Comparative analysis of semantic localization accuracies between adult and pediatric DICOM CT images.

Duncan Robertson[a], Sayan D. Pathak[b], Antonio Criminisi[a], Steve White[b], David Haynor[c], Oliver Chen[b] and Khan Siddiqui[b]

[a]Microsoft Research Labs, JJ Thomson Ave, Cambridge, Cambridgeshire, UK CB3 0FB
[b]Microsoft Health Solutions Group R&D, 1 Microsoft Way, Redmond WA, USA 98052
[c]Dept. of Radiology, University of Washington, Seattle WA, USA 98195

ABSTRACT

Existing literature describes a variety of techniques for semantic annotation of DICOM CT images, *i.e.* the automatic detection and localization of anatomical structures. Semantic annotation facilitates enhanced image navigation, linkage of DICOM image content and non-image clinical data, content-based image retrieval, and image registration. A key challenge for semantic annotation algorithms is inter-patient variability. However, while the algorithms described in published literature have been shown to cope adequately with the variability in test sets comprising adult CT scans, the problem presented by the even greater variability in pediatric anatomy has received very little attention. Most existing semantic annotation algorithms can only be extended to work on scans of both adult and pediatric patients by adapting parameters heuristically in light of patient size. In contrast, our approach, which uses *random regression forests* ('RRF'), learns an implicit model of scale variation automatically using training data. In consequence, anatomical structures can be localized accurately in both adult and pediatric CT studies without the need for parameter adaptation or additional information about patient scale. We show how the RRF algorithm is able to learn scale invariance from a combined training set containing a mixture of pediatric and adult scans. Resulting localization accuracy for both adult and pediatric data remains comparable with that obtained using RRFs trained and tested using only adult data.

Keywords: DICOM, RADLEX, Semantic, Tagging, Classification, Pediatrics

1. INTRODUCTION

Improving productivity in healthcare depends increasingly on technological innovation, with medical informatics playing an important role in improving the efficiency of patient care. In our previous work, we showed that the *random regression forest* (RRF) algorithm can be used automatically to detect and localize anatomical structures in DICOM CT images, which considerably facilitates efficient image navigation within our radiological image viewing software.[1,2] Other efficiency-driven applications include (i) the automatic linkage of DICOM image content and non-image clinical data,[3] (ii) content based image retrieval (where semantic image labels can be used to increase the proportion of relevant search results) and (iii) image registration, which is also greatly enhanced using these labels as priors.[4] While many of the authors who have described applications for the automated analysis of medical images have focused exclusively on adult anatomy,[2,5] considerable benefit could also be derived from automated analysis of pediatric CT scans. However, achieving robustness to large changes in scale is hard. Consequently, semantic annotation techniques that are well adapted for adult anatomy may perform badly on pediatric data without heuristic adaptation of parameters in light of patient size. For example, popular image annotation approaches may involve the registration of size-specific atlases,[6] the application of a size-specific sequence of filters/classifiers,[5,7,8] or modeling of the scale variation in a multi-scale representation often involving empirically tuned models to deal with scale variation between adult and pediatric anatomies. Ideally, an algorithm for automatic semantic annotation of CT images should be able to localize anatomical structures without significant variation in accuracy irrespective of whether the images are from adult or child, provided comparable anatomical entities are present in the patients (which is true for most of the human anatomy

Further author information: (Send correspondence to S. P.)
S.P.: E-mail: sayanpa@microsoft.com, Telephone: 1 425 538 7386

after birth). No additional parameter adaptation or additional information about patient size should be required. This paper uses a multivariate RRF algorithm for efficient, automatic detection and localization of anatomical structures within DICOM CT scans of both adult and pediatric patients. Regression forests are similar to the better known classification forests but are trained to predict continuous outputs, *e.g.* the positions of the faces of bounding boxes associated with the anatomical structures of interest. This paper shows that an RRF trained on adult data performs well on adult data but gives significantly less accurate localization in the pediatric case. However, the RRFs ability to learn from data enables it to perform equally well for adult and pediatric data when the training set is extended to include representative pediatric data. We show that the RRF is capable of learning an implicit model of scale variation directly from training data.

Outline. Section 2 summarizes our RRF-based anatomy bounding box detection algorithm. Section 3 introduces the error measures used for bounding box aided navigational efficacy evaluation. Section 4 describes our evaluation of the robustness of the organ detection algorithm and its ability to enable automated image navigation. Finally we summarize key insights in section 5.

2. ALGORITHM: HIERARCHICAL REGRESSION FOR ORGAN LOCALIZATION

This section briefly summarizes our algorithm for the automatic localization of anatomical structures in volumetric CT scans. For a full explanation please refer to Criminisi et. al.[2, 9]

Mathematical notation. Vectors are represented in boldface (*e.g.* $\mathbf{v}$), matrices as teletype capitals (*e.g.* $\mathtt{\Lambda}$) and sets in calligraphic style (*e.g.* $\mathcal{S}$). The position of a voxel in a CT volume is denoted $\mathbf{v} = (v_x, v_y, v_z)$.

The labeled database. The anatomical structures we wish to train the RRF to recognize are $\mathcal{C}$ ={ `abdomen, heart, left kidney, right kidney, liver, left lung, right lung, spleen, thorax`}. We are given a database of DICOM CT scans that have been manually annotated with 3D bounding boxes tightly drawn around the structures of interest (see fig. 1a). The bounding box for an organ $c \in \mathcal{C}$ is parameterized as a 6-vector $\mathbf{b}_c = (b_c^{\mathtt{L}}, b_c^{\mathtt{R}}, b_c^{\mathtt{A}}, b_c^{\mathtt{P}}, b_c^{\mathtt{H}}, b_c^{\mathtt{F}})$ where each element represents the position (in mm) of one axis-aligned face*. The scans exhibit large variability in image cropping, resolution, scanner type, and use of contrast agents, and the patients have a wide variety of medical conditions and body shapes (see fig. 2). Additionally, the database includes pediatric patients exhibiting considerable variation in size (see fig. 3) . Images are not pre-registered or normalized in any way. The goal is to localize anatomic structures of interest accurately and automatically, despite such large variability.

2.1 Problem parameterization and regression forest learning

Key to our algorithm is the idea that *all* voxels in a test CT volume contribute with varying confidence to estimating the position of the position of *all* anatomical structures' bounding boxes (see fig. 1b,c). Intuitively, some distinct voxel clusters (*e.g.* ribs or vertebrae) may predict the position of an organ (*e.g.* the heart) with high confidence. Thus, at detection time, those clusters should be used as landmarks for the localization of those structures. Our aim is to learn to cluster voxels together based on their appearance, their spatial context and their confidence in predicting position and size of all anatomical structures. We tackle this simultaneous feature selection and parameter regression task with a multi-class random regression forest (see fig. 4), *i.e.* an ensemble of regression trees trained to predict the location and size of all structures simultaneously.

*Superscripts follow standard radiological orientation convention: `L` = left, `R` = right, `A` = anterior, `P` = posterior, `H` = head, `F` = foot.

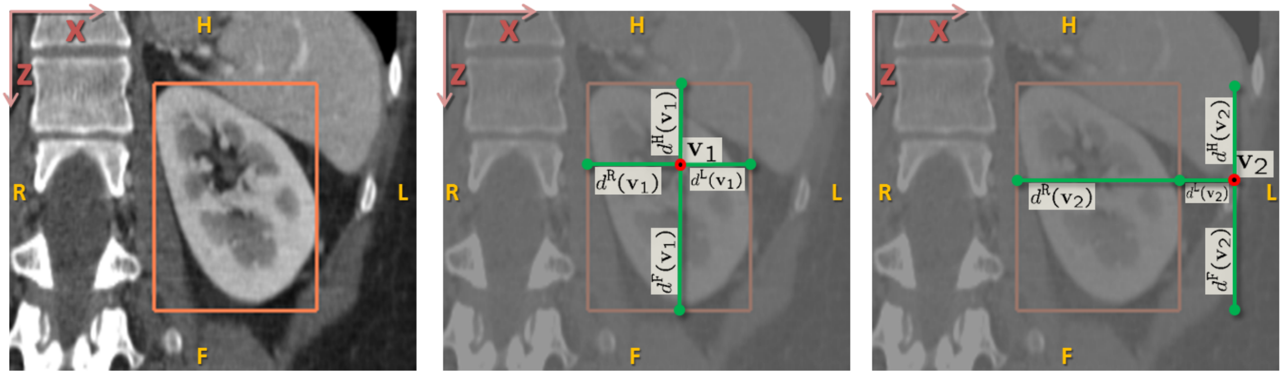
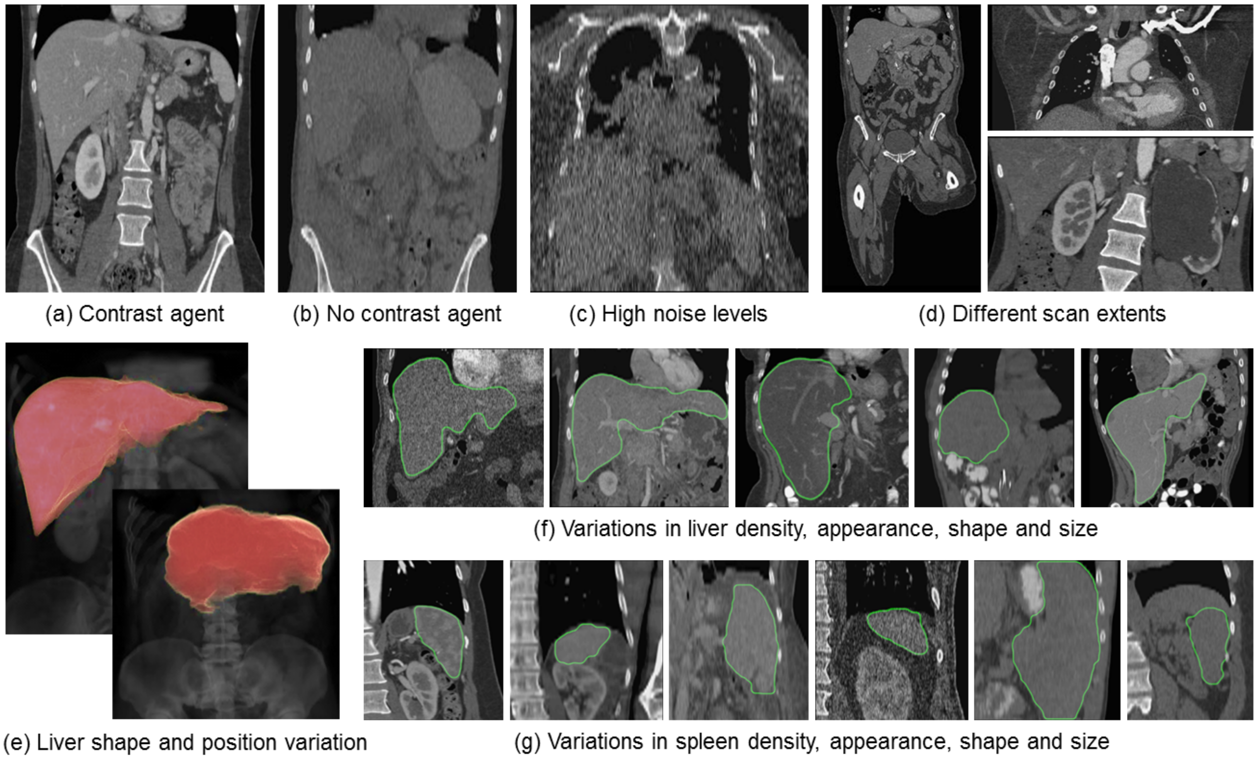

X
Z
H
R
L
F
$\mathbf{v}_1$
$d^{\mathrm{H}}(\mathbf{v}_1)$
$d^{\mathrm{R}}(\mathbf{v}_1)$
$d^{\mathrm{L}}(\mathbf{v}_1)$
$d^{\mathrm{F}}(\mathbf{v}_1)$
$\mathbf{v}_2$
$d^{\mathrm{H}}(\mathbf{v}_2)$
$d^{\mathrm{R}}(\mathbf{v}_2)$
$d^{\mathrm{L}}(\mathbf{v}_2)$
$d^{\mathrm{F}}(\mathbf{v}_2)$
(a)
(b)
(c)

Figure 1. **Problem parameterization. (a)** A coronal view of a left kidney and the associated ground-truth bounding box (in orange). **(b,c)** *Every* voxel $\mathbf{v}_i$ in the volume votes for the position of the six walls of each organ's 3D bounding box via 6 relative, offset displacements $d^{\mathrm{k}}(\mathbf{v}_i)$ in the three orthogonal planes along x, y and z axes.


(a) Contrast agent

(b) No contrast agent

(c) High noise levels

(d) Different scan extents

(f) Variations in liver density, appearance, shape and size

(e) Liver shape and position variation

(g) Variations in spleen density, appearance, shape and size

Figure 2. **Variability in our labeled database. (a,b,c)** Variability in appearance due to presence of contrast agent, or noise. **(d)** Difference in image geometry due to acquisition parameters and possible anomalies. **(e)** Volumetric renderings of liver and spine to illustrate large changes in their relative position and in the liver shape. **(f,g)** Mid-coronal views of liver and spleen across different scans in our database to illustrate their variability. All views are metrically and photometrically normalized to aid comparison.

2.1.1 Forest training

The training process constructs multiple regression trees and decides at each node how to best split the incoming voxels. We are given a subset of labeled CT volumes (the training set), and the associated ground-truth organ bounding boxes (fig. 1a). The size of the forest T is fixed and all trees are trained in parallel. Each voxel is pushed through each of the trees starting at the root. Each split node applies the following binary test $\xi_j > f(\mathbf{v};\boldsymbol{\theta}_j) > \tau_j$

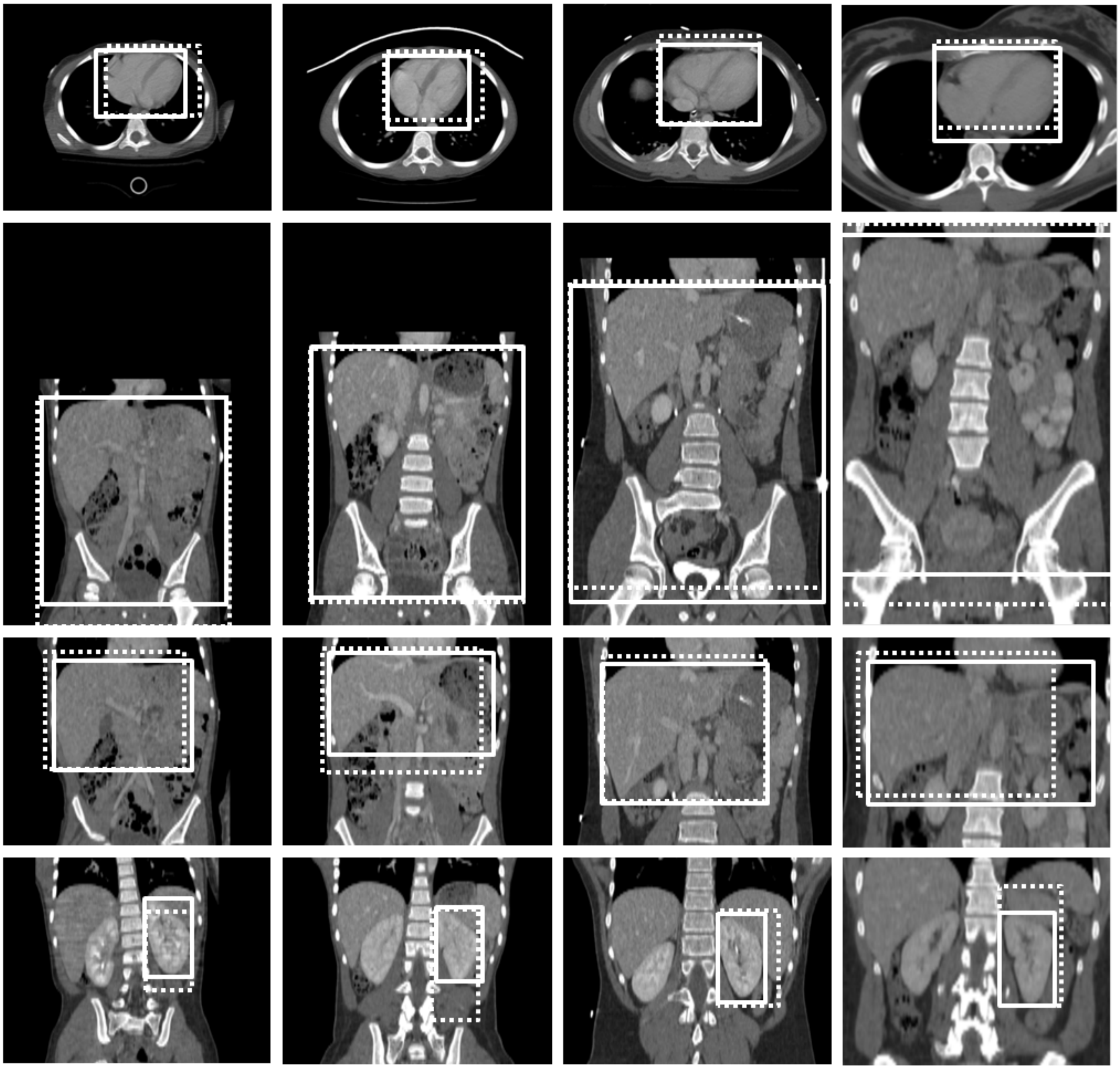


Figure 3. **Variability in organ scales our labeled database across age groups.** The columns correspond to scans of (i) a 2-5 year old patient, (ii) a 6-11 year old patient, (iii) 12-17 year old patient, and (iv) an adult patient. Rows contain (i) an axial slice intersecting the center of the manually annotated bounding box for the heart, and coronal slices intersecting the center of the manually annotated bounding boxes for the (ii) abdomen, (iii) liver, and (iv) left kidney. Manually annotated bounding boxes are shown using solid lines; those detected automatically using a single RRF trained using a combination of adult and pediatric data are shown using dashed lines. All views are metrically and photometrically normalized to aid comparison.

and based on the result sends the voxel to the left (if $f(.)$ falls between the two thresholds) or right child node. $f(.)$ denotes the feature response computed for the voxel $\mathbf{v}$. The parameters $\boldsymbol{\theta}_j$ represent the visual feature which applies to the j^{th} node. Our visual features are similar to those in,[6, 7, 9] *i.e.* they are mean intensity or intensity differences over displaced, asymmetric cuboidal regions. These features are efficient and capture spatial context. The feature response is $f(\mathbf{v}; \boldsymbol{\theta}_j) = |F_1|^{-1} \sum_{\mathbf{q} \in F_1} I(\mathbf{q}) - |F_2|^{-1} \sum_{\mathbf{q} \in F_2} I(\mathbf{q})$; with F_i indicating 3D box regions and

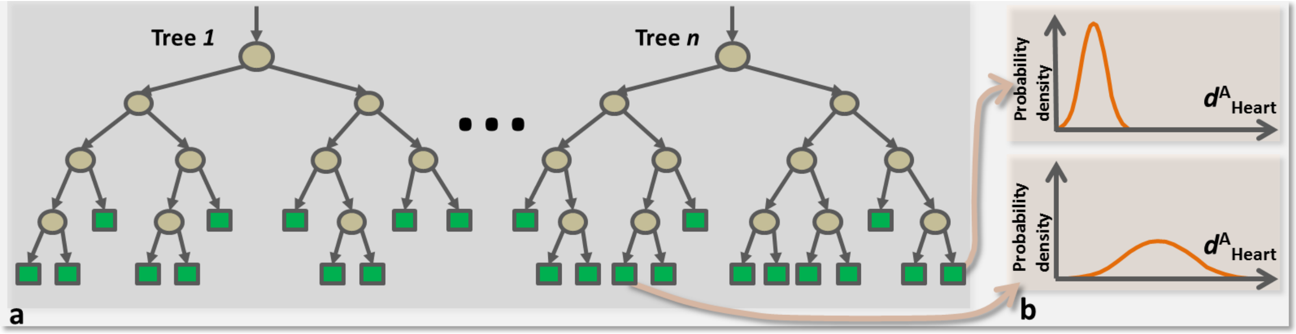


Figure 4. **A regression forest** is an ensemble of different regression trees. Each leaf contains a distribution for the continuous output variable/s. Leaves have associated different degrees of confidence (illustrated by the "peakiness" of distributions). During testing each text voxel is "pushed" through each tree starting at the root until it reaches a leaf node. The corresponding prediction is read at the leaves.

I the intensity. F_2 can be the empty set for unary features. Randomness is injected by making only a random subset of all features available at each node. This technique has been shown to increase the generalization of tree-based predictors.[8] Next we discuss how to optimize each node.

Node optimization. Each voxel $\mathbf{v}$ in each training volume is associated with an offset $\mathbf{d}_c(\mathbf{v})$ with respect to the bounding box $\mathbf{b}_c$ for each class $c \in \mathcal{C}$ (see fig. 1b,c). Such offset is denoted: $\mathbf{d}_c(\mathbf{v}) = (d_c^{\mathtt{L}}, d_c^{\mathtt{R}}, d_c^{\mathtt{A}}, d_c^{\mathtt{P}}, d_c^{\mathtt{H}}, d_c^{\mathtt{F}}) \in R^6$, with $\mathbf{b}_c(\mathbf{v}) = \hat{\mathbf{v}} - \mathbf{d}_c(\mathbf{v})$ and $\hat{\mathbf{v}} = (v_x, v_x, v_y, v_y, v_z, v_z)$. As with training classification trees, node optimization is driven by maximizing an information gain measure, defined as: $IG = H(\mathcal{S}) - \sum_{i=\{\mathtt{L},\mathtt{R}\}} \omega_i H(\mathcal{S}_i)$ where H denotes entropy, $\mathcal{S}$ is the set of training points reaching a node and $\mathtt{L}, \mathtt{R}$ denote the left and right children and $\omega_i = |\mathcal{S}_i|/|\mathcal{S}|$. In classification problems, the entropy is defined over distributions of discrete class labels. In regression instead we measure the purity of the probability density of the real-valued predictions. For a single class c we model the distribution of the vector $\mathbf{d}_c$ at each node as a multivariate Gaussian; *i.e.* $p(\mathbf{d}_c) = \mathcal{N}(\mathbf{d}_c; \overline{\mathbf{d}_c}, \Lambda_c)$, with the matrix Λ_c encoding the covariance of $\mathbf{d}_c$ for all points in $\mathcal{S}$. The differential entropy of a multivariate Gaussian can be shown to be $H(\mathcal{S}) = \frac{n}{2}(1 + \log(2\pi)) + \frac{1}{2}\log|\Lambda_c(\mathcal{S})|$ with n the number of dimensions ($n = 6$ in our case). Algebraic manipulation yields the following regression information gain: $IG = \log|\Lambda_c(\mathcal{S})| - \sum_{i=\{\mathtt{L},\mathtt{R}\}} \omega_i \log|\Lambda_c(\mathcal{S}_i)|$. In order to handle simultaneously all $|\mathcal{C}| = 9$ anatomical structures the information gain is adapted to: $IG = \sum_{c\in\mathcal{C}} \left(\log|\Lambda_c(\mathcal{S})| - \sum_{i=\{\mathtt{L},\mathtt{R}\}} \omega_i \log|\Lambda_c(\mathcal{S}_i)|\right)$ which is readily rewritten as

$$IG = \log|\Gamma(\mathcal{S})| - \sum_{i=\{\mathtt{L},\mathtt{R}\}} \omega_i \log|\Gamma(\mathcal{S}_i)|, \quad \text{with } \Gamma = \text{diag}\left(\Lambda_1, \cdots, \Lambda_c, \cdots, \Lambda_{|\mathcal{C}|}\right). \tag{1}$$

Maximizing (1) encourages minimizing the determinant of the $6|\mathcal{C}| \times 6|\mathcal{C}|$ covariance matrix Γ, thus decreasing the uncertainty in the probabilistic vote cast by each cluster of voxels on each organ pose. Node growing stops when IG is below a fixed threshold, too few points reach the node or a maximum tree depth D is reached (here $D = 7$). After training, the j^{th} split node remains associated with the feature $\boldsymbol{\theta}_j$ and thresholds ξ_j, τ_j. At each leaf node we store the learned mean $\overline{\mathbf{d}}$ (with $\mathbf{d} = (\mathbf{d}_1, \cdots, \mathbf{d}_c, \cdots, \mathbf{d}_{|\mathcal{C}|})$) and covariance Γ (fig. 4b).

2.2 Forest testing

Given a previously unseen CT volume $\mathcal{V}$, test voxels are sampled in the same manner as at training time. Each test voxel $\mathbf{v} \in \mathcal{V}$ is pushed through each tree starting at the root and the corresponding sequence of tests applied. The voxel stops when it reaches its leaf node $l(\mathbf{v})$, with l indexing leaves across the whole forest. The stored distribution $p(\mathbf{d}_c|l)$ for class c also defines the posterior for the absolute bounding box position: $p(\mathbf{b}_c|l)$ since $\overline{\mathbf{b}}_c(v) = \hat{\mathbf{v}} - \overline{\mathbf{d}}_c(v)$. The posterior probability for $\mathbf{b}_c$ is now given by

$$p(\mathbf{b}_c) = \sum_{t=0}^{T} \sum_{l \in \tilde{\mathcal{L}_\sqcup}} p(\mathbf{b}_c|l)p(l) \tag{2}$$

$\tilde{\mathcal{L}}_{\sqcup}$ is a subset of the leaves of tree t. We select $\tilde{\mathcal{L}}_{\sqcup}$ as the set of leaves which have the smallest uncertainty (for each class c) and contain 75% of all test voxels. Finally $p(l)$ is simply the proportion of voxels arriving at leaf l.

Organ localization. The final prediction $\tilde{\mathbf{b}_c}$ for the absolute position of the c^{th} organ is given by:

$$\tilde{\mathbf{b}_c} = \arg\max\, p(\mathbf{b}_c) \tag{3}$$

Under the assumption of uncorrelated output predictions for bounding box faces, it is convenient to represent the posterior probability $p(\mathbf{b}_c)$ as six 1D histograms, one per face. We aggregate evidence into these histograms from the leaf distributions $p(\mathbf{b}_c|l)$. Then $\tilde{\mathbf{b}_c}$ is determined by finding the histogram maxima. Furthermore, we can derive a measure of the confidence of this prediction by fitting a 6D Gaussian with diagonal covariance matrix $\tilde{\Lambda}$ to the histograms in the vicinity of $\tilde{\mathbf{b}_c}$. A useful measure of the confidence of the prediction is then given by $|\tilde{\Lambda}|^{-1/2}$.

Organ detection. An organ is declared present in the scan if the prediction confidence is greater than β. The parameter β is tuned to achieve the desired trade-off between the relative proportions of false positive and the false negative detections.

3. VALIDATION AND VERIFICATION

To facilitate evaluation of the localization accuracy of the RRF algorithm, we use two different validation measures. These are described below:

Measure 1: Bounding wall prediction error. The output of our semantic labeling algorithm is a set of predicted bounding box locations. In many applications, the detected bounding box is used to localize an image sub-volume where the organ of interest is likely to be located. Hence, we compare the positions of detected bounding box faces with ground truth. Fig. 5(a) illustrates the errors associated with a detected bounding box for an illustrative 2D detection example (normally we use 3D data). Here error is defined as the absolute difference between predicted and annotated (ground truth) face positions. In validation we use a set $\mathcal{T}$ of CT scans (independent of the set used to train the system) which have, as for the training set, a set of ground truth organ bounding boxes labels. For each CT scan $t \in \mathcal{T}$ we test the results of the RRF bounding boxes $\tilde{\mathbf{b}}_{t,c}$ against the scan's ground truth data $\mathbf{g}_{t,c}$.

$$\mathbf{e}_c = \frac{1}{|\mathcal{T}|} \sum_{t\in\mathcal{T}} |\tilde{\mathbf{b}}_{t,c} - \mathbf{g}_{t,c}| \tag{4}$$

where $\mathbf{e}_c$ represents the 6 component mean absolute error vector derived from all images in a test set. A standard deviation measure can be arrived at using the same error measure $\boldsymbol{\epsilon}_c = \tilde{\mathbf{b}}_{t,c} - \mathbf{g}_{t,c}$

Measure 2: Centroid-hit error. An important use case for our algorithm is as a navigational assistance tool for use in radiological image viewing software. When the user wishes to navigate to a certain anatomical structure in a CT scan, the application performs a Multi-Planar Rendering (MPR) of the image volume with the three cross-sectional planes centered at the centroid of the detected bounding box. To determine whether the MPR views contain the selected structure, we test to see whether the centroid of the detected bounding box falls within the ground-truth bounding box (schematically represented by fig. 5(b)). However, we expect that in some cases the detected centroid may lie outside the ground truth box. Fig. 5(c) shows one such situation where the detected box is taller compared to the ground-truth bounding box. This leads to an error in the prediction along the vertical dimension even though the horizontal prediction falls within the ground-truth box. User testing indicates that when two of the three centroid coordinates fall within the true bounding box bounds, the navigational assistance tool is still beneficial to productivity. Therefore our centroid hit error test measures the percentage of detected structures for which 2 or 3 of the centroid coordinates fall within the ground truth bounding box bounds.

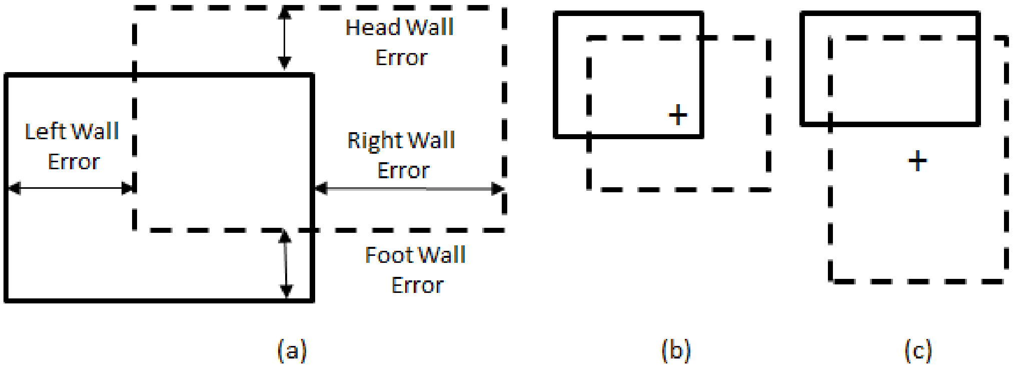


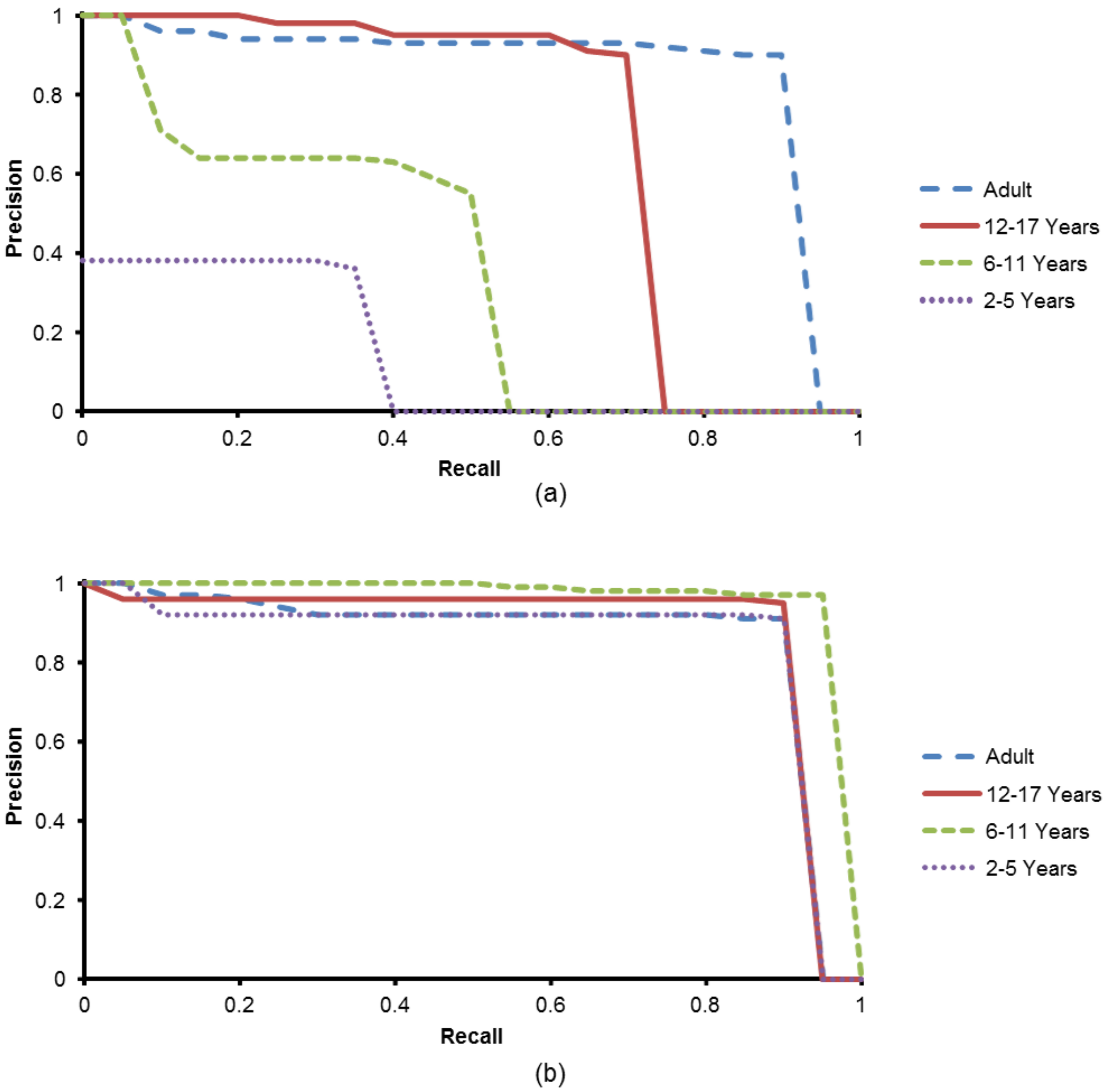
Figure 5. **Error measures.** **(a)** 2-D schematic depiction of the 4 errors associated with the position of each wall in the predicted bounding box (dotted line) as compared to ground-truth box (solid). **(b)** Centroid of the predicted bounding box falls inside the ground-truth bounding box. **(c)** Centroid of the predicted bounding box falls outside the ground-truth bounding box (solid).

4. RESULTS AND DISCUSSION

In this section, we demonstrate that an RRF can learn invariance to the scale variability (and other forms of variability) that is inherent in pediatric CT data. Our experimental dataset comprises 120 adult and 118 pediatric DICOM CT scans. The 3D bounding boxes of the anatomical structures of interest have been annotated manually in each scan. Both the adult and pediatric sets were randomly partitioned into training and test subsets, with one third of the scans being used for test. Since there is a relatively large variation in organ scale in pediatric population, we further sub-divide the pediatric test set into three sub-categories based on the age groups used for school enrollment in the United States: 2-5 years, 6-11 years, and 12-17 years. In what follows, we compare the localization accuracy of RRFs trained using three different combinations of training data: (i) adult only, (ii) pediatric only, and (iii) adult + pediatric data in combination. Each RRF was trained according to the procedure described previously and comprised 5 trees (provided there were more than three trees, results were found to be not very sensitive to the number of trees).

4.1 Precision recall

A useful means of characterizing the localization performance of our semantic labeling algorithm is to plot *precision-recall* curves. In this context, *precision* refers to the proportion of organs that were correctly detected, and *recall* refers to proportion of reported detections that were correct. Here, a correct detection is considered to be a detection for which the centroid of the predicted organ bounding box is contained by the ground truth bounding box. The plot shows how these quantities vary as the detection confidence threshold β is varied. As a first step in demonstrating that RRF can learn scale invariance, we measure the performance of an RRF trained using only adult data. Fig. 6a shows resulting precision-recall curves for both adult and pediatric test sets. As expected for a regression forest trained on adult data, performance for adult test data is good: average precision remains high until a recall value of approximately 0.9 is reached and the area under the curve is close to 1. In contrast, performance on pediatric data is much worse. The implication is that a significant component of the variability in the pediatric data was not modeled by a regression forest trained using only adult data. However, it is interesting to note that the area under the curve increases with the age of the patients in the test set. This is as expected, since older pediatric patients are more similar to the adults in the training set. For comparison, Fig. 6b shows corresponding precision-recall curves obtained using an RRF trained using the combination of adult + pediatric training data. Now the area under all curves is close to 1. This RRF gives good performance for both pediatric and adult test data, and performance for the pediatric data is significantly improved. This demonstrates that our approach is able to learn scale invariance effectively from training data.



(a)

(b)

Figure 6. **Precision-Recall curves for the 4 population groups using (a) adult and (b) combination adult and pediatric training data**. The curves show how precision and recall change as the detection confidence threshold is varied. By extending the training set to include pediatric scans, localization accuracy for pediatric test data has been significantly improved without a significant reduction in localization accuracy for adult data.

4.2 Accuracy evaluation

Table 1 shows mean localization error for RRFs trained using the three training sets. Here mean absolute error is computed across all bounding box faces and all detected organ instances. However, so that results can be compared meaningfully across multiple combinations of training and test sets, we report results obtained after tuning the detection confidence threshold parameter β tuned to give a consistent recall of 0.5[†]. Note that no result is reported for the case when the RRF trained using only adult training data is applied to the 2-5 years age group because the maximum recall was less than 0.5 in this case. The table shows that an RRF trained using only adult data gives a mean absolute error of 11.3 mm when applied to adult test data. This result is

[†] Localization errors are computed for true positive detections only, since it is not possible to compute them for false positive detections of anatomical structures that are not present within the scan extent. We report results at a constant recall to avoid giving unfair advantage to regressors that give a high proportion of false positive detections.

RRF training set	2-5 Years	6-11 Years	12-17 Years	Adult
Adult only	Fail	40.4	14	11.3
Pediatric only	8.8	8.1	11.4	17.8
Adult + Pediatric	10.3	7.4	11.2	12.1

Table 1. Mean bounding box localization errors in mm for different combinations of training and test data.

	2-5 Years		6-11 Years		12-17 Years		Adult	
organ	*mean*	*std*	*mean*	*std*	*mean*	*std*	*mean*	*std*
Abdomen	12.4	6.7	6.5	4.7	9.5	9.2	9.7	7.9
Heart	8.4	4.9	8.6	5.9	9.8	9.6	12.9	9.5
L. Kidney	7.8	5.9	8.5	5.8	10.4	9.7	12.3	8.2
R. Kidney	12.2	10.8	10.7	11	13.2	14.1	12.3	9.6
Liver	11.3	6.2	9.6	8	11.2	9.3	13.8	11.9
L. Lung	8.9	7.1	8.2	6	9.4	7.5	13.7	10.1
R. Lung	9.9	6.2	8	5.8	13.6	16.8	10.8	8.7
Spleen	11.2	7.2	9	7.6	11.4	11	13.9	11.7
Thorax	11.1	7.8	7.8	5.6	17.2	12	19.6	12.1

Table 2. Bounding box localization errors (mean and standard deviation, in mm) for an RRF trained using combined adult and pediatric data.

comparable with those presented in previous work on semantic annotation of adult CT scans.[1] However, the adult RRF performs poorly for pediatric scans (mean error 40.4 mm for the 6-11 age group). Similarly, an RRF trained using only pediatric data performs badly on the adult test set (mean error 17.8 mm). Interestingly, the pediatric RRF generalizes somewhat better to adult test data than the adult RRF does to pediatric data. This is presumably because the pediatric training set includes some high school age patients with nearly adult body shapes, wheras the adult training set contains no patients with body shapes like those of younger children. Finally, we see that the RRF trained using the combination of adult and pediatric data performs well for all age groups. Localization accuracy for the adult test set is comparable with that obtained by the RRF trained on only adult data; localization accuracy for pediatric test set is comparable with that obtained by the RRF trained using only pediatric data.

For the RRF trained using a combination of adult and pediatric data, localization errors for the individual anatomical structures of interest are reported seperately in table 2. Here, accuracy is computed with the detection confidence threshold tuned so as to give a higher recall value of 0.85 – more typical for our navigational use case. These figures illustrate that performance is comparably good for a wide range of anatomical structures, which include smaller organs such as the kidney and large scale anatomical regions such as the abdomen. Note the localization error is approximately constant over the various age groups. Furthermore a single RRF can recognize anatomical structures in scans of patients in multiple age groups without the need for algorithm parameters to be adapted heuristically in light of patient size.

Discussion. That a single RRF can give provide localization accuracy for patients of a variety of ages is a consequence of the ability of the training alogrithm to learn scale variability from training data. The information gain metric used for node optimization during training means that the nodes of the decision trees learn to cluster voxel samples that make similar predictions. For instance, one node in a decision tree might tend to partition samples from adult and teenage scans from those of younger children – so that different branches of the trained tree may be devoted to localizing anatomical structures for different patient age groups. The well-behaved generalization properties of decision forests[2] allow us to handle a wide range of anatomical variability.

4.3 Improvement in navigation

We have also evaluated the effectiveness of our semantic labeling algorithm within the navigational assistance application described above.[1] Here we use an RRF trained using combination of adult and pediatric training data so that navigational assistance can be provided without the need for information about the size of the patient.

Qualitative results. Figure 3 shows detected and ground truth bounding box positions for representative scans selected from each of the four age groups represented in our test set, and four different anatomical structures. Detected bounding boxes are visually in quite close agreement with ground truth for patients in all age groups and all structures of interest.

Quantitative results. Table 3 shows the centroid hit measures for the various structures of interest. For the majority of anatomical structures and patient age groups the centroid hit test scores 100%, which implies that two of three MPR views would contain the structure of interest. No significant differences are seen across the age range suggesting that the image navigation tool should provide a good user experience regardless of patient age.

	Abdomen	Heart	L. Kidney	R. Kidney	Liver	L.Lung	R.Lung	Spleen	Thorax
2-5 Years									
all axes	100	89	100	100	100	100	100	100	100
x-axis	100	89	100	100	100	100	100	100	100
y-axis	100	89	89	89	100	100	100	89	100
z-axis	100	89	89	89	100	100	100	89	100
6-11 Years									
x-axis	93	100	100	100	100	100	100	93	100
y-axis	93	100	100	100	100	100	100	100	100
z-axis	93	100	100	100	100	100	100	93	100
12-17 Years									
all axes	93	100	100	100	100	100	100	100	100
x-axis	93	100	100	100	100	100	100	93	100
y-axis	93	100	100	100	100	100	100	100	100
z-axis	93	100	100	100	100	100	100	93	100
Adult									
all axes	93	100	100	100	100	100	100	100	100
x-axis	93	100	100	100	100	100	100	93	100
y-axis	93	100	100	100	100	100	100	100	100
z-axis	93	100	100	100	100	100	100	93	100

Table 3. **Percentage of correct organ localizations** using the centroid-hit measure.

5. CONCLUSION

Lack of robustness, slow performance, and high error rates have been major barriers to rolling out fully automated solutions for automtic semantic labeling of DICOM CT images. In previous work, we have demonstrated the efficacy of the RRF algorithm as a means of efficient and robust semantic labeling of DICOM CT scans of adult patients. In this work, we have shown that the RRF algorithm is robut to the considerable additional inter-patient variability exhibited within pediatric CT data. The RRF algorithm can capture an implicit model of scale variability by learning directly from training data. In consequence, a single RRF can be used for semantic labelling of both adult and pediatric and CT scans without the need for heuristic adaptaion of algorithm parameters in light of patient scale. This means that the application of RRFs for semantic labelling in a clinical setting is increasingly feasible.

REFERENCES

[1] S. Pathak, A. Criminisi, S. White, I. Munasinghe, B. Sparks, D. Robertson, and K. Siddiqui, "Automatic semantic annotation and validation of anatomy in DICOM CT images," in *SPIE Medical Imaging*, **7967**, 2011.

[2] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning," Tech. Rep. MS-TR-2011-114, Microsoft Research, 2011.

[3] S. Pathak, W. Kim, I. Munasinghe, A. Criminisi, S. White, and K. Siddiqui, "Linking DICOM pixel data with radiology reports using automatic semantic annotation," in *SPIE Medical Imaging*, 2012.

[4] E. Konokoglu, A. Criminisi, S. Pathak, D. Robertson, S. White, and K. Siddiqui, "Robust linear regression of CT images using random regression forests," in *SPIE Medical Imaging*, **7962**, 2011.

[5] Y. Zheng, B. Georgescu, and D. Comaniciu, "Marginal space learning for efficient detection of 2d/3d anatomical structures in medical images," in *IPMI '09: Proc. of the 21st Intl Conference on Information Processing in Medical Imaging*, 2009.

[6] J. Gall and V. Lempitsky, "Class-specific Hough forest for object detection," in *IEEE CVPR*, (Miami), 2009.

[7] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context.," in *IJCV*, 2009.

[8] L. Breiman, "Random forests," Tech. Rep. TR567, UC Berkeley, 1999.

[9] A. Criminisi, J. Shotton, D. Robertson, and E. Konokoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in *Medical Computer Vision 2010: Recognition Techniques and Applications in Medical Imaging*, 2010.