Learning Interstitial Lung Diseases CT Patterns from Reports Keywords

José Ramos[1,2], Thessa Kockelkorn[2], Bram van Ginneken[3], Max A. Viergever[2], Jan Grutters[4], Rui Ramos[1], and Aurélio Campilho[1]

[1] Instituto de Engenharia Biomédica, Faculdade de Engenharia da Universidade do Porto, Portugal
[2] Image Sciences Institute, UMC Utrecht, The Netherlands
[3] Diagnostic Image Analysis Group, Department of Radiology, Radboud University Nijmegen Medical Centre, The Netherlands
[4] Department of Pulmonology, St Antonius Ziekenhuis Nieuwegein, The Netherlands

Abstract. The interpretation of CT exams from patients with interstitial lung diseases depends on the correct assessment of associated CT patterns. Computer aided diagnosis systems often study the automatic identification of CT patterns, using the division of the lung in volumes of interest and the use of supervised classification. Despite moderate success, this approach has been hampered by the shortage of medical annotations available to research groups. We propose a new method that collects exams that contain CT patterns through the presence of keywords in radiology reports, to learn pattern models using a multiple instance learning algorithm. We compared our approach to the traditional use of volumes of interest annotations for six interstitial lung diseases patterns. The results show our approach performed comparatively in four of the studied patterns, and poorly for the other two. The results suggest that under certain conditions learning CT patterns from radiology reports is possible, which could foster developments in computer aided diagnosis systems.

1 Introduction

Interstitial Lung Diseases (ILD) are a set of more than 100 lung disorders that affect the lung interstitium. Although jointly grouped, the several ILD sub-types have different treatments and prognoses, and hence an accurate sub-type identification is important for the disease management. ILD diagnosis normally requires the analysis of the patient thorax CT, in which radiologists detect, locate and characterize visual patterns (also denominated abnormalities or textures). These patterns constitute the basis for radiologists' conclusions, and are important elements in the diagnosis process.

ILD Computer Aided Diagnosis (CAD) systems have studied the automatic detection of CT patterns [6]. A popular approach is the division of the previously segmented lung in Volumes of Interest (VOI) and the use of supervised classifiers to automatically label VOI into patterns types. Uchiyama et al. [8]

used six specifically designed measures and artificial neural networks to classify each region of interest (ROI) into one of 7 categories. Sluimer et al. [7] used a set of Gaussian and Laplacian filters and a nearest neighbor classifier for 6 pattern classes. Similar approaches were used in other pattern classification systems [2] [1] [10].

The major limitation of supervised classification for CT pattern detection is that it requires a set of radiologists to annotate a sufficient number of representative VOIs for each possible pattern. Because of the large number of patterns and the variability in their presentation, and since most research groups have limited medical support, the construction of a representative dataset is difficult to accomplish. Consequently, most of the previous research on VOI classification is either focused on a limited number of patterns [1] [8] [10] or pattern superclasses which encompass different visually-like patterns, as hyperlucency or high attenuation pattern [2] [7]. These simplifications reduce the ability of CAD systems to significantly represent an ILD CT.

This paper presents a new method for building VOI classifiers that does not depend on radiologists' annotations. Our approach first automatically detects exams containing the target pattern through the presence of keywords in the radiology report associated with the exam. It subsequently employs a Multiple Instance Learning (MIL) algorithm to locate the pattern VOI in the respective exam. Since radiology reports are normally present in most hospital PACS, our approach could ease the construction of VOI classifiers, and broaden its application to a large number of patterns.

In this paper we will present and analyze our approach for VOI classification of 6 classes of ILD patterns, representing each VOI by its mean Hounsfield Unit (HU). Although the primitive model for VOI representation undermines the performance, its simplicity will be useful for analyzing the potential and limitations of the method.

MIL focuses on the problem of inferring concepts (which materialize as regions in feature space) from sets of positive and negative bags. A bag is positive if it contains at least one example of the concept, and negative if no concept examples are present. MIL is based on the premise that concept regions have a high concentration of examples from positive bags, and a small concentration of negative bags examples. This is the principle behind the Diverse Density measure introduced by Maron et al. [3], and the subsequent evolution expectation maximization diverse density [11] on which the method used in this paper is based. MIL has already been applied to region classification from labels at an image level (normally denominated weak labels) by considering each image as a bag and each region an example [3]. The use of labels automatically collected from associated text has also been described before for photo/video stocks [9]. This paper studies the same approach for lung CT analysis and radiology reports.

We showed in a previous paper [4] that radiology reports can be used as image labels to guide a manifold learning algorithm for content based image retrieval. This paper applies the same idea to a different problem, the classification of CT patterns.

2 Materials and Methods

2.1 Dataset

All exams used in our approach were retrieved from a collection of 1110 scans from 253 patients with one type of ILD. From these CT scans all radiology reports were collected. Scans with no reports or on which the segmentation process failed were discarded. Scans were acquired on a Siemens SOMATOM Sensation Cardiac 64 or a Siemens EMOTION DUO (Siemens Healthcare, Erlangen, Germany) with varied acquisition parameters. Reports are in Portuguese.

2.2 System Description

Overview This section describes our learning approach to design a CT pattern classifier from exams selected by the presence of keywords in radiology reports. Each pattern classifier is built independently of all others in an one-against-all configuration, and consequently the method description refers to examples containing/not containing the target pattern as positive/negative.

Our method consists of 4 consecutive stages: 1) Selection of exams from a dataset using a list of keywords; 2) Transformation of exams into bags of VOI; 3) Removal of noisy bags; 4) Learning the pattern model using MIL. Each stage is described in detail in the next sections.

Exam Collection The positive/negative exams were automatically extracted from the dataset described in section 2.1, by detecting in the radiology reports the presence of terms related to the patterns category. The terms lists were previously assembled by a radiologist and include multi-words expressions (eg. ground glass), inflections (eg. nodule, nodules, nodular), variations in hyphenation (eg. groundglass, ground-glass) and typographical errors. An English translation of the main terms is in Annex A. The list of table 1 presents the number of scans that contain the pattern terms, and the number of scans that only contain terms from that pattern.

A positive exam set is collected from the exams set that *only* contain the target term, while a negative exam set is collected from the exams set that *do not* contain the target term. The system selects 50 positive and 50 negative exams, randomly without replacement. If the number of positive exams is smaller than 50, all available exams are considered. Since only the nodular pattern has more than 50 positive exams, only for this pattern does the positive exams set vary. The negative exams set varies in all patterns.

Exams into Bags of VOI The second stage transforms all CT scans previously selected into bags of VOI. Our system was based in the infrastructure described in [7], and required three consecutive steps: Resizing, segmentation and division in VOI. This architecture is common in ILD CAD analysis systems [6].

Table 1. Number of exams present in the dataset per lung pathological tissue for hyperlucency (HL), ground glass (GG), honeycombing (HC), crazy paving (CP), consolidation (Cons.) and nodular (Nod.) patterns. The total number of exams in the dataset is 1110.

Tissue	HL	GG	HC	CP	Cons.	Nod.
# exams containing pattern	263	281	427	89	332	669
# exams only containing pattern	37	25	77	8	23	166

In the first step all scans are resized to $128 \times 128 \times 128$ voxels using a nearest neighbor interpolation. Segmentation uses the algorithm described in [5]. It is composed of two region growing processes, one to segment the main airways from a seed point, and a second starting from the lowest HU values in the bronchi that segments the parenchyma. The final phase divides the segmented lungs into VOI using the algorithm detailed in [2]. It uses seed points based on local maxima or minima with a minimum distance of 5 voxels. A volume growing algorithm is then applied until volumes collide based in an acceptance rule that equally takes into account the distance to the seed point and the difference in HU values to the VOI mean. Division based in homogeneous regions was previously shown to improve over a square grid [6]. In this preliminary work we will only use the mean of the HU value to describe each VOI.

Figure 1 a) contains the HU distributions for all points from hyperlucency positive bags (Positive) and hyperlucency negative bags (Negative). It can be observed a skewness of the positive set distribution to lower HU values.

Removing Noisy Exams Selecting exams based on the presence of terms in radiology reports is inherently noisy. The presence of the term might be to indicate its negation ("absence of nodules") or the image abnormality might be too small to be significant ("insignificant fibrosis in upper lobe"). The non presence of terms also does not imply its absence. In fact it is frequent that follow-up exams do not describe the findings present, but only pinpoint evolutions. The noisy nature of this process is visible in fig.1 b) where the distribution of VOI HU values on each bag containing hyperlucency terms is represented.

The third stage is responsible for removing bags which have a VOI distribution incoherent with the global VOI distribution, that is with the distribution of all VOI of all bags. The detection of incoherent bags is based on the Kolmogorov-Smirnov test between each bag points, with the distribution of all positive points (D_{pos}) and with the distribution of all negative points (D_{neg}). The positive bags for which the probability of belonging to D_{pos} is smaller than to belonging to D_{neg} are removed. Analogously negative bags for which the probability to belong to D_{pos} is larger than belonging to D_{neg} are also removed.

This process is exemplified in fig. 1 for hyperlucency terms. Fig. 1 a) and fig. 1 b) presents, respectively, the global and positive bags distributions prior to this stage, and fig. 1 c) and d) the global and positive bag distributions after this

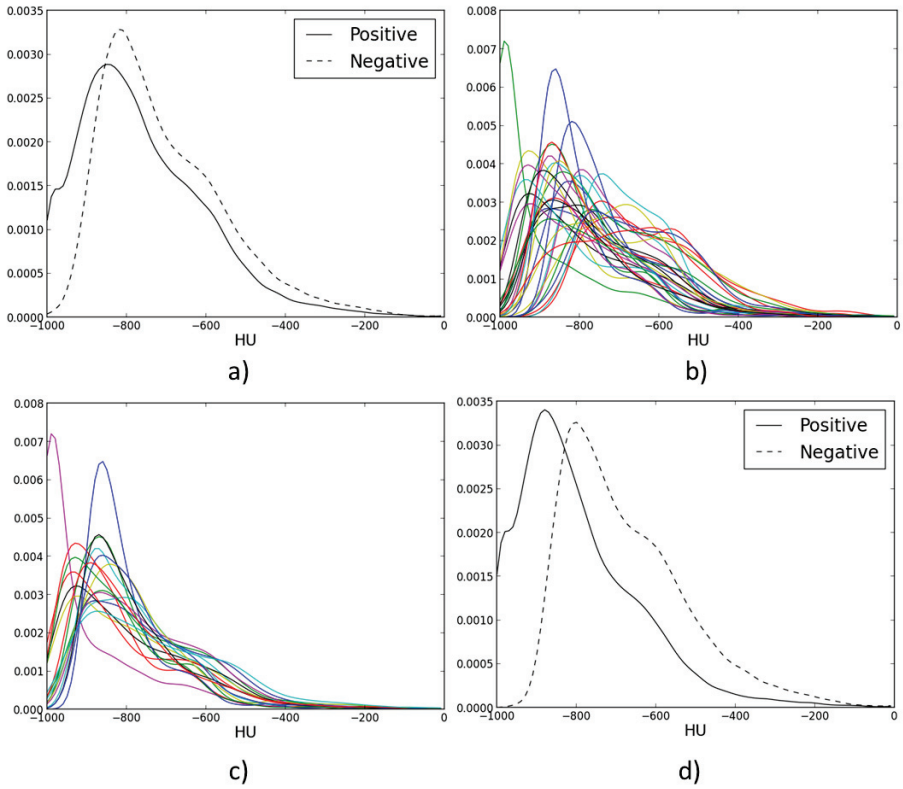

Fig. 1. Kernel density estimation for the probability densities functions (PDF) of the HU mean values of VOI before and after the removal of noisy exams for hyperlucency bags. a) PDF density from all VOI from positive and negative exams; b) PDF from each positive bag before the removal of noisy bags; c) PDF from each positive bag after the removal of noisy bags; d) PDF density from all VOI from positive and negative exams after removing noisy bags.

stage. It can be observed that the positive bags with higher HU are removed, and that the final positive distribution has been skewed to lower HU values.

Learning Patterns Representations through Multiple Instance Learning The final step uses a MIL algorithm to, from the set of positive and negative bags, infer a model of the radiological pattern. In this preliminary work the outcome of this stage will be the probability of the VOI containing the pattern. The MIL algorithm is based on the work described in [9], which is in itself based on the expectation maximization diverse density method described in [11].

The MIL method requires a VOI description x (which contains the HU mean of the VOI as described in section 2.2) and a binary label t which indicates the

presence/absence of the respective pattern in the VOI. Each sample has an associated binary label w which indicates if the bag it belongs to is positive/negative. The goal of the method is to determine the posterior probability $p(t|x)$. Since variable t is not observed, it has to be inferred from $p(t|x,w)$ which requires the knowledge of $p(x|t)$ which requires the knowledge of t. The EM algorithm is applied to this problem by alternating between:

1) Calculate the likelihood of each pathology $p^n(x|t)$ weighted by the probability that the VOI contains the positive pattern $p^{n-1}(t=1|x,w)$. We used a weighted Gaussian kernel density estimation:

$$p^n(x|t=1) = \sum_{i=0}^{N} p^{n-1}(t_i=1|x_i,w_i).K(x_i,x) \tag{1}$$

and,

$$p^n(x|t=0) = \sum_{i=0}^{N} (1-p^{n-1}(t_i=1|x_i,w_i)).K(x_i,x) \tag{2}$$

where K is a Gaussian kernel, n the iteration and N the number of VOI.

2) From the model of each pathology $p^n(x|t)$, determine the likelihood that each VOI is abnormal.

$$p^n(t=1|x,w) = \frac{p^n(x|t=1).p(t=1|w)}{(1-p(t=1|w))p^n(x|t=0) + p(t=1|w).p(x|t=1)^n} \tag{3}$$

The learning algorithm stops when $p^n(t=1|x,w)$ converges to a stable solution. $p^0(t|x,w)$ is initialized as $p(t|w)$.

The value of $p(t=1|w=1)$ can be considered as a prior knowledge on the average percentage of abnormal VOI in the positive examples and is set to a constant a, while $p(t=0|w=1) = 1-a$. Under the same reasoning $p(t=1|w=0) = 0$ and $p(t=0|w=0) = 1$.

From the obtained likelihood $p(x|t)$ the final posterior distribution can be estimated through Bayes' rule:

$$p(t=1|x) = \frac{p(x|t=1).p(t=1)}{p(x|t=0).(1-p(t=1)) + p(x|t=1).p(t=1)} \tag{4}$$

where $p(t=1)$ is a predefined value.

The definition of a influences the estimation of $p(t|x)$, specially for feature space regions where there is no clear majority in concentration of positive/negative bags. A large value of a gives a larger importance to positive examples, and the opposite for a small a. In the limit for $a=0$ no VOI will be considered positive, and for $a=1$ all VOI in positive bags are considered positive.

Fig. 2 contains the evolution of the $p(t|x,w)$ and $p(x|t)$ for the hyperlucency pattern. It is visible the progressive concentration of $p(x|t)$ to regions where the concentration of positive VOI is higher than negative VOI.

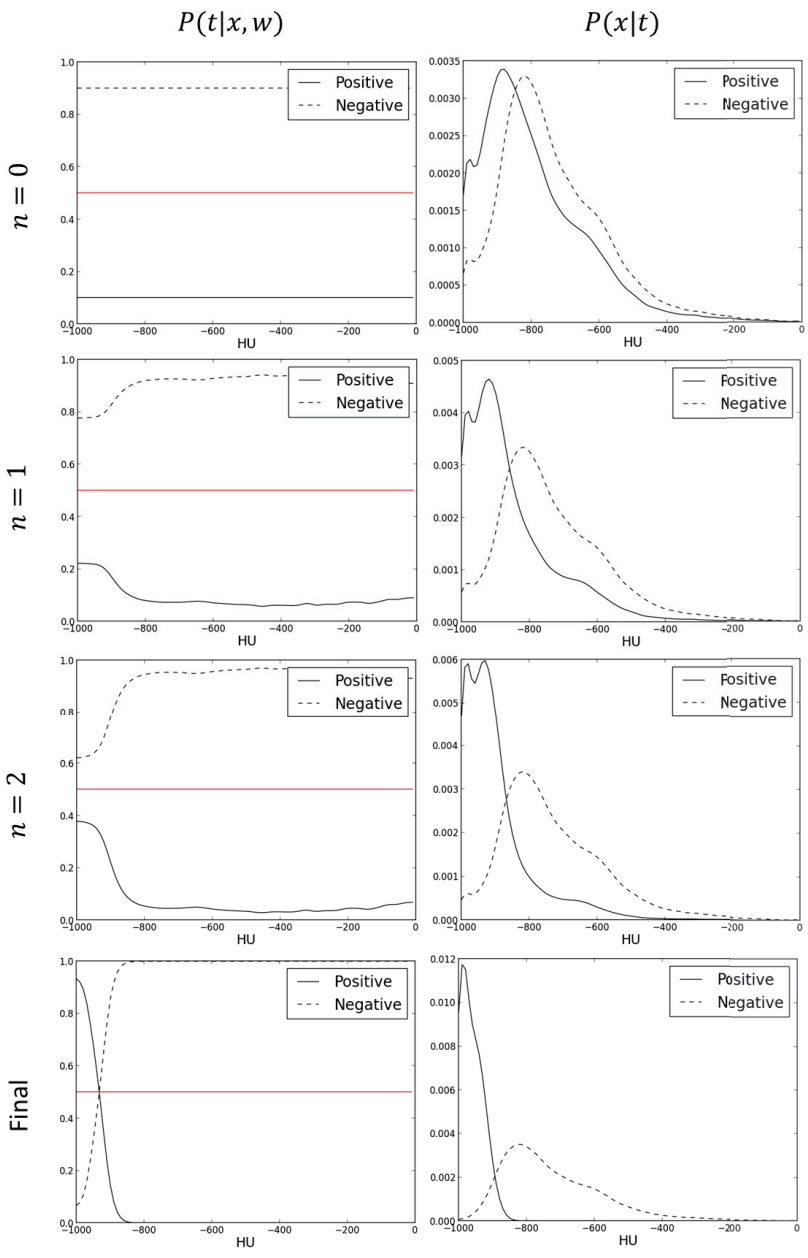

Fig. 2. Evolution of the the posterior probability (left column) and the likelihood (right column) for the described MIL algorithm and the hyperlucency pattern.

3 Experiments and Results

3.1 Evaluation Dataset

A second database of 24 clinical dose thoracic CT scans was used to evaluate the performance of our method. They were acquired between April 2004 and March 2010 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 was acquired in spiral mode and reconstructed to 512×512 or 768×768 matrices. Section spacing was between 0.8 and 5.0 mm with 0.0 - 1.0 mm overlap. No contrast material was used.

The manual annotation system described in [2] was used by an intern radiologist to mark each abnormal VOI into one of 7 classes: Hyperlucency, consolidation, honeycombing, ground glass, crazy paving, non-specific interstitial pneumonia pattern and Nodular pattern. In addition, he indicated all VOIs which contained more than one type of tissue as inhomogeneous. All remaining VOIs were labeled as normal lung tissue.

3.2 Results

The dataset described in the previous section was used to build the true posterior probability for each studied pattern ($p_{true}(t = 1|x)$), and this distribution was compared to the $p(t = 1|x)$ obtained from our method in terms of their ability to differentiate the VOI in the evaluation dataset. We considered $p(t = 1) = p(t = 0) = 0.5$. The binary classifier is a simple threshold on the posterior probability.

Table 2 presents the Area Under the Receiving Operation Curve (AUC) for true and estimated posterior probabilities. The AUC was estimated from the specificity/sensitivity for a progressively larger discriminant threshold. The posterior probability was estimated 20 times for different sets of randomly picked exams, and for different values of a. Table 2 presents the mean AUC, and in parenthesis the standard deviation. It can be observed that our approach has a slightly inferior performance to the true distribution for hyperlucency, ground glass, crazy paving and honeycombing. For consolidation and nodular patterns, the AUC is much lower. It can also be observed that there are slight differences in performance depending on the value considered for the parameter a. The standard deviation is small for all patterns, with the exception of consolidation.

Fig. 3 presents in the first column the initial positive/negative distributions for honeycombing, crazy paving, consolidation and nodular patterns. In the second column the true posterior distribution ($p_{true}(t = 1|x)$), and one of the estimated posterior distributions for different values of a. For the first two patterns it can be observed that our approach produces a similar posterior distribution to the VOI annotations. For the last two patterns the two distributions are very different. In all cases it can be observed that the value of a influences the final outcome of the distribution.

Table 2. Area under the ROC curve for hyperlucency (HL), ground glass (GG), honeycombing (HC), crazy paving (CP), consolidation (Cons.) and nodular (Nod.), from the posterior distribution obtained from VOI annotations p_{true}, and estimated by our method $p(t = 1|x)$ for different values of a. The $p(t = 1|x)$ was estimated 20 times for different sets of randomly picked exams. The table presents the mean AUC value for all experiments and in parenthesis the standard deviation.

	HL	GG	HC	CP	Cons.	Nod.
$p_{true}(t = 1\|x)$	0.72	0.75	0.78	0.78	0.85	0.72
$p(t = 1\|x)$ $a = 0.1$	0.67	0.72	0.77	0.77	0.39	0.32
	(0.003)	(0.003)	(0.001)	(0.003)	(0.18)	(0.008)
$p(t = 1\|x)$ $a = 0.5$	0.71	0.72	0.78	0.77	0.52	0.28
	(0.004)	(0.006)	(0.001)	(0.003)	(0.25)	(0.003)
$p(t = 1\|x)$ $a = 0.9$	0.71	0.72	0.77	0.77	0.51	0.29
	(0.005)	(0.003)	(0.008)	(0.004)	(0.25)	(0.013)

4 Discussion

This paper compared the performance of our approach, which learns representations of lung patterns through the detection of keywords in radiology reports, with the conventional use of VOI annotations.

From the analysis of table 2 it can be observed that our approach performed slightly worst for ground glass, honeycombing, crazy paving and hyperlucency and poorly for consolidation and nodular patterns. The observation of fig. 3 shows that the resulting posterior probability is close to the true distribution for honeycombing and crazy paving, and very different for consolidation and nodular pattern.

The results suggest that under certain conditions, patterns can be approximately modeled without the use of medical annotations. This is surprising in our case because of the different nature of the train and evaluation datasets, the primitive model chosen for the pattern representation and the low agreement between medical doctors on the annotation of such patterns [2]. Although the posterior distribution is not exactly matched, the deviations do not affect the general performance, as they mainly exist for sparse regions and don't affect the posterior probability decisively.

There are nevertheless limitations. The performance of the results is affected, although not dramatically, by the choice of a, and our approach fails in the estimation of the nodular and consolidation patterns. Moreover the collection of exams that exclusively contain the target pattern might bias the model.

The poor results of the last two patterns seems to originate on a deficient choice of the initial exam sets, as the initial positive exams distribution in fig. 3, contrarily to what was expected, has a higher percentage of low HU values. The large values of standard deviation for consolidation point that this behavior is not consistent, and depends on the randomly chosen negative exams. This suggests that the initial exam set was severely distorted by noisy exams.

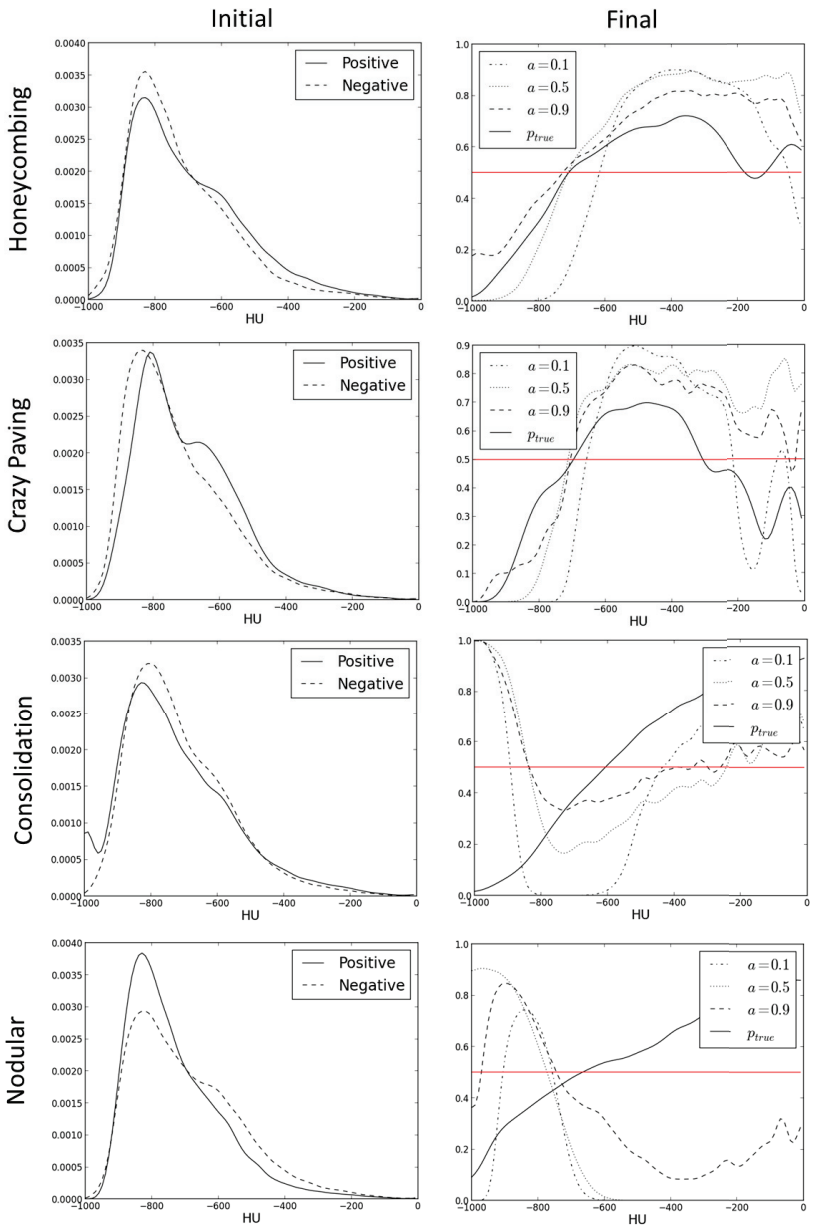

Fig. 3. Initial HU distributions for all samples from the positive/negative set (first column), and final estimated posterior distributions $p(t|x)$ and $p_{true}(t|x)$ for different values of a (second column).

There could be several reasons for this set distortion. For the nodular pattern it is possible that in the positive set are included exams with solitary nodules, which are a lot more common and would also contain the nodular keyword. This is suggested by the large number of exams containing the nodule keywords. More advanced information extraction techniques might be required for this particular finding. In the case of consolidation there might exist cross-talk with other patterns of high HU value in the negative exams set, which justifies the large values for the standard deviation. VOI models which include additional features could improve the performance of the consolidation pattern.

These limitations suggest that our approach can't be directly applied to all findings. Specific image/text approaches or manually picked exams might be required for an acceptable performance. Nevertheless our system seems to greatly reduce the amount of required medical intervention.

5 Conclusion

In this paper we presented a new approach for classifying representations of lung CT patterns without a dataset of annotated VOI by radiologists. Our approach used keywords detected in radiology reports to select exams that contain the target pattern, and subsequently used a multiple instance algorithm to single out the abnormal VOI.

We evaluated the proposed method by comparing its performance against the traditional use of VOI annotations for 6 patterns related to ILD diagnosis. Our approach had a comparable performance for four of the considered patterns, and failed for two patterns due to a deficient exam collection. The presented experiments suggest that our approach is effective if the initial exam set is not excessively cluttered with noisy exams which are discordant with the true pattern model. Despite these limitations results suggest that our approach greatly eases the design of VOI classification systems. This could broaden its application to a larger number of patterns, beyond the point where the use of traditional VOI annotations is feasible, hence improving CAD systems CT representations.

Future improvements include the evaluation of our approach in more elaborate VOI models with additional features, and the application of more advanced information extraction/natural language processing techniques.

6 Acknowledgements

We would like to thank Prof. Dr. Isabel Ramos and Dr. António Morais for their support. This work was funded by Fundação para a Ciência e Tecnologia (FCT) under the Programa Operacional de Potencial Humano (POPH) program through the grant SFRH / BD / 40864 / 2007.

A List of Keywords

These lists are translations from the original Portuguese terms, and for brevity do not contain terms inflections, variations in hyphenation, and typographical errors. **Hyperlucency** - (emphysema, hiper transparent, cyst, pneumotocele, pneumotorax); **Honeycombing** - (honeycombing, fibrosis); **Ground Glass** - (ground glass); **Crazy Paving** - (crazy paving); **Consolidation** - (consolidation, condensation, opacity); **Nodular** - (nodule, micronodule, multinodular).

References

1. Depeursinge, A., Vargas, A., Platon, A., Geissbuhler, A., Poletti, P., Muller, H.: 3D case based retrieval for interstitial lung diseases. In: Medical Content-Based Retrieval for Clinical Decision Support, pp. 39–48. Springer (2010)
2. Kockelkorn, T.T.J.P., de Jong, P.A., Gietema, H.A., Grutters, J.C., Prokop, M., van Ginneken, B.: Interactive annotation of textures in thoracic CT scans. In: Proceedings of SPIE Medical Imaging 2010: Computer-Aided Diagnosis. vol. 7624, pp. 76240X–8. SPIE, San Diego, California, USA (Mar 2010)
3. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: Proceedings of the Fifteenth International Conference on Machine Learning. pp. 341–349. ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998)
4. Ramos, J., Kockelkorn, T., van Ginneken, B., Viergever, M.A., Ramos, R., Campilho, A.: Supervised content based image retrieval using radiology reports. In: Proceedings of the 9th international conference on Image Analysis and Recognition - Volume Part II. pp. 249–258. ICIAR'12, Springer-Verlag, Berlin, Heidelberg (2012)
5. Sluimer, I., Prokop, M., van Ginneken, B.: Toward automated segmentation of the pathological lung in CT. IEEE Transactions on Medical Imaging 24(8), 1025–1038 (Aug 2005)
6. Sluimer, I., Schilham, A., Prokop, M., van Ginneken, B.: Computer analysis of computed tomography scans of the lung: a survey. IEEE Transactions on Medical Imaging 25(4), 385–405 (Apr 2006)
7. Sluimer, I.C., Prokop, M., Hartmann, I., van Ginneken, B.: Automated classification of hyperlucency, fibrosis, ground glass, solid, and focal lesions in high-resolution CT of the lung. Medical Physics 33(7), 2610–2620 (Jul 2006)
8. Uchiyama, Y., Katsuragawa, S., Abe, H., Shiraishi, J., Li, F., Li, Q., Zhang, C.T., Suzuki, K., Doi, K.: Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography. Medical Physics 30(9), 2440–2454 (2003)
9. Ulges, A., Schulze, C., Keysers, D., Breuel, T.: Identifying relevant frames in weakly labeled videos for training concept detectors. In: Proceedings of the 2008 international conference on Content-based image and video retrieval. pp. 9–16. CIVR '08, ACM, New York, NY, USA (2008)
10. Uppaluri, R., Hoffman, E.A., Sonka, M., Hartley, P.G., Hunninghae, G.W., McLennan, G.: Computer recognition of regional lung disease patterns. Am. J. Respir. Crit. Care Med. 160(2), 648–654 (Aug 1999)
11. Zhang, Q., Goldman, S.A.: EM-DD: an improved multiple-instance learning technique. In: Advances in Neural Information Processing Systems. pp. 1073–1080. MIT Press (2001)