlin/libsvm*) was sued in the experiment. The parameter γthat defines the spread of the radial function was set to be 5.0 and parameter C that defines the trade-off between the classifier accuracy and the margin (the generation) to be 3.0. We use the radial based function kernel for the SVM classifier when Gabor, GLCM and SLF features were applied and the histogram intersection kernel for LBP/CLBP histograms. With the random forest classifier, the number of trees was chosen as 300 and the number of variables to be randomly selected from the available set of variables was selected as 20. For the 2D HeLa data set, we randomly split it into training and testing sets, each time with 20% of each class's images reserved for testing while the rest for training. The classification accuracy results reported in Table 1 are the average accuracies from 100 runs, such that each run used a random split of the data to training and testing sets.

Classifier	Gabor	LBP	CLBP	GLCM	SLF
RF	73%	72%	85.3%	72%	84%
SVM	82.4%	71.9%	71.6%	78.4%	83.8%
MLP	80%	65.2%	58.5%	86.5%	85.4%

TABLE 1: Performances of three classifiers using different features.

Then we proceeded the experiment with the proposed two-stage hybrid classifier system. The first stage consists of five SVM ensembles which use different sets of features (Gabor, LBP, CLBP, GLCM and SLF). Each base SVM classifier ensemble is trained using the entire training set of the corresponding feature, for example, an LBP feature is used to train 10 binary SVMs. Each binary SVM classifier in a feature specific ensemble was trained to act as a subcellular location detector, outputting a high posterior probability if its corresponding feature is present and a low posterior probability otherwise. During classification, a test instance feature is sent to the 10 base SVM classifiers that estimate the posterior probabilities, with the largest one among the base SVMs indicating the class label. Then 3-out-of-5 majority voting is applied to the output labels from the five SVM ensemble to decide a class label if there is a consensus or reject otherwise. Here the ``consensus'' criterion k=3 acts like a threshold to split the instances into two partitions. In another words, the SVM classifier ensemble collectively labels the multiple feature instances for a give testing HeLa image as belonging or not to any of the 10 categories, while it rejects them from the remaining categories, i.e. no decision is taken about these latter categories. Using a holdout experiment with 80% of data were used for training while the remaining for testing, the first stage accuracy approximates 98% with rejection rate 48%.

The second stage of classifier ensemble consists of 5x3 = 15 multi-class classifiers, which are neural network (NN) classifier, multi-class support vector machine (SVM), and Random Forest classifier, with the five different features. All the base classifiers are simultaneously trained with stage 1 ensemble. During classification, the rejected instances from stage 1 ensemble is passed to the stage 2. Similar to stage 1, k-out-of-15 majority voting is applied to the output labels from the 15 classifiers to decide a class label if there is a consensus or reject otherwise, while k can be controlled to yield varying rejection rate. The overall classification accuracy is defined as the number of correctly classified samples from both stage 1 and stage 2 over the total number of samples tested. From the same holdout experiment with 80% of data for training while the remaining for testing, the second stage accuracy is above 96% with rejection rate 21%, as shown in Figure 7. We also compared different rejection rates between 6% and 42% from stage 2 by varying k in the k-out-of-15 majority voting, yielding the classification accuracies as illustrated in Figure 8. It seems that rejection rate larger than 35% will not bring any more improvement for the classification performance. The corresponding box plot for the comparison is given in Figure 9.

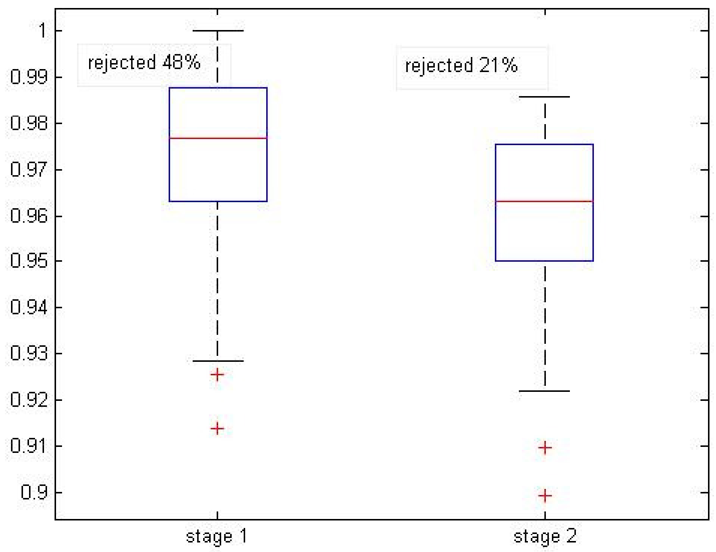


FIGURE 7: Comparison of the final accuracy from stage 2 with overall rejection rate 21% and the first stage accuracy with rejection rate 48%. resulting from holdout experiment with 80% of data were used for training while the remaining for testing. The results were from the average of 100 tests.

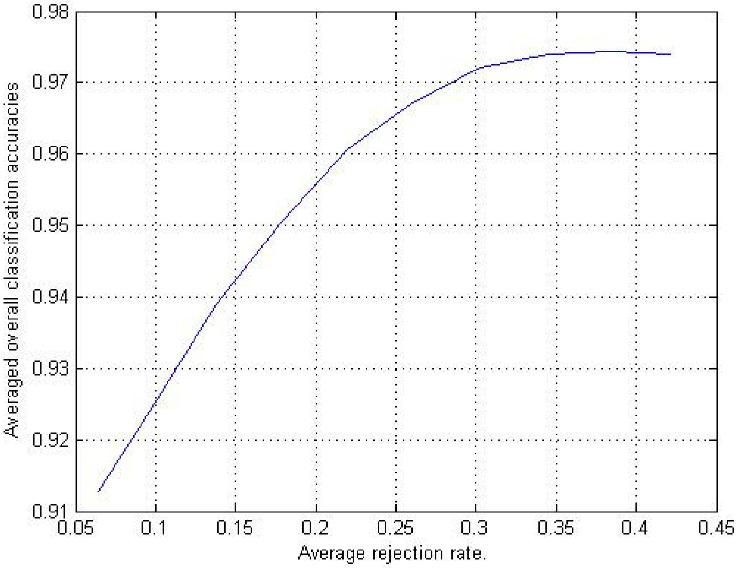


FIGURE 8: Overall accuracies with 10 varying rejection rates in the second stage

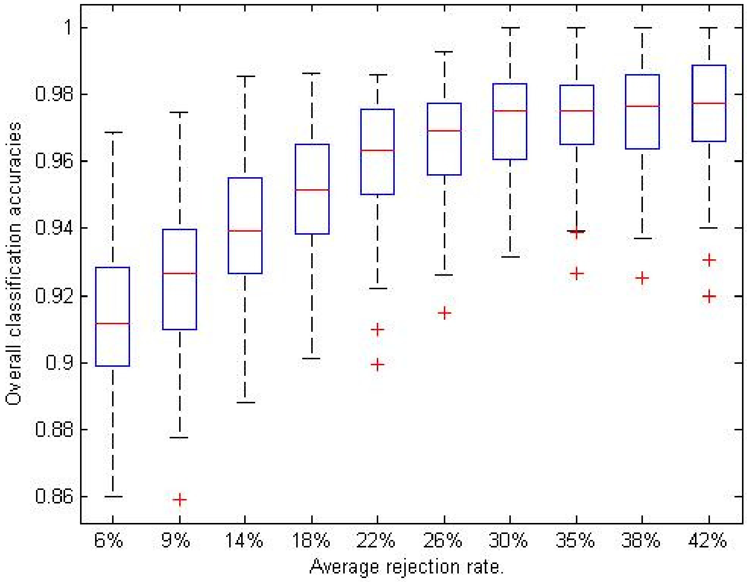


FIGURE 9: Boxplots of classification performances from the different classifiers resulting from holdout experiment with 80% of data were used for training while the remaining for testing.

The confusion matrices that summarize the details of a special situation with rejection rate 6\% is given the following Table 2. For the total number of 187 testing samples, the 10-by-10 matrix displays the number of correct and incorrect predictions made by the hybrid classification system compared with the actual classifications in the test data. It is obvious that among the 10 classes, Actin Filaments type is the easiest to be correctly classified while the Endosome and Golgi_gpp are the difficult categories. This is consistent with previous observations regarding the different degree of difficulties to distinguish the 10 type of subcellular locations [6-7] .

%	1	2	3	4	5	6	7	8	9	10	Accuracy
1	20	0	0	0	0	0	0	0	0	0	100%
2	0	12	1	0	0	2	1	1	0	0	70.6%
3	0	1	15	0	1	0	2	1	0	1	71.4%
4	0	0	0	16	2	1	0	0	0	1	80%
5	0	1	0	1	14	1	1	1	1	0	70%
6	0	3	0	0	1	14	0	0	0	0	77.8%
7	0	1	1	0	0	0	16	1	0	0	84.2%
8	0	1	1	0	0	1	1	11	0	0	73.3%
9	0	0	0	0	1	1	0	0	16	0	88.9%
10	0	0	1	0	0	0	0	1	0	17	89.5%

TABLE 2: Confusion matrix for test set with overall rejection rate 6%. (1: ActinFilaments, 2: Endosome, 3: ER, 4: Golgi_gia, 5: Golgi_gpp, 6: Lysosome, 7: Microtubules, 8: Mitochondria, 9: Nucleolus, 10: Nucleus)

5. CONSLUSION & FUTURE WORK

Automated identification of sub-cellular organelles is important when characterizing newly discovered genes or genes with an unknown function. In this paper, a two-stage multiple classifier system was proposed with rejection strategies for subcellular phenotype images classification. Rather than simply pursuing classification accuracy, we emphasized reject option in order to minimize the cost of misclassifications while secure high classification reliability. The two-stage method used a serial approach where the second classifier ensemble is only responsible for the patterns rejected by the first classifier ensemble. The first stage ensemble consits of binary SVMs with different features, including texture features local binary patterns (LBP), Gabor filtering and Gray Level Co-occurrence Matrix (GLCM), together with Subcellular Location Features (SLF). The first stage ensemble was trained in parallel with the second which is composed of multiple layer perceptron, multi-class support vector machine (SVM), and the Random Forest classifier. During classification, the cascade of classifier ensembles receives a plurality of samples corresponding to different features. The first stage classifier ensemble generates classifications for each of the samples as well as a confidence score associated with the classifications. If the confidence score for a received sample is above a threshold associated with the ensemble, then it absorbs the sample. Otherwise, the classifier ensemble rejects the sample, and such sample is directed to a subsequent classifier ensemble within the cascade. A high classification accuracy 96% is obtained with rejection rate 21% for the 2D HeLa cells from the exploitation of the complementary strengths of feature construction and classifiers decision fusion.

6. ACKNOWLEDGMENTS

The project is funded by China Jiangsu Provincial Natural Science Foundation Intelligent Bioimages Analysis, Retrieval and Management (BK2009146).

7. REFERENCES

1. J. Davis, M. Kakar, and C. Lim. "Controlling protein compartmentalization to overcome disease". *Pharm Res.* **24(1)**: pp.17—27, 2007..

2. N. Orlov, J. Johnston, T. Macura, L. Shamir and I.Goldberg, "Computer Vision for Microscopy Applications. Vision Systems: Segmentation and Pattern Recognition", Edited by: Goro Obinata and Ashish Dutta, pp.546, I-Tech, Vienna, Austria, June 2007

3. H. Peng, "Bioimage informatics: a new area of engineering biology". *Bioinformatics*, **24(17)**: pp. 1827—36, 2008.

4. E.J. Roques and R.F. Murphy RF. "Objective Evaluation of Differences in Protein Subcellular Distribution", *Traffic*, **3**, Pages 61 – 65, 2002.

5. M.V. Boland and R.F. Murphy, "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells". *Bioinformatics*, **17**(12): pp.1213—23, 2001.

6. M.V. Boland, M. Markey and R.F. Murphy, "Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images". *Cytometry*, **33**: pp. 366-375, 1998.

7. K. Huang and R.F. Murphy,. "Boosting accuracy of automated classification of fluorescence microscope images for location proteomics". *BMC Bioinformatics*, **5**: 78, 2004.

8. A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, RF., Murphy and J. Kovacevic, "A multiresolution approach to automated classification of protein subcellular location images". *Bioinformatics*, **8**: 210, 2007.

9. L.Nanni, A. Lumini, Y. Lin, C. Hsu, and C. Lin, "Fusion of systems for automated cell phenotype image classification". *Expert Systems with Applications*, **37**: pp. 1556-1562, 2010.

10. N.A. Hamilton, R.S. Pantelic, K. Hanson and R.D.Teasdale. "Fast automated cell phenotype image classification". *Bioinformatics*, **8**: pp. 110, 2007.

11. B. Zhang, "Classification of Subcellular Phenotype Images by Decision Templates for Classifier Ensemble". *International Conference on Computational Models for Life Sciences* (CMLS-09), AIP Conf. Proc. **1210**, pp.13-22, 2009.

12. T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(7): pp.971-987, 2002.

13. L. Wolf, T. Hassner and Y. Taigman, "Descriptor Based Methods in the Wild". *Faces in Real-Life Images workshop at the European Conference on Computer Vision (ECCV)*, Oct 2008.

14. T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(12): pp.2037-2041, 2006.

15. G. Zhang, X. Huang, S.Z. Li, Y. Wang, and X. Wu, "Boosting Local Binary Pattern (LBP)-Based Face Recognition". In *Proc. Advances in Biometric Person Authentication: 5th Chinese Conference on Biometric Recognition, SINOBIOMETRICS* 2004, Guangzhou, China. pp. 179-186, 2005.

16. Z. Guo, L. Zhang and D. Zhang "A Completed Modeling of Local Binary Pattern Operator for Texture Classification". accepted for *IEEE Trans Image Process*., preprint, 2010.

17. B. Manjunath and W. Ma, "Texture Features for Browsing and Retrieval of Image Data". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **18**(8):, pp.837—842, 1996.

18. R. Haralick "Statistical and Structural Approaches to Texture",. *Proceedings of the IEEE*, **67**(5)} pp. 786-804, 1979.

19. L.I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms". Wiley-Interscience., (2004).

20. R.M. Nosofsky, T.J. Palmeri and S.C. McKinley, "Rule-Plus-Exception Model of Classification Learning". *Psychological Review*, **101**, pp.53-79, 1994.

21. R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification". *Journal of Machine Learning Research*, **5**: pp. 101-141, 2004.

22. N. Holden and A.A. Freitas "A Hybrid PSO/ACO Algorithm for Discovering Classification Rules in Data Mining", *Journal of Artificial Evolution and Applications*, **2008**, Article ID 316145, 11 pages, 2008.

23. D.M.J Tax and R.P.W. Duin, "Growing a multi-class classifier with a reject option", *Pattern Recognition Letters*, **29**: pp. 1565-1570, 2008.

24. C.K.Chow, "On optimum recognition error and reject tradeoff". *IEEE Trans. Inf. Theory*, **IT-16** (1), 41–46, 1970.

25. N.Giusti, F. Masulli, F., Sperduti, "A Theoretical and Experimental Analysis of a Two-Stage System for Classification". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **24**: pp. 893–904,2002.

26. P. Pudil, J. Novovicova, S. Blaha, J. Kittler, "Multistage Pattern Recognition with Reject Option". In: *Proc. 11th IAPR Int. Conf. on Pattern Recognition*, **2**: pp.92-95, 1992.

27. G. Fumera and F. Roli. Support Vector Machines with Embedded Reject Option, *Int. Workshop on Pattern Recognition with Support Vector Machines* (SVM2002), Springer, Niagara Falls, Canada, p.68-82, 2002.

28. R.P.W. Duin and D.M.J. Tax, "Classifier conditional posterior probabilities". In: Amin, A., Dori, D., Pudil, P., Freeman, H. (eds.): Advances in Pattern Recognition. Lecture Notes in Computer Science 1451, Springer, Berlin, 611-619, 1998.

29. L. Lam and C.Y. Suen, "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance", *IEEE Transactions on Systems, Man, and Cybernetics -Part A: Systems and Human*, **27**: pp.553-568, 1997.

30. J. Shawe-Taylor and N. Cristianini, "Kernel methods for pattern analysis". Cambridge University Press (2004).

31. C.-W. Hsu and C.-J. Lin, "A comparison on methods for multi-class support vector machines". *IEEE Transactions on Neural Networks*, **13**: pp.415—425, 2002.

32. R.O. Duda, P.E. Hart and D.G. Stork,D.G. "Pattern classification", Second Edition, John Wiley and Sons, New York, (2001).

33. S. Maji, A.C. Berg, and J. Malik, . "Classification Using Intersection Kernel Support Vector Machines is efficient". In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)* ,Anchorage, Alaska, pp. 1-8, 2008.

34. L. Breiman, "Random Forests". *Machine Learning*, **45,** pp. 5–32, 2001.