Lung CT analysis and retrieval as a diagnostic aid

Henning Müller, Samuel Marquis, Gilles Cohen, Antoine Geissbuhler

Service of Medical Informatics, University and Hospitals of Geneva

Abstract

Image retrieval is currently a very active research field due to the large amount of visual data being produced in most modern hospitals. Most often, the goal is to aid the diagnostic process. Unfortunately, only very few medical image retrieval systems are currently used in clinical routine. One of the application domains for image retrieval is the analysis and retrieval of lung CTs. A first user study in the United States shows that these systems allow improving the diagnostic quality.

This article describes our approach to an aid for lung CT diagnostics. The analysis incorporates several steps and the goal is to automate the process as much as possible. Thus, several automatic steps are proposed from a selection of the most characteristic slides, to an automatic segmentation of the lung tissue and a classification on the segmented area into diagnostic classes. Feedback to the MD will be given in the form of marked regions in the images that appear to be different from the norm of healthy tissue. We are currently working on a small set of training images with marked and annotated regions but a larger set of images for the evaluation of our algorithm is in work. For this reason, the article does not contain any quantitative evaluation.

For several tasks we use existing open source software such as Weka and itk. This allows an easy reproduction of the search results.

Keywords

Content-based image retrieval, high-resolution lung CT, diagnostic aid, classification

1. Introduction

Content-based image retrieval has been an extremely active domain in the fields of computer vision and images processing for more than 20 years [1]. In the medical field, this domain is also starting to be very active, as an increasing amount of visual data is being produced in hospitals and made available in digital form [2,3]. General medical image retrieval in PACS-like databases is in this context very different from specialized retrieval in a very focused domain. In the medical field, the main goal is the use as a diagnostic aid. Current medical use is on the retrieval of tumour shapes [4] as well as on histological images [5], and in several other fields (dermatology, pathology). Another domain where textures play a very important role in the diagnostic process is the analysis of high-resolution lung CTs [6]. In [7], a user test shows that an image retrieval system can improve the diagnostic quality significantly, especially for less experienced radiologists. Still, most of these systems either rely on an extremely large amount of interaction with the user, which makes them hard to introduce into a clinical context, or they are too broad to be used as a diagnostic aid in a specialized domain.

This article details a solution for helping with the interpretation of high-resolution lung CTs, which is a domain where diagnostics are fairly hard especially for non-chest specialists. The diagnostic result strongly depends on the overall texture of the lung tissue, so automatic analysis seems possible. Our project also tries to limit the direct interaction with the user and performs as many tasks as possible in an automatic fashion, so a minimum of time is needed to operate the system and get responses for feedback.

Section 2 describes the various steps of the analysis and retrieval process and their degree of automation. Section 3 explains our current results and section 4 discusses our findings and the future work that is planned within the project.

2. Steps for a diagnostic aid on lung CT interpretation

This section describes the various steps that are necessary for a complete diagnostic aid system for lung diagnostics and their degree of automation.

2.1 Generation of a test database and acquisition of representative samples

The first and most important part is the creation of a database of thin-section lung CTs. This database needs to include healthy cases as well as pathologic cases. Characteristic regions need to be marked by a radiologist to allow us learning the characteristics of a certain pathology with respect to healthy tissue. Currently, we only have a fairly small database of 10 series containing around 50-60 images per series. Several regions are marked in the images to represent the following classes of tissue for the further classification step:

- Healthy tissue;
- Emphysema;
- Micronodules;
- Macronodules;
- Interstitial syndrome;
- Alveolary syndrome;
- Ground glass attenuation;
- Opacities.

It is important to note that prototypically healthy regions must also be annotated by the radiologist sufficiently so that a classifier can get a good idea of healthy tissue. Other system in the literature often use only pathologic classes for the classification but the first step in the diagnostic process is to find out whether the tissue is abnormal or not. We are currently creating a larger database projected to contain at least 100 series of 50-60 images that will allow us a better representation of these classes.
In Figure 1 you can see a screenshot of our tool for image annotation. It generates a simple XML file containing the regions of interests as a set of points (outline) and a label for each outline. These files are then fed into the system along with the images at the training step.

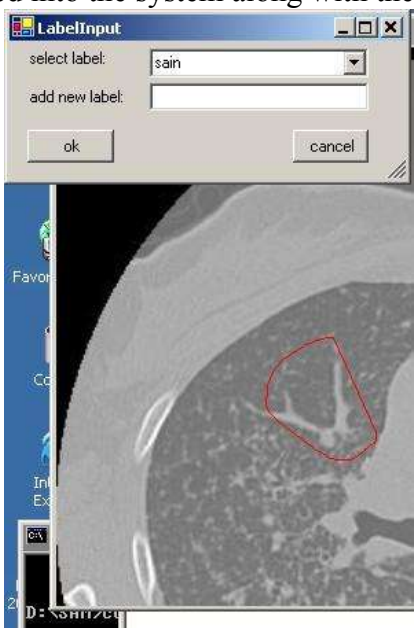

Figure 1: A screenshot of our utility for the annotation of image regions.

2.2 Analysis of blocks of lung tissue

In Figure 3 you can see the partitioning of the lungs into smaller blocks (currently of size 16x16 pixels) for further detailed texture analysis. A block is taken into account if it is by more than 80% inside of an area marked by the expert or automatically segmented by the

system. These small lung blocks are stored as references together with the original image. This avoids artifacts of the filters that can occur due to missing border pixels because we can take into account the entire block environment.

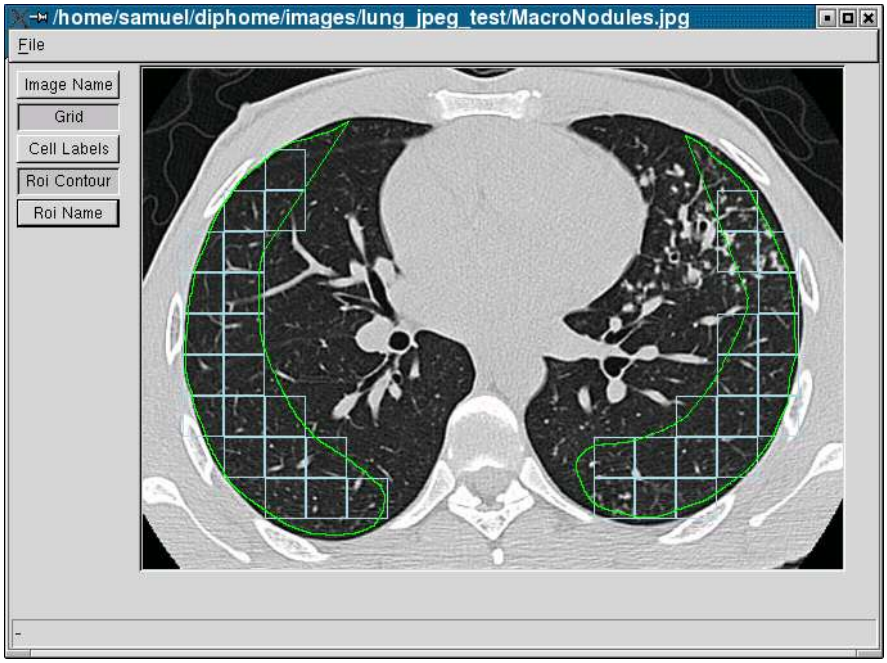

Figure 3: The partitioning of the lung tissue into small blocks for classification and feature extraction.

The framework is designed to facilitate finding the optimal size of these blocks for analysis and classification, which has not yet been explored. From each block we extract and store the following visual features:

- average grey level and standard deviation of the grey levels in the block;
- grey level histogram using 16 grey levels;
- features derived from co-occurrence matrices (four directions, two distances);
- responses of Gabor filters in four directions and at three scales;
- Tamura texture features.

The small number of grey levels in the histogram is sufficient for this kind of classification as has been shown in image retrieval applications, where small numbers of grey levels perform often better for retrieval.

2.3 Training

Then, the features together with the region label become a sample in a classification problem [8]. Based on the acquired training data, the weights of the features for classification are calculated. To develop an optimal classification strategy, several classifiers are tested and their performance is evaluated on the currently available data. An open source utility that allows us to compare several classifiers is Weka[1], which has also the advantage of being able to connect directly to the feature database (in mySQL). We can, for instance, perform cross-validation using various classifiers to get an idea of how discriminant our features are. Weka also performs an automatic attribute selection.

While Weka is an external tool, we have also included *libsvm* [9] into the framework, a convenient and easy-to-use Support Vector Machine (SVM) classifier, which can perform cross-validation. We also plan to integrate *torch* [10] to discover connections between the various features and the classes of our system. *Torch* is a fast data-mining tool. In the final version of our classification software, only the best classifier will be integrated but as for the testing phase we need several to find out the optimal solution and especially the best configuration of the classifyer.

[1] http://www.cs.waikato.ac.nz/~ml/weka/

2.3 Lung segmentation as data preparation for classification

When submitting a new image for analysis and as diagnostic aid, we concentrate on the part of the image that we are interested in, the lung tissue. While manual region selection of the image is still possible (using the tool described in 2.1 – only without a label), automatic segmentation is desired to minimize user interaction in the final diagnostic aid step. To this aim we use an algorithm similar to [11] to find an optimal threshold for lung tissue segmentation, which works on DICOM images having a full 12-bit resolution (or more) as well as on the jpeg images from our radiology teaching file. As basis for the segmentation we also use itk[2]. In Figure 2 you can see a lung CT, its segmented version and a view of the outline discovered by the software; the outline is stored in an xml file as above and fed into the system to create blocks and classify them.

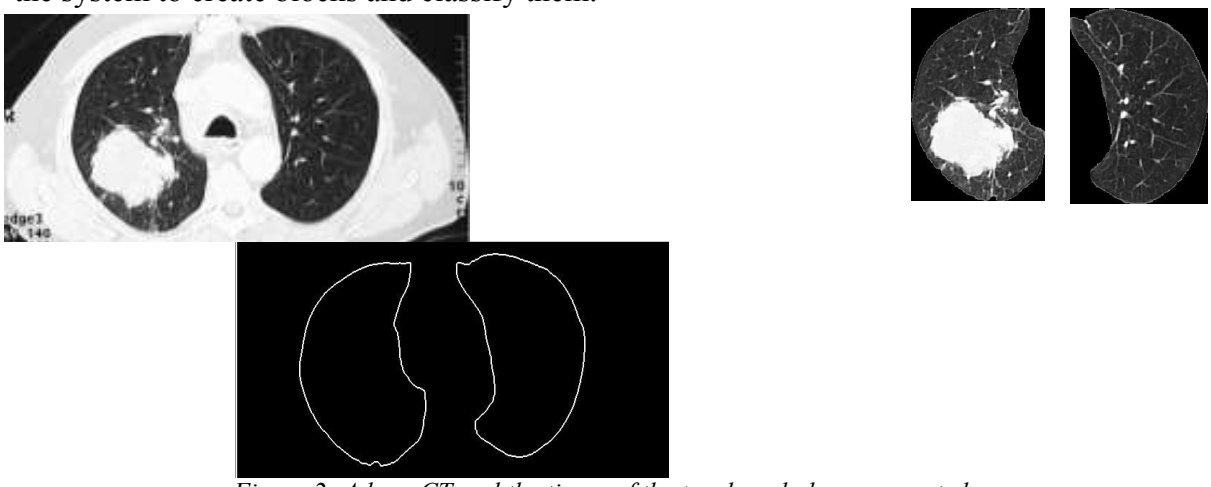

Figure 2: A lung CT and the tissue of the two lung halves segmented.

For the final texture classification we do not plan to take into account the entire lung tissue but rather the diagnostically interesting part which is the outside part of the lung with less vessels that can change the texture strongly and introduce noise for the classification. The inside part with the vessels is automatically removed from further analysis.

2.4 Classification of lung blocks

For the classification step of a lung image, a partitioning of the image into blocks is performed, and the system extracts the features of each block in the (manually or automatically) marked regions corresponding to lung tissue. This time, the samples are created by the block's features and have no label attached to them. The integrated classifier (currently *libsvm*) now performs the classification and attaches a label to each block of the lung tissue. Goal is to find the correct label for each block. Tuning the classifier is a difficult part of the problem. We use the parameters (kernel type, gamma factor, stop criterion, etc.) which performed best at cross-validation (see step 2.3).

2.5 Slice selection from a CT lung volume

This step is currently not implemented and will most likely be the last part to be done as its development is not crucial. Goal of this step is basically to perform the task of the medical doctor to find the slice(s) that best characterize the disease. Once a large database of labeled tissue samples is created it will be fairly easy to process the entire volume slice by slice and select those slices with the largest part of the tissue being marked as non-healthy for further inspection by a medical doctor. To select several slices, we can simply give the system a combination of a maximum number of slices and a threshold for percentage of tissue marked unhealthy. The selected slices can also be marked in the volume data by highlighting the volume part that contains the most pathologic blocks.

[2] http://www.itk.org/

2.6 Result presentation to the medical doctor

The goal of the results presentation is not to take the decision for of the radiologist but rather to highlight parts of the lung tissue that were classified as pathologic. This is planned by highlighting the background in colored shades instead of the grey scales into the parts of the images that were classified as pathologic. Each color presents one of the classes that we are detecting in the classification step. Currently, we only present the results in a 2D view and one image at a time, meaning that the slices with the largest pathogenic parts are taken and displayed. It can also be imagined to present the results in a 3D, where the entire pathogenic area over several slices can be highlighted within the volume.
Retrieval of similar cases from the reference database will enable the MD to judge the quality of the system response and verify his diagnosis. This is easily possible through a simple image retrieval application that uses the features from the currently active case and comparing them with features stored and labeled for past cases in the database.

3. Results

We currently have a framework in place that allows us to acquire the knowledge from the radiologists in the form of marked regions and their annotations within the images. Our database is still small but a much larger number of cases are planned for better evaluation. All the acquired data with labels is used to train our classifiers. This means that with a growing number of judged cases from the radiologist, the system is expected to perform better. The lung segmentation phase works reliably and stable as well as the partitioning of the lung tissue into small blocks and the feature extraction.
All these steps work in a completely automated fashion. The medical doctor can simply feed a volume of lung CT images into the system, the images are segmented and portioned into smaller blocks automatically. These blocks are then classified and unhealthy tissue is marked in a different color in the images so the medical doctor has a feedback for regions that he needs to inspect further.
We currently run the framework on a simply desktop computer with a Pentium IV processor with 2,8 GHz and 1 MB of RAM. On this computer, the segmentation takes around 5 seconds per slice and the subsequent cutting into blocks, feature extraction and classification another 2 seconds. Thus the analysis of a single slice is almost interactive whereas an entire volume takes a few minutes before the results are displayed.
We still need to experiment more with the classification part and maybe also with the features that we extract from the images to obtain an optimal feature set for classification.
The current framework is by now a research tool, designed to ease experimentation of features, classifiers, parameters, etc. The final system will probably discard a lot of these options, be much simpler and focus more on the user interface and the results display to the user.

4 Conclusions

This article presents a framework to aid the diagnostic process for lung diseases using lung CTs as input. The domain has shown its potential in other tests and our current cross validations also show good results. The various steps of the diagnostic process are performed in a completely automatic way. Abnormalities are highlighted in the original images by a change of color.
Once we have a larger database accessible, a quantitative evaluation is needed to evaluate the quality of our algorithms and show the usefulness of the application in a clinical environment. Many parameters will need to be optimized, from the feature extraction phase to the training step and the classifiers employed. We also need to think about the optimal

size of the blocks of lung tissue and whether we should rather take overlapping blocks to avoid misclassifying small parts of the texture and reduce false positives.
Lung CT analysis has shown its usefulness in practice by improving the diagnostic quality especially of non-experts. Now it is important to create reference databases and evaluate the many visual descriptors and techniques available to create a robust framework for routine use that needs to have as many steps of the process in an automatic way as possible.
Several questions still need to be resolved before routine use, for example the handling of other available data on the patients. The age can play an important role for the texture of the lung tissue. For the classification we need to integrate all these data into the framework.
The possibility to compare the images with annotated cases from the reference databases is expected to further increase acceptance of the technology because the system does not make a decision by itself but rather point out interesting areas of the lung tissue and gives evidence on these areas by supplying similar past cases.

References

[1] AWM. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12) pp 1349-1380, 2000.

[2] ·H Müller, N Michoux, D Bandon, A Geissbuhler, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, *International Journal of Medical Informatics*, **73**, pp 1-23, 2004.

[3] TM Lehmann, MO Güld, C Thies, B Fischer, K Spitzer, D Keysers, H Ney, M Kohnen, H Schubert, BB Wein, Content-based image retrieval in medical applications, *Methods of Information in Medicine*,**43**, pp 354-361, 2004.

[4] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, Z. Protopapas, Fast and effective retrieval of medical tumor shapes, *IEEE Transactions on Knowledge and Data Engineering,* **10**(6) 889—904, 1998.

[5] LHY Tang, R Hanka, HHS Ip, A review of intelligent content-based indexing and browsing of medical images, *Health Informatics Journal* **5**, 40—49, 1998.

[6] C.-R. Shyu, CE Brodley, AC Kak, A Kosaka, AM Aisen, LS Broderick, ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases, *Computer Vision and Image Understanding* **75** (1—2), pp. 111—132, 1999.

[7] AM Aisen, LS Broderick, H Winer-Muram, CE Brodley, AC Kak, C Pavlopoulou, J Dy, CR Shyu, A Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, *Radiology*, **228**, pp. 265-270, 2003.

[8] A K Jain, R P W Dvi, J Mao, Statistical Pattern Recognition: A Review, IEEE Transactions onf Pattern Analysis and Machine Intelligence, **22** (1) pp. 4-37, 2000.

[9] CC Chang, CJ Lin, libsvm, a library for support vector machines, Technical report, available with software at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[10]R Collobert, S Bengio, J Mariéthoz. Torch: a modular machine learning software library. *Technical Report* IDIAP-RR 02-46, IDIAP, 2002.

[11]S Hu, EA Hoffman, JM Reinhardt, Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Transactions on Medical Imaging*, **20**(6), pp. 490-498, 2001.

7. Address for correspondence

Henning Müller

University and Hospitals of Geneva,

Service of Medical Informatics

24, rue Micheli-du-Crest,

CH-1211 Geneva 14, Switzerland

henning.mueller@sim.hcuge.ch

tel ++ 41 22 372 6175, fax ++41 22 372 8680