Context Selective Decision Forests and their application to Lung Segmentation in CT Images

Albert Montillo[1]

[1] GE Global Research Center, Niskayuna, NY, USA montillo@ge.com

Abstract. This work addresses the challenging problem of segmenting the lungs in CT scans. We propose the context selective decision forest (CSDF) as a new discriminative classifier which augments the state of the art entangled decision forest, resulting in higher prediction accuracy and greater generalization for the clinic. Our main contribution is two-fold. First we propose to select the context used for each organ to that context which tends to be present in clinical scans. Second we propose to selectively add labels to our ground truth training data such that the classifier will learn a distinctive appearance and spatial location model for each class. This enables more effective use of context and improves segmentation accuracy. We assess our probabilistic segmentation technique using our labeled database of 110 subjects, and the LOLA11 database of 55 subjects with varying pathology. Quantitative comparisons with state of the art algorithms demonstrate comparable accuracy with superior computational efficiency.

Keywords: selective context, decision forests, CT, lung segmentation.

1 Introduction

Accurate lung segmentation in computed tomography (CT) is an important task that is not well solved. CT tends to be the modality of choice because of its high resolution and good signal to noise ratio. Accurate segmentation enables the detection and quantification of lung abnormalities and lung properties including: interstitial lung disease, emphysema, nodule detection, and lung volume estimation. Due to the importance of this task, it has received significant attention in the research community, however most methods developed to date do not meet the requirements for clinical adoption. Routine clinical use demands a method which: provides a very accurate segmentation, that requires little time to generate (e.g. seconds not hours), works across the variety of CT protocols used in the clinic, handles the variety of pathologies observed in the clinic, and is fully automated thus producing a segmentation which is not dependent on operator skill.

The existing approaches can be broadly grouped into four categories: conventional or rule based methods, atlas segmentation methods, model based methods, and supervised voxel classification methods. Rule based methods, such as [1], apply a sequence of ad hoc heuristics. They typically use one or more fixed thresholds, user selected seed points (e.g. in main bronchi) followed by region growing, connected

component labeling and morphological operations to fill in holes in the segmentation from dense pathology in the otherwise low density lung parenchyma. Due to the simple nature of the rules, conventional methods tend to be fast, requiring only one or two minutes to segment the lungs on a modern desktop, yet they are also brittle. They can fail to segment large portions of the lungs or the lungs entirely [2] depending on how early in the rigid sequence of steps, the anatomy violates the assumptions of the heuristics

Atlas based methods, such as [3], align a manually segmented reference scan to a novel test scan in order to transfer the manual segmentation from the source to the target. These methods tend to employ similarity metrics, such as mutual information, to allow alignment despite slight differences in scan protocol. These methods are less brittle than conventional methods because they nearly always produce some physiologically plausible segmentation, and they tend to include more lung pathology in the lungs. However, when only one atlas is used, the segmentation is biased to the anatomy of the reference subject. To overcome these problems, multi-atlas methods have been introduced, as in [2], which align multiple manually segmented scans to the test scan. Each provides a predicted voxel label which are integrated, for example, by maximum vote. Through bias reduction, the multi-atlas methods tend to outperform single atlas. However, since each atlas is individually registered, the methods can be slow, requiring two hours to segment a single scan. In addition, the atlases do not tend to agree well at lung borders yielding performance there below that of even the conventional methods.

Model based methods fit a probabilistic lung model to the likely lung location. Various methods to construct the lung model have been proposed. In [4], an active shape model is fit to the image, and the final boundary is refined through graph optimization. The method requires the corresponding left or right mean shape model to be manually placed in each individual data set to segment that lung. The authors report that an additional limitation is its reliance on images acquired at total lung capacity. In [5], a probabilistic shape model is constructed via PPCA, and then the model is fit to the image based on the predictions of the voxel classification method whose limitations we uncover and resolve in this paper.

Supervised voxel classification methods use a training database of lung scans with each voxel assigned an anatomical class label, such as left lung or right lung, to build an oracle which predicts the most likely label at each voxel in a novel test scan. In [6], a voxel classifier called the Entangled Decision Forest (EDF) was proposed and applied to segment 12 organs in volumetric CT scans including the lungs. The method is fast, segmenting the CT data in less than one minute on a standard high end desktop. However, the EDF uses semantic context that may not be present in the data at the clinic, in which case performance can be suboptimal.

In this paper, we propose the context selective decision forest (CSDF) which improves the EDF by regulating the use of context. We show how this extension improves test accuracy. The **two main contributions** are as follows. First, to regulate the incorporation of context, we restrict it to that part of the large field of view training scans which are apt to be present in scans observed in the clinic. Since we have the lung labels for the training data, we can do this in an automated fashion. Second, we add new labels to the training data to force the classifier to learn to label

more structures. This decreases the overlap of the learned appearance of neighboring structures, lessoning their confusion and improving overall segmentation accuracy.

2   Materials

2.1 Training and cross-validation database

To train our method and to quantitatively evaluate it via cross-validation, we use a database which consists of 110 large field of view CT scans in which each voxel has an intensity and is manually assigned one of 3 labels {*non-lung, left lung, right lung* }. The database includes a wide range of pathologies as well as healthy subjects. All major scanner vendors and a range of image acquisition protocols including contrast and non-contrast are represented. Also included is a range of ages, weights, heights and both genders.

2.2 LOLA11 database

We also evaluate our method on a separate database called the LOLA 2011 lung segmentation challenge database [7]. This database consists of 55 CT scans containing only the thoracic region. Each voxel has an intensity, however ground truth is not provided. We use this database: (1) to qualitatively evaluate the generalization ability of our method, and (2) to provide an independent, quantitative assessment of its performance as measured by the challenge organizers.

3   Methods

3.1 Decision forest background

We begin with a brief review of randomized decision forests [8,9]. A decision forest is an ensemble of T decision trees. During *training*, the data (Fig. 1), consists of the set of data points from all training images, $S = \{v_i, l_i\}_1^N$. Each data point, s_i, consists of the voxel position, v_i, and its label, l_i. Tree t_i receives the full set S, and its root node selects a test to split S into two subsets to maximize information gain. A test consists of a feature (e.g. an image feature) and a feature response threshold. The left and right child nodes receive their respective subsets of S and the process is repeated at each child node to grow the next level of the tree. Growth stops when one or more stopping criteria, such as minimal information gain or a maximum tree depth occur. Each tree is unique because each tree node selects a random subset of the features and thresholds to try. During *testing*, the data (Fig. 1) consists of the voxel positions in a test image. The voxels are routed to one leaf in each tree by applying the test (selected during training) which is stored in each node. The test is applied to the voxel in the
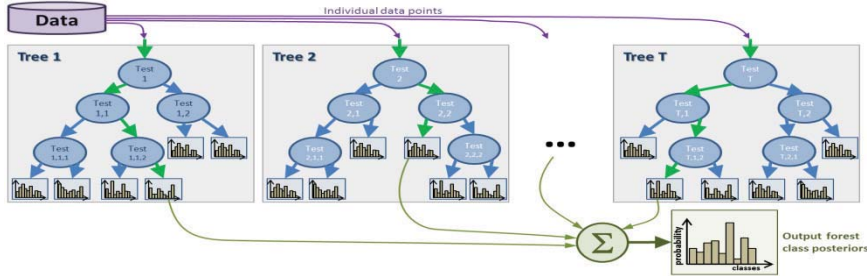
Fig. 1. Decision forest overview. During training, multiple trees are grown, using all training data for each tree. During testing, to classify a voxel, it is initialized at the root node of each tree, and recursively sent left or right (red arrows) according to binary tests stored at each node. The voxel is classified using the average of the T posterior label distributions, with one coming from the leaf reached in each of the T trees.

test image. The test result guides the voxel to the left or right child node, and this is repeated until a leaf node is reached. An empirical distribution over classes learned from the training data is stored at each leaf. The voxel is classified by averaging the class distributions from the set of leaves it reached. The following section describes the features we use to define the node tests of our decision forest.

3.2 Context rich, long-range visual features

It has been shown [10] that to classify a voxel at a given location anatomical context from regions up to 200mm away are often very helpful. Therefore, we do not use traditional features such as Haar wavelets whose range is too short. Instead we construct two types of long-range, context-rich features. The first capture "appearance context", the latter capture "semantic context" [6]. This will be explained next.

Appearance context features. We construct intensity features that can be computed in constant time regardless of size, using an integral image. They are spatially defined by (1) their position, $\mathbf{x}$, centered on the voxel to be labeled (Fig. 2a), and (2) one or two rectangular probe regions, $\mathbf{R_1}$ and $\mathbf{R_2}$, offset from $\mathbf{x}$ by displacements Δ_1 and Δ_2 which can be up to 200mm in each dimension (x,y,z). We construct two categories of intensity features. The first category consists of the mean CT intensity at a probed region, $\mathbf{R_1}$ (Fig 2a, left), while the second consists of the difference in the mean intensity at probed regions, $\mathbf{R_1}$ and $\mathbf{R_2}$ (Fig 2a, right). These are defined as follows:

$$f_{Intensity}\left(\mathbf{x}; \Delta_1, \mathbf{R_1}\right) = \overline{I}\left(\mathbf{R_1}\left(\mathbf{x}+\Delta_1\right)\right) \qquad (1)$$

$$f_{IntensityDiff}\left(\mathbf{x}; \Delta_1, \mathbf{R_1}, \Delta_2, \mathbf{R_2}\right) = \overline{I}\left(\mathbf{R_1}\left(\mathbf{x}+\Delta_1\right)\right) - \overline{I}\left(\mathbf{R_2}\left(\mathbf{x}+\Delta_2\right)\right) \qquad (2)$$

During training, the features to try at each node are parameterized by dimensions of $\mathbf{R_1}$ and $\mathbf{R_2}$, offsets Δ_1 and Δ_2 and an intensity threshold α. These parameters are chosen randomly to define the intensity test: $f(.) > \alpha$. Once training has finished, the
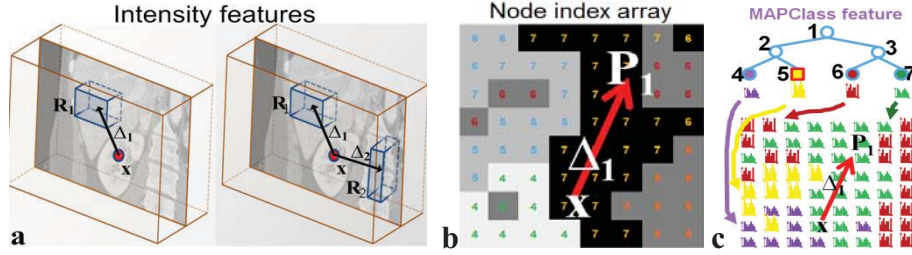
Fig. 2. Intensity and `MAPClass` features. (a) Intensity features measure image information from regions offset from the voxel to classify at **x**. (b) `MAPClass` feature retrieves the label that the classifier currently predicts at location **P₁** offset from **x**. Implementation-wise, we maintain a node index array which associates with each voxel the current tree node ID (represented by the number in each voxel). (c, top) This allows us to determine the current label posterior in the tree for the voxel at location **P₁**. (c, bottom) Conceptually, the tree induces a vector image of class posteriors which we used when developing the `MAPClass` and `TopNClasses` features.

max information gain node test along with its optimal features are frozen and stored within the node for later use during testing.

Semantic context entanglement features. During testing on novel images, we exploit the confident voxel label predictions (peaked distributions) that can be found using just early levels of the forest to aid the labelling of nearby voxels. This provides semantic context similar to auto-context [11,12], but does so within the same forest. We define four types of long range entanglement features to help train the node currently being grown using knowledge learned in already trained split nodes of the forest. Two features (`MAPClass` and `TopNClasses`) are based on the posterior class distribution of the nodes corresponding to probed voxels, and two (`NodeDescendant` and `AncestorNodePair`) are based on the location of the nodes within the trees.

We construct ***MAPClass entanglement features*** which use the maximum a posteriori label of a neighboring voxel at **P₁** in order to reduce uncertainty about the label at **x** (Fig 2b). When such semantic context is helpful to classify the voxel at **x,** the feature yields high information gain and may become the winning feature for the node during tree growth. `MAPClass` tests whether the MAP class in the posterior of a probed voxel $\mathbf{P_1} = \mathbf{x} + \mathbf{\Delta_1}$ is equal to a particular class, C:

$$f_{MAPClass}(\mathbf{x}; \mathbf{\Delta_1}, \mathbf{P_1}, C) = \begin{cases} \arg\max_c p(c; n_{\mathbf{p_1}}) = C & 1 \\ \text{otherwise} & 0 \end{cases} \tag{3}$$

where $p(c; n_{\mathbf{p_1}})$ is the posterior class distribution of the node of **P₁**. This posterior can be retrieved from the tree because we (1) train and test voxels in breadth first fashion and (2) maintain an association between voxels and the tree node ID at which they reside while moving down the tree. This association is a node index array (Fig 2b).

3.3 Context selectivity

In this section we describe our first contribution. Context selectivity is the embedding of prior knowledge about the variability of scan protocols used in the clinic in the training of the classifier. For any organ, O, denote the set of anatomical structures located within a small distance d from O's boundary as A. A subset of the structures in A, denoted as A_{incl}, are apt to be present in clinical scans and should be included as potential sources of context for the classifier to segment O. Meanwhile the residual set, $A_{excl} = \{ A - A_{incl} \}$ though located within the same distance d from O may not always be included in clinical scans and thus should be excluded as a source of context for the classifier to segment O. Therefore even though all of A may be present in a training database, restricting what context is used for each organ to A_{incl} makes the training more generalizable to scans observed in the clinic. There are many ways in which this general, yet powerful, idea can be applied. Consider for example the LOLA11 challenge database in which the lung scans cover only the lungs and few if any structures (e.g. liver , neck) inferior or superior to the lungs. Directional anisotropy in the use of context can be imposed during training by making the context unavailable (constant "zero" feature response) beyond a small margin of slices (~5mm) inferior and superior to the extent of the lungs in the z (through image plane) direction. Such context selectivity is organ dependent and can be imposed on the training because we have the ground truth voxel labels for the training data. The constant zero feature response for unavailable context (A_{excl}) will not yield any information gain and therefore the classifier will choose features from A_{incl} over A_{excl}.

3.4 Add labels to refine context and improve classification

In this section we describe our second contribution. For the EDF, the learned model of an anatomical structure (its signature) consists of the learned appearance and appearance variation throughout the structure, the learned appearance of neighboring structures, and the learned probabilistic shape of the structure [6]. Adding labels to the ground truth forces the classifier to label more classes. Rather than making the classification task harder, it can actually make the classification easier and more accurate when labels are added that split a given label into two or more sub-labels each of which having sharper intensity and location probability distribution functions. This general idea can be applied for lung segmentation. For example Fig. 3a shows a ground truth labeling overlaid on an intensity image. This example shows the original ground truth labeling with 3 labeled "structures": the right lung (green), left lung (blue) and non-lung which is rendered transparently so as to show the underlying intensity image. The non-lung class contains both voxels in the body which are non-lung as well as voxels exterior to the body. Splitting the non-lung label into two labels: "exterior to the body" and "body, non-lung" replaces the non-lung model that has bright and dark intensities and large spatial extent, with two more distinctive
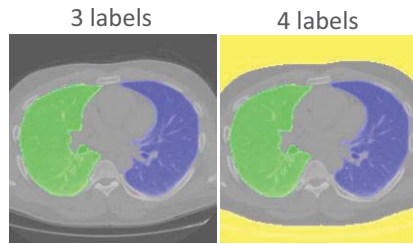
Fig. 3: Ground truth labels in our database. (a) 3 labels: right lung (green), left lung (blue) and non-lung (transparent) and (b) 4 labels: right lung (green), left lung (blue), "exterior to the body" (yellow) and "body, non-lung" (transparent).

classes. The "exterior to the body" class is composed almost entirely of dark pixels located far from the lungs, while "body, non-lung" class is almost entirely bright and near the lungs. For lung segmentation, the additional labels decrease the overlap of the learned appearance of neighboring structures to the lungs: "body, non-lung" is nearby but brighter, "exterior to the body" is similar in appearance but far away. This lessens confusion and improves overall segmentation accuracy.

3.5 Relabeling and our overall algorithm

An additional benefit of adding labels is that it makes error correction simple and efficient. For example when a small number of voxels are labeled "exterior to the body" yet are surrounded by the lung labels (Fig 4, column 3), then these noisy voxel labels can be effectively cleaned up by relabeling them as the closest lung label (Fig 4 column 4). We employ a simple iterative relabeling scheme with a maximum of 20 iterations and stop relabeling when the number of relabeled voxels in an iteration reaches zero or stops decreasing.

The overall steps of our algorithm are as follows:
1) Preprocessing: normalize intensities to Hounsfield units, normalization of one orientation "degree of freedom": upside down or not using image tags
2) Training the context selective decision forest (CSDF) on the training data
3) Applying CSDF to the test data
4) Relabeling

4 Results

4.1 Qualitative results including the impact of each of our contributions

The classifier modifications proposed in this paper enable the classifier to achieve a visually accurate segmentation of organs throughout the 55 test volumes in
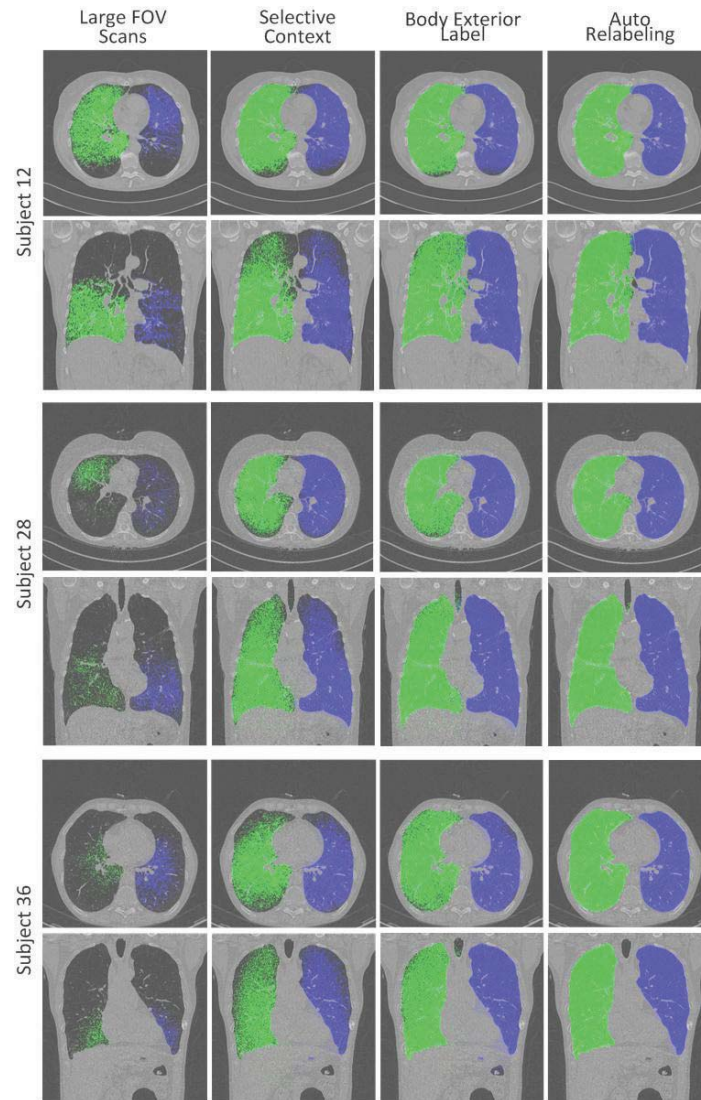
Fig. 4. Effect of our contributions on the LOLA11 database. (Columns 1,2) Training with selective context (column 2) provides marked improvement over free ranging context (column 1) when scans have limited FOV. (Columns 2,3) Training with the body exterior labeled separately (column 3) allows the classifier to segment a much greater extent of the lungs than when exterior and body share the same label (column 2). (Columns 3,4) Relabeling exterior labels in the body as the closest lung improves the final segmentation (column 4).

the LOLA11 database. When we train an EDF classifier on the 110 subjects in our own database using unregulated context, the classifier chooses context which tends to

be beyond the range of the FOV of the scans in the LOLA11 database. For example, appearance and semantic context from structures superior to the lungs (parts of the neck and head) and inferior to the lungs (parts of the liver and spleen) are used. By restricting the slice range of the scan that the classifier may use for context in each training scan, the classifier instead chooses context in the center of the thoracic region (heart), anterior (chest wall) and posterior (back, spine). This context is readily available in both our database and in the LOLA11 database. As a result, the performance in the LOLA11 database is greatly improved. In Fig. 4, column 1 shows the decidedly poor segmentation using unregulated context; while column 2 shows the performance improvement for 3 different subjects. While this performance is better there are still regions where lungs segmentation can be improved (e.g. superior lung subject 12). Splitting the non-lung class into "body non-lung" and "body exterior" provides yet another level of marked performance improvement. This can be seen by comparing the segmentations in column 2 and 3. The most similar appearing label to the lungs is the air filled "body exterior", but the classifier learns it has a spatially distinct location from the lungs. Our last contribution is to re-label the few "exterior non-lung" voxels that are connected to a lung as the closest lung. Column 4 shows the accurate segmentation result after relabeling.

As a segmentation challenge database, LOLA11 contains a wide range of pathologies that can cause many other segmentation algorithms difficulties. Our method handles the challenges well. Fig. 5 illustrates how the method accurately segments lungs despite large variations in shape and scale, including gross lung deformity (Subject 5), underdeveloped lungs or pathology affecting one lung (Subjects 44, 20), and very large lungs (Subject 37). Fig. 6 shows how the CSDF also segments several types of dense pathology in the lung properly, including diffusive dense pathology (yellow circled region in Subject 31). In (Subject 24), a small amount of dense pathology is omitted but could be included using standard techniques, such morphological operations for 3D hole filling and costal surface convexity detection and filling via convex hull as presented in [2].

4.2 Quantitative results

For a quantitative analysis, first, we measured the CSDF segmentation accuracy across our 110 subject labeled database using 5 fold cross validation and the overlap percentage, which is 100 times the average class Jaccard similarity coefficient [13]. The metric is the ratio of the intersection size (of ground truth and predicted labels) divided by the size of their union. This metric has been used widely in the literature and can be mathematically converted into the Dice coefficient.

Table 1 shows the overlap percentage for the left and right lungs for four variations of the CSDF method. The first row shows the accuracy when we restrict context by using only appearance context but not semantic context features (section 3.2). In the second row, we see the improvement for both lungs when we allow semantic context (entanglement) features. In the third row, we add the "body exterior" label and
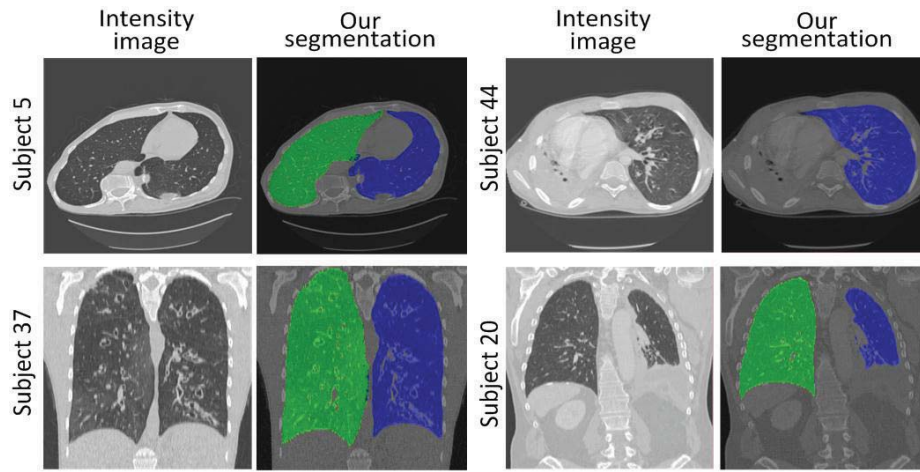
Fig. 5. Our method handles large variations in shape and scale. (LOLA11 Subject 5) Accurate lung segmentation despite gross lung deformity. (Subjects 44, 20) Lungs are properly segmented even when one lung is underdeveloped, while our method also automatically handles very large lungs as in (Subject 37).
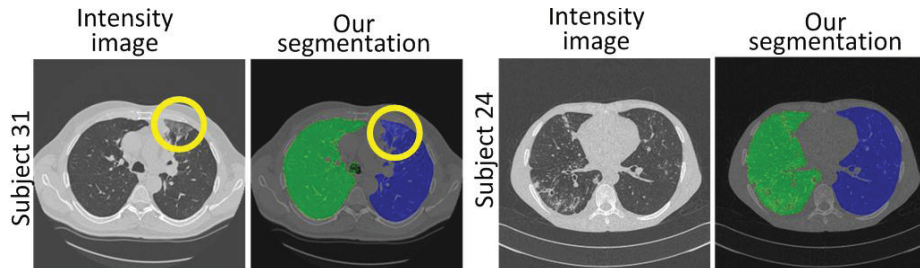
Fig. 6. Several types of dense pathology in the lung are properly segmented. (LOLA11 Subject 31) Diffusive dense pathology (yellow circled region) is accurately included in the lung. (Subject 24) Some dense pathology is omitted but could be included using standard techniques.

observe another improvement in the overlap percentages for both lungs. Finally, we modify the CSDF by training at 4x downsampling rather than 8x (typical setting). This improves the overlap to 94.6% (right lung) and 94.0% (left lung)..

Second, our segmentations on the 55 subjects in the LOLA11 Challenge were independently evaluated by the challenge organizers. Our method achieved a similarly high overlap of 95.1% (right lung) and 95.2% (left lung) as shown in Table 2.

4.3 Efficiency considerations

The parallel implementation of our voxel classification based method segments both lungs simultaneously in **43 seconds** per volume, where a typical volume is

| Experiment | Jaccard Overlap Percentage | |
	Right Lung	Left Lung
No Body Exterior label No semantic context features	93.70	92.76
No Body Exterior label With semantic context features	93.74	92.94
With Body Exterior label With semantic context features	94.17	93.17
With Body Exterior label With semantic context features Training at 4x	**94.56**	**94.02**

Table 1. Lung segmentation results for 4 experiments from 5 fold cross-validation in our 110 subject database. Best accuracy in bold. (rows 1,2) Entanglement [6] improves segmentation accuracy. (Rows 2,3) Adding the body exterior label further improves accuracy over entanglement. (Rows 3,4) Training at 4x downsampling rather than 8x further improves accuracy for both left and right lungs.

Organ	mean	SD	min	Q1	median	Q3	Max
L lung	**0.952**	0.117	0.116	0.965	0.974	0.978	0.987
R Lung	**0.951**	0.132	0	0.964	0.974	0.977	0.987

Table 2. Results of lung segmentation for the 55 scans in LOLA11. Accuracy is measured using the Jaccard overlap (multiply by 100 for %). Overall score is 0.952

256x256x250 (after 2x downsampling in each dimension) using a standard Intel Xeon 2.4GHz computer (8 core) with 16GB RAM running Win7 x64. A very good, coarse labeling (after 8x downsampling) can be achieved in **1.7 seconds**. Training on the 110 volumes, which need only be done once, requires 8 hours for 8 trees to depth 21.

5 Discussion

Our fully automated quantitative results in 43 seconds with an overlap percentage in excess of 94% compare favorably with those reported in the literature. [5] reports an overlap percentage of 74% (Dice index of 0.85). The atlas method in [3] yields an overlap percentage of 82%, while the region growing method in [1] yields 88.5%. In [4] the model based method yields an overlap accuracy of 94% (Dice 0.97) on a database with 10 subjects. However this method requires user input to position the left and right lung model on each individual image. Run time is roughly 4 minutes to segment both lungs. A hybrid conventional+multi-atlas method [2] yielded 95% overlap however the method's run time is quite variable requiring between 1 minute and 2 hours per subject using 2x downsampled volumes.

Currently a limitation of the CSDF is suboptimal performance on small diameter DFOV reconstructed images. A potential solution is to re-use our selective context idea (section 3) to limit the context used by the decision forest in the training data to exclude regions outside the cylinder that just encloses the lungs in the training data, because these regions are not reconstructed in small diameter DFOV images.

6 Conclusions

In conclusion, in [6] we showed that using semantic context improves segmentation. Here we show that regulated use of context provides even better results. When the use of context is regulated, such as when it is restricted to that part of the training scans which are apt to be present in scans observed in the clinic, then the classification results improve dramatically as they did for the LOLA11 database (Fig. 4). We also show how adding labels that force the classifier to label more structures can decrease the overlap of the learned appearance of neighboring structures, lessen their confusion and thereby improve segmentation accuracy. Lastly, we showed that for lung segmentation, relabeling is a simple but effective approach for improving lung segmentation (Fig. 4).

References

1. Sun, X., Zhang, H., Duan, H.: 3D computerized segmentation of lung volume with computed tomography, Acad Radiol, 13(6), 670-7 (2006)
2. van Rikxoort, E., de Hoop, B., Viergever, M., Prokop, M., van Ginneken, B.: Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Med Phys 36, 2934–47 (2009)
3. Sluimer, I., Prokop, M., van Ginneken, B.: Towards automated segmentation of the pathological lung in CT, IEEE Trans Med Imaging, 24(8) 1025-38 (2005)
4. Sun, S., McLennan, G., Hoffman, E., Beichel, R.: Model-Based Segmentation of Pathological Lungs in Volumetric CT Data, In: MICCAI-PIA Workshop (2010)
5. Iglesias, J., Konukoglu, E., Montillo, A., Tu, Z., Criminisi, A.: Combining Generative & Discriminative Models for Semantic Segmentation of CT Scans via Active Learning. In: Proc. of Info. Proc. in Medical Imaging, (2011)
6. Montillo, A., Shotton, J., Winn, J., Iglesias, J., Metaxas, D., Criminisi, A.: Entangled decision forests and their application for semantic segmentation of CT images. In: Proc. of Info. Proc. in Medical Imaging, (2011)
7. LOLA11 DATABASE, http://lola11.com (2011)
8. Amit. Y., and Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation, 9(7):1545–88 (1997)
9. Breiman, L.: Random Forests. Machine Learning, 45(1):5–32 (2001)
10. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E. Regression Forests for Efficient Anatomy Detection and Localization in CT Scans, In: MICCAI-MCV Workshop (2010)
11. Shotton, J., Johnson, M., and Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: Proc. of CVPR, 1-8. (2008)
12. Tu, Z., and Bai, X.: Auto-context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 32(10), 1744-57. (2010)
13. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. Int. J. Comp. Vision, 88(2), 303-38 (2010)