Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Dissimilarity-based classification in the absence of local ground truth: Application to the diagnostic interpretation of chest radiographs

Y. Arzhaeva[a,*], D.M.J. Tax[b], B. van Ginneken[a]

[a]Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100 Q.S.4.300, 3584 CX Utrecht, the Netherlands
[b]Information and Communication Theory Group, Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

ARTICLE INFO

Article history:
Received 22 July 2008
Accepted 13 January 2009

Keywords:
Dissimilarity-based classification
Weakly labeled data
Computer-aided diagnosis
Medical image analysis
Chest radiography

ABSTRACT

In this paper classification on dissimilarity representations is applied to medical imaging data with the task of discrimination between normal images and images with signs of disease. We show that dissimilarity-based classification is a beneficial approach in dealing with weakly labeled data, i.e. when the location of disease in an image is unknown and therefore local feature-based classifiers cannot be trained. A modification to the standard dissimilarity-based approach is proposed that makes a dissimilarity measure multi-valued, hence, able to retain more information. A multi-valued dissimilarity between an image and a prototype becomes an image representation vector in classification. Several classification outputs with respect to different prototypes are further integrated into a final image decision. Both standard and proposed methods are evaluated on data sets of chest radiographs with textural abnormalities and compared to several feature-based region classification approaches applied to the same data. On a tuberculosis data set the multi-valued dissimilarity-based classification performs as well as the best region classification method applied to the fully labeled data, with an area under the receiver operating characteristic (ROC) curve (A_z) of 0.82. The standard dissimilarity-based classification yields $A_z = 0.80$. On a data set with interstitial abnormalities both dissimilarity-based approaches achieve $A_z = 0.98$ which is closely behind the best region classification method.

1. Introduction

Computer-aided diagnosis (CAD) is an important pattern recognition application. Statistical and structural pattern recognition methods as well as artificial neural networks have been employed in the diagnostic interpretation of medical images of different modalities and organs. The choice of a classification method for a particular application can be influenced by many factors, among them the availability of well-annotated training data. In this paper we consider weakly labeled medical images with diffuse local textural abnormalities and the task of distinguishing them from images without abnormalities.

In object recognition, weakly labeled data are often defined as images labeled only according to the presence or absence of the objects of interest. For the diagnostic interpretation of medical images, data are weakly labeled in the sense that the absence or presence of disease in an image is known, however, the location of a lesion and its precise delineation are not available. This is, in fact, a common situation in practice because manual annotation of lesions is laborious or even impracticable. Ill-defined diffuse abnormal changes in the local textural appearance of an organ are a clear example in case. Manual segmentation of textural abnormalities is unreliable due to high inter-observer variability. However, texture features extracted from small local patches are potentially very informative in this case. In this work we show that local information alone, without local labels, can give good discriminatory results.

To detect images with abnormalities, we address the absence of local ground truth for training by combining local texture features extracted from a large number of regions of interest (ROIs). In the context of object recognition, a similar classification problem was considered in [1]. The authors focused on combining local information as well and developed generative and discriminative approaches to the task. They showed that the generative model gave a higher classification accuracy but required some fully labeled images for initialization. The discriminative model was considerably less accurate than the generative one. In this work we have chosen a very different approach that does not require any parametric modeling or careful initialization.

We assume that normal images of the same organ bear more similarity to each other with respect to their textural appearance

[*] Corresponding author.
E-mail addresses: yulia@isi.uu.nl (Y. Arzhaeva), d.m.j.tax@tudelft.nl (D.M.J. Tax), bram@isi.uu.nl (B. van Ginneken).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.01.016

than normal images and images with abnormalities. We also assume that images with the same type of abnormalities are more similar to each other than to normal images. Therefore we propose to reflect the common nature of images belonging to the same class by using dissimilarity representations. In [2] this is defined as the representation of objects by their pairwise comparisons instead of feature vectors. A pairwise comparison is done by computing a measure of dissimilarity, or distance, between two objects. To construct a classifier on dissimilarities one represents each training object as a vector of distances to a set of prototype objects. The standard dissimilarity-based classification is described in [2,3]. In this paper we propose a modification to this strategy.

In the standard approach, a single dissimilarity measure is computed between two images, a test image and a prototype, thus reducing the abundance of local textural information to one quantity. Since an image in our approach is represented by a set of texture features each of which is computed at multiple locations, we propose to retain more information by computing dissimilarities for each feature separately. Then, instead of combining feature-based comparisons into one value as in [4,5], we use them as a vector to train a classifier. That allows us to build as many classifiers as there are prototypes, and to classify each test image several times. Subsequently, we combine the outputs of all classifiers into one posterior probability value.

Experimentally, we will focus on two specific data sets. These data sets have already been used in other research papers which enables us to compare our results with previously reported ones. Both data sets present challenging ill-defined textural abnormalities in chest radiographs. The first one is a database from a tuberculosis (TB) mass screening program, the second database contains images with interstitial lung disease (ILD). In [6–9] one or both of them were used as test data for algorithms to distinguish between normal and abnormal images. In the two best performing classification schemes applied in [8,9] local labeling was used for training. That allowed the supervised classification of ROIs as normal or abnormal, and the subsequent integration of local decisions into a decision about the whole radiograph. Local labeling, provided it is correct, provides more information for training the system, and hence, such systems are potentially more powerful than ones where local labeling is unavailable. Although the labeling implemented in [8,9] might not have been a perfect ground truth, we will use their results as benchmarks for our study.

The paper is organized as follows: Section 2 introduces the dissimilarity representation and classification in dissimilarity space, as well as dissimilarity measures we intend to use. The proposed classification approach is also explained in this section. In Section 3 we compare the results of different classification strategies applied to medical images. We discuss the results and methods in Section 4. Section 5 draws conclusions.

2. Dissimilarity representations

In statistical pattern recognition objects are usually described by feature vectors. When we consider images described by sets of texture features extracted from a large number of patches, the description of the whole image becomes extremely high-dimensional and therefore inefficient for learning. Dissimilarities provide a convenient alternative for an image representation. Moreover, the proximity-based representation is a natural way of describing the class of similar objects. A profound discussion on this subject can be found in [2,3], and we borrow notation from that work.

If T is a training set of size n, and R is a set of prototype objects of size $r, R=\{p_1,\dots,p_r\}$, then any $x, x\in T$, is represented by a vector of dissimilarities $D(x,R)=\{d(x,p_1),\dots,d(x,p_r)\}$, where d is a dissimilarity measure. Thereby any traditional classifier operating on a feature space can be built on the $n\times r$ dissimilarity matrix $D(T,R)$. Usually, R is a subset of T, or the same set as T. Objects in R can be randomly selected from T, or selected using a systematic approach (see [10] for a discussion on possible approaches). A test set S of s objects is also described in terms of their distances to R, i.e. by $s\times r$ dissimilarity matrix $D(S,R)$.

Defining a discriminative dissimilarity measure is as difficult as defining good features in traditional feature-based classification. It is logical to demand that the measure is non-negative. Another natural requirement for dissimilarities is to be relatively small for similar objects. To ensure that, is desirable for the measure to satisfy the triangle inequality condition, $d(x,y)\leqslant d(x,z)+d(z,y)$, for all x,y,z, or else the compactness of dissimilarity representations might be violated [2]. If the measure is also symmetric and definite, it becomes a metric. Metrics are preferred as measures of dissimilarity because many classification methods work in metric spaces. In this paper we consider only measures that are metrics.

For clarity, we use the terms "feature" and "feature vector" only for original measurements extracted from an object. In our case, these are texture measurements computed from a large number of image ROIs. In the dissimilarity-based classification framework, a vector of dissimilarities constitutes an image representation and is passed to a classifier.

2.1. Dissimilarity measures

Let us introduce several common measures suitable for the task of image classification. In the context of image retrieval images are often characterized by multi-dimensional histograms of their features. An example of such features is the distribution of pixel intensities in an image and in filtered versions thereof. In this study, dedicated texture features are extracted from numerous and uniformly placed ROIs in images. Similarly to pixel intensity histograms we can build one- or multi-dimensional histograms to estimate the probability density of these features or the density of their joint distribution. Several non-parametric measures of dissimilarities between two histograms $h=\{h(i)\}$ and $k=\{k(i)\}$, i being a bin index, will be experimentally investigated in this paper.

Minkowski, or l_p, distance:

$$d_p(h,k)=\left(\sum_i |h(i)-k(i)|^p\right)^{1/p}. \tag{1}$$

For $p=1$, this becomes the city block distance, and for $p=2$, the Euclidean distance.

χ^2 statistics:

$$d_{\chi^2}(h,k)=\sum_i \frac{(h(i)-m(i))^2}{m(i)}, \tag{2}$$

where $m(i)=(h(i)+k(i))/2$. This measure calculates how unlikely it is that both histograms represent the same distribution.

Jeffrey divergence:

$$d_J(h,k)=\sum_i \left(h(i)\log\frac{h(i)}{m(i)}+k(i)\log\frac{k(i)}{m(i)}\right), \tag{3}$$

where again $m(i)=(h(i)+k(i))/2$. The Jeffrey divergence is a modification of the Kullback–Leibler divergence [11] and is numerically stable, symmetric and robust with respect to noise and the size of histogram bins [4].

Match distance:

$$d_M(h,k)=\sum_i |\hat{h}(i)-\hat{k}(i)|, \tag{4}$$

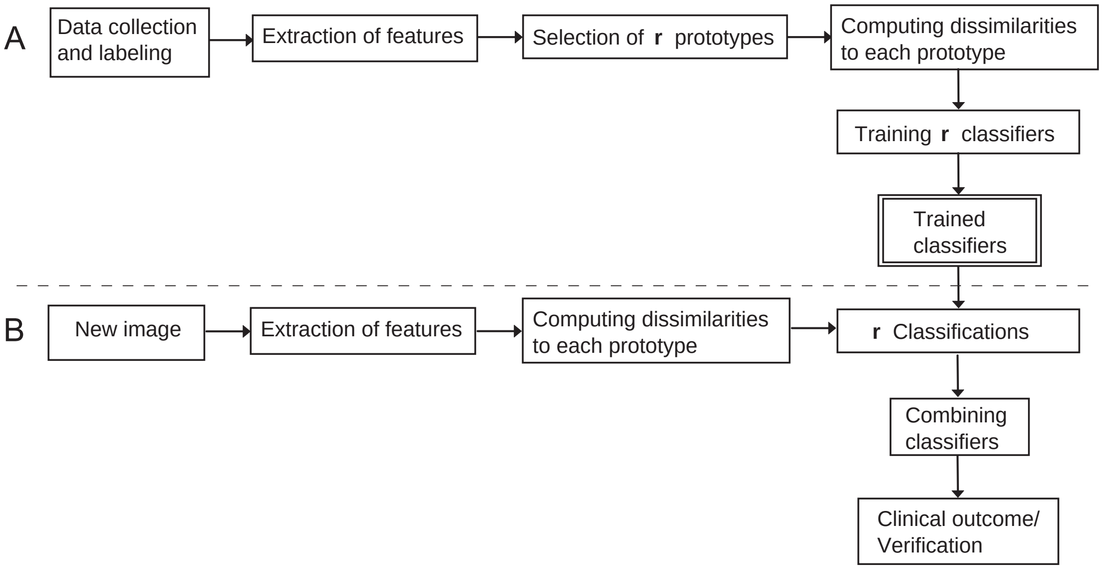

A
Data collection and labeling
Extraction of features
Selection of **r** prototypes
Computing dissimilarities to each prototype
Training **r** classifiers
Trained classifiers
B
New image
Extraction of features
Computing dissimilarities to each prototype
r Classifications
Combining classifiers
Clinical outcome/ Verification

Fig. 1. Flow chart of the proposed approach. (A) Training phase. (B) Testing phase.

where $\hat{h}(i)=\sum_{j\leqslant i}h(j)$ and $\hat{k}(i)=\sum_{j\leqslant i}k(j)$ are cumulative histograms of h and k, respectively.

Kolmogorov–Smirnov distance:

$$d_{KS}(h,k)=\max_i(|\hat{h}(i)-\hat{k}(i)|), \tag{5}$$

where again $\hat{h}(i)$ and $\hat{k}(i)$ are cumulative histograms. The match and Kolmogorov–Smirnov distances are only defined for one-dimensional histograms, because the ordering relation $j\leqslant i$ is arbitrary in more than one dimension [12].

A one-dimensional histogram is obtained by a suitable binning of the range of feature values. However, we do not apply binning for multidimensional joint feature distributions in order to avoid sparse and unstable histograms. Instead, we use a clustering algorithm such as k-means to partition the feature space into a fixed number of bins [12].

A dissimilarity between two images x and y can be expressed as the dissimilarity of their joint feature distributions, $d(x,y)=d(h,k)$, where h and k are the multi-dimensional histograms of the features of x and y, respectively. Dissimilarity measures from Eqs. (1) to (3) are used for this purpose in the experimental part of this study. This is by no means an exhaustive list of suitable dissimilarity measures (see [2,12] for more). Besides comparing their joint feature distributions, a dissimilarity between two images can be computed by combining independently evaluated comparisons of individual feature distributions. Just as in [4], we use the Minkowski norm of order 1 to combine them, $d(x,y)=\sum_f d(h_f,k_f)$, where $d(h_f,k_f)$ is any of the measures from Eqs. (1) to (5) computed for the histograms h_f and k_f of the feature f of the images x and y, respectively.

2.2. Proposed approach

We propose, however, not to combine the individual comparisons $d(h_f,k_f)$ into one value but to construct a vector $D(x,y)=\{d(h_f,k_f)\}$ and use it as a new image representation. The main difference between this approach and the standard dissimilarity-based classification lies in how an image is represented through its comparisons with the prototypes. In the standard approach, each element of the image representation vector $D(x,R)$ expresses the dissimilarity of the image x to a different prototype from the set R. In our representation, each element of the vector $D(x,p_k)$ is a dissimilarity between the image x and the same prototype p_k, $p_k\in R$, computed with respect to a different image characteristic. Here those characteristics are the distributions of various features extracted from both images.

Thus, we can obtain as many image representations as we have prototypes, each representation vector having the same dimensionality as the set of original features. With r prototype images, r representations are obtained for each training image, and consequently r classifiers can be trained. A test image, subsequently, can be classified r times using its prototype-bound representations. We suggest combining the outputs of all classifiers to obtain a final solution. In Fig. 1 the training and testing phases of the proposed approach are schematically depicted.

The combination of classifiers benefits from complementary information provided by different image representations. Various fixed, trainable and adaptable combiners have been described in the literature (see [13] for references). In the absence of a large pool of training data we opt for a fixed combination rule and leave the exploration of trainable schemes for further research. It is shown in [14] that the sum (or average) rule outperforms other fixed rules (such as the voting and product rules) for combining classifiers that use different representations of the patterns to be classified. The sum rule proved to be less sensitive to the errors of individual classifiers. In this paper we combine the image posterior probabilities resulting from different classifiers with the sum rule:

$$P(c|x)=\frac{1}{r}\sum_{k=0}^{r-1}P_k(c|x), \tag{6}$$

where $P(c|x)$ is a posterior probability that the image x belongs to a class $c, c=\{0,1\}$, and $P_k(c|x)$ is a posterior probability yielded by the classifier k.

3. Comparative experiments

In this section we apply both standard and proposed approaches of Section 2 to the classification of two sets of medical images exhibiting textural abnormalities. Additionally, we compare the dissimilarity-based methods to a region classification strategy adapted to weakly labeled data. The classification task is discrimination between normal images (of healthy subjects) and images

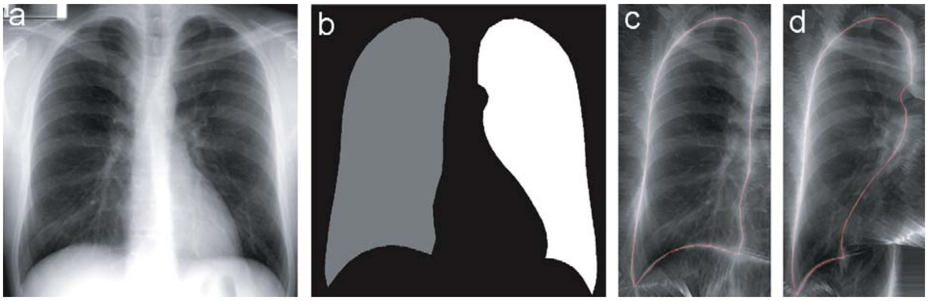


Fig. 2. Data preprocessing steps are shown on an example chest radiograph: (a) the original radiograph; (b) the lung mask obtained from the lung segmentation, with distinct mask values for the right and left lung fields; (c) the right lung, delineated, with its exterior substituted by corresponding pixel values from the lung inside; (d) the left lung, delineated and flipped, with its exterior substituted by corresponding pixel values from the lung inside.

containing disease (we refer to such images as abnormal throughout the paper).

3.1. Data

The TB database was collected from a TB screening program in the Netherlands. Posterior–anterior (PA) chest radiographs were digitized to 932 by 932 pixels and 12-bit intensity. More technical details can be found in [9]. The data set used in our experiments contains 241 normal cases and 147 abnormal cases with textural abnormalities. These cases were selected from a larger database by exclusion of images with non-textural abnormalities as well as images with artifacts (e.g. clothing artifacts). The same subset was used in [9]. The ground truth for the images was set by two radiologists. The image was considered abnormal if one of them judged the image to be abnormal.

The ILD database consists of 100 normal and 100 abnormal PA chest radiographs obtained from the daily clinical practice of the University of Chicago hospitals [15]. The abnormal radiographs exhibited various ILD and were selected on the basis of radiological findings, clinical and computed tomography data, biopsy and the consensus of the panel of experienced radiologists. Each normal case was chosen based on the consensus of the same panel. The radiographs were digitized to 2000 by 2000 pixels.

In both data sets the lung fields were segmented from the rest of the image using the Active Shape Model algorithm, description of which can be found elsewhere (e.g. in [16]). The segmentation was performed with the same settings of parameters as in [17]. Prior to feature extraction the resolution of images in both data sets was subsampled to 700 by 700. In Figs. 2(a) and (b) an example radiograph from the ILD database is shown, together with its lung mask obtained from the lung segmentation.

3.2. Texture features

In order to extract discriminative texture features the images are filtered with a multiscale filter bank of Gaussian derivatives, and the moments of histograms are calculated from regions in the derived images. Using multiple scales enables the characterization of texture elements of different sizes, and the analysis of local histograms considers the texture primitives regardless of their spacial distribution. This is a general approach to texture characterization [18,19]. The histogram moments were successfully used for automatic detection of textural abnormalities in chest radiographs [9,20] and for texture analysis in thoracic computed tomography scans [21].

Prior to filtering the image, pixel values in the lung fields are mirrored outside the lungs symmetrically with respect to the lung borders. Namely, for each pixel outside the lungs, the pixel value is substituted by its counterpart inside the lungs with the nearest pixel on the lung border as a center of symmetry. This prevents a major distortion in the filter output near the lung borders which is normally caused by a large difference in appearance inside and outside the lungs. Additionally, the left lung is flipped to resemble the right lung in orientation of various anatomical and texture elements. Figs. 2(c) and 2(d) illustrate the mirroring and flipping preprocessing steps.

The lung fields are subdivided into overlapping ROIs. We use an 8 by 8 pixel spacing to define the centers of circular ROIs, each of which have a radius of 32 pixels. The number of ROIs per radiograph ranges from 1400 to 4100 approximately depending on the size of individual lungs. Radiographs are filtered with Gaussian derivatives of orders 0, 1 and 2 at five scales, $\sigma = 1, 2, 4, 8, 16$ (illustrated in Fig. 3). Then four central moments of the pixel intensity distribution, namely, the mean, standard deviation, skewness and kurtosis are calculated from each ROI in the original and filtered images, amounting to 124 features in total. These are the same features that were successfully used in the classification of small regions in [22] and in [20] for localization of interstitial abnormalities.

In those works local texture features were complemented by two position features, namely *x* and *y* coordinates of the ROI centers relative to the center of the mass of the lung field. For this study we have assumed that the histograms of ROI locations covering lung fields uniformly would not be informative attributes in distinguishing between normal and abnormal lungs. For the multi-dimensional histograms of joint feature distributions adding position features could have resulted in the mistaken estimation of two abnormal images as dissimilar when their abnormalities were located in different lung regions. However, in practice, including position features brings minor improvements to dissimilarity classification results for both approaches. This could possibly be explained by an observation made in [9] that the spacial distribution of abnormal areas in the TB and ILD databases is not uniform. TB is known to often affect the upper lung fields, while interstitial abnormalities are more likely to occur in the lower lung fields.

The original features are normalized. In the set of prototypes, each feature is translated and scaled to have zero mean and unit standard deviation. Then the same normalization parameters are applied to feature vectors in the rest of the images.

3.3. Histograms

Before applying dissimilarity classification methods, the probability densities of local texture features have to be estimated. We represent the probability density of individual features by a histogram with 128 bins. The bin partitioning is fixed on a set of prototype

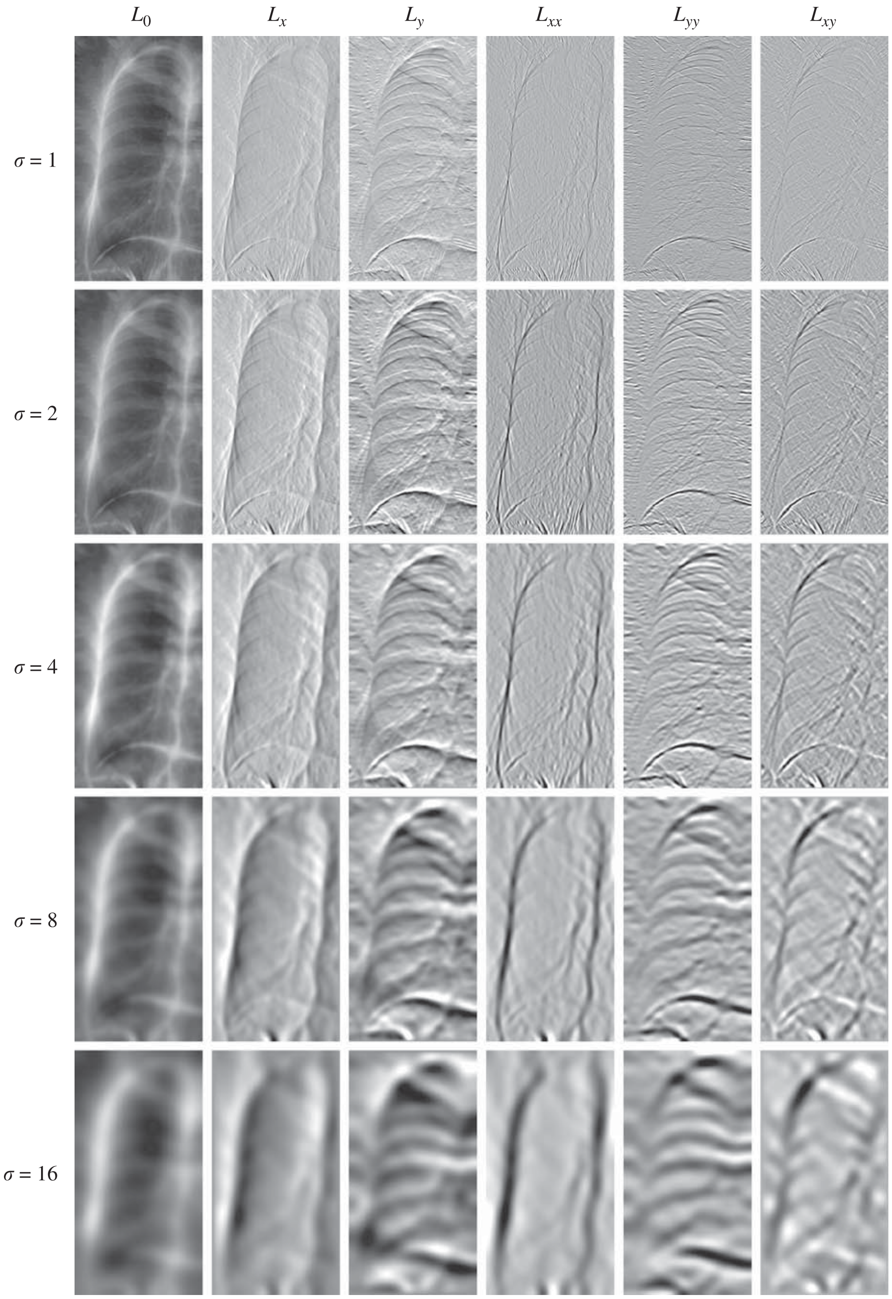


Fig. 3. Illustration of the 30 filtered images for the input image of a right lung from Fig. 2(c). The input image is convolved with Gaussian derivatives of orders 0, 1 and 2 at five scales. Each row shows the resulting images, L_0, L_x, L_y, L_{xx}, L_{yy} and L_{xy}, at one scale.

images. Namely, the range of possible values of each feature is estimated and split into 128 equal intervals.

To construct the multi-dimensional histogram of the joint distribution of features we first run a k-means algorithm with 128 clusters on the combined distribution of ROIs from all the prototypes. Then, for any image in a database, each feature vector is assigned to the closest cluster in the partitioned feature space.

3.4. Comparison with standard dissimilarity classification

In both databases we randomly selected 10 normal and 10 abnormal radiographs to serve as prototype images, and computed dissimilarities between each prototype and all the images in a database (including the prototypes themselves). For use with the standard dissimilarity-based classification approach each image was represented by a 20-dimensional vector computed with every appropriate dissimilarity measure described in Section 2.1. For the proposed multi-valued dissimilarity-based method each image was described by 20 126-dimensional image representation vectors computed using every dissimilarity measure for one-dimensional histograms from Eqs. (1) to (4).

The classification experiments were conducted by means of cross-validation. Each database, with exclusion of the prototypes, was divided into four folds, with equal amounts of normal and abnormal images in each fold. Classification was performed 4 times, each time with a different fold as a test set and the other three folds together as a training set. The prototype images were always appended to the

Table 1
The performances of the standard dissimilarity-based classification applied to the TB and ILD data sets for the dissimilarity measures described in Section 2.1.

Dissimilarity measure	TB data set	ILD data set
City block	0.715 (0.052)	0.936 (0.05)
Euclidean	0.693 (0.069)	0.924 (0.037)
χ^2 statistics	0.719 (0.034)	0.929 (0.047)
Jeffrey divergence	0.724 (0.033)	0.931 (0.04)
Combined city block	0.771 (0.04)	0.971 (0.014)
Combined Euclidean	0.767 (0.042)	0.976 (0.016)
Combined χ^2	0.797 (0.044)	0.974 (0.018)
Combined Jeffrey	0.798 (0.041)	0.974 (0.019)
Combined match distance	0.747 (0.08)	0.964 (0.014)
Combined Kolmogorov–Smirnov	0.748 (0.067)	0.966 (0.015)

The table presents the average and standard deviation of the areas under the ROC curve.

Table 2
The performances of the proposed multi-valued dissimilarity-based classification applied to the TB and ILD data sets for different dissimilarity measures.

Dissimilarity measure	TB data set	ILD data set
City block	0.820 (0.026)	0.974 (0.011)
Euclidean	0.817 (0.012)	0.970 (0.008)
χ^2 statistics	0.812 (0.020)	0.975 (0.010)
Jeffrey divergence	0.817 (0.018)	0.974 (0.014)
Match distance	0.793 (0.038)	0.961 (0.033)
Kolmogorov–Smirnov distance	0.775 (0.025)	0.978 (0.011)

The table presents the average and standard deviation of the areas under the ROC curve.

training set. We estimated the classification performances of both methods by means of receiver operating characteristic (ROC) analysis [23]. The ROC curve plots the sensitivity of a classifier against its 1-specificity at varying confidence thresholds. A_z, the area under the ROC curve, was used as a classification performance measure.

For both dissimilarity-based approaches we compared the linear discriminant analysis (LDA), quadratic discriminant analysis, k-nearest neighbor classifier ($k = 15$), and support vector machine (radial basis function kernel, the kernel parameter $g = 1.0$ and penalty parameter $C = 1.0$). Details of these classifiers can be found elsewhere, i.e. in [24]. We found that the LDA performed considerably better than the other classifiers did with the same fixed test, training and prototype sets. We think that one of possible causes for this is the simplicity of the LDA classifier. Another likely explanation could be that the dissimilarity measures are based on the summation over many components, and therefore tend to be normally distributed. Hence, normal density-based classifiers, such as the linear and quadratic discriminants, should perform well in dissimilarity spaces, as was already observed in [3]. Moreover, since the LDA is a linearly weighted combination of dissimilarities, it is less sensitive to errors caused by some individual dissimilarities. The quadratic discriminant analysis might have performed well in our experiments had we had more training samples to accurately estimate the class covariance matrices.

For the LDA, the number of features might still be too large relatively to the number of training samples in the multi-valued dissimilarity-based experiment, especially in the ILD database. In an attempt to resolve this we applied principal component analysis (PCA) retaining 99% of variance to the image representation vectors. This improved the multi-valued dissimilarity classification performance on the ILD data.

Table 1 displays the results of the standard dissimilarity-based classification for all the dissimilarity measures under consideration. The results of the proposed method for different dissimilarity measures are presented in Table 2. Note, that the results in Table 2 on the ILD database were obtained by application of PCA while no PCA was applied to the TB data. In both tables A_z values are averaged over the folds and accompanied by the standard deviation. Examples of radiographs, correctly classified or misclassified by the proposed method, are given in Fig. 4.

The image in Fig. 4(d) is the instructive example of a situation where the dissimilarity-based methods can fail. Subtle abnormalities with a small size relative to the whole area of interest are unlikely to be discernibly reflected in global measures, such as histograms over the whole lung fields. On the other hand, the normal image shown in Fig. 4(b) exhibits an enlarged perihilar region in the upper and middle right lung with a bright and pronounced texture pattern. We assume that the contribution of that pattern into global histograms caused the misclassification of this image as abnormal by our system. Correct classification of the perihilar region is also difficult for a region-based classification because of its bright and pattern-rich manifestation and large variability [20].

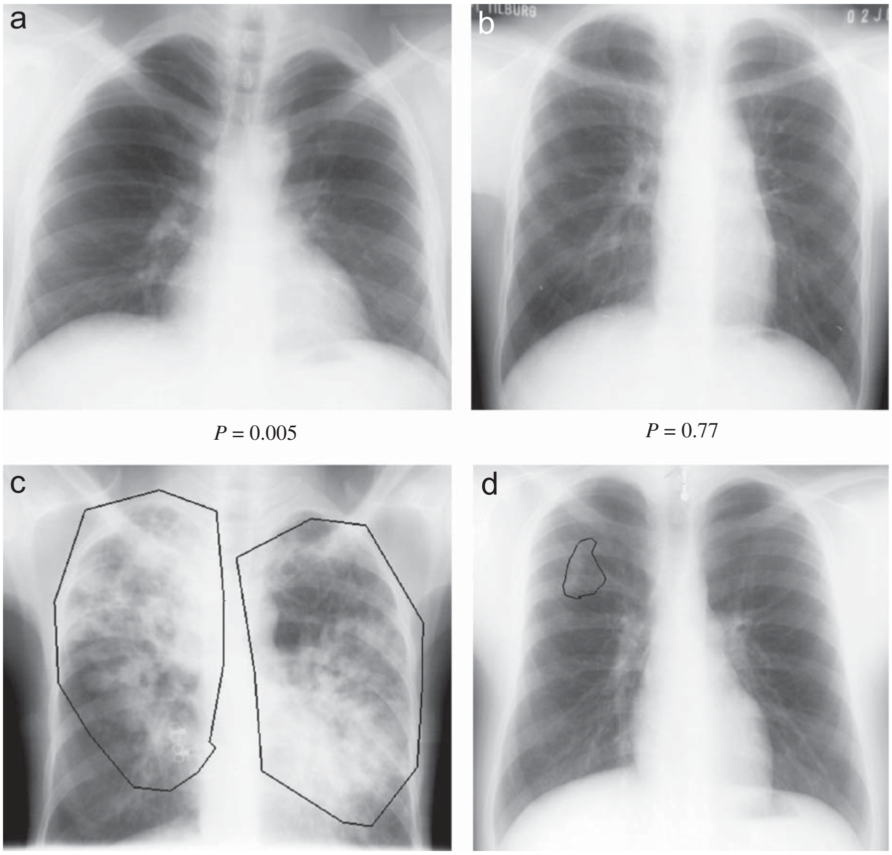
Overall, the standard dissimilarity-based classification performed as well as the multi-valued approach on the ILD data. On the TB data the standard dissimilarity-based classification were less accurate than our modification. From Table 1 we can also conclude that the performance of the standard dissimilarity-based method largely depended on what type of dissimilarity measures were employed. All combined measures were superior to the measures computed between multi-dimensional histograms. It is likely that 128 clusters in the 126-dimensional feature space was a rather coarse histogram partitioning which made histograms of different classes less distinctive; with 128 clusters we got from 11 to 32 entries per histogram bin on average. Given a fixed number of samples, a considerable increase in the number of bins could lead to a histogram sparseness which, in turn, could make a histogram less discriminative as well. The one-dimensional histogram binning did not have that problem and produced, possibly, more discriminatory histograms.

Concerning the utilized dissimilarity measures, the χ^2 statistics and Jeffrey divergence were consistently successful in both classification approaches. The match and Kolmogorov–Smirnov distances were the least stable of all the measures. All the measures performed better in the multi-valued approach than in the standard approach when applied to the TB data. In the application to the ILD data, the Euclidean and match distances showed slightly better performances in the standard settings. The differences in performance between the measures were rather small in both approaches applied to the ILD data. On the contrary, when applied to the TB data, the χ^2 statistics and Jeffrey divergence performed considerably better than the other combined measures in the standard approach. The performance differences between measures are less striking in the multi-valued approach, with four measures showing nearly identical results.

3.5. Comparison with region classification

To put the dissimilarity-based approaches in perspective we compare them with another strategy to deal with weakly labeled data. It is based on a naive assumption that every pixel in an abnormal image is abnormal. Hence, every region extracted from the lung fields as described in Section 3.2 is labeled according to the radiograph it belongs to. In this way the absence of local labels is circumvented. The image classification task then can be considered as a region classification and subsequent fusion of regional posterior probabilities. Such an approach was first proposed in [25].

In the region classification experiment the same division of the data into folds was applied. The images used as prototypes in the dissimilarity-bases experiments were added to the training sets of each fold. Each ROI was described by a 126-dimensional normalized feature vector consisting of 124 texture features and two position features (see Sections 3.2 and 3.3). For practical reasons we randomly



$P = 0.005$

$P = 0.77$

$P = 0.996$

$P = 0.07$

Fig. 4. Two normal, (a) and (b), and two abnormal, (c) and (d), cases from the TB database. Abnormalities are roughly outlined on images (c) and (d). The proposed dissimilarity-based method with the city block dissimilarity measure found image (a) most normal and image (b) most abnormal of all normal images. Image (c) was found most abnormal and image (d) most normal of all abnormal images. Estimated probabilities of being abnormal are indicated below each image.

Table 3
Comparison of different classification strategies in terms of A_z.

Study or method	TB data set	ILD data set
van Ginneken [9]	0.820 (0.022)	0.986 (0.006)
Ishida [8]	N.a.	0.976 (0.012)
Naive region classification	0.786 (0.035)	0.962 (0.031)
Standard dissimilarity-based approach (best result)	0.798 (0.041)	0.976 (0.016)
Multi-valued dissimilarity-based approach (best result)	0.820 (0.026)	0.978 (0.011)

selected 20% of the ROIs from each training image to train a classifier. As we already saw in [20], the LDA was a good choice of a classifier to discern between normal and abnormal ROIs. The classification yielded the posterior probability of being abnormal for each ROI in the image. The overall image decision was obtained by integrating the regional posterior probabilities using the 90% percentile rule.

The classification performance was evaluated in terms of A_z and averaged over the folds. We achieved $A_z = 0.786$ with a standard deviation of 0.035 on the TB database, and $A_z = 0.962$ with a standard deviation of 0.031 on the ILD database.

4. Discussion

The merits of the two dissimilarity-based methods in this particular application can be evaluated in the light of previous research. In Table 3 the results of different classification strategies applied to the TB and ILD databases are given. The first two rows present the results from the studies where, as mentioned in Section 1, ROIs were labeled and classified, and then the results of region classification were integrated into an overall image decision. The third row holds the results of region classification cf. Section 3.5. In the last two rows the results of the two dissimilarity-based approaches are presented, taking the best results from Tables 1 and 2.

It is interesting to note that the best and worst approaches for both data sets are region classification techniques. The best approach described in [9] used the manual annotation of lesions in the images. Although it was a rough annotation, it was obviously a better ground truth than the naive region abnormality assumption employed as a labeling strategy in Section 3.5. It should be noticed, however, that the gap between the best and worst classification results is relatively small. The multi-valued dissimilarity-based approach shows the same result as the best performing region classification on the TB data. On the ILD data both dissimilarity-based methods perform similarly to the region-based classification described in [8]. This comparison gives reason to hypothesize that the dissimilarity-based approaches are advantageous in dealing with weakly labeled textural data because they are capable of achieving results close to or even equal to those obtained with data labeled as fully as possible.

We suppose that the dissimilarity-based methods might not be equally useful in dealing with weakly labeled images with other types of abnormalities, e.g. in detecting chest radiographs with lung nodules. Similar to the abnormality in Fig. 4(d), many lung nodules are too small to possibly make a difference in a global measure, unless such a measure is a dedicated lung nodule filter of some sort. A more important consideration, however, is how well- or ill-defined a type of abnormality is. From the beginning, the application of the dissimilarity-based methods was motivated by the difficulty or impracticability of obtaining the local ground truth. When a reliable local ground truth is available to train a CAD system, we would suggest to use detection algorithms that can directly employ local information.

Next we discuss how the results of the standard dissimilarity-based approach compared with its multi-valued modification. The former used one-dimensional histograms of individual feature distributions in order to produce combined dissimilarity measures, while the latter used comparisons between feature histograms directly in the image representation vectors. Both strategies yielded comparable classification performances, which is not unexpected since image posterior probabilities are conveyed through a weighted sum of feature dissimilarities by both approaches.

The LDA, applied in the standard approach, results in the linear combination of dissimilarities to all the prototypes, where a dissimilarity to each prototype is, in turn, a linear combination of individual feature dissimilarities. In the multi-valued approach, the LDA first yields the linear combination of individual feature dissimilarities to one prototype. Then, by averaging the LDA results over all the prototypes in the classifier combination phase, we obtain the same type of additive solution as in the standard approach. It seems that the order of application of the LDA is what makes the proposed multi-valued approach slightly more accurate than the standard classification on dissimilarities. This assumption conforms with our initial idea that a classifier applied to feature-based dissimilarities rather than to image-based dissimilarities should benefit from more information.

Regarding the dissimilarity measures, it was noted in Section 3.4 that the χ^2 statistics and Jeffrey divergence measures showed comparable classification results in both approaches on both the TB and ILD data sets. While the city block and Euclidean distances could be considered as general-purpose measures in the Euclidean space, the χ^2 statistics and Jeffrey divergence are dedicated measures for probability distributions originating from statistical and information theory, respectively. That might explain their reliable performance in classifying features based on comparisons between distributions. The match and Kolmogorov–Smirnov distances are special distance measures for cumulative histograms and are known to produce better results with finer binning [12]. In our experiments they performed less satisfactorily than the other measures, with the exception of the Kolmogorov–Smirnov distance yielding the best result on the ILD data in the multi-valued approach. They were also less stable, generally, exhibiting some of the largest standard deviations. We may only hypothesize that with finer histogram binning and more samples, the results of these two measures could improve.

It was beyond the scope of this paper to investigate the use of prototype selection methods. In [10] it is argued that the systematic selection of prototypes in general does better than the random selection. With the random selection of prototypes we have already achieved results that are closely comparable with those obtained in the presence of local ground truth. It could be an intriguing future study to investigate whether adding prototype selection notably improves the classification performance. Achieving better results on a weakly labeled data set than on fully labeled data would indicate either unsatisfactory local ground truth used by the feature-based approaches in classification of regions, or some room for improvement in the region classification method itself.

5. Conclusion

In conclusion, we successfully applied a dissimilarity-based classification approach to weakly labeled chest radiographs with textural abnormalities. The obtained results were similar to those obtained by feature-based methods on fully labeled data. Our proposed modification to the standard dissimilarity-based approach was preferable in the classification of TB, while both dissimilarity-based methods performed equally well in the classification of interstitial abnormalities. The application of these techniques to other weakly labeled image data is of interest for future research.

Acknowledgments

This research was supported by the Dutch Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs.

References

[1] C. Bishop, I. Ulusoy, Object recognition via local patch labelling, in: J. Winkler, N. Lawrence, M. Niranjan (Eds.), Workshop on Machine Learning, 2004, pp. 1–21.

[2] E. Pekalska, Dissimilarity representations in pattern recognition, Ph.D. Thesis, Delft University, the Netherlands, 2005.

[3] E. Pekalska, R. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognition Letters 23 (2002) 943–956.

[4] J. Puzicha, T. Hofmann, J. Buhmann, Non-parametric similarity measures for unsupervised texture segmentation and image retrieval, in: Proceedings of Computer Vision and Pattern Recognition, IEEE Computer Society, 1997, pp. 267–272.

[5] D. Guru, B. Kiranagu, Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns, Pattern Recognition 38 (2005) 151–156.

[6] S. Katsuragawa, K. Doi, H. MacMahon, Image feature analysis and computer-aided diagnosis in digital radiography: classification of normal and abnormal lungs with interstitial lung disease in chest radiographs, Medical Physics 16 (1) (1989) 38–44.

[7] T. Ishida, S. Katsuragawa, K. Ashizawa, H. MacMahon, K. Doi, Artificial neural networks in chest radiographs: detection and characterization of interstitial lung disease, in: Proceedings of the SPIE, vol. 3034, 1997, pp. 931–937.

[8] T. Ishida, S. Katsuragawa, K. Ashizawa, H. MacMahon, K. Doi, Application of artificial neural networks for quantitative analysis of image data in chest radiographs for detection of interstitial lung disease, Journal of Digital Imaging 11 (4) (1998) 182–192.

[9] B. van Ginneken, S. Katsuragawa, B.M. ter Haar Romeny, K. Doi, M.A. Viergever, Automatic detection of abnormalities in chest radiographs using local texture analysis, IEEE Transactions on Medical Imaging 21 (2) (2002) 139–149.

[10] E. Pekalska, R. Duin, P. Paclik, Prototype selection for dissimilarity-based classifiers, Pattern Recognition 39 (2006) 189–208.

[11] Mathworld ⟨http://mathworld.wolfram.com/⟩.

[12] Y. Rubner, C. Tomasi, L.J. Guibas, The Earth mover's distance as a metric for image retrieval, International Journal of Computer Vision 40 (2) (2000) 99–121.

[13] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 4–37.

[14] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.

[15] S. Katsuragawa, K. Doi, H. MacMahon, Image feature analysis and computer-aided diagnosis in digital radiography: detection and characterization of interstitial lung disease in digital chest radiographs, Medical Physics 15 (3) (1988) 311–319.

[16] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models—their training and application, Computer Vision and Image Understanding 61 (1) (1995) 38–59.

[17] B. van Ginneken, M. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, Medical Image Analysis 10 (1) (2006) 19–40.

[18] M. Unser, M. Eden, Multi-resolution feature extraction and selection for texture segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 717–728.

[19] B. van Ginneken, B. ter Haar Romeny, Multi-scale texture classification from generalized locally orderless images, Pattern Recognition 36 (2002) 899–911.

[20] Y. Arzhaeva, M. Prokop, D.M.J. Tax, P.A. de Jong, C.M. Schaefer-Prokop, B. van Ginneken, Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography, Medical Physics 34 (12) (2007) 4798–4809.

[21] I.C. Sluimer, P.F. van Waes, M.A. Viergever, B. van Ginneken, Computer-aided diagnosis in high-resolution CT of the lungs, Medical Physics 30 (12) (2003) 3081–3090.

[22] Y. Arzhaeva, D.M.J. Tax, B. van Ginneken, Improving computer-aided diagnosis of interstitial disease in chest radiographs by combining one-class and two-class classifiers, in: J.M. Reinhardt, J.P.W. Pluim (Eds.), Proceedings of the SPIE, vol. 6144, 2006, p. 614458.

[23] C. Metz, ROC methodology in radiologic imaging, Investigative Radiology 21 (9) (1986) 720–733.

[24] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley, New York, 2001.

[25] M. Loog, B. van Ginneken, M. Nielsen, Detection of interstitial lung disease in PA chest radiographs, in: M. Jaffe, M. Flynn (Eds.), SPIE Medical Imaging: Physics of Medical Imaging, vol. 5368, 2004, pp. 848–855.

About the Author—YULIA ARZHAEVA received her M.Sc. degree from the Belorussian State University, Minsk, Belarus, specializing in Mathematics in 1998. She is currently working on her Ph.D. thesis on Computer-Aided Diagnosis of Interstitial Lung Disease under supervision of B. van Ginneken.

About the Author—DAVID M.J. TAX studied physics at the University of Nijmegen, the Netherlands, in 1996, and received Master degree with the thesis "Learning of structure by Many-take-all Neural Networks". After that he had his Ph.D. at the Delft University of Technology in the Pattern Recognition group, under the supervision of R.P.W. Duin. In 2001 he promoted with the thesis "One-class classification". After working for two years as a Marie Curie Fellow in the Intelligent Data Analysis group in Berlin, at present he is a post doc in the Information and Communication Theory group at the Delft University of Technology. His main research interest is in the learning and development of outlier detection algorithms and systems, using techniques from machine learning and pattern recognition. In particular, the problems concerning the representation of data, simple and elegant classifiers and the evaluation of methods have focus.

About the Author—BRAM VAN GINNEKEN is Associate Professor at the Image Sciences Institute where he is leading the Computer-Aided Diagnosis group. He studied Physics at the Eindhoven University of Technology and at Utrecht University. In March 2001, he obtained his Ph.D. at the Image Sciences Institute on Computer-Aided Diagnosis in Chest Radiography. He has (co-)authored over 35 journal publications. He is Associate Editor of IEEE Transactions on Medical Imaging, member of the Editorial Board of Medical Image Analysis and member of the program committees of the Image Processing and the Computer-Aided Diagnosis conferences of SPIE Medical Imaging.