Exploring the limits of CT image intensity for discriminating lung tumours and the atelectasis

Vassili Kovalev[a], Maria Petrou[b] and Serguey Khoruzhik[c] *

[a]Biomedical Image Analysis Group, United Institute of Informatics Problems,
Belarus National Academy of Sciences, Surganova, 6, 220012 Minsk, Belarus
[b]Communications and Signal Processing Group, Department of Electrical and Electronic Engineering,
Imperial College, Exhibition Road, South Kensington, London SW7 2AZ, UK
[c]Department of Diagnostic Imaging, N.N. Alexandrov Research Institute of Oncology
and Medical Radiology, P.O. Lesnoj-2, 223040 Minsk, Belarus

Abstract. Computed tomography (CT) is the primary modality for imaging the lung cancer patients. However, the differentiation of lung atelectasis and malignant tumours is hardly possible due to a very similar visual appearance. In this paper, we explore the limits of the usefulness of CT image intensity information for discriminating the lung atelectasis and tumour regions. The statistical significance of intensity differences as a function of voxel sample size was assessed on CT scans of 40 lung cancer patients using unpaired *t*-test. The classification accuracy was evaluated with the help of Hierarchical Clustering, Support Vector Machines, and Random Forests methods using 44000 training and test sets of voxels sampled at random. Visualisation of sample data vectors was performed using a Multidimensional Scaling technique. The Hierarchical Clustering algorithm was found to be the best suited for segmentation purposes with its potential segmentation accuracy of 4.1 mm for 2D and 2.0 mm for 3D cases.

1 Introduction

The atelectasis term denotes the collapse of all or part of a lung due to bronchial plugging or the chest cavity being opened to atmospheric pressure. This can happen when the vacuum between the lung and chest wall is broken, allowing the lung to collapse within the chest (e.g., pneumothorax), when the lung is compressed by masses in the chest, or when an airway is blocked, leading to slow absorption of the distal air into the blood without replenishment. In this work we were dealing with the bronchial compression caused by lung cancer tumours, the most common cause of the atelectasis.

Computed tomography (CT) is the primary modality for imaging lung cancer patients. However, on CT scans the lung regions with the atelectasis and malignant tumours have quite similar attenuation values. Therefore the visual discrimination and separation of the atelectasis and tumours is hardly possible [1], [2]. Yet an accurate tumour segmentation is strongly necessary by the following two reasons. First, the correct tumour localisation, segmentation, and precise measurement of tumour diameter play a crucial role in the therapy planning and choosing suitable surgery technique. Second, if the radiation therapy is prescribed, an exact tumour border is required for precise targeting and accurate delivery of the ionising radiation exactly to the tumour but not to the surrounding tissue [3].

This work should be considered as a part of more general project the ultimate goal of which is to develop methods and software solutions for interactive discrimination and segmentation of cancerous tissue from the atelectasis. In this paper, we exploring the limits of the usefulness of CT image intensity information alone for differentiation of lung atelectasis and malignant tumours using statistical and pattern recognition methods. All the analyses were performed intra-subjectly (eg, the training and test sets of image voxels were sampled from the *same* patient). This is because we were mainly interested in performing differentiation of atelectasis and tumour regions in a gradient-like manner [4] for each particular patient but *not* in the evaluating existing inter-subject distinctions.

2 Materials

In this study we used 40 CT images of the chest of patients with lung cancer and atelectasis of a portion of the lung as diagnosed by a qualified radiologist and confirmed histologically. Thirty seven of them were males and remaining three were females. The age of patients ranged from 41 to 80 years with the mean value of 61.1 and STD of 9.2 years.

CT scanning was performed on a multi-slice Volume Zoom Siemens scanner with the standard clinical kV and mA settings during the one-breath hold. The voxel size was equal to 0.68 mm in the axial image plane with the slice thickness equal to the inter-slice distance of 7 mm. No intravenous contrast agent was administered before the collection of scan data. Typical example of original CT image slice and mean intensity values are shown in Fig. 1.

*vassili.kovalev@gmail.com (corresponding author), maria.petrou@imperial.ac.uk, skhoruzhik@nld.by

Atelectasis and tumour regions were segmented manually in each image slice of each patient by a radiologist using institutional software package called Voligator. Depending on the particular patient, the atelectasis occupied from 4 to 27 axial slices. A narrow manually-adjusted intensity window was used during the outlining atelectasis and tumour borders as it helps to slightly accentuate the intensity difference between the two region types. Generally, the tumours were located in the lung more centrally (ie, closer to the mediastinum) while the atelectasis often was located more peripherally. It was also noticed that the tumour in-plane contours tend to have more bulging shape while the atelectasis regions were rather bounded by "piecewise-straight" lines. Finally, it should be pointed out that the manual segmentation was performed intuitively, solely based on an extensive clinical experience. Nevertheless, in some cases the correct delineation of the "true" boundaries may not be guarantied due to the above-mentioned reasons.

3 Methods

The approach followed in this study was to sub-sample image voxels from two types of lung regions at random and to evaluate the significance of the intensity differences as a function of the sample size. This was done for each patient separately. The voxels were sampled *without* replacement. The training and test sets do not overlap. In order to ease the interpretability of the results, the sample sizes were selected so that they correspond to the number of voxels in square-shaped image slice patches with the side size of 3, 4, ..., 10, 15, 20, and 30 voxels that is 9, 16, ..., 100, 225, 400, and 900 sample voxels respectively. This does not means that the analysis methodology we developing is 2D-oriented, though. All statistical and pattern recognition analyses described in this work were performed using R, a language and environment for statistical computing which is available for free [5]. The atelectasis and tumour classes were compared by various ways to eliminate possible bias of one singe method. First, the significance of intensity differences between the two classes was assessed statistically using a two-tailed unpaired *t*-test with the significance level of *t*-statistics set to $p < 0.05$. The resultant *t*-values, which depend on the degree of freedom (sample size) were converted into $z-$scores to enable direct comparison of statistical significance obtained in different experiments as well as to calculate the mean significance scores over all 40 patients correctly. For each patient and each sample size the procedure consisting of random voxel sub-sampling and performing *t*-test was replicated 100 times in order to obtain reliable results.

At the second stage, the atelectasis and tumour voxel samples (ie, the vectors of voxels sorted in descending order and treated as features) were clustered using Hierarchical Clustering [6], Support Vector Machines [7], and Random Forests [8] methods. For each sample size and each patient the classifiers were trained on a training sets containing 10 atelectasis and 10 tumour samples and tested on the datasets of the same size. Training and test sets were sampled independently. There was no voxels included in both training and test sets simultaneously. The three classifiers were run on exactly the same data. Each test was replicated 100 times in order to obtain statistically representative estimates of the classification accuracy. The classification accuracy was corrected for agreement by chance using the kappa statistic as implemented in `classAgreement` function provided with R. For two classes this particularly means that the minimal accuracy value was 0 but not 50%. The corrected classification accuracy was used as a measure of the dissimilarity of two lung regions as well as the basic value for estimating possible image segmentation accuracy. The total number of performed classification tests was: 40 patients $\times$ 11 sample sizes $\times$ 3 methods $\times$ 100 replications = 132000.

Finally, the Cailliez′s version [9] of Multidimensional Scaling (MDS) method [10] was utilized for reducing the dimensionality of the space of voxel sample vectors down to two dimensions in order to produce conventional voxel sample scatterplots. Note that the multidimensional scaling provides an approximate solution, which is mostly suitable for visual examination of overlapping object classes. The total computational time (including all overheads) for 132000 classification tests and 40 $\times$ 11 = 440 runs of the multidimensional scaling algorithm was 109 minutes on a personal computer equipped with 1.86 GHz Intel Core 2 Duo processor.

4 Results

Results of statistical assessment of the significance of intensity differences between the atelectasis and tumour regions of lung CT scans of 40 patients are reported in Fig. 2. As it can be seen from the figure, the fraction of significantly different voxel samples and the mean significance scores varied considerably depending on the patient. For instance, for one patient the percentage of significantly different samples exceeds notable 60% already on 9 voxels and achieves 100% with the sample size as little as 36 voxels (see the left panel of Fig. 2) while in other it starts close to zero with 9 voxels and finishes at about 10% only. Similarly, for some patients the mean $z-$score achieves the significance threshold $z > 1.96$ which is equivalent to $p < 0.05$ on the sample size of 9–25 voxels (see the right panel of Fig. 2)

while for others these values remain insignificantly low even on reasonably large samples consisting of 400-900 voxels.

On the contrary, the voxel sample classification results demonstrate much more consistent behaviour (see Fig. 3). As it can be revealed from the figure, a very useful property of the classification approach for separating the atelectasis and tumour regions is that the results are converged to 90–100% of the classification accuracy for relatively large samples *in each* patient. As for the comparative efficiency of the three classification methods, it is easy to see from Fig. 3 that the Hierarchical Clustering algorithm outperforms both SVM and Random Forests for each voxel sample size. Moreover, in case of Hierarchical Clustering, the classification accuracy corrected for the agreement by chance starts from the value above 50% almost for each patient and achieves 90% on the sample size of 225 voxels for all 40 patients except for 2 outliers. The mean and standard deviation values of the classification accuracy computed over 40 patients (see the bottom right plot of Fig. 3) make the superiority of Hierarchical Clustering method evident and renders other two as almost identical in the voxel sample classification task. Considering that the one possible segmentation technique could be based on a direct voxel sample classification using sliding window of suitable size, the mean accuracy threshold should be set to a reasonably high value, say 95%. If so, the minimal sample size should be set to approximately 100-200 voxels. This corresponds to the window size of about 12×12 voxels (ie, the half window size is 4.1 mm) for 2D and less than $6\times6\times6$ voxels (2.0 mm) for 3D case.

Fig. 4 shows an example of separation of voxel samples in multidimensional space. It should be stressed that this figure illustrates one single case out of 440 and thus may not be as conclusive as the results represented in Fig.3.

5 Conclusions

In this work we have documented results of patient-wise assessment of CT image intensity differences between the lung atelectasis and malignant tumour regions. Our results suggest that it is unlikely that the use of statistical significance scores for separating lung atelectasis and tumour regions would produce good quality discrimination for all the patients. However, the recent clustering algorithms demonstrate some encouraging classification accuracy on the CT intensity samples consisting of few hundred voxels. The Hierarchical Clustering method was found to be better suited for highlighting the border between the two types of regions in a gradient-like way [4] comparing to SVM and Random Forests classifiers. This is in agreement with other studies where classes overlap in feature space substantially (eg, [11], [4]). The voxel sample classification accuracy potentially allows to reliably discriminate atelectasis and tumour regions using relatively small sliding window of 12×12 voxels (ie, the half window size is 4.1 mm) in 2D and no more than $6\times6\times6$ voxels (2.0 mm) in 3D case.

Acknowledgements

This work was supported by the International Joint Project grant 2006/R1 from British Royal Society .

References

1. J. Mira, G. Fullerton, J. Ezekiel et al. “Evaluation of computed tomography numbers for treatment planning of lung cancer.” *Int J Radiat Oncol Biol Phys.* **8(9)**, pp. 1625–1628, 1982.
2. J.-M. Kuhnigk, H. Hahn, M. Hindennach et al. “Lung lobe segmentation by anatomy-guided 3D watershed transform.” In M. Sonka & J. Fitzpatrick (editors), *Medical Imaging 2003*, volume 5032, pp. 1482–1490. SPIE, USA, May 2003.
3. P. Bowden & F. Richard. “Measurement of lung tumor volumes using three-dimensional computer planning software.” *Int J Radiat Oncol Biol Phys.* **53(3)**, pp. 566–573, 2002.
4. V. Kovalev & M. Petrou. “The classification gradient.” In *Int Conf on Pattern Recognition (ICPR’06)*, volume 3, pp. 830–833. IEEE Comp Society, Hong Kong, 20-24 August 2006.
5. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
6. L. Breiman, J. Friedman, R. Olshen et al. *Classification and Regression Trees*. Chapman & Hall, New York, 1984. 384 p.
7. C. Cortes & V. Vapnik. “Support-vector network.” *Machine Learning* **20**, pp. 1–25, 1995.
8. L. Breiman. “Random forests.” *Machine Learning* **45(1)**, pp. 5–32, 2001.
9. F. Cailliez. “The analytical solution of the additive constant problem.” *Psychometrika* **48**, pp. 343–349, 1983.
10. T. F. Cox & M. A. A. Cox. *Multidimensional Scaling*. Monographs on Statistics and Applied Probability 88. Chapman & Hall/CRC Press, London, second edition, 1994. 309 p.
11. V. Kovalev, N. Harder, B. Neumann et al. “Feature selection for evaluating fluorescence microscopy images in genome-wide cell screens.” In *Computer Vision and Pattern Recognition (CVPR’06)*, pp. 276–283. IEEE Comp Society, New York, 17-22 June 2006.

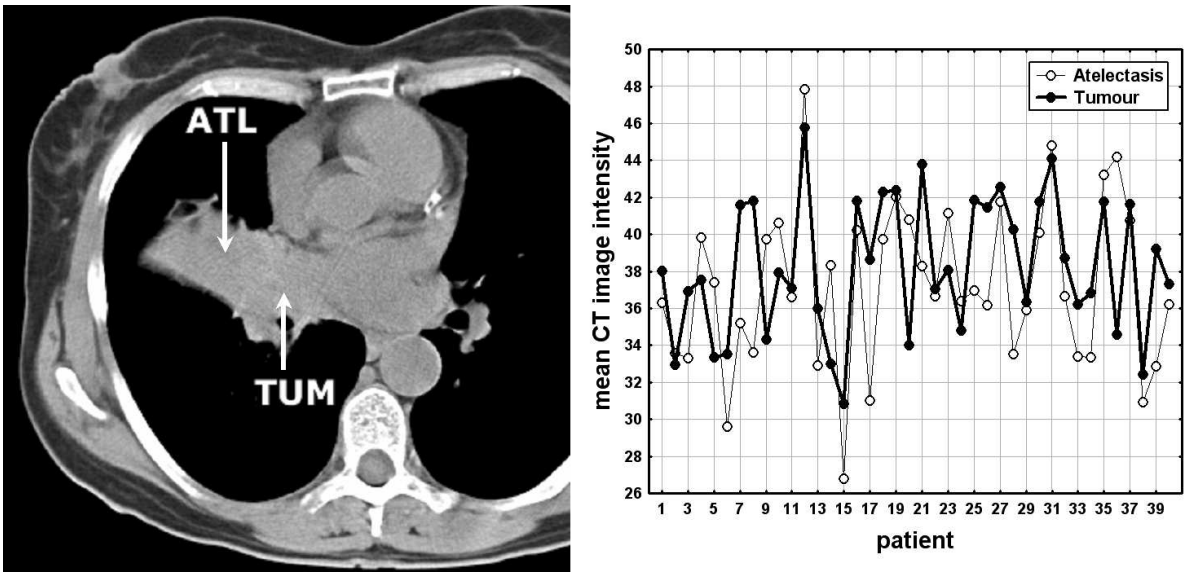

Figure 1. Example slice of typical lung CT image with atelectasis (ATL) and tumour (TUM) regions (left panel) and plots of mean intensity values of the atelectasis and tumour regions for 40 patients (right panel).

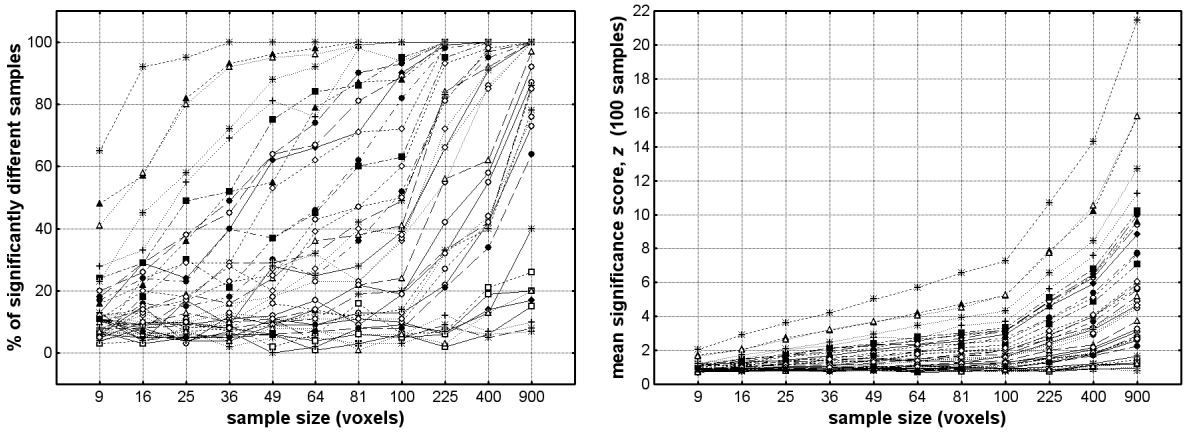

Figure 2. Significance of the intensity differences of lung atelectasis and tumour voxel samples for 40 patients (curves) as a function of the voxel sample size. Left panel: percentage of voxel samples for which the intensity difference is statistically significant at $p < 0.05$. Right panel: the mean value of significance score z. In both occasions image voxels were sampled from atelectasis and tumour regions at random and each measurement is replicated 100 times.

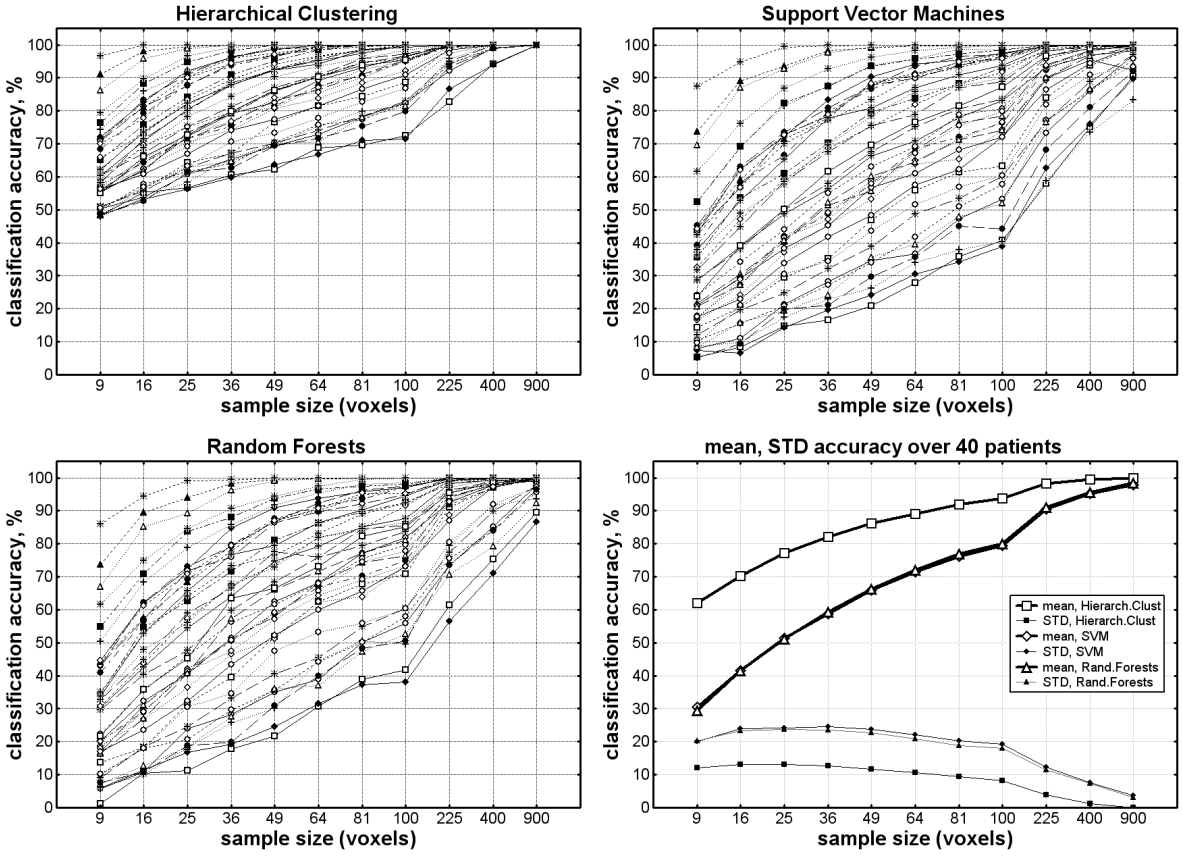

Figure 3. Dependence of the classification accuracy on sample size of lung atelectasis and tumour voxels for 40 patients (curves) when using Hierarchical Clustering (top left plot), Support Vector Machines (top right plot), and Random Forests (bottom left plot) clustering methods. Each test was replicated 100 times for the reliability of results. The mean and standard deviation accuracy computed over 40 patients is given on the bottom right panel.

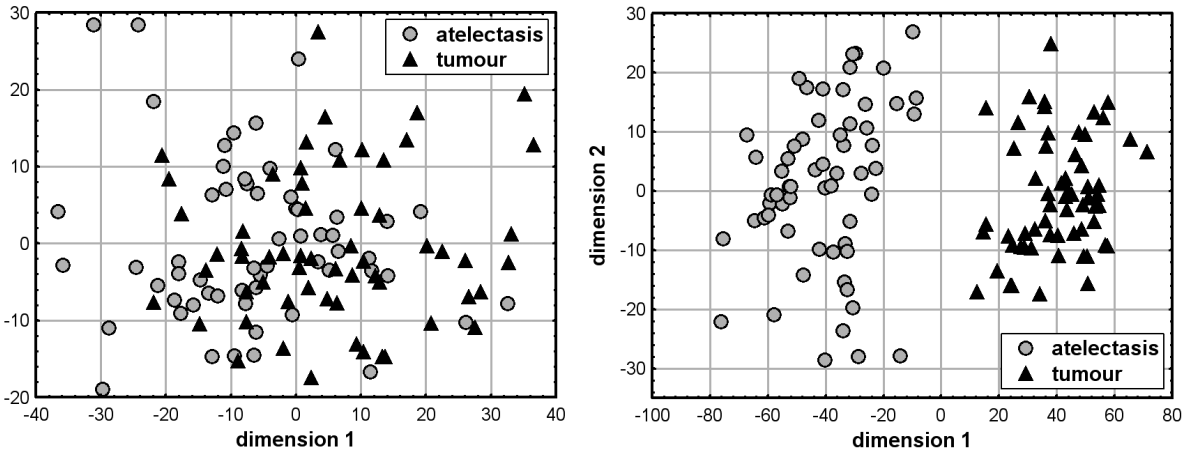

Figure 4. Example of separation of 120 image voxel vectors (dots) sampled from the lung atelectasis and tumour regions of a patient with the sample size of 9 (left panel) and 400 (right panel) voxels. In both cases the sample points spanning the N-dimensional space (N is equal to 9 and 400 respectively) are projected to a plane with two conditional dimensions using the Multidimensional Scaling method. The scatterplots represent two replications taken at random.