21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Interactive classification of lung tissue in CT scans by combining prior and interactively obtained training data: a simulation study

Thessa T.J.P. Kockelkorn
Image Sciences Institute, UMC Utrecht, Utrecht, The Netherlands

Rui Ramos
Instituto de Engenharia Biomédica, Faculdade de Engenharia da Universidade do Porto, Portugal

José Ramos
Instituto de Engenharia Biomédica, Faculdade de Engenharia da Universidade do Porto, Portugal

Clara I. Sánchez
Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Pim A. de Jong
Department of Radiology, UMC Utrecht, Utrecht, The Netherlands

Cornelia Schaefer-Prokop
Department of Radiology, Meander Medical Centre, Amersfoort, The Netherlands

Jan C. Grutters
Department of Pulmonology, St Antonius Ziekenhuis Nieuwegein, The Netherlands

Max A. Viergever
Image Sciences Institute, UMC Utrecht, Utrecht, The Netherlands

Bram van Ginneken
Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Abstract

We describe an interactive system for classification of normal and seven types of abnormal lung tissue in CT scans from interstitial lung disease patients, using training data from previously annotated scans and annotations by the observer in the scan under investigation. We compared seven different interactive annotation strategies using different combinations of both types of training data, in order to minimize user effort in the interactive annotation process. The lungs in all scans were divided into roughly spherical volumes of interest (VOIs). An observer labeled all VOIs in 21 thoracic CT scans. Leave-one-scan-out experiments that simulated slice-by-slice interactive annotation sessions were performed. The best results were obtained with a strategy in which the simulated user decides for each slice whether to use a classifier trained on pooled data from prior scans or a classifier trained on data from the current scan. In this approach, the labels of 88% of all VOIs were predicted correctly, meaning that only 12% of all labels needed to be changed by the simulated user.

1. Introduction

The term interstitial lung disease (ILD) comprises a group of inflammatory and fibrotic lung diseases that mainly affect the tissue and space around the air sacs of the lungs. They have distinct, but nevertheless partially overlapping imaging features. Diagnosis is made using an interdisciplinary approach, in which CT scans are of pivotal importance. Since the group of diseases varies substantially in terms of treatability and prognosis, it is important to make the correct diagnosis. Analysis of imaging features substantially varies even among experienced radiologists and computerized diagnosis of interstitial lung disease is therefore of great interest.

978-4-9906441-1-6

Analysis of both the types and spatial distribution of the pathological textures in the CT scan is used for diagnosing an interstitial lung disease, as well as for assessment of treatment response and disease progression.[1] Annotation of all present textures is a prerequisite for an accurate quantitative analysis; however, this is a non-trivial task in 3D for both human experts and computer systems. On the one hand, manual delineation of textures in complete scans is too laborious to use in clinical practice or in research. On the other hand, a proven fully automatic solution does not exist. We therefore developed a system for interactive annotation of 3D volumes of interest (VOIs), applied to scans of ILD patients.[2] This system uses a classifier that is trained continuously on data entered by the observer in the scan under consideration. In this work we investigate how annotations from previously seen scans can be employed to decrease user effort in the interactive annotation process.

2. Materials

For this project, 21 clinical dose thoracic CT scans with submillimeter resolution were used. Scans were acquired between April 2004 and March 2010 at the St Antonius Ziekenhuis Nieuwegein, the Netherlands, on a Philips Mx8000 IDT or a Philips Brilliance iCT scanner (Philips Medical Systems, Best, The Netherlands). Scans were taken at full inspiration with patients in supine position. Data were acquired in spiral mode and reconstructed to 512 × 512 or 768 × 768 matrices. No contrast material was used.

3. Methods

Figure 1 gives a schematic overview of the manual and the simulated interactive and annotation process. First, lungs were segmented using a hybrid segmentation algorithm.[3] Lung volumes were divided into roughly spherical volumes of interest containing only one type of texture, using the algorithm described in Ref. 2. All scans were annotated by a radiologist, who indicated all VOIs containing the following abnormalities: decreased density, consolidation, honeycombing, ground glass, crazy paving, non-specific interstitial pneumonia (NSIP) pattern and nodular pattern. In addition, all VOIs which contained more than one type of pattern were indicated as inhomogeneous. These were not used in further analysis. All remaining VOIs were labeled as normal lung tissue. These annotations were used as ground truth in all experiments. In total, 42,185 VOIs were included, on average 2,009 per scan.

For each VOI, 36 features were calculated. Scans were filtered using Gaussian, Laplacian and gradient

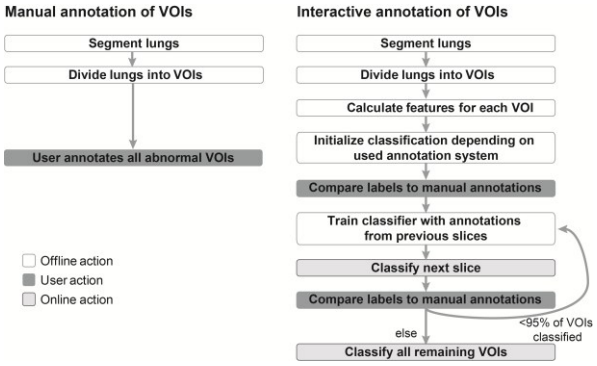

Figure 1. Schematic overview of steps in the manual and interactive annotation processes

magnitude filters. Each filter was applied at three scales (σ = 1, 2, and 4 voxels). In each filtered image, the mean, standard deviation, kurtosis and skew of the CT densities per VOI were used as features. These rotationally invariant features were chosen since the textures that are to be detected do not have a specific orientation.

3.1 Experiments

Simulation experiments were conducted to compare the different annotation systems, depicted in Figure 2. Experiments were done using a leave-one-scan-out approach: the VOIs of each scan were once used as test data, while the annotations of the other 20 scans were used as training data.

In all systems, annotation was done per slice, yielding several annotation rounds per CT scan, until 95% of all VOIs had been labeled. All remaining VOIs were classified in the final round. After each round, the classifier was retrained, using the results of all previous classification rounds. In each round, the outcome of the classification process was compared to the manual annotations of the radiologist. The number of VOIs that were classified incorrectly, and therefore needed relabeling, was taken as the performance measure.

In the first system, (Figure 2a), all VOIs were classified automatically using only training data from previous scans ($c_{ensemble}$). To this end, data from each scan was used to train a separate k-nearest neighbor (kNN) classifier with k = 101, yielding 20 classifiers. Since not all scans contained all abnormal tissue types, 200 random samples from other scans were added to the training data for each absent category. All VOIs were classified using voting: each VOI received the label chosen by the majority of classifiers. In the first round, all classifiers received the same weight. In following rounds, classifiers were weighted using their accuracy on the VOIs classified previously. In this way, better classifiers were favored.

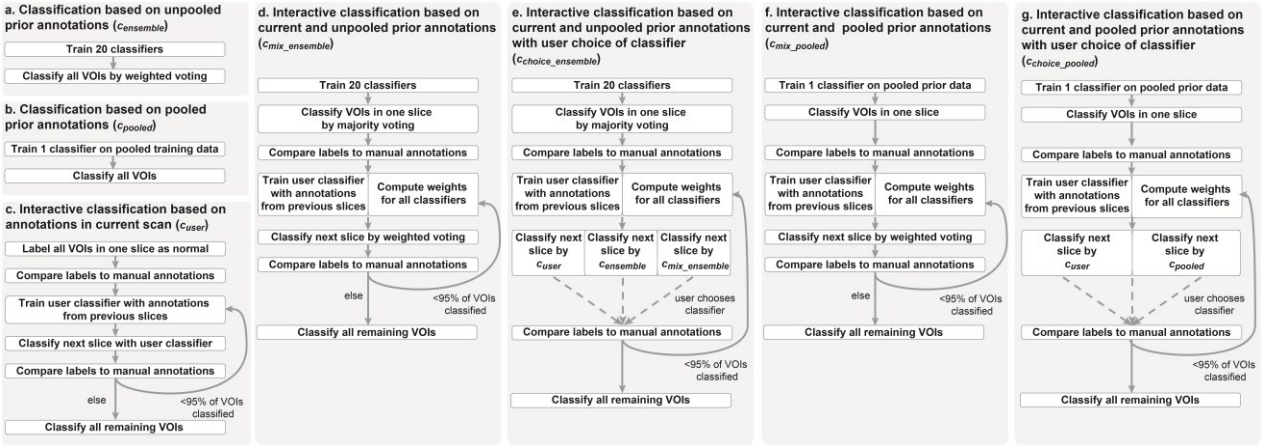

Figure 2. Schematic overview of the different simulation systems. a. $c_{ensemble}$: ensemble of kNN classifiers with k = 101, each trained with data from one previously annotated scan. b. c_{pooled}: kNN classifier with k = 101, trained on pooled data from all previously annotated scans. c. c_{user}: kNN classifier with k = 7, trained on annotations in the current scan. d. $c_{mix\_ensemble}$: *weighted combination of c_{user} and $c_{ensemble}$*. e. $c_{choice\_ensemble}$: *system in which the user can choose between the classification results of c_{user}, $c_{ensemble}$* and $c_{mix\_ensemble}$ in each slice. f. c_{mix_pooled}: *weighted combination of c_{user} and c_{pooled}*. g. c_{choice_pooled}: *system in which the user can choose between the classification results of c_{user} and c_{pooled}* in each slice.

In the second system (Figure 2b), all VOIs were classified automatically using one kNN classifier (k = 101) trained on the pooled training data of all prior scans (c_{pooled}).

In the third protocol (Figure 2c), all VOIs were annotated using only training data from the current scan. In the first slice, no training data was available. Therefore, the best guess is to label all VOIs as normal tissue, since this is the category with the largest prior probability. In the second round, the annotations of the first round were used to train a kNN classifier with k = 7 (c_{user}). In each following round, annotations from the previous round were added to the training data, which was expected to increase the accuracy of c_{user}.

In the fourth protocol (Figure 2d), c_{user} was added to the ensemble of previous classifiers $c_{ensemble}$, yielding $c_{mix\_ensemble}$. Classification was done by weighted voting based on the following formulas:

$$w_{cx} = \mathrm{acc}_{cx} \times \mathrm{percentage}_{unclassified}$$
$$w_{user} = \mathrm{acc}_{user} \times \mathrm{percentage}_{classified} \times 10$$

Each individual classifier c_x trained on a previously annotated scan was weighted according to its accuracy on VOIs in the current scan that had been classified in previous rounds (acc_{cx}). In addition, it was weighted according to the percentage of the scan that had not been classified so far ($\mathrm{percentage}_{unclassified}$), thus decreasing its influence during the interactive annotation process. The weight of c_{user} consisted of its accuracy on VOIs classified so far (acc_{user}), the percentage of VOIs that had been classified so far ($\mathrm{percentage}_{classified}$) and a factor 10. This factor was determined in pilot experiments. In this way, c_{user} gained influence during the annotation of the scan and was favored with respect to the other classifiers, since it was supposed to yield better classification results than each individual classifier trained with data from one other scan.

In the fifth protocol (Figure 2e), the simulated user could choose between the classification results of $c_{ensemble}$, c_{user} and $c_{mix\_ensemble}$ in each annotation round ($c_{choice\_ensemble}$). In this way, he could choose the results that best match his annotations, thereby decreasing the number of manual corrections.

In the sixth protocol (Figure 2f), c_{user} was combined with c_{pooled}, which resulted in c_{mix_pooled}. Classification was done by the classifier with the largest weight as calculated by the following formulas:

$$w_{pooled} = \mathrm{acc}_{pooled} \times \mathrm{percentage}_{unclassified}$$
$$w_{user} = \mathrm{acc}_{user} \times \mathrm{percentage}_{classified} \times 2$$

The classifier trained on pooled prior data was weighted according to its accuracy on VOIs encountered in the current scan so far (acc_{pooled}) and the percentage of unannotated VOIs. The weight of the interactive classifier was determined by its accuracy on all annotated VOIs in the current scan and the percentage of annotated VOIs. The factor 2 was added to increase the influence of c_{user} and was determined in pilot experiments.

In the last protocol (Figure 2g), the simulated user could choose between the classification results of c_{pooled} and c_{user} in each annotation round (c_{choice_pooled}). Please note that the classification results of c_{mix_pooled} for one slice were exactly the same as the results of either c_{pooled} or c_{user}.

5. Results

0.64% of all VOIs were annotated as inhomogeneous and hence excluded from further analysis.

Only using $c_{ensemble}$ gives an overall accuracy of 74%, meaning that a user would have to change the label of on average 26% of all VOIs in a scan to obtain a complete annotation of all lung tissue. Accuracies for individual scans ranged from 46% to 100%. The results of c_{pooled} were better, with an overall accuracy of 82% (range: 62-98%). Using the interactive approach c_{user} with data from the current scan only gives an accuracy of 85% (range: 63-100%). $c_{mix\_ensemble}$ yields an accuracy of 83% (range: 58-100%). The system based on $c_{choice\_ensemble}$ resulted in 87% (range: 66-100%) of the labels receiving the correct label. Combining c_{user} and c_{pooled} into c_{mix_pooled} gave an accuracy of 85% (range: 64-99%). The best results were obtained by c_{choice_pooled}, which lets the simulated user choose between c_{user} and c_{pooled} in each round; this resulted in an accuracy of 88% (range: 72-100%).

6. Discussion

We presented several ways for interactive annotation of normal and abnormal textures in the lungs of ILD patients, using prior and interactively obtained training data. All strategies were tested using simulation software. The approach in which the user could choose between the results of a classifier trained on data from the current scan only and results from a classifier trained on pooled training data from previous scans resulted in the smallest number of classification errors. In this case, only 12% of all VOIs were misclassified and needed relabeling.

Fully automatic classification of lung tissue is a non-trivial task. Differences between the training scans and the scan under consideration, in for example scanning protocol or degree of inspiration by the patient, decrease the accuracy of such a system. In addition, interobserver and intraobserver variability are well-known issues in lung texture annotation[4,5]. Individual observers may label textures in different ways and even the same observer may interpret the pattern in a given VOI differently at different time points. Therefore, user interaction is needed to obtain a final classification result that is completely satisfactory.

We have developed such an interactive classification system, based only on data from the current scan, which is tailored to the specifics of the scan under investigation and the radiologist performing the annotations. This interactive system starts untrained, but it learns quickly and soon outperforms the classification based on previous data only. By combining the both systems, the strengths of the system trained on pooled prior data, namely the large size of its training set, can be combined with the strengths of the system using only data from the current scan, namely its adaptability and specificity. By offering the user a choice between the different classification strategies at slice level, the number of individual VOIs requiring relabeling is decreased. In this way, user effort and time investment can be reduced.

A limitation of the present work is that it is a simulation study. We plan to repeat the experiments with human observers in order to test if the predicted reduction in user effort will indeed occur. In addition, we aim to let the annotation software decide which classifier to use for each slice, which would further reduce user interaction.

The major advantage of the current system is the reduction of the percentage of VOIs requiring relabeling, which enables annotation of large datasets. These datasets can then be used as training data for classification of textures in new scans. At patient level, complete annotation of scans may be used for quantitative monitoring of lung parenchyma pathologies in ILD patients, making it possible to track disease progression and to assess the effectiveness of medication. Finally, detailed analysis of abnormal textures may be beneficial in making a specific diagnosis for patients with ILD.

7. References

[1] Z.A. Aziz, A.U. Wells, E.D. Bateman, S.J. Copley, S.R. Desai, J.C. Grutters, D.G. Milne, G.D. Phillips, D. Smallwood, J. Wiggins, M.L. Wilsher and D.M. Hansell. Interstitial lung disease: effects of thin-section CT on clinical decision making. *Radiology*, 238(2), 725–733, February 2006.

[2] T.T.J.P. Kockelkorn, P.A. de Jong, H A. Gietema, J.C. Grutters, M. Prokop and B. van Ginneken. Interactive annotation of textures in thoracic CT scans, in: *SPIE Medical Imaging,* vol. 7624, p. 76240X, February 2010.

[3] E.M. van Rikxoort, B. de Hoop,M.A. Viergever, M. Prokop and B. van Ginneken. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical Physics* 36(7), 2934–2947, July 2009.

[4] 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. *Medical Physics,* 33(7), 2610–2620, July 2006.

[5] R. Uppaluri, E.A. Hoffman, M. Sonka, P.G. Hartley, G.W. Hunninghake and G. McLennan. Computer recognition of regional lung disease patterns. *American Journal of Respiratory and Critical Care Medicine*, 160(2), 648–654, August 1999.