9

Information Fusion in a High Dimensional Feature Space for Robust Computer Aided Diagnosis using Digital Mammograms

Saurabh Prasad, Lori M. Bruce and John E. Ball
Geosystems Research Institute
Department of Electrical and Computer Engineering
Mississippi State University
U.S.A.

1. Introduction

Most end-to-end Computer Aided Diagnosis (CAD) systems follow a three step approach – (1) Image enhancement and segmentation, (2) Feature extraction, and, (3) Classification. Although the current state-of-the-art in image enhancement and segmentation can now very accurately identify regions of interest for feature extraction, these methods typically result in very high dimensional feature spaces. Although high dimensional feature spaces can potentially provide more discrimination information for a more robust classification of images, they are typically a double-edge sword. They can adversely affect the performance of classification systems because a large feature space dimensionality necessitates a large training database to accurately model the statistics of class features (e.g. benign versus malignant classes). Further, high dimensional feature spaces affect the generalization capacity of the classification system – that is, when feature spaces are high dimensional, it is possible that the classifier can learn decision boundaries from the training data very accurately, but it's tolerance to slight variations in the statistics of test data will be poor. In other words, the system will potentially over-fit the decision boundaries to the training data. Although the extent of this reduction in generalization capacity of the classifier depends on the classification algorithm being employed, practically all classification algorithms are prone to this problem, also known as the Hughe's phenomenon in the pattern classification community.

Current state-of-the-art in mammography based CAD systems extract a wide-variety of features (for example, texture based features, shape features etc.) from the mammograms. The resulting feature spaces generated for classification are hence typically very high dimensional. To mitigate the ill-effects of high dimensionality and limited training data, most classification systems employ dimensionality reduction techniques. Algorithms such as Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) are used in the pattern classification community for dimensionality reduction. They have their own limitations – PCA is not optimal for classification tasks [Prasad *et al*, October 2008], since it is

only designed to reduce features, not optimize them for classification, and LDA requires sufficient training data to be available for learning the projections. Stepwise Linear Discriminant Analysis (S-LDA) is a more popular technique that reduces the dimensionality of the feature space by selecting a smaller subset of features from the original feature space, and then further reduces the dimensionality of this subset by an optimizing linear transformation based on the LDA algorithm. One key limitation of this approach is that it is by design sub-optimal at best, because it does not utilize all available features for estimating the final "optimizing" transformation.

In this chapter, we present a robust multi-classifier decision fusion framework that employs a divide-and-conquer approach for alleviating the affects of high dimensionality of feature vectors extracted from digital mammograms. After appropriate pre-processing of the digital mammograms (including contrast enhancement), the core region of interest is segmented, and various features are extracted from the segmented images (including morphological features, statistical features, texture features, etc.). The resulting high dimensional feature space is partitioned into multiple smaller sized subspaces, and a bank of classifiers (a multi-classifier system) is employed to perform feature optimization and classification in each partition separately. Finally, a decision fusion system merges decisions from each classifier in the bank into a single decision. Unlike previous approaches which discard potentially useful features for alleviating the over-dimensionality problem, the proposed method employs all available features for classification. The performance of the resulting CAD system is reported using overall classification accuracy, specificity and sensitivity. Results show that the proposed system results in significantly better classification performance as compared to previously employed techniques.

The outline of this chapter is as follows. In section 2, we provide an overview of mammography based CAD systems, and on common image processing methods employed in such systems. We also provide a description of the feature extraction and optimization method employed in this work, as well as conventional classification approaches and the proposed classification framework. In section 3, we present experimental results demonstrating the benefits of the proposed system, and conclude this chapter with a summary of results and future directions.

2. Computer Aided Diagnosis using Digital Mammography

Digital mammography uses x-rays to project structures in the 3D female breast onto a 2D image [Egan, 1988]. The primary use for digital mammography is for screening for breast cancer. According to the American Cancer Society (ACS), breast cancer is the leading type of cancer in women and the second most fatal type of cancer in women [ACS, 2005].

2.1 Importance of Mammography

Mammography is important for many reasons. First, early detection can increase survival rates [Rangayan, 2005] and decrease the probability that cancer cells are able to infiltrate

Based on "A multi-classifier and decision fusion framework for robust classification of mammographic masses," by Saurabh Prasad, Lori M. Bruce, John E. Ball which appeared in the 30th Annual International Conference of the ,Engineering in Medicine and Biology Society, 2008. © 2008IEEE

other parts of the body [Voegeli, 1989]. The survival rate of breast-cancer patients is inversely related to the tumor size and to the number of auxiliary lymph nodes that are found with malignant cells [Lille, 1992]. Mammography can detect cancer years before physical symptoms occur [Tabar *et al*, 2001]. Often, early detection finds a minimal cancer, and the cure rate approaches 95 percent [Egan, 1991].

2.2 Challenges in Digital-Mammography Image Analysis

Digital-mammographic analysis is a very difficult problem because of the complexity of digital mammograms, poor contrast, and in many cases, a lack of a clearly defined mass border [Hughes, 1968]. First, a digital mammogram is a 2D image of a 3D and highly complex structure. Due to the complex 3D nature of breasts and the view used when taking the mammogram, some tumors may be partially obscured. Second, it is well-known that digital mammograms often have poor contrast, especially in dense breasts [Egan, 1988]. Many specialized methods have been developed to denoise and enhance digital mammographic images. Third, according to Egan, there is a wide variability in the appearance of mammograms of different patients [Egan, 1988]. Adipose tissue has a higher concentration of low-atomic-number elements, such as hydrogen and carbon, and therefore low-energy x-rays are attenuated less in adipose tissue than other tissues. The amount of adipose tissue present in the breast is important, because it provides good contrast with other structures in the breast [Egan, 1988]. Furthermore, the size of malignant lesions that can be detected depends on the adipose content of the breasts, and breasts with more fat provide the ability to detect smaller tumors than breasts with a paucity of fat [Egan, 1988]. Fourth, cancerous masses often have a similar appearance to some benign masses, or even normal fibroglandular breast tissue [Peters *et al*, 1989], and many masses have irregular or obscured borders [Egan, 1988], [Tabar *et al*, 1985]. Furthermore, most masses do not fit a distinct border type, but have mixtures of border types [Egan, 1988], [Peters *et al*, 1989]. Mass borders are a primary method radiologists use to identify benign versus malignant masses [Ball, May 2007].

2.3 Conventional Computer-Aided-Diagnosis (CAD) Systems

Many digital mammography CAD systems have been previously developed, both as experimental models in academia, and marketed and FDA-approved products for breast-cancer detection. CAD systems serve a variety of purposes, including: (1) providing a prompting system for radiologists to help them locate suspicious areas in the mammogram, (2) providing a second opinion, by analyzing the mammogram and deciding if a mass is malignant or benign, and (3) providing mammographic image enhancement, where digital mammograms can be enhanced for noise removal or to provide better contrast with the overall goal to allow better radiologist interpretation.

Qian *et al* [Qian *et al*, October 1995] and Huo *et al* [Huo *et al*, 2002] show that radiologists can benefit from the aid of CAD systems. Sahiner *et al* demonstrated that a good CAD system is comparable to expert radiologists [Sahiner *et al,* December 2001]. Burhenne *et al* show that CAD can potentially reduce radiologists' false negative rate, which is the rate at which a radiologist falsely detects a malignant mass as benign [Burhenne *et al,* 2000].

2.4 Database Description

For this study, a subset of the DDSM database consisting of 60 test cases is selected. This data subset contains 30 benign and 30 malignant cases, where 17 of the 30 malignant cases are spiculated, and none of the benign cases are spiculated. Table lists the benign and malignant cases in the study dataset.

Benign Cases	Malignant Cases
1305, 1370, 1371, 1372, 1379, 1387, 1389, 1394, 1397, 1408, 1432, 1442, 1443, 1445, 1447, 1453, 1459, 1498, 1512, 1518, 1519, 1554, 1556, 1560, 1566, 1607, 1615, 1679, 1682, 1691	1112, 1132, 1134, 1144, 1147, 1149, 1155, 1156, 1157, 1163, 1168, 1171, 1182, 1215, 1237, 1252, 1262, 1263, 1404, 1468, 1533, 1534, 1537, 1558, 1574, 1622, 1665, 1726, 1827, 1999

Table 1. Listing of case numbers for the DDSM database

Examination of figures 1 – 4 shows that the DDSM-database cases chosen comprise a relatively difficult database in terms of preprocessing, segmentation, and classification. Fig. 1 shows the ACR density-rating distribution for the test dataset. The ratings are assigned by radiologists who worked on the DDSM database [Heath *et al*, 1998]. The density rating is an integer from one to four, where a density rating of one is the least dense, and a density rating of four is the densest. In general, the higher the density rating, the more difficult segmentation may be because of increased paucity of adipose tissue, which provides mammographic contrast. Fig. 1 shows that most of the malignant tumors have density values two and three, while the benign tumors have density three and four. Also, there are mammograms for each density category, from one to four. Therefore, density is probably not a good feature for distinguishing malignant from benign cases. Fig. 2 shows the ACR Bi–RADS assessment distribution. This figure indicates that this dataset is a difficult one, since a large number of the benign cases appear in category 3. Figure 3 shows the subtlety-ratings distribution [Heath *et al*, 1998]. A subtlety rating of one indicates least subtle, while a rating of five is the most subtle. Please note that this subtlety-rating system is unique to the DDSM database, and is different from other subtlety-rating systems used in digital mammography. A case with a DDSM subtlety rating of N is N times more subtle than a case with a subtlety rating of one. There are 4 benign and 15 malignant mammograms whose subtlety is 5 (the most subtle). There are 24 out of 30 benign cases and 27 out of 30 malignant cases which have a subtlety rating from 3 to 5, as shown in Figure 3. This figure shows the database is difficult.

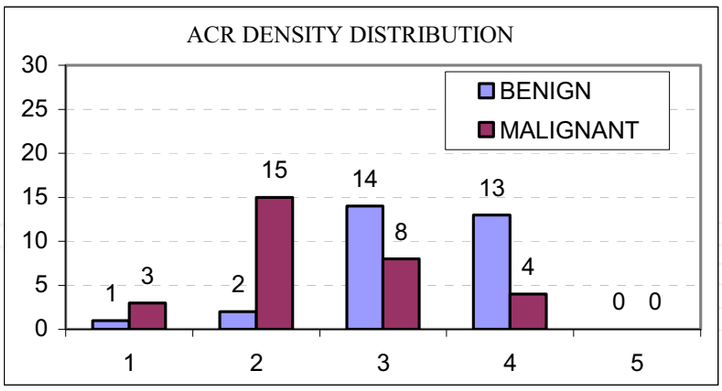

Fig. 1. Case study breast density distributions. The least dense is 1 and the most is dense is 5.

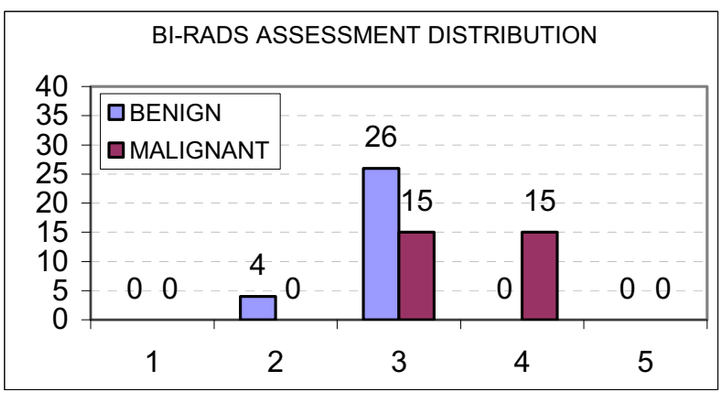

Fig. 2. Case study Bi-RADS assessment category distribution

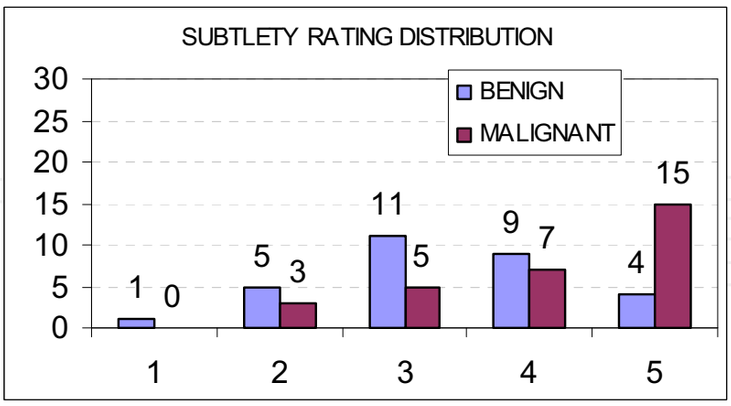

Fig. 3. Case study DDSM subtlety rating distribution. Ratings range from 1 (least subtle) to 5 (most subtle).

It is well-known that mass shape plays a very important factor in mammographic analysis. Fig. 4 shows the shape distribution. The shape keywords are defined in Table . The shapes

are assessed by experienced radiologists. As expected, almost all irregularly shaped masses are malignant, while oval and round shapes are benign.

Keyword	Meaning	Keyword	Meaning
Oval	Shaped like an oval.	Lobulated	Shaped like breast lobes.
Irregular	Irregular shape.	Round	Circularly shaped.

Table 2. Shape keywords and their meanings

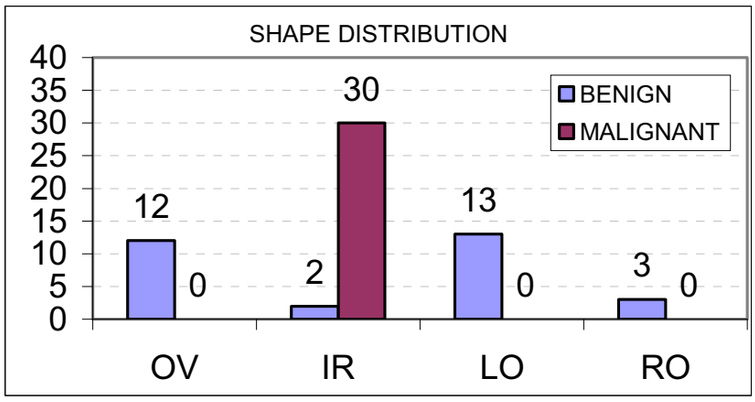

Fig. 4. Case study shape distribution. Legend: OV= Oval, IR = Irregular, LO = Lobulated, RO = Round.

2.5 Image Processing – Enhancement and Segmentation

The Catarious segmentation method (CSM) is implemented as described in [Catarious, June 2004], [Catarious, August 2004]. The original mammogram is passed through an unsharp-masking filter. The algorithm uses the DDSM supplied Region of Interest (ROI) containing a suspicious region and creates a new, square ROI sized approximately 112.6 mm^2 (245^2 pixels) centered around the seed point. The interior pixels of the square ROI are then extracted from the unsharp masked image. A circular region of radius 16 mm is selected as the initial segmentation, where interior pixels are segmented, and pixels outside of the circle are considered outside of the segmentation. Fisher's LDA (FLDA) is used to select a graylevel threshold for selection of the new boundary, which is the connected region containing the ROI centroid. Boundary constraints are applied according to the Catarious algorithm, such as interior pixels on each ray emanating from the center must not have gaps of more than d pixels, and border pixels must be within a specified city-block distance n of their immediate neighbors. This analysis is performed in the polar transform. If the stopping criterion (the boundary does not change from the previous iteration) is met, then the segmentation stops; otherwise, a new iteration is performed. It is discovered that the CSM could get in a loop wherein there will not be a state during which the segmentation does not change, and no proof of convergence is given in [Catarious, June 2004], [Catarious, August 2004]. Thus, a 40-iteration limit is added to the segmentation method. For the segmentation

stage, n=2, d=3, and the unsharp-masking weighting is 0.9, which are the original parameters proposed in [Catarious, June 2004], [Catarious, August 2004]. The CSM system performed very well on Catarious' database, which is why this method was chosen as a comparison method. It provided a "successful" segmentation method against which to compare the proposed method, and also had a straightforward implementation. More details about this segmentation technique can be found in [Catarious, June 2004], [Catarious, August 2004].

2.6 Feature Extraction

This section discusses the various types of features which are typically used in mammography analysis [Ball, May 2007]. The type of features include patient information, statistical features from the pixel graylevel values, textural features, normalized radial-length (NRL) feature, based on the segmentation boundary, and morphological features based on the size and shape of the segmentation.

Patient-Information Features

It is well-known that age is the most important feature in mammographic analysis [Laws, January 1980 a], [Laws, January 1980 b]. Therefore, in this study, age will be included as a feature. The other patient information described in the previous sections will not be used in this dissertation's study.

Statistical Features

Statistical features are derived from the segmentation boundary or an extension of the segmentation boundary. The statistical features are graylevel mean value, graylevel standard deviation, and graylevel standard-deviation ratio. The standard-deviation ratio is the ratio of the graylevel standard deviation in a region from one to 200 pixels outside of the segmentation boundary to the standard deviation of pixels inside the segmentation boundary.

Textural Features

Textural features can capture important characteristics of a small area, and thus may be beneficial in image segmentation. The textural features include graylevel co-occurrence-matrix (GLCM) features derived from the RBST image, GLCM features extracted from the segmentation boundary, and Laws texture features [Ball, May 2007]. The GLCM textural features include energy, variance, correlation, inertia, inverse difference moment, and entropy [Laws, January 1980 a], [Laws, January 1980 b]. The Laws texture features include features derived from 2D Laws texture kernels. Each of these is explained in detail below.

GLCM Textural Features

In order to calculate the GLCM texture features efficiently, the image is first quantized using N bins, resulting in a quantized image whose pixel values elements are in $\{0,\cdots,N-1\}$. In this dissertation, the value for the number of quantized bins is set to N=*20.* Distances, d, and direction angles, θ, are selected and the associated relative-frequency matrix $P=[P_{i,j}]$ is calculated, where i and j are the pixel intensities. Following [Laws, January 1980 a], [Laws,

January 1980 b], the following equations are used to calculate the P matrix values for the distance d and the angles $\theta \in \{0°, 45°, 90°, 135°\}$:

$$P(i,j,d,0°) = \left\{ \begin{array}{c} ((k,l),(m,n)) \in I \\ \forall \, |k-m| = 0, \; |l-n| = d \\ I(k,l) = i, \, I(m,n) = j \end{array} \right\} \tag{1}$$

$$P(i,j,d,45°) = \left\{ \begin{array}{c} ((k,l),(m,n)) \in I \\ \forall (k-m = d, l-n = -d) \; or \; (k-m = -d, l-n = d) \\ I(k,l) = i, \, I(m,n) = j \end{array} \right\} \tag{2}$$

$$P(i,j,d,90°) = \left\{ \begin{array}{c} ((k,l),(m,n)) \in I \\ \forall \, |k-m| = d, \; |l-n| = 0 \\ I(k,l) = i, \, I(m,n) = j \end{array} \right\} \tag{3}$$

$$P(i,j,d,135°) = \left\{ \begin{array}{c} ((k,l),(m,n)) \in I \\ \forall (k-m = d, l-n = d) \; or \; (k-m = -d, l-n = -d) \\ I(k,l) = i, \, I(m,n) = j \end{array} \right\} \tag{4}$$

where {} denotes the number of elements that satisfy the conditions in {}, and the pixel coordinates of image I are (k,l) and (m,n).

The GLCM matrices are created from these matrices by normalizing the matrices. The matrices are normalized by dividing by R_θ for a given angle θ. That is,

$$P_{i,j} = \frac{p_{i,j}}{R_\theta}. \tag{5}$$

The values used for R_θ are given in Table 3, where n_x and n_y are the number of columns and the number of rows, respectively, in the image, and d is the user specified distance.

θ in degrees	R_θ
0°	$2(n_x - d)n_y$
45°,135°	$2(n_x - d)(n_y - d)$
90°	$2n_x(n_y - d)$

Table 3. Normalizing values

Several of the GLCM features require marginal probability mass functions (PMF). The marginal PMF's, p_x and p_y, are defined by the following formulas:

$$p_x = \sum_{j=1}^{N} P(i,j)/R, \tag{6}$$

and

$$p_y = \sum_{i=1}^{N} P(i,j)/R, \tag{7}$$

where N is the number of graylevels in the quantized image, and R is the appropriate R_θ value.

The GLCM texture features are calculated as follows. The GLCM energy feature is a measure of the image uniformity. A region with little graylevel change will have a low-energy value,

$$f_{ENERGY} = \sum_i \sum_j \left[p(i,j)\right]^2. \tag{8}$$

The GLCM variance is a measure of the spread of the elements in the matrix. The feature is calculated from

$$f_{VARIANCE} = \sum_i \sum_j \left(i - \mu\right)^2 p(i,j), \tag{9}$$

where μ is the mean value of the elements in *p*. The GLCM correlation measure quantifies the quantized graylevel linear dependence. The GLCM correlation is calculated from

$$f_{CORRELATION} = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}, \tag{10}$$

where μ_x, μ_y, σ_x, and σ_y are the mean values and the standard deviations of the marginal probability matrices, p_x and p_y, respectively. The entropy feature is defined as

$$f_{ENTROPY} = -\sum_i \sum_j p(i,j) \log\{p(i,j)\}, \tag{11}$$

with the convention that log(0) = 0. Entropy defines a measure of randomness, and a very uniform area will have very low entropy.

Laws Textural Features

Laws [Laws, January 1980 a], [Laws, January 1980 b] developed a set of one-dimensional (1D) and 2D kernels which are designed to extract small-scale textural information from images. The 1D Laws kernels are given by

$$\begin{aligned} &L5 = [\ 1\ \ 4\ \ 6\ \ 4\ \ 1\ ] \qquad E5 = [\ -1\ \ -2\ \ 0\ \ 2\ \ 1\ ] \\ &S5 = [\ -1\ \ 0\ \ 2\ \ 0\ \ -1\ ] \qquad W5 = [\ -1\ \ 2\ \ 0\ \ -2\ \ 1\ ] \\ &R5 = [\ 1\ \ -4\ \ 6\ \ -4\ \ 1\ ] \end{aligned}, \tag{12}$$

where the first letter stands for Level, Edge, Spot, Wave, and Ripple, respectively [Laws, January 1980 a], [Laws, January 1980 b]. The 2D Laws textures are sized $[5 \times 5]$, and are created by multiplying the 1D Laws kernel as a row vector by the second kernel as a column vector. Since there are five 1D Laws kernels, there will be 25 2D kernels. As an example, the L5L5 kernel is given by

$$L5L5 = \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}, \tag{13}$$

To create the Laws texture features, an image is convolved with each of the Laws 2D kernels. Each image is then normalized by dividing pixel-by-pixel with the *L5L5* image, and is termed I_m for each Laws texture kernel $m \in \{1,2,\cdots,25\}$. For each Laws texture and each pixel in the convolved image, the following equation is used to calculate the texture for that pixel for the *m*-th texture:

$$L_m(x,y) = \sum_{j=-D}^{D} \sum_{k=-D}^{D} I_m(x+j, y+k), \tag{14}$$

where the texture mask size is (2*D*+1) by (2*D*+1). As suggested by Laws [Laws, January 1980 a], the variable *D* is selected as 7, thus equation **Error! Reference source not found.** uses a $[15 \times 15]$, region centered at each pixel for analysis. Note that some applications take the absolute value before summing in this formula.

Morphological Features

Morphological features are derived from the shape characteristics of the segmented region. These features include the area, axis ratio, box ratio, circularity, convex-hull area, eccentricity, equivalent diameter, extent, extent ratio, major-axis length, minor-axis length, perimeter length, solidity, and width-to-height ratio.

The area is the number of pixels in the segmented region. The axis ratio is the ratio of the major-axis length to the minor-axis length. The box ratio is the ratio of the area to the product of the height times the width, where the height and width are defined by the bounding box of the region. Circularity is defined as the product of 4π times the region area divided by the square of the perimeter length in pixels. The convex-hull area is the size of the convex hull of the region, where the convex hull *H* of an arbitrary set *S* is the smallest convex set containing *S*. The equivalent diameter is defined as the diameter of a circle which has the same area as the segmented region. The extent feature is calculated as the area divided by the bounding-box area. The extent ratio is defined as max(*height*, *width*) / min(*height*, *width*). The major- and minor-axis lengths are defined as the length in pixels of the major and minor axes of the ellipse that has the same normalized second central moments as the region, respectively. The perimeter length is the number of pixels on the region perimeter. The solidity feature is defined as the region area divided by the convex-hull area. The width-to-height ratio is defined as *width* / *height*.

Normalized Radial-Length (NRL) Features

The normalized radial-length (NRL) features are derived from a normalized version of the radial-length measure [Agatheeswaran, December 2004]. First, the border pixels of the segmentation are extracted and the centroid of the segmentation region, (c_x, c_y), is calculated. Assume there are N_B border pixels and that the coordinate of the *k*-th border pixel is (x_k, y_k). For each border pixel, the Euclidean distance between the pixel and the centroid is calculated as $D_k = \sqrt{(x_k - c_x)^2 + (y_k - c_y)^2}$. The largest distance, D_{MAX}, is calculated, and the normalized distance is calculated by dividing the pixel Euclidean distance by the maximum distance: $NRL_k = D_k / D_{MAX}$. Table 4 summarizes the NRL features, and was adapted from [Agatheeswaran, December 2004].

Feature	Description
Entropy	A measure of the randomness of the NRL vector values. $f_{NRL\_ENTROPY} = -\sum_{j=1}^{N_B} p_j \log_2(p_j),$ where $\log_2(0) = 0$ and p_j is the PMF of NRL values. This PDF is estimated with a 256 bin histogram.
Length	The length of the NRL distance vector. $f_{NRL\_LENGTH} = N_B$
Mean	The mean value of the NRL distances. $f_{NRL\_MEAN} = -\frac{1}{N_B}\sum_{k=1}^{N_B} NRL_k$
Roughness	A measure of border roughness. $f_{NRL\_MEAN} = \left[\frac{L}{N_B}\right]\sum_{j=1}^{[N_B/L]} R_j$ where the roughness parameter R_j is given by $R_j = \sum_{i=j}^{L+j} \left\lvert NRL_i - NRL_{i+1} \right\rvert$ with $j \in \{1, \cdots, [N_B / L]\}$.
Standard deviation	The standard deviation of the NRL vector. $f_{STD\_NRL} = \frac{1}{N_B - 1}\sum_{j=1}^{N_B} \left[NRL_k - f_{NRL\_MEAN}\right]^2$ Note: Some of the literature uses a biased definition for this feature, i.e. they divide by N_B and not by $N_B - 1$. This analysis uses the unbiased estimator.
Zero crossing count	The number of times the NRL distance crosses over the NRL mean.

Table 4. NRL feature descriptions

Source: Adapted from [Agatheeswaran, December 2004].

Feature Type and source [1]	Features	Num. Features[4,5]	References
Patient age (DDSM)	Patient Age	1	[Heath *et al*, 1998]
Morphological (SB)	Area, Axis ratio, Box ratio, Circularity, Convex hull area, Eccentricity, Equivalent diameter, Extent, Extent ratio[2], Major axis length, Minor axis length, Perimeter length, Solidity, Width to height ratio	14	[Catarious, June 2004], [Catarious, August 2004]
Statistical (SB)	Graylevel mean, Graylevel std. dev, Graylevel std. dev. ratio[3]	3	[Cheng et al, August 2004]
NRL (SB)	Entropy, Length, Mean, Roughness, Std. dev., Zero crossing count	6	[Cheng et al, August 2004]
GLCM (SB)	(Note [4]) Energy, Variance, Correlation, Inertia, Inverse Difference Moment, Entropy	144	[Haralick *et al*, 1973], [Agatheeswaran, December 2004], [Cheng et al, August 2004]
GLCM (RBST)	(Note [5]) Energy, Variance, Correlation, Inertia, Inverse Difference Moment, Entropy	864	[Sahiner *et al*, April 1998], [Haralick *et al*, 1973], [Agatheeswaran, December 2004], [Cheng et al, August 2004]

Table 5. Feature list

Notes:
[1] This denotes the region from which the features were extracted. DDSM=DDSM database (there is no region, as the patient age is part of the database). SB=segmentation boundary. NRL stands for Normalized Radial Length. RBST=Rubber Band Straightening Transform. GLCM stands for gray level co-occurrence matrix. GLCM is also known as spatial gray level dependence (SGLD). [2] The extent ratio is max(length, height) / min(length, height). [3] the Gray level std. dev. ratio is the ratio of the std. dev. of the gray levels inside the segmentation to the std. dev. of gray levels outside the segmentation boundary and within 200 pixels of the segmentation boundary. [4] The GLCM SB features are calculated at distances d={1,2,4,6,8,10} and directions θ={0°,45°,90°,135°}. There will be a total of 6 GLCM features x 6 distances x 4 angles for 144 GLCM features. [5] The RBST features are the same features as the GLCM SB features. The RBST uses a parameter k to choose how many pixels before and after are used to create the normal vector to the spiculation boundary. The RBST features are calculated for distances k={2,4,6,8,10,12} For each value of k, there will be 144 features generated, thus there are 864 = 144 x 6 RBST features.

2.7 Feature Optimization and Statistical Pattern Classification

Conventional mammography based CAD systems employ a single classifier system to classify and label mammograms as either malignant or benign. This is typically preceeded by a feature reduction and optimization step, in an attempt to reduce the size of training

data required to model the feature spaces, and to improve classification performance and generalization ability (ability to account for variability in statistics of training and testing data) of the classifier. Some common feature optimization techniques include PCA, LDA and S-LDA. These optimization techniques are at best sub-optimal. Recently, a mathematical argument was presented [Prasad *et al*, October 2008] showing that in many situations, a PCA projection can reduce the class separation in the projected space, thereby deteriorating the classification performance.

Stepwise Linear Discriminant Analysis (S-LDA)

LDA and it's variants are preferred linear feature optimization strategies, but when operating in high dimensional feature spaces, LDA can potentially fail if the amount of available training data is insufficient to estimate the within and between class scatter matrices. S-LDA employs a forward selection, backward rejection procedure to select a very small subset of features, upon which LDA is performed for feature optimization. Hence, by definition, it is sub-optimal, since it does not include a majority of the features for feature optimization. Since S-LDA will be used as a baseline feature optimization strategy, a brief description of S-LDA follows.

Stepwise linear discriminant analysis is a version of LDA which is a compromise for a full search of the feature space. In systems with a large feature vector size, an exhaustive search for the optimal solution is generally not feasible. Stepwise LDA (SLDA) with forward selection and backward rejection can be used for feature optimization. SLDA requires a discrimination metric to decide which features are better-suited to the given task, such as separating two different tumor types in medical imagery. Commonly used methods include receiver operating characteristics (ROC) area under the curve [Ball, May 2007], which is known as A_Z, Bhattacharyya Distance (BD), and Jeffries-Matusita distance (JMD).

In this work, A_Z is used as the discrimination metric, but theoretically, any appropriate metric could be used. The forward-selection procedure starts by calculating A_Z values for each feature separately, using one class as the target and all others as the non-target. The A_Z values are sorted in descending order. The feature with the highest A_Z value gets placed into a feature vector, and the ROC area, A_{Z_BEST} , is set to A_{Z1}. The second-best feature is then appended to the feature vector and A_{Z2} is computed. The second-best feature is only retained if $A_{Z2} > A_{Z\_BEST}$. In this case, A_{Z_BEST} is set to A_{Z2}. The third-best feature is then appended to the feature vector and A_{Z3} is computed. The third best feature is only retained if $A_{Z3} > A_{Z\_BEST}$. This process is continued until all the individual features are examined, or until the maximum number of features allowed is reached. The maximum number of the resulting features is determined by the minimum number of training signatures for a class. As a rule of thumb, for every five to ten training signatures, one feature can be added. Therefore, to keep five to ten features, there needs to be at least 25 to 50 training signatures for each class.

Next, backwards rejection is performed. Assume at this stage that there are b best features selected in the feature vector, and the best ROC area is A_{Z_BEST}. If $b = 1$, then no features may be removed, and the process halts. If $b > 1$, then the first feature is removed, and the ROC area A_{Z1}' is calculated. If $A_{Z1}' > A_{Z\_BEST}$, then the first feature is removed, and A_{Z_BEST} is set to A_{Z1}'. This process continues until all features have been removed and the ROC area has been recalculated. At the end of the procedure, there is a feature vector which contains the set of best features, the best A_Z value found, and the weighting coefficients.

The advantage of using SLDA is that it can produce very good results, even when individual features may not have very high A_Z values. Disadvantages of SLDA are (1) an exhaustive search is not performed, (2) a large percentage of (potentially useful) features are discarded and not considered for the classification task, and, (3) features near the end of the feature vector with tie scores to features earlier in the vector may not be chosen.

The Multi-Classifier Decision Fusion (MCDF) Framework

In recent work, Prasad *et al* [Prasad *et al*, May 2008] proposed a new divide-and-conquer paradigm for classification in high-dimensional feature spaces as pertaining to hyperspectral classification in remote sensing tasks. In this chapter, we show that this framework can be extended to other high dimensional feature spaces (features extracted from mammograms in this case), for robust classification, even when the amount of training data available is insufficient to model statistics of the classes in the high-dimensional feature space.

Figure 5 illustrates the proposed divide-and-conquer framework for mammogram classification. The algorithm is as follows. Find a suitable partition of the feature space, i.e., identify appropriate subspaces (each of a much smaller dimension). Perform "local" classification in each subspace. Finally, employ a suitable decision fusion scheme to merge the local decisions into a final malignant/benign decision per mammogram image. In our work with hyperspectral imagery, we found that the correlation structure of the feature space was approximately block-diagonal. This permitted the use of a correlation or mutual

Fig. 5. The proposed MCDF framework. Training data is employed to learn an appropriate feature grouping, feature pre-processing (optimization) and class-conditional statistics. To classify mammograms as malignant/benign, the feature extraction, feature grouping and pre-processing is followed by independent classification. Each class label/posterior proabability is then combined using a decision fusion mechanism.

information based metric in the partitioning of the corresponding feature space into multiple contiguous subspaces [Prasad *et al*, May 2008]. However, unlike hyperspectral data, where the feature space comprises of reflectance values over a continuum of wavelengths, features extracted from mammogram images typically do not possess a standard correlation structure to them. This is primarily because these features are created by concatenating various different kinds of quantities, such as morphological characteristics, texture information, patient history etc. Hence, in an attempt to define a suitable partition of the feature space derived from mammogram images, we break up the feature space into small groups, each comprised of m adjacent features, where m is a small integer valued number, determined experimentally. In previous work, Ball [Ball, May 2007] found that when doing a forward selection and backward rejection of mammography features, patient age was always selected as an important feature in the final feature selection. Hence, in this work, patient age was injected into each partition/subspace generated above to strengthen each local classifier.

Since each subspace is of a much smaller dimensionality than the dimension of the original feature space, a suitable preprocessing (such as LDA) may prove beneficial before making the local classification decisions. After creating multiple subspaces as described above, an LDA based pre-processing is performed in each subspace. The benefits of LDA based pre-processing are well known and documented in the pattern classification literature [Fukunaga, 1990]. Employing LDA in the proposed setup indeed further strengthens each classifier by improving class separation in the LDA projected space. Since the dimension of each subspace is small as compared to that of the original feature space, LDA based dimensionality reduction at the local subspace level is going to be well conditioned, even when a single LDA projection over the original feature space is ill conditioned. After LDA based pre-processing, a classifier is allocated to each subspace. The multi-classifier system is hence essentially a bank of classifiers that make "local" decisions in the partitioned subspaces. These can be parametric classifiers such as maximum likelihood classifiers, or non parametric classifiers such as k nearest neighbors classifiers, neural network based classifiers etc. In this work, we use quadratic maximum likelihood classifiers [Fukunaga, 1990]. These classifiers assume Gaussian class distributions for the i'th class, $p(x/w_i) \sim N(\mu_i, \Sigma_i)$. Assuming equal priors, the class membership function for such a classifier is given by

$$p(w_i \mid x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln\left|\Sigma_i\right|. \tag{15}$$

Here, w_i is the class label, x is the feature vector in the subspace, μ_i and Σ_i are the mean vector and covariance matrix of the i'th class respectively. Local classification decisions from each subspace are finally merged (fused) into a single class label (malignant or benign) per mammogram using an appropriate decision fusion rule. Decision fusion can occur either at the class label level (hard fusion), or at the posterior probability level (soft fusion). We test our system with decision fusion at both of these levels.

In hard decision fusion, we arrive at a final classification decision based on a vote over individual class labels (hard decisions) from each subspace. Unlike soft fusion based techniques, the overall classification of majority voting (MV) based fusion is not very sensitive to inaccurate estimates of posterior probabilities. However, in situations where posterior probabilities can be accurately estimated, soft fusion methods are likely to provide stable and accurate classification. A form of majority voting that incorporates a non uniform weight assignment [Prasad *et al*, May 2008] is given by:

$$w = \arg\max_{i \in \{1,2...C\}} N(i)$$
$$\text{where, } N(i) = \sum_{j=1}^{n} \alpha_j I(w_j = i) \quad , \tag{16}$$

where α_j is the confidence score / weight (e.g., training accuracies) for the *j*'th classifier, *I* is the indicator function, *w* is the class label from one of the *C* possible classes for the test pixel, *j* is the classifier index, *n* is the number of subspaces / classifiers, and *N(i)* is the number of times class *i* was detected in the bank of classifiers. A popular soft decision fusion strategy – the Linear Opinion Pool (LOP) uses the individual posterior probabilities of each classifier (*j* = 1, 2,… *n*), $p_j(w_i/x)$ to estimate a global class membership function:

$$C(w_i \mid x) = \sum_{j=1}^{n} \alpha_j p_j(w_i \mid x)$$
$$w = \arg\max_{i \in \{1,2...C\}} C(w_i \mid x) \tag{17}$$

This is essentially a weighted average of posteriors across the classifier bank. In this work, uniform weights are assigned to decisions from the bank of classifiers, although, theoretically, non-uniform weight assignments can be made using "training accuracy assessment" [Prasad *et al*, May 2008].

In addition to resolving the over-dimensionality and small-sample-size problems, the MCDF framework provides another advantage – Irrespective of whether we use hard or soft decision fusion, it provides a natural framework to fuse information from different modalities (in this case, different types of physical features extracted from mammograms), and hence allows simultaneous exploitation of a diverse variety of information.

3. Experimental Setup and Classification Performance

Classification experiments were conducted in the proposed framework using the database described in section 2.4, using a leave-one-out (i.e. N-fold cross validation) testing methodology [Fukunaga, 1990] for unbiased accuracy estimates. Under this scheme, a mammogram is sequestered for testing, while the system (all sub-components, i.e., Multi-classifier system, LDA etc.) is trained on the features extracted from the remaining mammograms. This is repeated in a round-robin fashion until all mammograms have been employed for testing. After segmenting the region of interest, features were extracted for each mammogram. After this stage, the multi-classifier decision fusion framework takes over. This was repeated per iteration of the leave-one-out scheme.

Table 6 depicts the results from the experimental setup described above. Results from a conventional stepwise LDA, single classifier system are also included as baseline (for comparison). Performance of a binary classification in CAD applications is typically quantified by: (1) Overall accuracy - proportion of correctly identified malignant and benign cases, (2) Sensitivity -proportion of true positives (true malignant cases) identified correctly, and, (3) Specificity - proportion of true negatives (true benign cases) identified correctly. These numbers, expressed in percentage are provided in Table 6. The 95% Confidence interval in estimation of the overall accuracy is also reported, in order to account for the

Table 6. Classification performance of the proposed system with the DDSM dataset. OA: Overall Accuracy; CI: 95% Confidence Interval; SE: Sensitivity; SP: Specificity (all expressed in percentage); *m*: partition size

Stepwise LDA (Baseline)					*MV based fusion (Proposed)*				*LOP based Fusion (Proposed)*			
OA	*CI*	*SE*	*SP*	*(m)*	*OA*	*CI*	*SE*	*SP*	*OA*	*CI*	*SE*	*SP*
82	4	80	83	**2**	85	3.8	87	83	85	3.8	87	83
				3	**90**	**3.2**	**90**	**90**	**88**	**3.4**	**90**	**87**
				4	85	3.8	83	87	85	3.8	83	87
				5	80	4.2	77	83	82	4.1	80	83
				6	85	3.8	83	87	83	3.9	83	83
				7	82	4.1	80	83	82	4.1	80	83
				8	82	4.1	80	83	82	4.1	80	83
				15	78	4.4	73	83	78	4.4	73	83

finite sample size. Classification performance is studied over a range of values of *m*, the size of each partition in the multi-classifier framework.

For the baseline Stepwise LDA, single-classifier system, the overall accuracy, sensitivity, and specificity were 82%, 80%, and 83%, respectively. These values are all higher for the proposed multi-classifier, decision fusion system for small window sizes (e.g., feature subsets have dimensionality of 2,3,4), regardless of whether MV and LOP based decision fusion is utilized. This improvement is highest for $m = 3$, where the overall accuracy, sensitivity, and specificity were each 90% when MV based fusion is used. If *m* is increased, these performance metrics start to drop for the proposed system, and as *m* is increased to 15, the classification accuracy and sensitivity of the system eventually fall below that of the baseline system.

4. Conclusion

To conclude, the proposed multi-classifier, decision fusion system significantly outperforms the baseline single-classifier based system for small partition sizes (*m*). By employing the proposed system, the overall accuracy, sensitivity and specificity of the binary classification task improve by as much as 10%. Hence, the multi-classifier, decision fusion framework promises robust classification of mammographic masses even though the dimensionality of feature vectors extracted from these mammograms is very high. In this study, the multi-classifier decision fusion approach proves to be very promising and certainly warrants future study. The proposed information fusion based approach also provides a natural framework for integrating different physical characteristics derived from the mammogram images (e.g., combining morphological, textural and statistical information). Additional patient information, if available, can also be added to the feature stream without over-burdening the classification system. In future work, we will explore the benefits of a nonlinear pre-processing of the feature space and an adaptive weight assignment based decision fusion system within the proposed framework.

5. References

Agatheeswaran, A., "*Analysis of the effects of JPEG2000 compression on texture features extracted from digital mammograms*". Masters Thesis in Electrical and Computer Engineering. Starkville, MS: Mississippi State University, pp. 20-37, 42-43, Dec. 2004.

Andolina, V.F., Lillé, S.L., and Willison, K.M., *Mammographic Imaging: A Practical Guide*. New York, NY: Lippincott Williams & Wilkins, 1992.

American Cancer Society, "*American Cancer Society: Breast Cancer Facts & Figures 2005-2006*," pp. 1-28, 2006. Available: http://www.cancer.org/downloads/STT/CAFF2005BrF.pdf.

Ball, J. *Three Stage level Set Segmentation of Mass Core, Periphery, and Spiculations for Automated Image Analysis of Digital Mammograms*, Ph.D. in Electrical Engineering. Starkville, Mississippi: Mississippi State University, May 2007.

Burhenne, L. J. W., et. al., "Potential Contribution of 164 Computer-aided Detection to the Sensitivity of Screening Mammography," *Radiology*, vol. 215, no. 2, pp. 554-562, 2000.

Catarious, D. M., "*A Computer-Aided Detection System for Mammographic Masses*." Ph.D. Dissertation in Biomedical Engineering. Durham, NC: Duke University, Aug. 2004.

Catarious, D.M., Baydush, A.H., and Floyd, C.E., Jr., "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system," *Medical Physics*, vol. 31, no. 6, pp. 1512-1520, Jun. 2004.

Cheng, H.D., et. al., "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognition*, vol. 39, no. 4, pp. 646-668, Apr. 2006.

Egan, R.L., *Breast Imaging: Diagnosis and Morphology of Breast Diseases*. Philadelphia, PA: W. B. Saunders Co., 1988.

Egan, R., "*The new age of breast care*," Administrative Radiology, p. 9, Sept. 1989.

Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

Haralick, R. M., Dinstein, I., and Shanmugam, K., "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 610-621, Nov. 1973.

Heath, M., et. al., "Current status of the Digital Database for Screening Mammography," in *Digital Mammography*, N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, Eds. Boston, MA: Kluwer Academic Publishers, pp. 457-460, 1998.

Hughes, G. "*On the mean accuracy of statistical pattern recognizers*," *IEEE Trans. on Information Theory,* vol. 14, no. 1, pp. 55-63, 1968.

Huo, Z., Giger, M.L, Vyborny, C.J., and Metz, C.E., "*Breast Cancer: Effectiveness of Computer-aided Diagnosis – Observer Study with Independent Database of Mammograms*," Radiology, vol. 224, no. 2, pp. 560-568, 2002.

Kundel H. L. and Dean, P.B., "Tumor Imaging," in *Image Processing Techniques for Tumor Detection*, R. N. Strickland, Ed. New York, NY: Marcel Dekker, Inc., pp. 1-18, 2002.

Laws, K., "*Textured Image Segmentation*." Ph.D. in Electrical Engineering. Los Angeles, CA: Image Processing Institute, University of Southern California, Jan. 1980 a.

Laws, K., "Rapid Texture Identification," *Proc. of the Image processing for missile guidance seminar*, San Diego, CA, pp. 376-380, Jan. 1980 b.

Lillé, S. L., "Background information and the need for screening," *in Mammographic Imaging: A Practical Guide*. New York, NY: Lippincott Williams & Wilkins, pp. 7-17, 1992.

Peters, M.E., Voegeli, D.R., and Scanlan, K.A., *Breast Imaging*. New York, NY: Churchhill Livingstone, 1989.

Prasad, S., Bruce, L. M., "Limitations of Principal Components Analysis for Hyperspectral Target Recognition," in *IEEE Geoscience and Remote Sensing Letters*, Vol. 5, Issue 4, pp 625-629, October 2008.

Prasad, S., Bruce, L. M., "Decision Fusion with Confidence based Weight Assignment for Hyperspectral Target Recognition," in *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 46, No. 5, May 2008.

Qian, W., Clarke, L.P, Baoyu, Z., Kallergi, M. and Clark, R., "Computer assisted diagnosis for digital mammography," *IEEE Engineering in Medicine and Biology Magazine,* vol. 14, no. 5, pp. 561-569, Sep.-Oct. 1995.

Rangayyan, R. M., "The Nature of Biomedical Images," in *Biomedical Image Analysis*, M. R. Neuman, Ed. Boca Raton, FL: CRC Press, pp. 22-27, 2005.

Sahiner, B., Petrick, N., Heang-Ping, C., Hadjiiski, L.M., Paramagul, C., Helvie, M.A., and Gurcan, M.N., "Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization," *IEEE Trans. on Medical Imaging*, vol. 20, no. 12, pp. 1275-1284, Dec. 2001.

Sahiner, B., Chan, H.-P., Petrick, N., Helvie, M.A., and Goodsitt, MM., "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics*, vol. 25, no. 4, pp. 516-526, Apr. 1998.

Tabár, L. and Dean, P.B., *Teaching Atlas of Mammography*, vol. 2nd revised ed. New York, NY: Georg Thieme Verlag, 1985.

Tabár, L., Vitak, B., Chen, H.-H.T., Yen, M.-F. , Duffy, S. W., and Smith, R.A., "*Beyond randomized controlled trials*," Cancer, vol. 91, no. 9, pp. 1724-1731, 2001.

Voegeli, D. R., "Mammographic Signs of Malignancy," in *Breast Imaging*, R. L. Eisenberg, Ed. New York, NY: Churchhill Livingstone, pp. 183-217, 1989.

Willison, K. M. "Breast anatomy and physiology," in *Mammographic Imaging: A Practical Guide New York*, NY: Lippincott Williams & Wilkins, pp. 119-161, 1992.

New Developments in Biomedical Engineering
Edited by Domenico Campolo

ISBN 978-953-7619-57-2
Hard cover, 714 pages
Publisher InTech
Published online 01, January, 2010
Published in print edition January, 2010

Biomedical Engineering is a highly interdisciplinary and well established discipline spanning across engineering, medicine and biology. A single definition of Biomedical Engineering is hardly unanimously accepted but it is often easier to identify what activities are included in it. This volume collects works on recent advances in Biomedical Engineering and provides a bird-view on a very broad field, ranging from purely theoretical frameworks to clinical applications and from diagnosis to treatment.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen