Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography

Yulia Arzhaeva[a)]
Images Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands

Mathias Prokop
Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands

David M. J. Tax
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Pim A. De Jong
Department of Radiology, Meander Medical Center, Amersfoort, The Netherlands

Cornelia M. Schaefer–Prokop
Department of Radiology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Bram van Ginneken
Images Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands

(Received 19 April 2007; revised 12 September 2007; accepted for publication 13 September 2007; published 26 November 2007)

A computer-aided detection (CAD) system is presented for the localization of interstitial lesions in chest radiographs. The system analyzes the complete lung fields using a two-class supervised pattern classification approach to distinguish between normal texture and texture affected by interstitial lung disease. Analysis is done pixel-wise and produces a probability map for an image where each pixel in the lung fields is assigned a probability of being abnormal. Interstitial lesions are often subtle and ill defined on x-rays and hence difficult to detect, even for expert radiologists. Therefore a new, semiautomatic method is proposed for setting a reference standard for training and evaluating the CAD system. The proposed method employs the fact that interstitial lesions are more distinct on a computed tomography (CT) scan than on a radiograph. Lesion outlines, manually drawn on coronal slices of a CT scan of the same patient, are automatically transformed to corresponding outlines on the chest x-ray, using manually indicated correspondences for a small set of anatomical landmarks. For the texture analysis, local structures are described by means of the multiscale Gaussian filter bank. The system performance is evaluated with ROC analysis on a database of digital chest radiographs containing 44 abnormal and 8 normal cases. The best performance is achieved for the linear discriminant and support vector machine classifiers, with an area under the ROC curve (A_z) of 0.78. Separate ROC curves are built for classification of abnormalities of different degrees of subtlety versus normal class. Here the best performance in terms of A_z is 0.90 for differentiation between obviously abnormal and normal pixels. The system is compared with two human observers, an expert chest radiologist and a chest radiologist in training, on evaluation of regions. Each lung field is divided in four regions, and the reference standard and the probability maps are converted into region scores. The system performance does not significantly differ from that of the observers, when the perihilar regions are excluded from evaluation, and reaches $A_z=0.85$ for the system, with $A_z=0.88$ for both observers. © *2007 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2795672]

Key words: computer-aided detection, interstitial lung disease, chest radiography, texture analysis

I. INTRODUCTION

Conventional chest radiography is an important diagnostic examination for a variety of lung disorders, including interstitial lung disease (ILD). In recent years, computer tomography (CT) has become the modality of choice for the diagnostics of ILD.[1] However, chest radiography remains the first and most common examination in clinical practice. In comparison to CT, it is simple to perform and inexpensive. Therefore, the role of chest radiography is to provide an initial detection of abnormalities and a preliminary diagnosis and to give a recommendation for a subsequent CT examination.[2] Techniques for automated detection and characterization of abnormalities in chest radiography have been developed for about two decades.[3,4] In recent studies,[5,6] computer-aided detection (CAD) systems for chest radiographs have been demonstrated to be potentially useful tools leading to more accurate diagnoses for various lung diseases, including detection of ILD. Currently, the results of computer analysis are considered to play a complementary role in clinical practice as a second opinion.

ILD, also known as diffuse parenchymal lung disease, is the common term for more than 200 types of disorders, which may cause considerable morbidity and mortality.[2] The interstitium of the lung is the tissue between the air sacs, and when the interstitium is damaged the "textural appearance" of the lung is changed in radiological images. Detection and differentiation of ILD is an exceptionally difficult task, even for an experienced chest radiologist. The key radiological finding is widespread or focal shadowing with specific underlying patterns. The majority of interstitial diseases exhibit a reticular, nodular, or ground-glass pattern,[7,8] or a combination of these. Whereas a large variation of abnormal patterns can represent one type of ILD, radiographs of patients with different types of ILD may look alike. Moreover, the difference between normal and abnormal texture patterns is ambiguous even for human experts which is revealed by high interobserver variability.[2,9] As a result, development of a CAD system for the detection of ILD in chest radiographs is an extremely challenging task.

The majority of works in this field[10–16] used an approach that could be roughly divided into three steps. First, regions of interest (ROIs) were manually or automatically selected within the lung fields. From each ROI a set of texture features was computed. Then classification was performed using rule-based or pattern recognition methods, and as the result of classification an "opinion" (a class label or probability of being normal/abnormal) about each ROI was obtained. Finally, probabilities over regions were fused to yield a conclusion for the whole image, determining whether it contained any interstitial abnormalities.

The CAD systems exploiting this method were evaluated using receiver operating characteristic (ROC) analysis, and showed high performances when evaluated either as a standalone system or as an assistant to radiologists in the task of discrimination between normal chest radiographs and radiographs that contained signs of interstitial abnormalities.[5,15,17] Although the classification of images was based on classification of regions, only in van Ginneken *et al.*[15] was the CAD performance on regions itself evaluated. It appeared there that the classification performance was poorer at region level than at image level (the area under ROC curve values ranged from 0.67–0.93 for different regions, and reached 0.99 for images). In other words, the CAD system could not always distinguish between samples of healthy lung texture and regions affected by ILD. In Ref. 15, the reference standard for a region was set visually by one radiologist and may therefore not be considered highly reliable. In other previous studies such an evaluation was not carried out at all, possibly because of the absence of a region-level reference standard to compare with.

The accurate delineation of abnormalities is an anticipated ability of such a CAD system. The distribution of ILD throughout the lung is often related to a type of ILD (e.g., extrinsic allergic alveolitis is commonly found in the upper lobes, whereas usual interstitial pneumonia is seen mainly in the lower lobes and the lung periphery), and this has important differential diagnostic implications.[18] In our opinion, the main obstacle that prevents the construction of a system that localizes ILD abnormalities in chest radiographs is the difficulty of obtaining a reliable local reference standard. Without this, a CAD system cannot be verified and in many cases cannot be properly trained. Well-defined lesions, e.g., tumors, nodules, and calcifications, can be manually segmented or pinpointed and often histologically proven to be a lesion. However, manual segmentation is less suited to the delineation of interstitial lesions due to their diffuse and ambiguous appearance. In this article, we use an alternative way to establish a more reliable reference standard for chest radiographs.

In our previous work,[19] we demonstrated with small-size peripheral regions from digitized chest radiographs that local analysis could yield a high classification performance. In this article, we present a CAD system for detection of ILD lesions in complete posterior-anterior (PA) chest radiographs. The system is trained and tested on a relatively small set of digital chest radiographs to show its ability to locate interstitial abnormalities. An innovative method for obtaining the local reference standard is presented. The reference standard is established by using a CT scan of the same patient to estimate the positions of interstitial disease in a chest radiograph and, consequently, label each pixel within the lung fields on a radiograph as either normal or affected by ILD.

The labels are used to train the CAD system on a subset of our x-ray database (training data). In the rest of the database (test data) the labels are used as the reference standard for evaluation of the system performance. When the trained system is applied to a new chest radiograph, a probability of being abnormal is assigned to each lung pixel. A CAD outcome, either as a color-coded probability map or as regional scores, can be presented to a radiologist as an assisting tool. The performance is evaluated by means of ROC analysis and compared with the performance of two human observers on the same set of data.

The article is organized as follows: Section II describes the data and the CAD system starting with the system outline and the methodology of data collection and continuing with detailing of different parts of the system. Section III gives details of experimental setup and evaluation methods. In Section IV the results are presented. They are discussed in Section V, and conclusions are drawn in Section VI

II. MATERIALS AND METHODS

II.A. System outline

From an engineering point of view, most CAD systems, including the one presented in this article, have a typical design relying on a combination of image-processing and pattern recognition or artificial intelligence techniques. In Fig. 1 the scheme of our system is depicted. In the training phase (Fig. 1 row A) image data that represent the diversity of ILD manifestations are collected, together with normal image data. Images with pathology are annotated, i.e., lesions are delineated and given a subtlety rank. These are

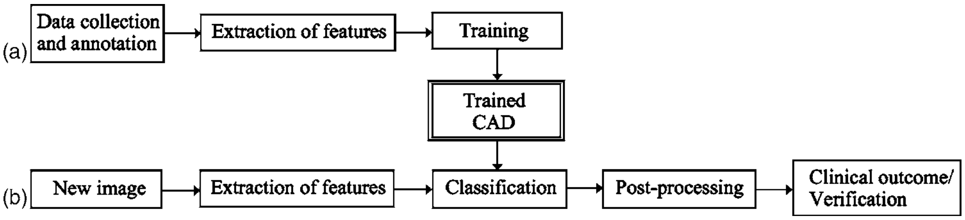

FIG. 1. Flow chart of the CAD system. (a) Training phase. (b) Testing phase.

training examples used to train the CAD system to distinguish between normal and abnormal patterns. To be able to train the system, pixels within the lung fields are represented by vectors of features computed from their neighborhood. After extraction of features a statistical decision model is constructed. The trained CAD system can yield an opinion about the presence of ILD lesions in a new image that was not included in the set of training examples. Pixels from a new chest radiograph, also represented by feature vectors, are classified according to the decision model and receive a probability of being abnormal (depicted in Fig. 1, row B). In the postprocessing stage a probability map is produced that accentuates areas with a high probability of being affected by ILD. If abnormality annotations (a reference standard) for this image are available, the outcome of the system can be verified.

II.B. Data set

For the study presented here we collected a number of digital PA chest radiographs from the picture archiving and communication system (PACS) of the University Medical Center Utrecht, the Netherlands. These images were acquired in a daily clinical practice between 2004 and 2006. Direct radiography units (Digital Diagnost, Philips Medical Systems) with a cesium iodine scintillator, a matrix of 3000 ×3000 pixel and 0.143 mm pixel size were used for acquisition. Images were exported to the PACS with 15 bits data depth. The PACS provides a single point of entry for all images and their associated data related to the same patient. In this study, patient data were treated in accordance with the Declaration of Helsinki. All patients registered in this system between 2004 and 2006 were considered, and among them chest radiographs were selected based on two criteria.

First, for a patient with a chest radiograph, a multislice chest CT scan, which was taken within one month before or after the x-ray examination, was also required. This time interval was chosen by an expert radiologist since the majority of interstitial diseases axe known to progress rather slowly.

Second, for an x-ray to be classed as normal, radiological reports associated with both images (x-ray and CT) were required to clearly indicate healthy lungs. For an x-ray to be classed as abnormal, either both reports or the CT report were required to refer to ILD or describe textures typical for ILD.

All CT images selected for the study were acquired on one of several multislice scanners (Philips Medical Systems, the Netherlands), namely, Brilliance-16P, Brilliance-40, Brilliance-64 and Mx8000 IDT 16, with standard parameters for high-resolution volumetric CT scanning. Collimation varied between 0.625 mm (40- and 64-slice scanners) and 0.75 mm (16-slice). Images of 0.9 mm thickness (40- and 64-slice) or 1 mm thickness (16-slice) were reconstructed every 0.7 mm. Exposure settings were 120 kVp and between 100 mAs and 170 mAs, depending on a scanner and patient size.

Normal and abnormal radiographs, selected in this manner, together with accompanying CT scans, were subsequently examined by an expert chest radiologist (Prokop, the second author, with more than 15 years of experience) who decided whether a chest x-ray and the corresponding multislice CT scan were normal (not containing interstitial abnormalities) or abnormal (containing signs of ILD), and whether the extent of abnormality was similar in both modalities. We included in the study the following types of chronic interstitial abnormalities: (a) focal pulmonary opacities, (b) diffuse interstitial reticular or linear changes, (c) diffuse nodular changes, (d) diffuse changes with increased parenchymal density, and (e) pulmonary diseases with cystic changes. Diffuse lung changes with decreased density were excluded.

From the 50 initially selected cases, six cases were excluded by the radiologist. They were excluded either because of an unclear diagnosis (the radiologist did not agree with the reported presence of ILD) or owing to a varying amount of disease manifestation on the CT scan and radiograph. The final set of abnormal images contained 44 cases. The average patient age was 58 (range 26–87 years, standard deviation 15 years). There is a higher prevalence of ILD in older patients. The gender distribution of patients was 21 males and 23 females.

After delineation of pathology in the abnormal images, eight normal radiographs were added to balance the total quantity of normal and abnormal tissue in the data. There were four males and four females, with an average age of 46 (range 20–81 years, standard deviation 19 years).

II.C. Reference standard

A multislice CT scan accompanying each chest radiograph is not only used to confirm a diagnosis but also to set up a reference standard on the corresponding radiograph. Thin section width and overlapping image reconstruction of multislice CT result in good quality two-dimensional (2D) image reformations in all directions. Moreover, such a 2D view is nonsuperimposed and has excellent contrast resolution. In Fig. 2, a one-voxel thick coronal plane of CT scan [Fig. 2(a)] is compared with a PA chest radiograph of the

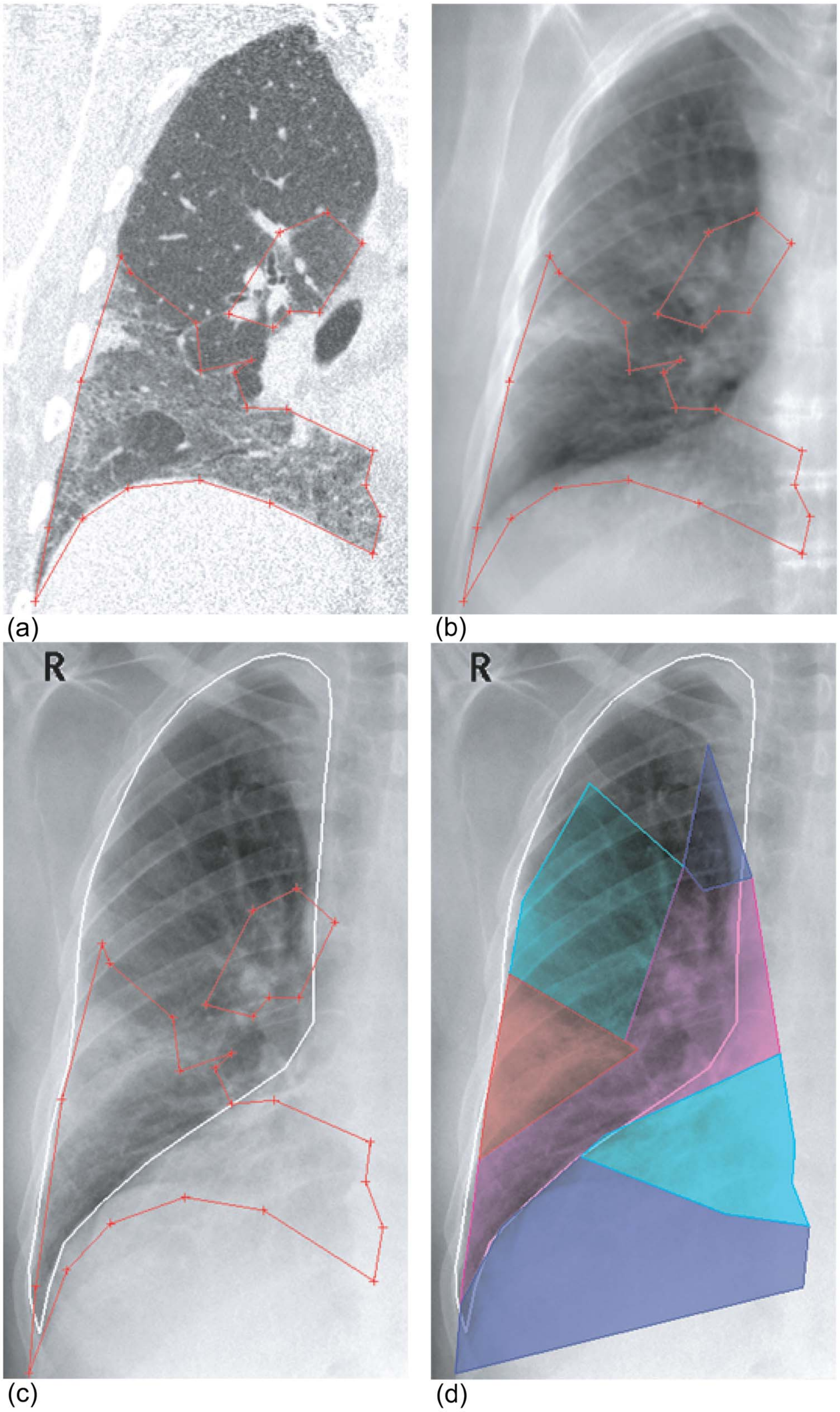

FIG. 2. An example of an interstitial lesion delineation. The borders between normal and abnormal texture are clearly visible on a CT slice (a) in contrast to an indistinct border on an x-ray (c). In order to segment all abnormal areas on an x-ray, several delineations have to be made on different CT slices that will combine into one or more lesion delineations on an x-ray. A combined projection of all lesion delineations made on CT slices is mapped onto an x-ray [in (d)]. An abnormality segmentation obtained in this way is divided into areas of different abnormality subtlety. The lung field is outlined in white. Areas lying within abnormality delineations but outside the lung fields are not considered in our system.

same patient [Fig. 2(c)]. Note that pathological areas stand out much clearer against the background lung tissue on the CT slice than on the radiograph. Manual, and even automatic (e.g., in Ref. 20), segmentation of interstitial abnormalities on CT slices is feasible whereas definite borders between pathological and normal lung tissue in conventional radiographs can barely be found. This point is clearly illustrated in Fig. 2.

The proposed method uses CT data as a superior gold standard and is based on delineation of abnormalities on coronal CT sections (the same orientation as the conventional chest radiograph), which are then transferred to the corresponding radiograph. Thus, we circumvent the inability to segment abnormalities directly on radiographs.

Note that the annotation of radiographs with use of CT scans of the same patients happens during the collection of training data in order to obtain a set of well-annotated radiographs that can be used to train a CAD system. Once the CAD system is trained, the analysis of a new radiograph does not require a supporting CT scan. For the test data we used the same method to establish a reliable reference standard in order to evaluate the system performance.

For each pair of an abnormal x-ray and a CT scan, interstitial abnormalities were manually delineated by the expert chest radiologist (MP) with a dedicated computer program built for this study. Delineations were performed on single 0.7 mm thick coronal slices selected at every 10 mm. In order to translate delineations to an x-ray, a mapping function is established between a coronal projection from the CT volume and the x-ray. The coronal projection is obtained for this purpose by averaging CT numbers in the coronal direction. In this way, the coronal projection approximates the radiograph [see Fig. 2(b)]. Deformation between the CT projection and the radiograph is found using radial basis functions as described in Ref. 21. This method requires a set of known corresponding points (control points). The same radiologist who segmented abnormalities indicated several (from 6 to 10, depending on the image) anatomically corresponding landmarks in the radiograph and CT projection to be used as control points. The mapping function is constructed based on the control points and applied to the vertices of the abnormality outlines. As a result, the corresponding outlines in the radiograph are obtained [see Fig. 2(c)]. Their shapes can be corrected manually, if deemed necessary. Superimposed outlines were replaced by their union. Usually, the final delineation of a lesion on the x-ray corresponds to a combination of several delineations made on different CT slices.

The radiologist made no corrections to outlines transferred to the radiographs. However, in 14 cases one or both lungs were completely affected by ILD, and per slice delineations were deemed unnecessary. In those cases, the lung boundaries in a radiograph were used to define an abnormal area. In ten cases the final delineations were slightly corrected for smoothness.

In Fig. 2(d) an example of a final outcome of the described segmentation procedure is presented. Additionally, the abnormal areas in each radiograph were divided into areas of different abnormality subtlety of disease. This was done by the same radiologist who also had defined the abnormality areas based on the CT findings. For subtlety assigning, no information from the CT scan was used. The expert radiologist's judgment was based on his visual assessment of the radiograph. Four levels of subtlety in detection of an interstitial abnormality are recognized: (a) Obvious (detection of abnormality is easy); (b) relatively obvious (detection is relatively easy); (c) subtle (detection is difficult), and (d) very subtle (detection is very difficult, almost impossible). These categories are further used to evaluate the performance of the CAD system and human observers on areas that differ in the visible amount of ILD signs.

II.D. Features

II.D.1. Segmentation

In order to analyze lung fields, they have to be segmented from the rest of the radiograph. For this study the lung fields were delineated manually. In spite of the existence and availability of automatic segmentation methods for PA chest radiographs on the research site (e.g., see Ref. 22), these supervised methods were previously trained with digitized films and appeared to perform imperfectly when applied to digital images.

II.D.2. Feature extraction

Since ILD mostly manifests itself in radiographs through a distortion of the normal appearance of the lung texture, it is important to extract discriminative texture features. A powerful method for local texture analysis is filtering the image with a multiscale filter bank of Gaussian derivatives and calculating the moments of histograms from regions in the derived images. Using multiple scales allows us to characterize texture elements of different sizes, and analysis of local histograms considers the texture primitives regardless of their spacial distribution. This is a general approach to texture characterization.[23,24] The histogram moments were successfully used for automatic detection of interstitial abnormalities in chest radiographs in Refs. 15 and 19, and by Sluimer *et al.*[25] for texture analysis in high resolution thoracic CT.

In our work, prior to filtering the image, pixel values are mirrored with respect to the lung borders. Namely, a pixel value outside the lungs is substituted for its counterpart inside the lungs with the nearest pixel on the lung contour as a center of symmetry. This step is taken to avoid 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. Next, the left lung is flipped to resemble the right lung. Chest radiographs are filtered with Gaussian derivatives of orders 0, 1, and 2 at five scales, $\sigma = 1, 2, 4, 8, 16$:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}\text{-zero-order Gaussian derivative,}$$

$$G_x(x,y), \quad G_y(x,y)\text{-first-order Gaussian derivatives,}$$

$$G_{xx}(x,y), \quad G_{xy}(x,y),$$

$$G_{yy}(x,y)\text{-second-order Gaussian derivatives.}$$

Four central moments of the pixel intensity distribution, namely, the mean, standard deviation, skewness, and kurtosis are calculated from a circular neighborhood of selected pixels (pixels of interest, POIs). These are pixels lying on a 10×10 grid within the lung fields. The features are computed from all the filtered images and from the original image. The radius of the neighborhood is chosen to be 128 pixels for the final system setup. Other radii (64, 96, and 160) appeared slightly less successful in pilot classification experiments. The chosen radius also approximates the sizes of regions in Ref. 19. Unlike Ref. 19, in this work the local analysis is performed on automatically selected regions that covered the lung fields completely. Two position features were added to the feature set, namely, the x and y coordinates of the POI relative to the center of mass of the lung containing it. The position features are scaled to have unit standard deviation per image. In total, 126 features are extracted for each POI.

II.E. Classification

In the next step, a soft classification of POIs is performed in the feature space. Prior to classification, features are normalized. In a training set, each feature is translated and scaled to have zero mean and unit standard deviation. Then the same normalization parameters are applied to feature vectors of test samples.

A supervised classification method is used, which means that a classification function (a classifier) is first trained on labeled samples from both normal and abnormal classes. There is a wide choice of classifiers available in the literature with no superior learning method overall (e.g. see Ref. 26 or 27). The type of problem and prior knowledge determine which classifier provides a better performance. Since no prior knowledge was available, we evaluated and compared several different classifiers. For convenience, we restricted ourselves to four different types of classifiers, namely, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), a k-nearest-neighbors classifier (k-NN), and a support vector machine (SVM).

Both LDA and QDA assume Gaussian distributions for the samples of each class. The LDA additionally assumes equal covariance matrices for each distribution. The k-NN is a nonparametric classifier, with a free parameter k that has to be found empirically. In the k-NN rule, the posterior probability for each of the classes is estimated by the fraction of training samples among the k nearest neighbors of a test sample that belong to that class. In this work, the fast implementation of the k-NN classifier by Arya and Mount[28] was used.

The SVM is a family of classifiers that has gained popularity in recent years. In the case of an ideal linear separability of a training set the SVM finds an optimal discriminative plane by maximizing the margin between the nearest samples, also known as support vectors, of both classes. For linear nonseparable data, the plane found by the SVM is a tradeoff, controlled by a penalty parameter between the classification error on the training set and margin maximization. A useful feature of the SVM is that this method can be kernelized if linear discriminants are not appropriate for a given data set. By mapping original feature vectors into a higher dimensional feature space and solving an SVM optimization problem there, a highly nonlinear classification function can be obtained in the original feature space. The success of the SVM for a particular classification problem depends on a correctly estimated penalty parameter and a suitable kernel function. For the SVM implementation, the LIBSVM library available at http://www.csie.ntu.edu.tw/~cjlin/libsvm was used.

II.F. Post-processing

The classification of an image results in the estimation of posterior probabilities for POIs. To convert this into a probability map we compute new pixel posterior probabilities by averaging posterior probabilities of neighboring POIs. For each pixel i in the lungs, including a POI, its posterior probability p_i is calculated as

$$p_i = \frac{1}{N_{R_i}} \sum_{r \in R_i} p_r^c,$$

where R_i is a neighborhood of i, p_r^c is an estimated by a soft classification posterior probability in a POI r, $r \in R_i$, and N_{R_i} is the number of all POIs lying in the neighborhood R_i. The neighborhood is defined in the same way as the one used to calculate the features.

III. EXPERIMENTS

III.A. Cross validation

All experiments were performed by cross validation. We randomly divided 52 images into four folds, with the condition that normal images and images containing abnormalities were equally spread among the folds (two normal and 11 pathological radiographs in each fold). The CAD system made four iterations. In each iteration a different fold was used as the test set and three other folds together as the training set. This setup guaranteed the optimal use of the available data, as well as an unbiased evaluation, because at no time did training and test sets contain samples originating from the same images.

III.B. Generation of training set

As mentioned in Sec. II C interstitial lesions were divided into four different categories of abnormality subtlety. A straightforward approach would be to use samples from all four categories to train the CAD system. However, pilot experiments showed that this approach would not give the best possible performance. The best performance was achieved when the system was trained with normal samples taken from both normal images and images containing ILD lesions, and abnormal samples from the "obvious" category. Approximately the same performance was reached when the abnormal class was represented by samples from both "obvious" and "relatively obvious" categories. Adding samples with less pronounced abnormalities worsened performance. Therefore, all results in this article were obtained with training sets that contained only normal and obviously abnormal samples. Approximately half of the normal samples in the training set in each fold came from normal training images, and the other half came from normal parts of abnormal training images. Note that for the evaluation of the system no selection of test samples was made. Normal samples from images containing some abnormality were included in the training set only if they were calculated from neighborhoods that did not overlap with any outlined lesions.

III.C. Choice of system parameters

The *k*-NN and SVM classifiers require some parameter tuning. For the *k*-NN, the parameter $k=39$ was chosen experimentally, with negligible differences in the system performance for the whole range of *k* between 25 and 45. For the SVM, we considered the radial basis function (RBF) kernel. The SVM requires a long tuning and training time, therefore the number of samples in the training set was reduced by random subsampling. The penalty and kernel parameters were found using a five-fold cross-validation grid search on the training set.

No feature selection was applied in the final experimental setup. In pilot experiments we found that the system did not gain in performance when run with a subset of features selected by the standard approaches, like sequential forward search and sequential backward search.

III.D. Evaluation

The evaluation of pixel classification was done by means of the ROC analysis.[29] The ROC curve plots the sensitivity against 1-specificity of a system at varying confidence thresholds. A_z, the area under an ROC curve, was used as a classification performance measure. Outcomes of all iterations of the cross-validation were analyzed together yielding a single ROC curve that estimated an overall system performance.

An individual ROC curve was computed at each of four levels of subtlety, with abnormal samples of that subtlety level as positives and all normal samples as negatives. Such analysis may yield better understanding of the relationship between the abnormality subtlety and the detection abilities of the system or humans. Additionally, a generalized ROC curve was calculated that considered all abnormal samples together as positives.

III.E. Observer study

An observer study was performed that closely resembles usual clinical routine. Each lung field was automatically divided into four equal-sized regions (see Fig. 3), altogether eight regions per radiograph. The observers were asked to diagnose each region separately, stating whether it contained any interstitial abnormalities.

The division did not accurately correspond to any anatomical landmarks but was guided by the notion that the top, perihilum, middle periphery, and bottom of a normal lung field exhibit different textural patterns. According to our division algorithm, the region around the hilum (perihilum) included lung pixels overlapping with a circle placed at the lungs' center of mass. The radius of the circle was separately chosen for the left and right lung, such that the overlap covered one quarter of the pixels of that lung. The rest of the lung field was horizontally divided into three equal-sized parts—the top, middle and bottom regions.

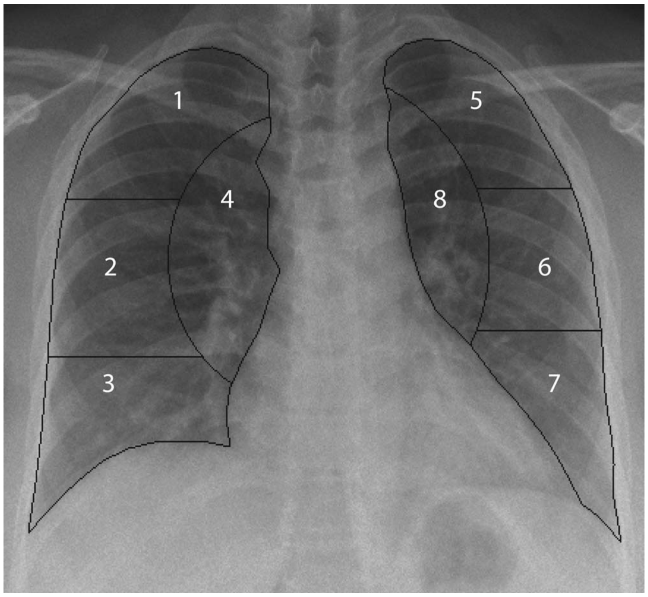

FIG. 3. An example of the lung fields automatically divided into eight regions for the observer study. The regions are (1) and (5)—top, (2) and (6)—middle periphery, (3) and (7)—bottom, and (4) and (8)—perihilum.

The observers assessed regions using the discrete scale of grades from 1 to 5, where 1 corresponded to "normal" (a region looked completely normal, or the amount of ILD was negligible, i.e., less than 10% of the region area) and 5 to "obviously abnormal" (more than 10% of the region area clearly contained interstitial abnormalities). The Grades 2, 3, and 4 corresponded to intermediate levels of observers' certainty whether a region contained interstitial abnormalities. The 10% threshold was a subjective choice of the MP. According to his experience, setting a lower threshold would cause a large interobserver variability.

In order to compare the human performance with the CAD system, region scores for the system were computed from the probability maps by averaging posterior probabilities within each region. The reference standard was also converted into region labels. A label for each region was determined based on the total amount of abnormal tissue present in that region. If less than 10% of the region area fell within any abnormality outline then the region was considered normal. Otherwise, the region was considered abnormal. Similarly to the pixel classification, a subtlety category was assigned to the region in order to evaluate the system performance on regions of different complexity. Such a category was assigned in accordance with the most frequent subtlety type in the region. For example, if the majority of abnormal pixels in the region had been previously ranked as "subtle" than the region received the label subtle. The quantitative distribution of regions among the normal class and different abnormality subtlety categories is shown in Table I.

With the reference standard obtained for each region, ROC analysis becomes possible. It is performed at each of

TABLE I. The numbers of normal regions and regions of different abnormality subtlety are represented in the first column. A distribution limited to peripheral regions only is presented in the second column.

Abnormality subtlety	All regions	Excluding perihilum
Normal	117	91
Obvious	119	74
Relatively Obvious	79	60
Subtle	75	62
Very Subtle	26	25
All categories	416	312

four levels of abnormality subtlety, similarly to the evaluation of pixel classification. Four ROC curves are computed, each considering a subset of abnormal regions that have a certain subtlety level as positives, and all normal regions as negatives. Thus, a number of false positives is the same for each of the curves, while a number of false negatives varies depending on which subset of abnormal images is considered. An overall ROC curve was also calculated considering regions of any abnormality subtlety as positives.

IV. RESULTS

For the pixel classification, the CAD system was trained and tested with each of four classifiers described in Sec. II E. The values of A_z for different classifiers and different levels of abnormality subtlety are listed in Table II. It is shown that the CAD outcomes are comparable for the LDA and SVM. Both classifiers outperform the QDA and k-NN. Later in this article we refer only to the system that uses the LDA, because the LDA is a simpler classifier than SVM. ROC curves measuring the performance of the system with the LDA are plotted in Fig. 4. The curves are clearly distinguished for the different degrees of the abnormality subtlety.

IV.A. Probability maps

The output of the CAD system is a probability map that assigns to each pixel in the lung fields a probability p to being abnormal, $0 \leqslant p \leqslant 1$. For each of the 52 radiographs in the data set such a map has been generated by the system. The map is visualized as a color-coded overlay, where a certain color corresponds to a articular probability range. In Fig. 5 examples of normal and abnormal radiographs and their probability maps are shown. The color-coded overlays in all maps are thresholded, i.e., pixels within the lung fields that have been assigned a posterior probability lower than a threshold value $p=0.15$ stay transparent. After sorting probability maps according to their mean probabilities, separately for normal and abnormal images, three examples from each image class have been randomly chosen from the lower, middle and upper parts of the mean probability range. Figures 5(a), 5(f), and 5(e) are the examples of one normal and two abnormal radiographs for which the system output matches the reference standard well. In Fig. 5(d) the findings of the system are misplaced, and the opacity in the bottom of the left lung is not found. The probability maps in Fig. 5(b) and 5(c) demonstrate the most common mistake that the system makes, namely, misclassification of the perihilar regions.

TABLE II. CAD performances for different classifiers in terms of the area under the ROC curve. The performance is estimated separately for different abnormality subtleties vs normal class, as well as for all abnormality types together vs normal class.

Abnormality subtlety	Classifier			
	LDA	QDA	kNN	SVM
Obvious	0.90	0.84	0.88	0.90
Relatively Obvious	0.77	0.71	0.73	0.77
Subtle	0.65	0.61	0.62	0.66
Very Subtle	0.58	0.52	0.55	0.59
All categories	0.78	0.73	0.76	0.78

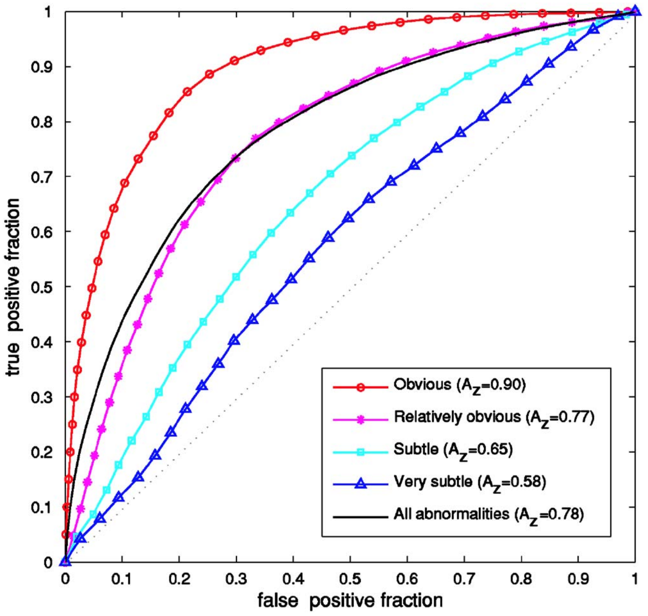

FIG. 4. ROC curves for the evaluation of pixel classification using the linear discriminant analysis. Curves are plotted for different abnormality subtleties vs normal class.

IV.B. CAD performance compared to human observers

We compared the system performance at region level to the performance of two human observers. One observer was an experienced chest radiologist (CSP, more than 15 years of experience) and one was a chest radiologist in training (PdJ). They were not involved in setting the reference standard for the data in this study. The observers were presented the same set of 52 PA chest radiographs. Normal and abnormal images were shuffled and presented in no particular order. The observers reviewed the radiographs using dedicated medical displays (Barco Medical Imaging Systems, Belgium), namely, MFGD 3220D (3MP, 10-bit, 2048 × 1536 native resolution), comparable to displays they would normally use in their clinical practice.

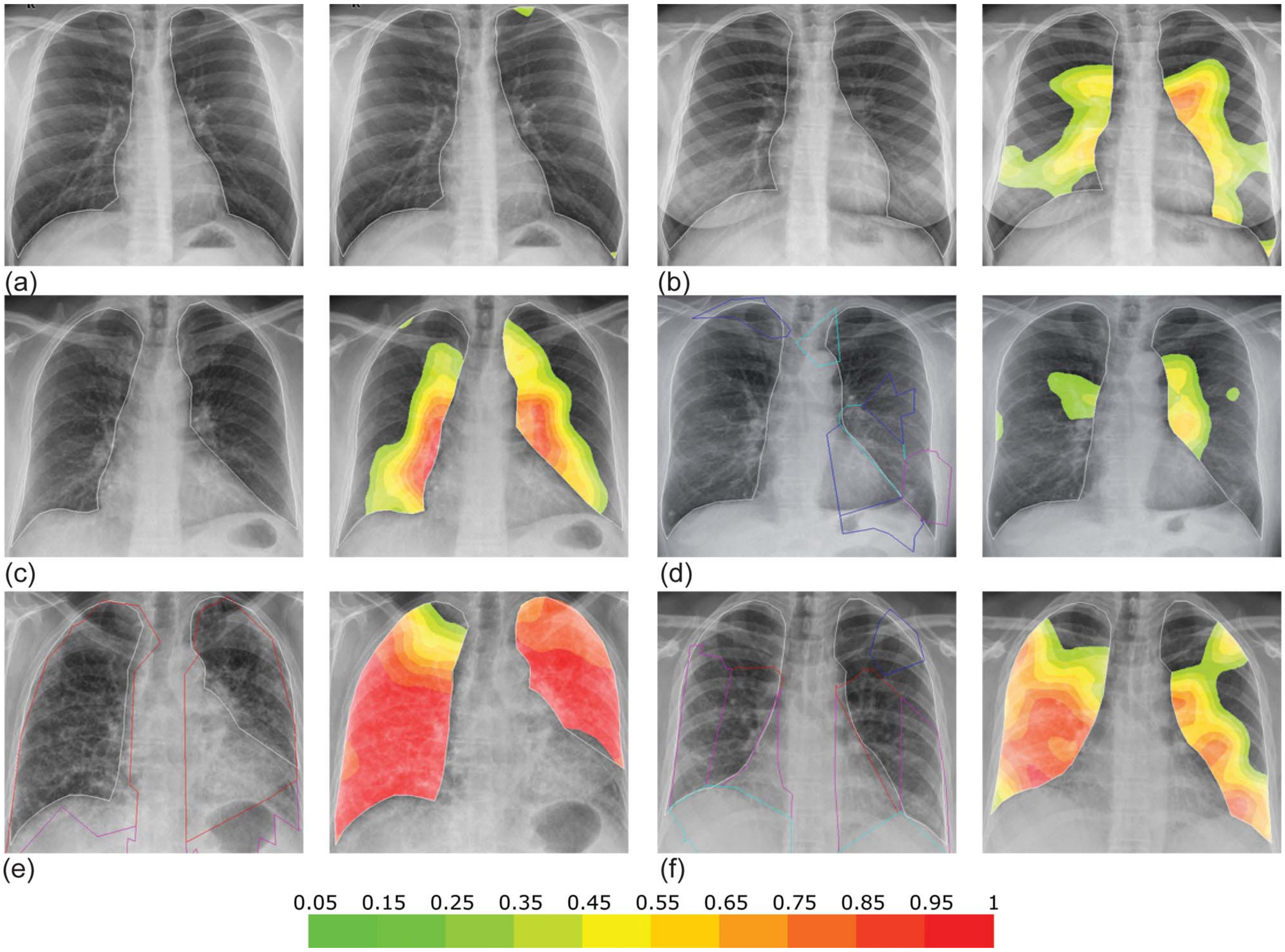

FIG. 5. The examples of chest radiographs and corresponding probability maps produced by the CAD system. The left column depicts an original radiograph with a reference standard for abnormal images. Lung contours are outlined in white. In the probability maps (the right column) only pixels with a posterior probability $p > 0.15$ are shown for convenience. The color bar explains the correspondence between probability ranges and colors.

The values of A_z for both observers and the CAD system are presented in Table III. The results listed in the first three columns were obtained from the evaluation of all eight lung subdivisions. In the last three columns the corresponding performances are shown for the case when the perihilar regions in each radiograph were excluded from evaluation. The statistical analysis described in Ref. 30 was applied to compare A_z of the system with that of each of the observers.

The last row in Table III demonstrates that both observers and the system did not perform significantly different in distinguishing abnormal regions from normal ones when the perihilar regions were excluded from evaluation. It is also

TABLE III. Performance on regions in terms of the area under the ROC curve. The first observer is an expert chest radiologist, the second observer is a chest radiologist in training. The performance is estimated separately for different levels of abnormality subtlety vs normal class, as well as for all abnormality types together vs normal class. Significantly different human performances are marked with an asterisk (a two-tailed test, at significance level of 5%) or a double asterisk (at significance level of 1%).

Abnormality subtlety	All regions			Excluding perihilum		
	CAD	First observer	Second observer	CAD	First observer	Second observer
Obvious	0.92	0.93	0.94	0.96	0.96	0.96
Relatively obvious	0.81	0.85	0.86	0.87	0.87	0.88
Subtle	0.67	0.80^{**}	0.81^{**}	0.73	0.83	0.85^{*}
Very subtle	0.67	0.76	0.71	0.74	0.78	0.69
All categories	0.80	0.86^{*}	0.87^{*}	0.85	0.88	0.88

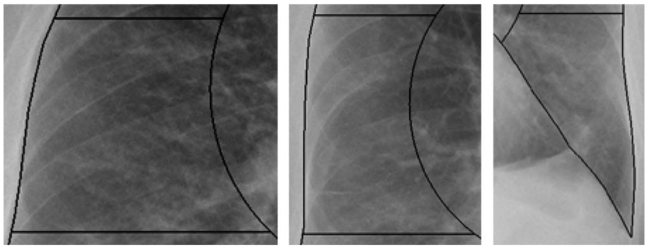

FIG. 6. Examples of regions where the CAD system and one of the observers disagreed, and either the system or an observer misclassified the region. On the left and right, regions with interstitial abnormalities are shown, the region shown in the middle is normal. The CAD system was correct about the regions on the left and in the middle and missed an abnormality in the region on the right.

shown that the poorer performance on the perihilar regions caused the system to be significantly worse ($P<0.05$) than the human observers. The humans were significantly better ($P<0.01$) with slightly abnormal regions throughout lungs, but this difference became smaller when only peripheral regions were considered—the second observer still performed significantly better ($P<0.05$) than the system, while the difference between the system and the first observer became insignificant. Statistical analysis for very slightly abnormal regions did not show any significant differences in performance. The latter may be attributed to a low sample size of very slightly abnormal regions (see Table I). Moreover, the second observer appeared to perform worse than the system when the perihilar regions were excluded. There were no significant differences in AUC values with obviously and relatively obviously abnormal regions.

V. DISCUSSION

We undertook this work in order to provide radiologists with a tool to assist them with the task of finding textural abnormalities in conventional chest radiographs. The complexity of the task of differentiation between normal lung tissue and areas affected by ILD is well illustrated in Figs. 5 and 6.

After training with the annotated data our system assigned a probability to be abnormal to each pixel within the lung fields on the radiograph. The ROC analysis showed that pixel classification results were not reliable for subtle and very subtle areas of abnormality (in Fig. 4). One of possible reasons for that is the informatively superior nature of our reference standard, which was obtained with use of CT known to be a more descriptive modality than conventional radiography. This reason is supported by our pilot experiments mentioned in Sec. II B, when adding subtly and very subtly abnormal pixels to the training set worsened the system performance.

In the future, an additional investigation should be undertaken to identify more powerful features to cope with subtle textural differences. Furthermore, the use of a larger training database might improve the classification of subtle abnormalities.

However, one should not be discouraged by the low system performance on very subtle abnormalities. The overall classification result was not bad ($A_z=0.78$), taken into account the superior reference font standard. And the pixel probability maps are still useful as graphic presentations of the output of the system, even if lesions are only partly found. Such maps give a general idea where, according to the system, abnormalities are situated, and can be conveniently consulted by radiologists as the second opinion.

Averaging posterior probabilities over regions improved the system classification performance (Table III) and made possible a comparison with human observers. Not only the CAD system but also the observers quite often interpreted radiological findings erroneously as it is seen from Table III. Still, both our observers performed significantly better than CAD on the subtle abnormalities. However, CAD was comparable or even better than the human observers on the very subtle lesions (when the perihilum was excluded), and also on the relatively obvious and obvious lesions. Moreover, the system and the observers relatively often made complementary mistakes, which means some regions were correctly classified by the CAD system and misclassified by one or both observers, and vice versa. This point is illustrated in Fig. 6, which shows examples of regions where the CAD output and the human opinion disagreed. It implies that even the output from an imperfect system might be used by a radiologist as an advisory vote, as long as the radiologist understands its limitations.

Evaluation of the utility of our probability maps or regions scores in improving the detection performance of humans is a pertinent continuation of this study.

The performance evaluation for different regions and different degrees of abnormality subtlety would be impossible without the local reference standard. Our system uses a new method to obtain a superior reference standard for the estimation of the position and extension of interstitial lesions in the lung fields. It is a semiautomatic method that involves manual segmentations on CT sections. Although the manual delineation of abnormalities on thin CT sections is more reliable than that on conventional radiographs, it still introduces inherent subjectivity into the pixel-based reference standard, especially near the boundaries of abnormal areas. We hypothesize that differences in segmentation, when performed by different radiologists, might be averaged out to a large extent in the final projection on to the radiograph. However, it is an open question how strong the influence of the manual part of our method could be on the system output. We leave this analysis for future research. One possible extension of our method is an automatic segmentation of lesions on CT sections. The starting point for that could be, for example, a method proposed in Ref. 20 that generates regions containing homogeneous texture.

Another potential source of inaccuracies in the reference standard is the mapping function, which might introduce errors while transferring abnormality outlines from a CT projection to a corresponding radiograph. We ensured the minimization of such errors by providing an auxiliary tool that enabled a radiologist to test the accuracy of matching between two images preliminary to performing segmentations. Based on the visual correspondence of test shapes, control points could be moved or added until a visually satisfactory matching was obtained. The fact that only a limited number of corrections were made to the obtained outlines suggests that errors possibly introduced by the mapping function were limited.

The system demonstrated in this article yields promising results but has considerable room for improvement in the perihilar regions of the lungs. This is the region where the bronchi and blood vessels enter the lung. It is a difficult area for texture analysis not only because of its bright and pattern-rich manifestation but also because its normal appearance has great individual variability. Moreover, our database does not contain many images with healthy perihilum since ILD frequently involves this area. Among 44 abnormal images, only five had a normal perihilum in one or both lungs. Taking into account that only six absolutely normal images were present in each training set we might, conclude that the training set may not be representative of normal varieties of the perihilum. Extension of the database toward inclusion of more normal representatives of the perihilum might improve the system performance even without further modification of the underlying algorithm. In addition, when a radiologist makes a decision as to whether a perihilum looks normal or not, he or she pays attention to its size and shape features along with the textural signs of abnormality. Perhaps, perihilum classification should become a separate component of an automated system and include perihilar shape and size analysis as well.

VI. CONCLUSION

In this article a computer-aided diagnosis system was presented for the task of the detection and localization of interstitial abnormalities in chest radiographs. The system was built using a supervised pattern recognition approach. As an output, the system produced a map of posterior probabilities, where each pixel inside the lung fields received a probability of being abnormal.

We collected and annotated a unique database of digital chest radiographs containing ILD abnormalities. A novel method was developed to define a reference standard on the abnormal radiographs. This method utilized a CT scan of the same patient and automatically translated manual delineations of abnormalities made on a subset of thin coronal sections to the corresponding radiograph.

Our CAD system employed local statistical features calculated from filtered images. The filters were the first and second order Gaussian derivatives at multiple scales. A linear discriminant classifier and support vector machine yielded the best classification performances. The evaluation was done by means of ROC analysis for different levels of abnormality subtlety. It was shown that the system was considerably better in distinguishing obviously abnormal pixels from normal ones than in distinguishing between very slightly abnormal and normal pixels. This is likely due to an overinformative nature of the reference standard that we compared our findings with.

The system performance was also compared with that of an expert radiologist and a radiologist in training. The system was shown to perform significantly worse than both observers on slightly abnormal regions and all abnormalities together, with no significant differences in the detection of obviously, relatively obviously, and very slightly abnormal regions. Moreover, the system was shown to approach the human performance in the detection of abnormalities when the perihilar regions were excluded from evaluation.

ACKNOWLEDGMENTS

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

[a] Author to whom all correspondence should be addressed. Electronic mail: yulia@isi.uu.nl

[1] E. Kazerooni, "High-resolution CT of the lungs," Am. J. Roentgenol. **177**, 501–519 (2001).

[2] British Thoracic Society, BTS guidelines on the diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults, Thorax **54**(Supplement 1), S24–S30 (1999).

[3] H. P. McAdams, E. Samei, J. Dobbins, III, G. D. Tourassi, and C. Ravin, "Recent advances in chest radiography," Radiology **241**, 663–683 (2006).

[4] B. van Ginneken, B. M. ter Haar Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: A survey," IEEE Trans. Med. Imaging **20**, 1228–1241 (2001).

[5] H. Abe, H. MacMahon, R. Engelmann, Q. Li, J. Shiraishi, S. Katsuragawa, A. Aoyama, T. Ishida, K. Ashizawa, C. E. Metz, and K. Doi, "Computer-aided diagnosis in chest radiography: Results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies," Radiographics **23**, 255–265 (2003).

[6] H. Abe, K. Ashizawa, F. Li, N. Matsuyama, A. Fukushima, J. Shiraishi, H. MacMahon, and K. Doi, "Artifical neural networks (ANNs) for differential diagnosis of interstitial lung disease: Results of a simulation test with actual clinical cases," Acad. Radiol. **11**, 29–37 (2004).

[7] M. B. Gotway, "Interstitial lung diseases: Imaging evaluation," Appl. Radiol. **29**, 31–46 (2000).

[8] W. T. Miller, Jr., "Chest radiographic evaluation of diffuse infltrative lung disease: Review of a dying art," Eur. J. Radiol. **44**, 182–197 (2002).

[9] S. P. Padley, D. M. Hansell, C. D. Flower, and P. Jennings, "Comparative accuracy of high resolution computed tomography and chest radiography in the diagnosis of chronic diffuse infiltrative lung disease," Clin. Radiol. **44**, 222–226 (1991).

[10] S. Katsuragawa, K. Doi, and 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," Med. Phys. **16**, 38–44 (1989).

[11] T. Ishida, S. Katsuragawa, T. Kobeyashi, H. MacMahon, and K. Doi, "Computerized analysis of interstitial disease in chest radiographs: Improvement of geometric-pattern feature analysis," Med. Phys. **24**, 915–924 (1997).

[12] S. Katsuragawa, K. Doi, H. MacMahon, L. Monnier-Cholley, T. Ishida, and T. Kobayashi, "Classification of normal and abnormal lungs with interstitial diseases by rule-based method and artificial neural networks," J. Digit Imaging **10**, 108–114 (1997).

[13] T. Ishida, S. Katsuragawa, K. Ashizawa, H. MacMahon, and K. Doi, "Application of artificial neural networks for quantitative analysis of image data in chest radiographs for detection of interstitial lung disease," J. Digit Imaging **11**, 182–192 (1998).

[14] S. Kido, S. Tamura, N. Nakamura, and C. Kuroda, "Interstitial lung disease: Evaluation of the performance of a computerized analysis systems versus observers," Comput. Med. Imaging Graph. **23**, 103–110 (1999).

[15] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever, "Automatic detection of abnormalities in chest radio-

graphs using local texture analysis," IEEE Trans. Med. Imaging **21**, 139–149 (2002).

[16] M. Long, B. van Ginneken, and M. Neilsen, "Detection of interstitial lung disease in PA chest radiographs," edited by M. Jaffe, and M. Flynn, SPIE Medical Imaging: Physics of Medical Imaging **5368**, 848–855 (2004).

[17] T. Ishida, S. Katsuragawa, K. Nakamura, K. Ashizawa, H. MacMahon, and K. Doi, "Computerized analysis of interstital lung disease on chest radiographs based on lung texture, geometric pattern features and artificial neural networks," Proc. SPIE **4684**, 1331–1338 (2002).

[18] W. Webb, N. Müller, and D. Naidich, *High resolution CT of the lung*, 3rd ed. (Lippincott, Williams, & Wilkins, Philidelphia, PA, 2001).

[19] Y. Arzhaeva, D. M. J. Tax, and B. van Ginneken, "Improving computer-aided diagnosis of interstitial disease in chest radio-graphs by combining one-class and two-class classifiers," edited by J. M. Reinhardt, and J. P. W. Pluim, Proc. SPIE **6144**, 614458 (2006).

[20] I. C. Sluimer, M. Prokop, I. Hartmann, and B. van Ginneken, "Automated classification of hyperlucency, fibrosis, ground glass, solid and focal lesions in high resolution ct of the lung," Med. Phys. **33**, 2160–2620 (2006).

[21] D. Ruprecht and H. Müller, "Image warping with scatterred data interpolation," IEEE Comput. Graphics Appl. **15**(2), 37–43 (1995).

[22] B. van Ginneken, M. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database," Med. Image Anal **10**(1), 19–40 (2006).

[23] M. Unser and M. Eden, "Multi-resolution feature extraction and selection for texture segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 717–728 (1989).

[24] B. van Ginneken and B. ter Haar Romeny, "Multi-scale texture classification from generalised locally orderless images," Pattern Recogn. **36**, 899–911 (2002).

[25] I. C. Sluimer, P. F. van Waes, M. A. Viergever, and B. van Ginneken, "Computer-aided diagnosis in high-resolution ct of the lungs," Med. Phys. **30**(12), 3081–3090 (2003).

[26] R. O. Duda, P. E. H. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (John Wiley and Sons, New York, 2001).

[27] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 4–37 (2000).

[28] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbour searching in fixed dimensions," J. ACM **45**(6), 891–923 (1998).

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

[30] J. A. Hanley and B. J. McNeil, "A methodology of comparing the areas under receiver operating characteristic curves derived from the same cases," Radiology **148**(3), 839–843 (1983).