Chapter 14
Anatomy Detection and Localization in 3D Medical Images

A. Criminisi, D. Robertson, O. Pauly, B. Glocker, E. Konukoglu, J. Shotton, D. Mateus, A. Martinez Möller, S.G. Nekolla, and N. Navab

This chapter discusses the use of regression forests for the automatic detection and simultaneous localization of multiple anatomical regions within computed tomography (CT) and magnetic resonance (MR) three-dimensional images. Important applications include: organ-specific tracking of radiation dose over time; selective retrieval of patient images from radiological database systems; semantic visual navigation; and the initialization of organ-specific image processing operations. We present a continuous parametrization of the anatomy localization problem, which allows it to be addressed effectively by multivariate random regression forests (Chap. 5). A single pass of our probabilistic algorithm enables the direct mapping from voxels to organ location and size, with training focusing on maximizing the confidence of output predictions. As a by-product, our method produces *salient anatomical landmarks*, *i.e.* automatically selected "anchor" regions which help localize organs of interest with high confidence. This chapter builds upon the work in [78, 286] and demonstrates the flexibility of forests in dealing with both CT and multi-channel MR scans. Quantitative validation is performed on two ground truth labeled datasets: (i) a database of 400 highly variable CT scans, and (ii) a database of 33 full-body, multi-channel MR scans. In both cases localization errors

A. Criminisi (✉) · B. Glocker · E. Konukoglu · J. Shotton
Microsoft Research Ltd., 7 J.J. Thomson Avenue, Cambridge, UK

D. Robertson
Redimension Ltd., Cambridge, UK

O. Pauly · D. Mateus
Institute of Biomathematics and Biometry, Helmholtz Zentrum München, München, Germany

A. Martinez Möller · S.G. Nekolla
Nuklearmedizin, Klinikum rechts der Isar, Technische Universität München, München, Germany

O. Pauly · D. Mateus · N. Navab
Computer Aided Medical Procedures, Technische Universität München, München, Germany

A. Criminisi, J. Shotton (eds.), *Decision Forests for Computer Vision and Medical Image Analysis*, Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-4929-3_14,

are reduced and results are more stable than those from more conventional atlas-based registration approaches. The simplicity of the regressor's context-rich visual features yield typical run-times of only 4 seconds per scan on a standard desktop. This anatomy recognition algorithm has now received FDA approval and is part of Caradigm's Amalga (www.caradigm.com).

14.1 Introduction

This chapter presents a new algorithm for the efficient detection and localization of anatomical structures ('organs') in CT and MR 3D images. A possible application is the automatic estimation of cumulative radiation dose being absorbed by the patient's individual organs during their lifetime, an important topic in modern radiology. Another application is the efficient retrieval of selected portions of patients' scans from radiological databases (PACS systems). When a physician wishes to inspect a particular organ, the ability to determine its position and extent automatically means it is not necessary to retrieve the entire scan (which could comprise gigabytes of data) but only a small region of interest, thus making economical use of the limited bandwidth. Other applications include single-click semantic navigation, automatic hyper-linking of textual radiological reports to the corresponding image regions, and the initialization of organ-specific image processing operations.

The main contribution of this work is a new parametrization of the anatomy localization task as a continuous multivariate parameter estimation problem. This is addressed effectively via non-linear regression, in the form of regression forests (see Chap. 5 and our previous work in [78, 286]). Our approach is fully probabilistic and, unlike previous techniques (*e.g.* [106, 422]), maximizes the confidence of output predictions. As a by-product, our method yields *salient anatomical landmarks*, *i.e.* automatically selected "anchor" regions that help localize organs of interest with high confidence. Our algorithm can localize both macroscopic anatomical regions[1] (*e.g.* abdomen, thorax, trunk, *etc.*) and smaller scale structures (*e.g.* heart, left adrenal gland, femoral neck, *etc.*) using a single model (*cf.* [108]).

The focus of our approach is both on accuracy of prediction and speed of execution, as we wish to achieve anatomy localization in seconds.

Regression Approach Regression algorithms [152] estimate functions which map input variables to *continuous* outputs.[2] The regression paradigm fits the anatomy localization task well. In fact, its goal is to learn the non-linear mapping from voxels *directly* to organ position and size. The work in [421] presents a thorough overview of regression techniques and demonstrate the superiority of boosted regression [115] with respect to *e.g.* kernel regression [380]. In contrast to

[1] DICOM tags for the anatomical region are often erroneous [147].

[2] As opposed to *classification* where the predicted variables are categorical.

the boosted regression approach, maximizing confidence of output prediction is integral to our forest approach. An empirical comparison between boosting, forests and cascades is found in [413].

Comparison with Classification-Based Approaches In [414] organ detection is achieved via a confidence maximizing sequential scheduling of multiple, organ-specific *classifiers*. In contrast, our single, tree-based regressor allows us to deal naturally with *multiple* anatomical structures simultaneously. As shown in the machine learning literature [371] this encourages feature sharing and, in turn better generalization. In [328] a sequence of PBT classifiers (first for salient slices, then for landmarks) are used. In contrast, our single forest regressor maps directly from voxels to organ locations and extents; latent, salient landmark regions are extracted as a by-product. In [77] the authors achieve localization of organ *centers* but fail to estimate the organ extent (similar to [117]). Here we present a more direct, continuous model which estimates the position of the walls of the bounding box containing each organ thus achieving simultaneous organ localization and extent estimation.

Comparison with Registration-Based Approaches Although atlas-based methods have enjoyed much popularity [106, 337, 408], their conceptual simplicity belies the technical difficulty inherent in achieving robust, inter-subject registration. Robustness may be improved by using multi-atlas techniques [170] but only at the expense of multiple registrations and hence increased computation time. Our algorithm incorporates atlas information implicitly, within a tree-based model. As shown in the results section, such model is more efficient than keeping around multiple atlases and can achieve anatomy localization in only a few seconds. Comparisons with global affine atlas registration methods (somewhat similar to ours in computational cost) show that our algorithm produces lower errors and more stable localization results. Next we describe details of our approach.

14.2 Organ Localization as Regression Task

This section presents mathematical notation, our parametrization of organ locations within a medical scan, and the formulation of the localization problem as a regression task.

Following the notation set out in Chap. 3, vectors are represented in boldface (*e.g.* $\mathbf{p}$), matrices as teletype capitals (*e.g.* $\mathtt{A}$), and sets in calligraphic style (*e.g.* $\mathcal{S}$). The position of a voxel in a volumetric image is denoted as a 3-vector $\mathbf{p} = (p_x, p_y, p_z)$. A voxel at position $\mathbf{p}$ is associated with a d-dimensional vector of feature responses which is denoted as $\mathbf{v}(\mathbf{p}) = (v_1, \ldots, v_i, \ldots, v_d) \in \mathbb{R}^d$.

14.2.1 Parametrization and Regression Formulation

A 3D patient scan is represented by the intensity function $J : \Omega \rightarrow \mathbb{R}$, where $\Omega \subset \mathbb{N}^3$ is the image domain. Given a set $\mathcal{C}$ of organs of interest, we propose

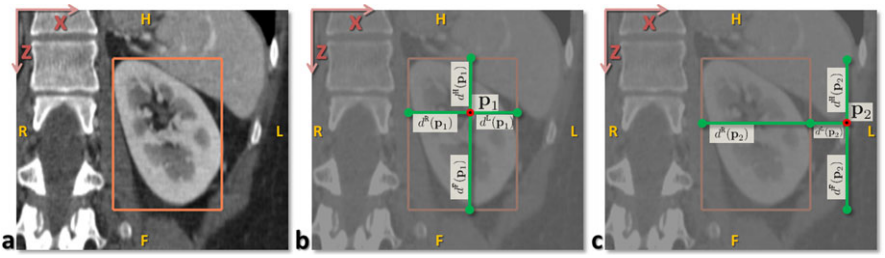

Fig. 14.1 Problem parametrization. (**a**) A 2D (coronal) view of a left kidney within a 3D CT scan, and the associated ground truth bounding box (*in orange*). (**b**, **c**) *Every* voxel $\mathbf{p}_i$ in the volume votes for the position of the six walls of each organ's 3D bounding box via six relative, offset displacements $d^{\mathrm{k}}(\mathbf{p}_i)$ in the three canonical directions x, y and z

to model their absolute positions within the patient scan by a set of 3D bounding boxes. Each bounding box $\mathbf{b}_c$ contains one organ $c \in \mathcal{C}$ and is parametrized as a 6-dimensional vector $\mathbf{b}_c = (b_c^{\mathtt{L}}, b_c^{\mathtt{R}}, b_c^{\mathtt{A}}, b_c^{\mathtt{P}}, b_c^{\mathtt{H}}, b_c^{\mathtt{F}})$. Each vector element represents the absolute position (in mm) of one axis-aligned face.[3] The goal of multiple organ localization is to estimate simultaneously the parameters of the different bounding boxes containing the organs of interest. Thus, the desired output is one six-dimensional vector $\mathbf{b}_c$ per organ, a total of $6 \times |\mathcal{C}|$ continuous parameters. In a probabilistic fashion, we aim at modeling the probability distribution $p(\mathbf{b}_c|\mathbf{v})$ for all $c \in \mathcal{C}$, so that given a previously unseen image and all per-voxel features $\{\mathbf{v}\}$, we can predict the location of all visible organs by estimating

$$\mathbf{b}_c^* = \arg\max_{\mathbf{b}_c} p(\mathbf{b}_c|\mathcal{V}) \quad \text{with } \mathcal{V} = \big\{\mathbf{v}(\mathbf{p}) \mid \mathbf{p} \in \Omega\big\}. \tag{14.1}$$

More generally, one could also define the regression over a single distribution $p(\mathbf{b}|\mathbf{v})$ with $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_c, \ldots, \mathbf{b}_{|C|}) \in \mathbb{R}^{6|C|}$. This would allow to model inter-organ location dependencies.

Key to our algorithm is the idea that *all* voxels in a test image contribute with varying degrees of confidence to the estimates of the positions of *all* organs. Formally, we propose a probabilistic regression strategy in which each voxel $\mathbf{p} \in \Omega$ votes for the *relative offsets* to all organs' bounding boxes. Thus, each voxel $\mathbf{p}$ in a medical scan is associated with an offset with respect to the bounding box, $\mathbf{b}_c$ for each organ $c \in \mathcal{C}$ (see Fig. 14.1b, c). Such offset is a function of both $\mathbf{p}$ and c as follows: $\mathbf{d}(\mathbf{p}; c) = \hat{\mathbf{p}} - \mathbf{b}_c$, with $\hat{\mathbf{p}} = (p_x, p_x, p_y, p_y, p_z, p_z)$. Therefore $\mathbf{d}(\mathbf{p}; c) \in \mathbb{R}^6$. Given a database of annotated scans, our goal becomes then to learn the conditional distribution of 3D displacements $p(\mathbf{d}|c, \mathbf{v})$ for each organ $c \in \mathcal{C}$.

Intuitively, some distinct voxel clusters (*e.g.* tips of ribs, or vertebrae) may predict the position of an organ (*e.g.* the heart) with high confidence. Ideally, at detection time those clusters would be automatically detected and used as landmarks for the localization of those organs. In contrast, clusters of voxels in larger regions

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

of texture-less tissue or even air should contribute little to the estimation of organ positions because of their higher prediction uncertainty. Therefore, our aim is to learn to cluster voxels based on their appearance, their spatial context and, above all, their confidence in predicting the position and size of all organs of interest. Note that this is different from the task of assigning a categorical label to each voxel (*i.e.* the classification approach in [77]). Here we wish to produce confident predictions of a small number of continuous localization parameters. The *latent* voxel clusters (think of them as some sort of predictive landmarks) are discovered automatically.

To tackle the simultaneous feature selection and parameter regression task, we use a multivariate random regression forest (Chap. 5); *i.e.* an ensemble of regression trees trained to predict the location and size of all organs simultaneously. Next we describe the details of our approach.

14.3 Regression Forests for Organ Localization

This section presents how to solve the organ localization problem using multivariate regression forests. First, we start by detailing the type of feature response we employ in order to capture the visual appearance and contextual information of each voxel within a medical scan. The feature responses form the input of the regression forest, while the absolute organ locations are the output.

14.3.1 Feature Responses for Application in CT and MR

Medical images are acquired using different physical principles (*e.g.* based on X-rays or magnetic resonance), and thus the images have very different visual appearance. Intensity values correspond to different physical properties, and image analysis applications need to be tailored with respect to a specific imaging modality. Motivated by the underlying imaging technique, we employ distinct visual features for the two cases of CT and MR images.

The feature vector $\mathbf{v}(\mathbf{p}) = (v_1, \ldots, v_i, \ldots, v_d) \in \mathbb{R}^d$ for a reference 3D voxel location $\mathbf{p}$ is a collection of mean intensity values over (possibly) displaced feature boxes, *i.e.*

$$v_i = \frac{1}{|\mathbf{F}_{\mathbf{p};i}|} \sum_{\mathbf{q} \in \mathbf{F}_{\mathbf{p};i}} J(\mathbf{q}) \tag{14.2}$$

where $J(\mathbf{q})$ denotes the image intensity at position $\mathbf{q}$ in the image, and $\mathbf{q} \in \mathbf{F}_{\mathbf{p};i}$ are the image points within the feature box. The box $\mathbf{F}_{\mathbf{p};i}$ is displaced relative to the reference point $\mathbf{p}$ (see Fig. 14.2). In theory, for each reference point we can determine an infinite number of such features. In practice we will randomly generate thousands of such features during the training phase, which is part of the feature selection process of decision forests. The types of feature used here are similar to those in [77, 117, 342], *i.e.* mean intensities over displaced, asymmetric cuboidal

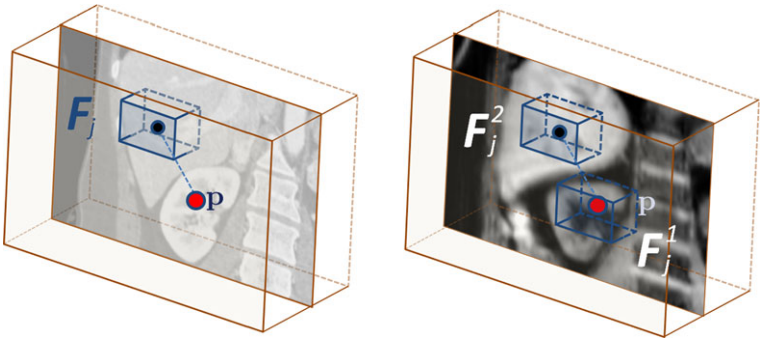

Fig. 14.2 Visual features. As shown *on the left*, voxel features in CT images are computed as the mean intensity over a 3D cuboid displaced with respect to the reference voxel position. *On the right*, voxel features in MR images are binary numbers encoding the difference between the mean intensity computed over two 3D cuboids

regions within the volume. These features are efficient to compute via integral images/volumes [389] and are able to capture spatial context.

Similar to Chap. 13 we define a selector function

$$\boldsymbol{\phi}(\mathbf{v}) = (v_i, v_j) \quad \text{with } i, j \in \{1, \ldots, d\} \tag{14.3}$$

thus, the tests applied by internal nodes will act upon scalar values computed as

$$f(\mathbf{p}; \boldsymbol{\phi}, \boldsymbol{\psi}) = \boldsymbol{\phi}\big(\mathbf{v}(\mathbf{p})\big) \cdot \boldsymbol{\psi}. \tag{14.4}$$

Visual Features in CT Computed tomography images are characterized by the fact that the intensity values directly indicate the tissue density (in Hounsfield units) at a particular location. So, it makes sense to use absolute intensity values to construct visual features. This is achieved readily by fixing $\boldsymbol{\psi} = (1, 0)$ for all split nodes.

Visual Features in MR In magnetic resonance images, we cannot rely on the absolute intensity values since there is no calibration between different scans. However, the relative intensity changes between different regions within the same scan can provide important visual cues. In the case of MR it makes sense to construct visual features which are invariant to global additive intensity biases and take into consideration image gradients. This is achieved here by fixing $\boldsymbol{\psi} = (1, -1)$ for all split nodes. This corresponds to taking the difference of mean intensities in two image regions. As detailed in [286], the feature boxes can also be chosen by using a predefined 3D pattern, and can be seen as a multi-scale 3D version of local binary patterns [274].

The set of feature vectors is a crucial component of the regression function, whose aim is to determine a functional mapping from the input feature space to the output space of organ bounding boxes. We will now describe how this mapping is learned.

14.3.2 Regression Forest Learning

Weak Learners The training process constructs each regression tree and decides at each node how to best split the incoming voxels. We are given a subset of all labeled volumes (the training set), and the associated ground truth organ bounding box positions (Fig. 14.1a). A subset of voxels in the training volumes is used for forest training. These training voxels are sampled on a regular grid within ± 10 cm of the center of each axial slice in the training volume. The size of the forest T is fixed and all trees are trained in parallel.

Each training voxel $\mathbf{p}$ is sent down each of the trees starting at the root. The jth split node applies an axis-aligned weak learner test $h(\mathbf{p}, \boldsymbol{\theta}_j)$ and based on the result sends the voxel to the left or right child node. The parameters $\boldsymbol{\theta}_j = (\boldsymbol{\phi}_j, \boldsymbol{\psi}, \tau_j)$ characterize the weak learners associated with the jth node. The corresponding weak learner is

$$h(\mathbf{p}, \boldsymbol{\theta}_j) = \left[f(\mathbf{p}; \boldsymbol{\phi}_j, \boldsymbol{\psi}) > \tau_j \right]. \tag{14.5}$$

For application in CT, τ is a learned scalar parameter. Instead, in MR we usually set $\tau = 0$. As usual the voxel $\mathbf{p}$ is sent to the right child node if h is true and to the left child node otherwise.

Objective Function Node optimization is driven by maximizing a *continuous* information gain measure, defined in general terms as

$$I(\mathcal{S}, \boldsymbol{\theta}) = H(\mathcal{S}) - \sum_{i=\{\mathtt{L},\mathtt{R}\}} \omega_i H\left(\mathcal{S}^i\right) \tag{14.6}$$

where H denotes entropy, $\mathcal{S}$ is the set of training points reaching a node, $\mathtt{L}$ and $\mathtt{R}$ denote the left and right sets generated from $\mathcal{S}$ through the split defined by parameters $\boldsymbol{\theta}$, and finally $\omega_i = |\mathcal{S}^i|/|\mathcal{S}|$.

For a given organ c we model the continuous conditional distribution of the 3D displacement $\mathbf{d}(\mathbf{p}; c)$ at each node as a multivariate Gaussian; *i.e.*

$$p(\mathbf{d} \mid c, \mathcal{S}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Lambda_c(\mathcal{S})|} e^{-\frac{1}{2}((\mathbf{d}-\overline{\mathbf{d}}_c(\mathcal{S}))^\top \Lambda_c(\mathcal{S})^{-1}(\mathbf{d}-\overline{\mathbf{d}}_c(\mathcal{S})))}, \tag{14.7}$$

with $N = 6$ and $\int_{\mathbb{R}^6} p(\mathbf{d}|c, \mathcal{S})\, d\mathbf{d} = 1$. The vector $\overline{\mathbf{d}}_c$ indicates the mean displacement and Λ_c the 6×6 covariance matrix of $\mathbf{d}$ for all points in the set $\mathcal{S}$.

In the context of organ localization, we incorporate an organ visibility prior in the objective function. In fact, due to surgery or image cropping a given organ may not be present or visible in a scan. For the set $\mathcal{S}$ this prior is defined as

$$p(c|\mathcal{S}) = n_c(\mathcal{S})/Z, \tag{14.8}$$

where $n_c(\mathcal{S})$ is the number of training voxels in the set $\mathcal{S}$ for which it is possible to compute the displacement $\mathbf{d}(\mathbf{p}; c)$; *i.e.* the training points in the set that come from

training volumes for which the organ c is present. Z is a normalization constant which ensures that $\sum_c p(c|\mathcal{S}) = 1$. Thus we can estimate the joint distribution for displacement and organ class as

$$p(\mathbf{d}, c|\mathcal{S}) = p(\mathbf{d}|c, \mathcal{S})p(c|\mathcal{S}). \tag{14.9}$$

Now, using the definition of the differential entropy of a Gaussian density and after some algebraic manipulation we obtain the following joint entropy for the node:

$$H(\mathbf{d}, c; \mathcal{S}) = H(c; \mathcal{S}) + \sum_c p(c|\mathcal{S}) \left(\frac{1}{2} \log\left((2\pi e)^N \left| \Lambda_c(\mathcal{S}) \right| \right) \right). \tag{14.10}$$

The joint information gain is

$$I(\mathcal{S}, \boldsymbol{\theta}) = H(\mathbf{d}, c; \mathcal{S}) - \sum_{i \in \{\mathrm{L},\mathrm{R}\}} \omega_i H\left(\mathbf{d}, c; \mathcal{S}^i\right), \tag{14.11}$$

which after some manipulation can be rewritten as

$$I(\mathcal{S}, \boldsymbol{\theta}) = I^{\mathtt{reg}}(\mathcal{S}, \boldsymbol{\theta}) + I^{\mathtt{cls}}(\mathcal{S}, \boldsymbol{\theta}) \tag{14.12}$$

where

$$I^{\mathtt{reg}}(\mathcal{S}, \boldsymbol{\theta}) = \frac{1}{2} \left(\sum_c p(c|\mathcal{S}) \log\left| \Lambda_c(\mathcal{S}) \right| - \sum_{i \in \{\mathrm{L},\mathrm{R}\}} \omega_i \sum_c p\left(c|\mathcal{S}^i\right) \log\left| \Lambda_c\left(\mathcal{S}^i\right) \right| \right) \tag{14.13}$$

with $\omega_i = |\mathcal{S}^i|/|\mathcal{S}|$, and

$$I^{\mathtt{cls}}(\mathcal{S}, \boldsymbol{\theta}) = H(c; \mathcal{S}) - \sum_{i \in \{\mathrm{L},R\}} \omega_i H\left(c; \mathcal{S}^i\right), \tag{14.14}$$

with $H(c; \mathcal{S})$ the standard Shannon entropy for categorical distributions.

We remember from Chap. 3 that optimizing the node parameters implies maximizing the information gain. Here nodes are trained via "randomized node optimization", as

$$\boldsymbol{\theta}_j = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}_j} I(\mathcal{S}, \boldsymbol{\theta}). \tag{14.15}$$

But maximizing (14.12) corresponds to minimizing the determinants of the 6×6 covariance matrices Λ_c (covariances defined over displacement random variables) associated with the $|\mathcal{C}|$ organs, where each organ's contribution is weighted by the associated prior probability for its visibility. This decreases the uncertainty in the probabilistic vote cast by each cluster of voxels on each organ pose. In our experiments we have found that this prior-driven organ weighting produces more balanced trees and has a noticeable effect on the accuracy of the results. In practice, the visibility prior favors a clustering of points of both similar locations but also corresponding to images with similar field-of-views.

Stopping Criterion Branching stops when the number of points reaching the node is smaller than a threshold $n_{\min}$, or a maximum tree depth D has been reached. After training, the jth split node remains associated with the parameters $\boldsymbol{\theta}_j$. At each leaf node we store the learned means $\overline{\mathbf{d}}_c$ and covariance matrices Λ_c, and the class priors $p(c)$.

This framework may be reformulated using non-parametric distributions, with pros and cons in terms of regularization and storage. We have found our parametric assumption not to be restrictive since the multi-modality of the input space is captured by our hierarchical piece-wise Gaussian model. However, under the simplifying assumption that bounding box face positions are uncorrelated (*i.e.* diagonal Λ_c), it is convenient to store at each leaf node learned 1D histograms over face offsets $p(\mathbf{d}|c;\mathcal{S})$.

Discussion Equation (14.12) is an information-theoretical way of maximizing the confidence of the desired continuous output *for all* organs, without going through intermediate voxel classification (as in [77] where positive and negative examples of organ centers are needed). Furthermore, this gain formulation enables testing different context models, for example imposing a *full* covariance Λ_c would allow correlations between all walls in each organ. One could also think of enabling correlations between different organs. Taken to the extreme, this might have undesirable overfitting consequences. On the other hand, assuming *diagonal* Λ_c matrices can lead to uncorrelated output predictions. Interesting models live in the middle ground, where some but not all correlations are enabled to capture *e.g.* class hierarchies or other forms of spatial context.

14.3.3 Regression Forest Prediction

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

$$p(\mathbf{b}_c|\mathbf{v}) = \sum_{t=0}^{T} \sum_{l \in \tilde{\mathcal{L}}_t} p(\mathbf{b}_c|\mathbf{v},l)\, p(l). \tag{14.16}$$

$\tilde{\mathcal{L}}_t$ is a subset of the leaves of tree t. We select $\tilde{\mathcal{L}}_t$ as the set of leaves corresponding to the 75 % of all test voxels which have the highest confidence (for each class c). Finally $p(l)$ is simply the proportion of samples arriving at leaf l. Note that here the leaf prediction model is a multivariate, probabilistic-*constant* model rather than the more flexible probabilistic-*linear* one used in Chap. 5.

Organ Localization The final prediction $\mathbf{b}_c^*$ for the absolute position of the cth organ is given by

$$\mathbf{b}_c^* = \arg\max_{\mathbf{b}_c} p(\mathbf{b}_c|\mathbf{v}). \tag{14.17}$$

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

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

14.4 Results, Comparisons and Validation

This section assesses the proposed algorithm for anatomy localization within 3D computed tomography and magnetic resonance scans in terms of accuracy, runtime speed and memory efficiency, and compares it to state of the art techniques.

14.4.1 Anatomy Localization in Computed Tomography Scans

The Labeled CT Database We wish to recognize the following 26 anatomical structures $\mathcal{C} = \{$abdomen, l./r. adrenal gland, l./r. clavicle, l./r. femoral neck, gall bladder, head of l./r. femur, heart, l./r. atrium of heart, l./r. ventricle of heart, l./r. kidney, liver, l./r. lung, l./r. scapula, spleen, stomach, thorax, thyroid gland$\}$. We are given a database of 400 scans which have been manually annotated with 3D bounding boxes tightly drawn around the structures of interest (see Fig. 14.1a).

The database comprises patients with a wide variety of medical conditions and body shapes and the scans exhibit large differences in image cropping, resolution, scanner type, and use of contrast agents (Fig. 14.3). Voxel sizes are ∼0.5–1.0 mm along x and y, and ∼1.0–5.0 mm along z. The images were not pre-registered.

A regression forest was trained using 318 volumes selected randomly from our 400-volume dataset. Organ localization accuracy was measured using the remaining 82 volumes, which contained a total of 1504 annotated organs of which 907 were fully visible within the scan. Only organs that are entirely contained in the volumes are used for training and test. Training and test volumes were downsampled using nearest neighbor interpolation. Integer downsampling factors were chosen so that

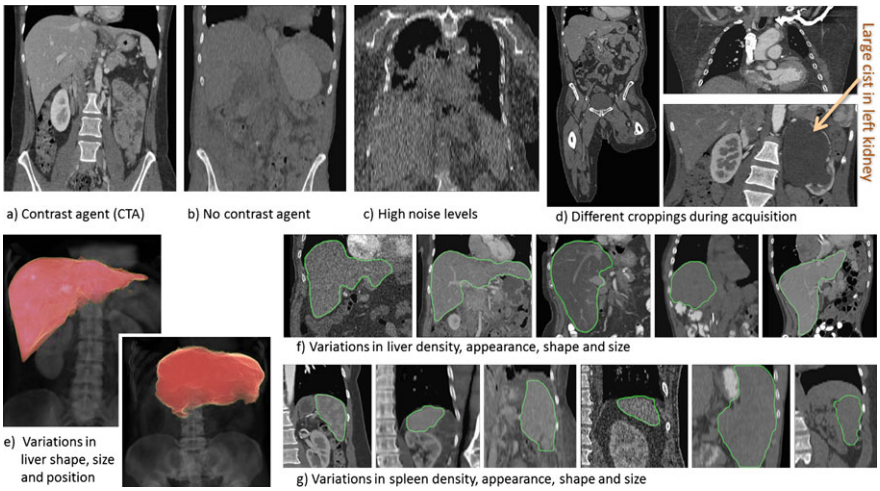

Fig. 14.3 Variability in our labeled CT database. (**a**, **b**, **c**) Variability in appearance due to presence of contrast agent, or noise. (**d**) Difference in image geometry due to acquisition parameters and possible anomalies. (**e**) Volumetric renderings of liver and spine to illustrate large 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 shape variability. (The organ outlines, drawn by hand, are highlighted in *green*). All CT scans are natively calibrated, both metrically and photometrically

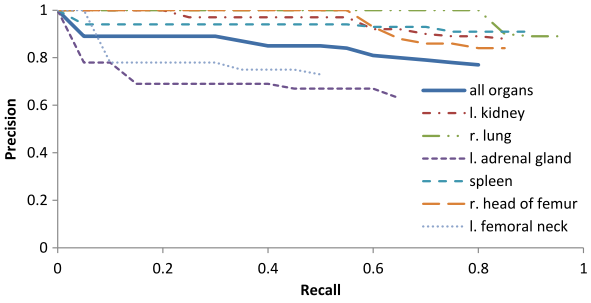

Fig. 14.4 Precision-recall curves for some representative organ classes and for all organ classes. The *curves* show how precision and recall change as the detection confidence threshold β is varied, both for all organs and for a representative group of individual organs (several organ classes are omitted to avoid clutter)

the resulting voxel pitch was as near as possible to 3 mm per voxel in the x, y, and z directions. Downsampling to this resolution reduces memory usage without noticeable reduction in accuracy.

Quantitative Evaluation To characterize the performance of the algorithm, *precision-recall* curves are plotted (Fig. 14.4). In this context *precision* refers to the proportion of organs that were correctly detected, and *recall* to the 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.[4] The plot shows how precision and recall vary as the detection confidence β is varied.

In the figure the average precision remains high until recall reaches approximately 80 %. Accuracy is best for larger organs; those with smaller size or greater positional variability are more challenging.

Table 14.1 shows mean localization errors, *i.e.* the absolute difference between predicted and ground truth bounding box face positions. Errors are averaged over all faces of the bounding boxes. Despite the large variability in our test data we obtain a mean error of only 13.5 mm, easily sufficient for our intended applications. Errors in the axial (z) direction are approximately the same as those in x and y despite significantly more crop variability in this direction. Consistently good results are obtained for different choices of training set as well as different training runs.

Computational Efficiency With our C# software running in a single thread, organ detection for a typical $30 \times 30 \times 60$ cm volume requires approximately 4 s of CPU time for a typical four-tree forest. Most of the time is spent aggregating offset distributions (which are represented as histograms) over salient leaves. However, significant speed-up could be achieved with relatively simple code optimizations, *e.g.* by using several cores in parallel for tree evaluation and histogram aggregation.

Comparison with Affine, Atlas-Based Registration An alternative strategy for anatomy localization is to align the input volume with a suitable *atlas*, *i.e.* a reference scan for which organ bounding box positions are known. Approximate bounding box positions in the input volume are then determined by using the computed atlas alignment transformation to map bounding box locations from the atlas into the input image.

Non-linear atlas registration (via non-rigid registration algorithms) can, in theory, provide the most accurate localization results. In practice, however, this approach is not robust to bad initialization and requires significantly greater computation times than the approach we describe here. Since speed is an important aspect of our work, here we chose to compare our results with those from comparably fast atlas-based algorithms, *i.e.* those based on global affine registration. This is a rather approximate approach because accuracy is limited by inter- and intra-subject variability in organ location and size. However, it is robust and its computation times are close to those of our method.

Instead of using a single atlas we use a multi-atlas approach due to its higher accuracy [170]. From the training set, five scans were selected to be used as atlases. The selected scans included three abdominal-thorax scans (one female, one male and one slightly overweight male), one thorax scan, and one whole body scan. This

[4]This metric is appropriate in light of our intended data retrieval and semantic navigation applications because the bounding box centroid would typically be used to select which coronal, axial, and sagittal slices to display to the user. If the ground truth bounding box contains the centroid of the predicted bounding box, then the selected slices will intersect the organ of interest.

Table 14.1 Regression forest results for CT. Bounding box localization errors in mm and associated standard deviations. The table compares results for our method with those for the Elastix and Simplex methods. Lowest errors for each class of organ are shown in bold. Our method gives lower errors for *all* organ classes

organ	Our method		Elastix		Simplex	
	mean	std	mean	std	mean	std
abdomen	**14.4**	13.4	34.6	74.2	27.6	36.5
l. adrenal gland	**11.7**	9.6	20.5	42.4	15.5	20.9
r. adrenal gland	**12.1**	9.9	22.2	45.0	18.2	29.6
l. clavicle	**19.1**	17.4	34.3	20.5	31.1	16.3
r. clavicle	**14.9**	11.6	39.0	44.3	24.1	13.9
l. femoral neck	**9.7**	7.5	38.3	78.5	16.1	15.4
r. femoral neck	**10.8**	8.3	38.4	82.3	17.3	17.7
gall bladder	**18.0**	15.0	28.1	54.5	23.2	26.6
l. head of femur	**10.6**	14.4	38.8	80.8	19.4	26.6
r. head of femur	**11.0**	15.7	39.6	84.9	19.1	28.4
heart	**13.4**	10.5	34.4	52.0	16.9	15.8
l. heart atrium	**11.5**	9.2	30.7	50.5	15.4	15.4
r. heart atrium	**12.6**	10.0	33.0	51.9	15.2	15.5
l. heart ventricle	**14.1**	12.3	35.9	51.7	18.1	16.7
r. heart ventricle	**14.9**	12.1	35.4	52.8	17.2	16.8
l. kidney	**13.6**	12.5	22.1	46.1	18.7	25.6
r. kidney	**16.1**	15.5	25.3	49.8	21.1	27.0
liver	**15.7**	14.5	26.9	53.3	23.2	30.4
l. lung	**12.9**	12.0	24.5	29.2	16.9	23.4
r. lung	**10.1**	10.1	25.0	27.2	16.0	21.7
l. scapula	**16.7**	15.7	50.9	54.1	33.1	20.1
r. scapula	**15.7**	12.0	44.4	41.2	22.7	12.4
spleen	**15.5**	14.7	29.0	46.6	23.0	22.8
stomach	**18.6**	15.8	27.6	48.9	22.8	23.4
thorax	**12.5**	11.5	36.5	37.4	25.3	35.1
thyroid gland	**11.6**	8.4	13.3	10.3	12.9	10.2
all organs	**13.5**	13.0	28.9	52.4	19.4	24.7

selection was representative of the overall distribution of image types in the dataset. All five atlases were registered to all the scans in the test set. For each test scan, the atlas that yielded the smallest registration cost was selected as the best one to represent that particular test scan. Registration was achieved using two different global affine registration algorithms. The first algorithm ('Elastix') is that implemented by the popular *Elastix* toolbox [184] and works by maximizing mutual information using stochastic gradient descent. The second algorithm ('Simplex') is our own im-

plementation and works by maximizing correlation-coefficient between the aligned images using the simplex method as the optimizer [265]. In each case parameters were optimized for best accuracy.

Resulting errors (computed on the same test set) are reported in Table 14.1. The atlas registration techniques give larger mean errors and error standard deviation (nearly double in the case of Elastix) compared to our approach. Furthermore, atlas registration requires between 90 s and 180 s per scan (*cf.* our algorithm runtime is $\sim$4 s for $T = 4$ trees, on a single CPU core).

Figure 14.5 further illustrates the difference in accuracy between the three approaches. For the atlas registration algorithms, the error distribution's larger tails suggest a less robust behavior. This is reflected in larger values of the error mean and standard deviation and is consistent with our visual inspection of the registrations. In fact, in about 30 % of cases the registration process got trapped in local minima and produced grossly inaccurate alignment. In those cases, results tend not to be improved by using a non-linear registration step (which tends not to help the registration algorithm to escape bad local minima, whilst increasing the runtime considerably).

Automatic Landmark Detection Figure 14.6 visualises the anatomical landmark regions that were automatically selected for organ localization. Given a trained regression tree and an input volume, we select one or two leaf nodes with high prediction confidence for a chosen organ class (*e.g.* left kidney). Then, for each sample arriving at the selected leaf nodes, we shade in green the cuboidal feature boxes used during weak learner evaluation. Those green regions represent some of the anatomical locations that were automatically selected and used to predict the location of the chosen organ. In this example, the bottom of the left lung and the top of the left pelvis are used to predict the position of the left kidney. Similarly, the bottom of the right lung is chosen to localize the right kidney. Such regions correspond to meaningful, visually distinct, anatomical landmarks that have been discovered in a completely unsupervised manner.

14.4.2 Anatomy Localization in Magnetic Resonance Scans

The Labeled MR Database As described in [286] we also have a database of 33 patients. For each patient we have available labeled MR Dixon 3D images [226]. This means that for each patient we have two image channels, a "water" channel J^w and a "fat" channel J^f. As these two channels are captured simultaneously, they are aligned to each other. In this application, we propose to use both J^w and J^f, *i.e.* to extract features in both channels. Just like in the CT database, an expert has annotated different anatomical structures with axis-aligned bounding boxes. Here we have annotated the following five anatomical structures: head, heart, l. lung, r. lung and liver.

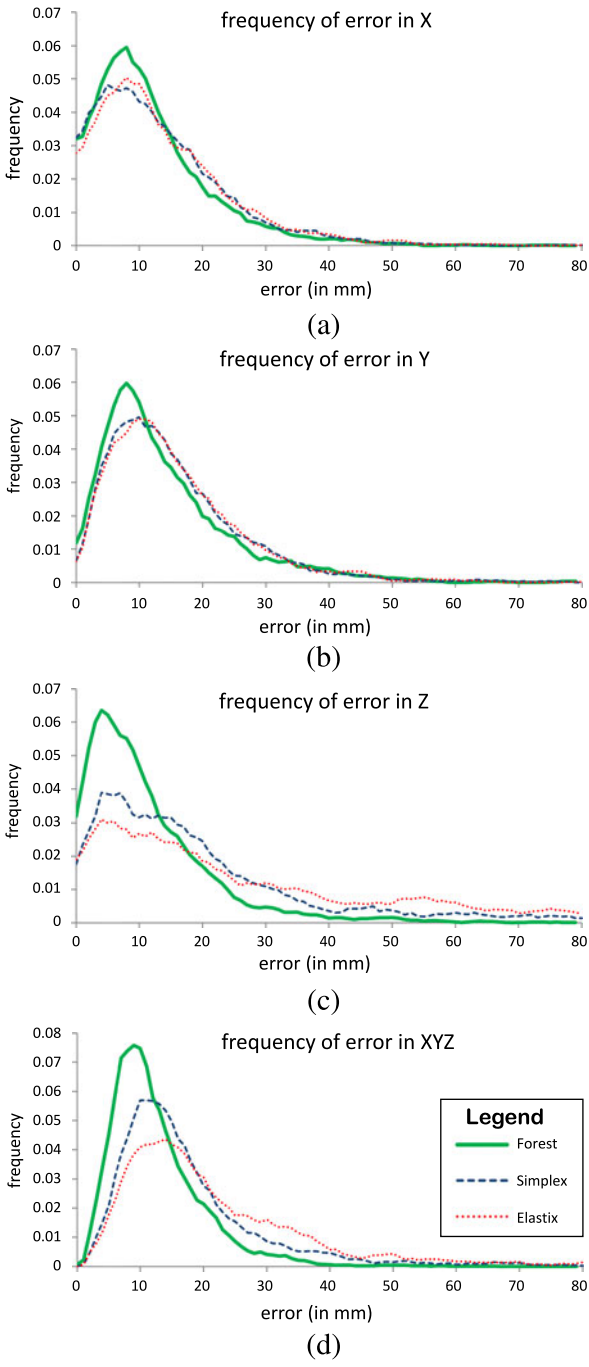

Fig. 14.5 Comparison with atlas-based registration. Distributions of bounding box localization errors for our algorithm ('Forest') and two atlas-based techniques ('Elastix' and 'Simplex'). Error distributions are shown separately for (**a**) left and right, (**b**) anterior and posterior, and (**c**) head and foot faces of the detected bounding boxes, and (**d**) averaged over all bounding box faces for each organ. The error distributions for the atlas techniques (particularly in plots (**c**) and (**d**)), have more probability mass in the tails, which is reflected by larger mean errors and error standard deviations

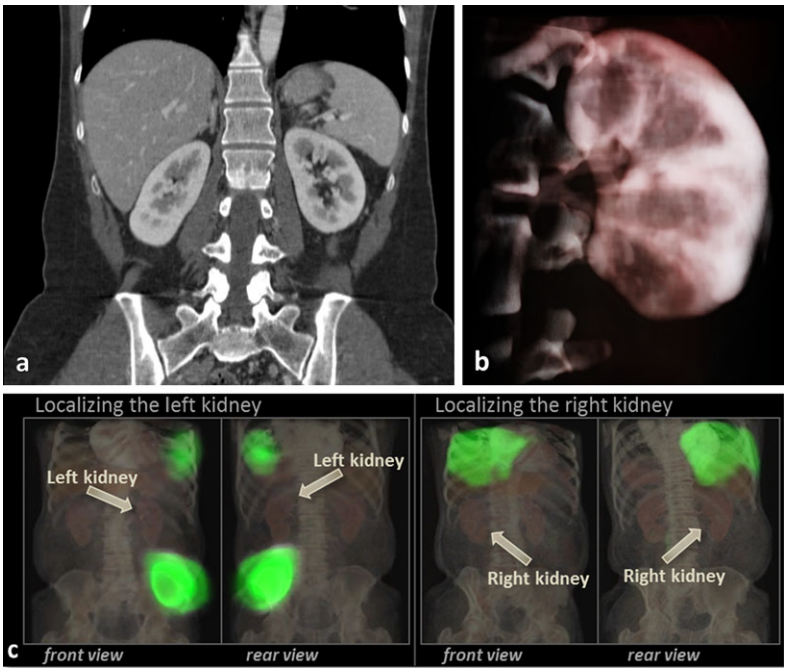

Fig. 14.6 Automatic discovery of salient anatomical landmark regions. (**a**) A test volume and (**b**) a 3D volume rendering of the left kidney's bounding box, as detected by our algorithm. (**c**) The highlighted *green regions* correspond to regions of the volume that were automatically selected as salient predictors of the position of the kidneys

Comparative Experiments When dealing with MR images we chose to implement both random forests and their special case, random ferns (see Chap. 9). Both are compared quantitatively with an atlas-based registration approach and the results shown in Table 14.2. The reported lower and upper bounds correspond to the best and worst results across different atlases. Of course, in practice, it is not possible to know which atlas yields best results for a specific test image, so we report the mean error achieved when averaging over the different atlas results. For further reading on the relationship between forests and ferns please refer to Chap. 9 and [80].

The table shows that both regression forests and regression ferns achieve an accuracy which is better than the best case atlas accuracy, while providing increased robustness (smaller standard deviation of errors). Taking a look at the localization error per organ, one can notice that the lowest error for our approach is achieved for the localization of the head, which is due to the fact that the head is surrounded by a lot of air which makes it easier to localize. While the heart shows the second lowest error, lungs and liver were more difficult to localize. This is mainly due to the high inter-patient variability of the shape of these organs. The best results were obtained with 14 ferns and six nodes for random ferns, six trees of depth 8 for regression forests. On a 64 Core Duo 2.4 GHz laptop running MATLAB the training/testing time on 20/13 patients is only 0.7/0.5 s for random ferns. Decision forests need

Table 14.2 Regression forest results for MR bounding box localization errors in mm and associated standard deviations. The table compares results for our method using random forest, random ferns and multi-atlas registration

organ	Random ferns		Random forests		Atlas lower bound		Atlas upper bound		Atlas mean	
	mean	std	mean	std	mean	std	mean	std	mean	std
head	**9.82**	8.07	10.02	8.15	18.00	14.45	70.25	34.23	35.10	13.17
l. lung	14.95	11.35	**14.78**	11.72	14.94	11.54	60.78	29.47	30.41	11.39
r. lung	**16.12**	11.73	16.20	12.14	15.02	13.69	63.95	30.13	29.85	12.62
liver	**18.69**	13.77	18.99	13.88	18.13	16.26	70.59	32.88	31.74	13.49
heart	**15.17**	11.70	15.28	11.89	13.31	11.03	60.38	28.90	29.82	12.23
all organs	**14.95**	11.33	15.06	11.55	15.88	13.40	65.19	31.12	31.38	12.58

25/1 s. Concerning atlas registration, each single affine registration needs 12.5 s. See [286] for more details.

14.5 Conclusion

Anatomy localization has been cast here as a non-linear regression problem where *all* voxel samples vote for the position of all anatomical structures. Location estimates are obtained by a multivariate regression forest algorithm that is shown to be more accurate and efficient than competing registration-based techniques. At the core of the algorithm is a new information-theoretic metric for regression tree learning which works by maximizing the confidence of the predictions over the position of all organs of interest, simultaneously. Such strategy produces accurate predictions as well as meaningful anatomical landmark regions.

Accuracy and efficiency have been assessed on a database of 400 diverse CT studies as well as on a database of 33 2-channel MR Dixon sequences. Our algorithm for anatomy detection and localization in 3D CT scans has now been validated by the FDA and has been approved for commercial use.

In more academic settings, the usefulness of our algorithm has been demonstrated in the context of systems for efficient visual navigation of 3D CT studies [283] and robust linear registration [193]. Another application where regression forests have been used is automatic vertebrae localization in arbitrary field-of-view CT scans [133]. Here, the regression part is used to provide robust initialization for a subsequent localization refinement stage based on a shape and appearance model. Similarly, one could employ the organ localization as a first step in an organ segmentation approach. Organ-specific algorithms could then be applied at the predicted organ location, removing the often necessary step of manual interaction.