Combining Generative and Discriminative Models for Semantic Segmentation of CT Scans via Active Learning

Juan Eugenio Iglesias[1,2], Ender Konukoglu[2], Albert Montillo[3,2], Zhuowen Tu[1], and Antonio Criminisi[2]

[1] University of California, Los Angeles, USA
jeiglesias@ucla.edu,zhuowen.tu@loni.ucla.edu,
[2] Microsoft Research, Cambridge, UK
enderk@microsoft.com,antcrim@microsoft.com
[3] GE Global Research Center, Niskayuna, NY USA montillo@ge.com

Abstract. This paper presents a new supervised learning framework for the efficient recognition and segmentation of anatomical structures in 3D computed tomography (CT), with as little training data as possible. Training supervised classifiers to recognize organs within CT scans requires a large number of manually delineated exemplar 3D images, which are very expensive to obtain. In this study, we borrow ideas from the field of active learning to optimally select a minimum subset of such images that yields accurate anatomy segmentation. The main contribution of this work is in designing a combined generative-discriminative model which: i) drives optimal selection of training data; and ii) increases segmentation accuracy. The optimal training set is constructed by finding unlabeled scans which maximize the disagreement between our two complementary probabilistic models, as measured by a modified version of the Jensen-Shannon divergence. Our algorithm is assessed on a database of 196 labeled clinical CT scans with high variability in resolution, anatomy, pathologies, etc. Quantitative evaluation shows that, compared with randomly selecting the scans to annotate, our method decreases the number of training images by up to 45%. Moreover, our generative model of body shape substantially increases segmentation accuracy when compared to either using the discriminative model alone or a generic smoothness prior (e.g. via a Markov Random Field).

1 Introduction

Large field-of-view CT scans are widely used in the diagnosis of systemic diseases, which affect several organs. Automatic segmentation of body structures has application in anomaly detection, disease assessment, change tracking, registration, navigation, and further organ-specific analysis. In this study, we present an algorithm for simultaneous segmentation of nine anatomical structures in clinical CT scans: `heart`, `liver`, `spleen`, `l/r pelvis`, `l/r kidney` and `l/r lung`. The segmentation task is cast as voxel-wise classification. In clinical CT this is challenging due to similar density in different organs, presence of contrast agents,

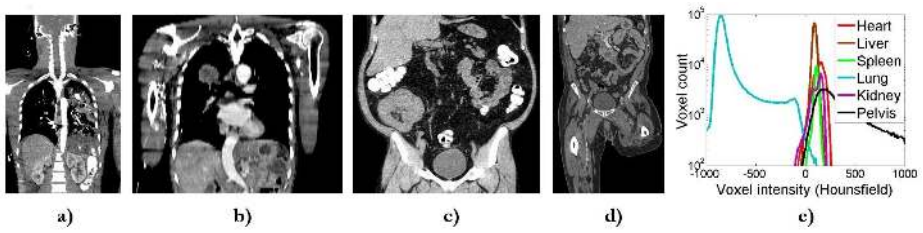

Fig. 1. Variability in our 196 clinical scans. a) Coronal view of a subject with pulmonary anomaly. b) Subject with lung tumor. c) Subject with oral contrast and abnormal kidney shape. d) Subject with amputated left leg. e) In a typical full-body scan, the tissue densities of different organs overlap considerably with one another. All these sources of variability make automatic anatomy segmentation challenging.

varying resolution and noise levels, variability in anatomy and small field of view (see Fig. 1). Thus, machine learning-based, supervised classifiers require a large amount of expensive, labeled data to generate high-quality segmentations. While that is the ultimate goal of our system, this study addresses the problem of finding the minimal sufficient training data within a pool of unlabeled CT scans. To do so we borrow ideas from the field of active learning.

Background. Active learning [1] studies the problem of training a robust supervised system with as little manual labeling as possible, i.e. which data, once labeled, yield the largest increase in accuracy? Most techniques are iterative. Starting from a pool of unlabeled data, a small set of samples is selected and labeled by a human expert. These enable training an initial classifier which in turn is used to select the most informative instances from the unlabeled pool, which often correspond to data that is under-represented in the current training set e.g. anatomical variations, pathologies, etc. Then, the expert provides ground truth for this new set and the classifier is updated, beginning a new iteration. Active learning suits the medical imaging domain well because unlabeled data are readily available whereas expert annotations are expensive.

There are two families of criteria for selecting samples to be annotated, based on: 1) maximizing a predefined informativeness measure (e.g. [2]); and 2) reducing the "version space" i.e. the set of all classification hypotheses consistent with the ground-truth. The most representative example of the latter is "query by committee" (QBC)[3], in which two or more classifiers sampled from the version space are used to classify an unlabeled sample. Manual labeling is then performed on the data for which the models outputs disagree. In co-testing [4], the members of the committee are trained on independent views of the data.

Contribution. Our framework combines two complementary models: a classifier that focuses on the appearance of the organs, and a generative model, which

captures organ relative location and thus global, probabilistic shape information. As in QBC and co-testing, the models are jointly trained iteratively. At each iteration, the classifier is tested on the unlabeled data and the output is evaluated under the generative model. The joint confidence of the probabilistic predictions is used to select the scans to be labeled next. As opposed to QBC and co-testing, our method: 1. focuses on increasing the prediction confidence of the two models; 2. does not require multiple views of the data; and 3. exploits long-range context via a generative model of anatomy. Additionally, we demonstrate how such shape prior increases the segmentation accuracy substantially.

Further related literature. Most approaches to joint segmentation of multiple structures rely on registration. Segmentation is achieved by deforming an atlas (or a set thereof) to the target scan. The atlas and deformation can be seen as a generative model of anatomy. Registration is computationally expensive, but works very well for organs such as the brain. However, it falters on structures whose relative locations can change substantially across subjects, which is the case in multi-organ segmentation. Post-processing is then required to refine the result [5–7]. On the other hand, Seifert et al. [8] make use of increasingly popular machine learning methods to approach the problem. They use a cascade of classifiers to detect key slices, then landmarks, and finally segment six organs.

2 Materials: the labeled CT database

Clinical CT scans from 196 different subjects acquired at different hospitals were used in this study. Most of the scans are challenging to segment due to variability in anatomy, acquisition protocols (resolution, filters, dose, etc), pathology, contrast agents, and even limb amputations (see Fig. 1). Most of the scans cover the thoracic or abdominal region, or both. Some also include the head or legs. Manual ground-truth segmentations were achieved via a semi-automatic approach [9]. All scans are resampled to $6mm$ isotropic resolution for faster calculations.

3 Methods

The proposed framework combines a discriminative and a generative model for obtaining high-quality segmentations and driving the selection of training images. Henceforth, we use the following notation: i indexes the scans in the training set, and $o \in \{1, 2, \ldots, N_o\}$ indexes the $N_o = 9$ organs of interest. Appending the `background` class produces an extended set with $N_c = N_o + 1$ classes, indexed by $c \in \{1, 2, \ldots, N_o, N_c\}$ (where $c = N_c$ corresponds to `background`). Subscripts d, g, and cog represent "discriminative", "generative" and "center of gravity".

3.1 Discriminative voxel classification

Following the work in [10], we have applied random forest classification [12] to the task of assigning organ class probabilities to all voxels of a previously unseen

CT scan. The classifier is based on box-shaped visual features. Details are out of the scope of this paper, and can be found in the referenced work. Any other type of probabilistic classifier could also be used in our framework.

3.2 Generative model of CT scans

Although random forests do capture some level of context, they fail at modeling the long-range spatial relationships between organs (see Fig.3b). We address this issue by introducing a generative graphical model which captures relative organ positions and organ shapes probabilistically.

The model. The graphical model used here is shown in Fig. 2a. Its relatively few parameters allow it to be trained with very few scans, making it possible to start selecting scans actively early in learning. The model represents a CT scan by a collection of organs and the background. Each organ is represented by its centroid location $\boldsymbol{k}_o$ and a probabilistic atlas of shape $A_o(\boldsymbol{r})$ such that the probability that the voxel at location $\boldsymbol{r}$ is inside the organ is $A_o(\boldsymbol{r} - \boldsymbol{k}_o)$.

There are two coordinate systems in the model: a reference frame in which the sets of centroids from the training dataset are jointly aligned and a physical coordinate system in which all CT scans are defined. The coordinates $\boldsymbol{k}'_o$ of the N_o centroids of a scan in the reference frame, stacked into a vector $\boldsymbol{x}'_{cog} = [\boldsymbol{k}'^t_1, \ldots, \boldsymbol{k}'^t_{N_o}]^t$, follow a multivariate Gaussian distribution: $\boldsymbol{x}'_{cog} \sim \mathcal{N}(\bar{\boldsymbol{x}}'_{cog}, \Sigma_{cog})$. These coordinates are mapped to the physical frame by a simple rigid transform $\boldsymbol{x}_{cog} = [\boldsymbol{k}^t_1, \ldots, \boldsymbol{k}^t_{N_o}]^t = s\boldsymbol{x}'_{cog} + \boldsymbol{t}$, where the scaling factor s is log-normal $\log s \sim \mathcal{N}(\overline{\log s}, \sigma^2_{\log s})$ and the translation $\boldsymbol{t}$ is improper uniform i.e. free. Finally, the organ probabilities for each of the V_{tot} voxels in the CT volume $\boldsymbol{p}_g = [p_{g,1}, \ldots, p_{g,Nc}]^t$ are obtained by integrating the organ atlases and the centroid model. Since the atlases represent binary tests "organ o vs. rest", $\boldsymbol{p}_g$ is given by the product rule (assuming independence between the atlases):

$$p_{g,o}(\boldsymbol{r}) = \frac{1}{Z_g(\boldsymbol{r})} A_o(\boldsymbol{r} - \boldsymbol{k}_o) \prod_{o'=1, o' \neq o}^{N_o} [1 - A_{o'}(\boldsymbol{r} - \boldsymbol{k}_{o'})], \quad o \in [1, \ldots, N_o]$$

$$p_{g,Nc}(\boldsymbol{r}) = \frac{1}{Z_g(\boldsymbol{r})} \prod_{o=1}^{N_o} [1 - A_o(\boldsymbol{r} - \boldsymbol{k}_o)] \quad \text{(for the background)}$$

where $Z_g(\boldsymbol{r})$ ensures that $\sum_c p_{g,c}(\boldsymbol{r}) = 1, \forall \boldsymbol{r}$. The background requires special treatment because it has neither a centroid nor a probabilistic atlas in the model.

Learning the model. The first step to learn the centroid distributions is to create the reference coordinate system in which the sets of centroids are aligned. For each training scan i, we have $\boldsymbol{x}'_{i,cog} = (\boldsymbol{x}_{i,cog} - \boldsymbol{t}_i)/s_i$, where the scalings s_i and translations $\boldsymbol{t}_i$ maximize the joint alignment of the centroid sets. This simple transform suffices to align the data because the scanned subject is always

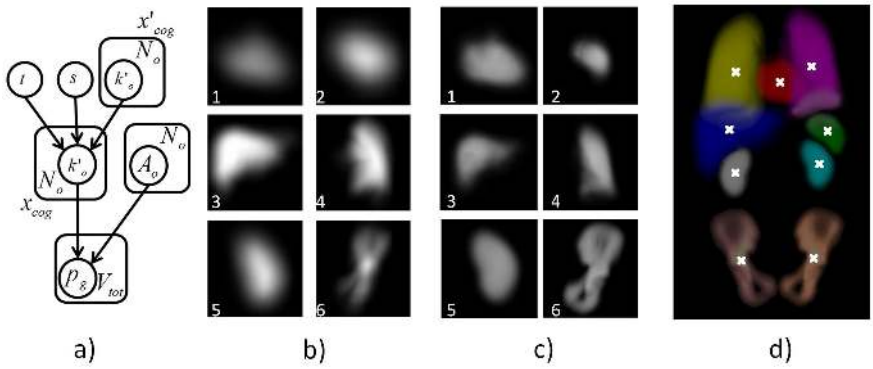

Fig. 2. Generative model: a) Graphical model. b) 3D rendering of probabilistic organ atlases, trained with 5 scans: 1.heart; 2.spleen; 3.liver; 4.left lung; 5.left kidney; and 6.left pelvis. c) Same atlases, trained on 20 scans. d) 3D rendering from a random sample of the complete model $\boldsymbol{p}_g$ (trained with 20 scans, centroids marked with crosses).

well aligned with the scanner bed. Moreover, the low number of parameters of the transform makes it very robust. The values of s_i and $\boldsymbol{t}_i$ are obtained through an iterative, maximal agreement Procrustes alignment algorithm[13]. Then, the parameters $\overline{\log s}$ and $\sigma^2_{\log s}$ are just the sample mean and variance of $\{\log s_i\}$.

From the coordinates in the reference frame $\boldsymbol{x}_{i,cog}$, we can estimate the parameters of the their multivariate Gaussian model. We use probabilistic principal component analysis (PPCA [14]) to deal with missing data (organs out of the field of view): $\boldsymbol{x}'_{cog} = \bar{\boldsymbol{x}}'_{cog} + \boldsymbol{\Phi}\boldsymbol{b} + \boldsymbol{\varepsilon}$, where Φ is the matrix with the orthonormal principal components, $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{diag}(\boldsymbol{\lambda}))$ is the shape vector, $\boldsymbol{\lambda}$ is a vector with the variances of the principal components, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\varepsilon}}\boldsymbol{I})$ is a noise vector with spherical covariance. The parameters $\bar{\boldsymbol{x}}'_{cog}$,$\Phi$,$\boldsymbol{\lambda}$,$\sigma^2_{\boldsymbol{\varepsilon}}$ are estimated by PPCA.

Finally, the probabilistic atlases are constructed independently for each organ as follows: 1. scaling each manually labeled organ mask by s_i, to be consistent with the Procrustes analysis; 2. aligning the centroids of the masks; 3. taking the voxel-wise average; and 4. blurring the result with a Gaussian kernel. Blurring accounts for the fact that the generative model is in general not perfectly aligned to the anatomy when testing an unseen test scan (see Section 3.3). Fig. 2b-c displays the probabilistic atlases of the organs of interest trained on different numbers of scans, in order to illustrate their evolution with the size of the training dataset. A random draw from the full generative model is displayed in Fig. 2d.

3.3 Segmenting previously unseen scans

This section describes how to achieve high-quality segmentations by combining the two models. This requires: 1. fitting the generative model to the classifier output; and 2. combining the models via Bayes' theorem.

1. The CT volumes are downsampled to 20mm isotropic resolution.
2. Assuming $\boldsymbol{b} = \mathbf{0}$ (mean centroid shape), do exhaustive search across s and t_z. We explore $t_z \in [t_{z,min}, t_{z,max}]$, with $t_{z,min}$ and $t_{z,max}$ such that translations with no overlap between the scan and the generative model are disregarded. For the scale s, we explore the interval $s \in [\exp(\overline{\log s} - 2.5\sigma_{\log s}), \exp(\overline{\log s} + 2.5\sigma_{\log s})]$. For each value of s, t_x and t_y are designed to match the x-y c.o.g. of the scan and the x-y c.o.g. of the centroids of the organs.
3. From the optimal point from the previous step, do coordinate descent on $\{s, \boldsymbol{b}, \boldsymbol{t}\}$, with $s \in [s_{min}, s_{max}]$ and $\boldsymbol{b}$ constrained to stay in the ellipsoid that embraces 95% of the probability mass of the Gaussian density function.
4. From the solution from the previous step, optimize the noise $\boldsymbol{\varepsilon}$ independently for each organ, searching a sphere of radius $\sqrt{\sigma_{\varepsilon}^2 [\chi_3^2]_{0.95}}$ around each centroid, where $[\chi_3^2]_{0.95}$ is the 95% quantile of a Chi-square distribution with 3 degrees of freedom.

Table 1. Algorithm to align the generative model with the output of the classifier.

Aligning the models. Given voxel-wise class probabilities from the forest classifier $\boldsymbol{p}_d(\boldsymbol{r}) = [p_{d,1}, \ldots, p_{d,N_c}]^t$, the parameter space of the generative model $(s, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{\varepsilon})$ is explored to maximize a similarity metric between the class distributions from the two models $\boldsymbol{p}_g(\boldsymbol{r})$ and $\boldsymbol{p}_d(\boldsymbol{r})$. Here, we minimize the Jensen-Shannon divergence $JS(P\|Q) = (1/2)[KL(P\|R) + (KL(Q\|R)]$, where $R = \frac{(P+Q)}{2}$ and $KL(P\|Q) = \sum_j P_j(\log P_j - \log Q_j)$ is the Kullback-Leibler divergence. Unlike KL, JS defines a bounded metric. The problem is then formulated as:

$$\boldsymbol{\theta}^* = \{s^*, \boldsymbol{b}^*, \boldsymbol{t}^*, \boldsymbol{\varepsilon}^*\} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{\boldsymbol{r}} JS(\boldsymbol{r}, \boldsymbol{\theta}) =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2V_{tot}} \sum_{\boldsymbol{r}} \sum_{c=1}^{N_c} \left(p_{g,c}(\boldsymbol{r}) \log \frac{2p_{g,c}(\boldsymbol{r})}{p_{g,c}(\boldsymbol{r}) + p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r})} + p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r}) \log \frac{2p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r})}{p_{g,c}(\boldsymbol{r}) + p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r})} \right) \quad (1)$$

where $\boldsymbol{\theta}^* = \{s^*, \boldsymbol{b}^*, \boldsymbol{t}^*, \boldsymbol{\varepsilon}^*\}$ are the optimal parameters and V_{tot} is the number of voxels in the scan. This step uses well-behaved probability maps (particularly $p_{g,c}^{\boldsymbol{\theta}}$, see Fig.2bc) rather than pixel intensities. This makes the JS divergence smooth, which, next to the low number of degrees of freedom of the generative model, makes the optimization fast and robust. First, exhaustive search at 20mm resolution is used to initialize the two most sensitive parameters: axial translation (t_z) and scale (s). From that point, coordinate descent is used to optimize $\boldsymbol{b}$ and refine $\boldsymbol{t}$ and s. Finally, the noise $\boldsymbol{\varepsilon}$ is optimized independently for each organ. The algorithm is detailed in Table 1 and illustrated in Fig. 3a-c.

Bayesian semantic segmentation. The aligned generative model can be interpreted as a location prior in a Bayesian framework. The posterior probability of label L at location $\boldsymbol{r}$ is therefore given by Bayes' theorem:

$$p[L(\boldsymbol{r}) = c] = \frac{p_{d,c}(\boldsymbol{r}) \cdot p_{g,c}^{\boldsymbol{\theta}^*}(\boldsymbol{r})}{Z_L(\boldsymbol{r})} \quad (2)$$

where the partition function $Z_L(\boldsymbol{r})$ ensures that the probabilities add to one. The final hard segmentation can be obtained as the voxel-wise MAP estimate of the

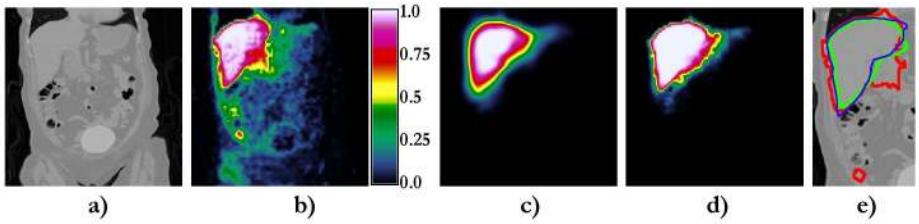

Fig. 3. Bayesian anatomy segmentation. a) Coronal slice of a sample test scan. b) Probability for the liver as output of the classifier (20 training scans). c) Aligned generative model. d) Posterior probability for liver as in Equation 2. The posterior map is much more accurate than in (b). e) Hard segmentations provided by the discriminative model alone (red) and the Bayesian model (green), as well as ground truth (blue). The generative spatial prior has a positive effect on the quality of the final segmentation (green). Please note that the segmentation (i.e. the MAP estimate) is not the 0.5-level of the probability map; that holds in binary classification, but not in our N_c-ary case.

class probabilities i.e. $\text{argmax}_c\ p[L(\boldsymbol{r}) = c]$. The effect of the generative model on the segmentation is illustrated in Fig. 3, where erroneous probability masses are removed from $p_{d,c}(\boldsymbol{r})$, improving the quality of the overall segmentation.

3.4 Training set construction through active learning

The second task at hand is that of building a minimal set of manually labeled CT images. Detailed algorithmic steps are presented in Table 2. In the spirit of QBC and co-testing, at each iteration an expert labels the scan which maximizes the disagreement between the discriminative and generative models, which is the one with largest potential of improving the overall segmentation accuracy. To remove the bias towards larger organs, desirable when aligning the models but detrimental in active learning, we use a modified version of the JS divergence to measure the disagreement. We adapt the concept of weighted entropy of a distribution [15]: $H_w = \sum_c u_c p_c \log p_c$, where class c has an utility cost u_c and probability p_c. Making the utilities inversely proportional to the average volumes of the organs $u_c = 1/V_c$ weighs them uniformly. We define $V_c = \sum_{\boldsymbol{r}} A_c(\boldsymbol{r}), c \in [1, \ldots, N_o]$, and V_{N_c} as the average volume of the background class in the training scans. Switching the order of $\sum_{\boldsymbol{r}}$ and $\sum_o$ in (1) gives the weighted JS divergence:

$$JS_w := \frac{1}{2} \sum_{c=1}^{N_c} \frac{1}{V_c} \sum_{\boldsymbol{r}} \left(p_{d,c}(\boldsymbol{r}) \log \frac{2p_{d,c}(\boldsymbol{r})}{p_{d,c}(\boldsymbol{r}) + p^{\boldsymbol{\theta}^*}_{g,c}(\boldsymbol{r})} + p^{\boldsymbol{\theta}^*}_{g,c}(\boldsymbol{r}) \log \frac{2p^{\boldsymbol{\theta}^*}_{g,c}(\boldsymbol{r})}{p_{d,c}(\boldsymbol{r}) + p^{\boldsymbol{\theta}^*}_{g,c}(\boldsymbol{r})} \right) \tag{3}$$

Another important component of the algorithm is the outlier rejection strategy. Here we identify as outliers unlabeled scans for which the JS_w measure is far away from the rest of the population using the local outlier factor (LOF [16]). LOF compares the density of the data around each point with the density at a number of nearest neighbors, computing an index which can be thresholded to detect outliers. Here, it is convenient to use an aggressive threshold (2.0) to be certain that all the outliers are detected, since the cost of leaving out informative inliers is lower than the toll of including outliers in the training set.

1. The generative and discriminative models are built starting with 2 labeled scans.
2. The remaining unlabeled scans are fed to the classifier, yielding multi-class probability maps for each voxel.
3. Align the generative model by minimizing JS in (1).
4. Compute disagreement via the weighted JS divergence as in (3).
5. Rejection of outlying scans via the local outlier factor (LOF) [16] on JS_w.
6. Select the unlabeled scan that maximizes JS_w and obtain its manual ground truth from a human expert.
7. Update the classifier and the generative model.
8. If the testing segmentation accuracy is satisfactory then stop. Otherwise, goto 2.

Table 2. Active learning for optimal construction of manually segmented database.

4 Experiments and results

Experimental setup. This section assesses two aspects of our work: i) the accuracy of our Bayesian segmentation approach versus the discriminative classifier alone; and ii) the validity of our database construction strategy as compared to alternative techniques. On the second task, five algorithms are evaluated:

A1. The proposed active learning approach.
A2. Same as 1, but randomly selecting scans from the unlabeled pool.
A3. Uncertainty sampling [2], in which the scan that maximizes the mean voxel entropy $H_{av} = (V_{tot}^{-1}) \sum_{\boldsymbol{r}} \sum_{c=1}^{N_c} p_{d,c}(\boldsymbol{r}) \log p_{d,c}(\boldsymbol{r})$ is selected. Our generative model is thus not used for data selection, but it is still used in segmentation.
A4. Same as 2, but the generative model is replaced by a generic Markov Random Field (MRF) prior in segmentation. Graph cuts [17] are used to minimize:

$$\xi[\Lambda(\boldsymbol{r})] = -\sum_{\boldsymbol{r}} \log p_{d,\Lambda(\boldsymbol{r})} + \gamma \sum_{(\boldsymbol{r}_i,\boldsymbol{r}_j)\in\mathcal{N}} \delta[\Lambda(\boldsymbol{r}_i) = \Lambda(\boldsymbol{r}_j)]$$

where $\mathcal{N}$ is a 6-neighborhood system and γ is the smoothness of the MRF.
A.5 Same setup as in 2, but without any generative model at all. For organ o, the hard segmentation is computed as the largest connected component of the binary volume $\text{argmax}_c\ p_{d,c}(\boldsymbol{r}) == o$ (i.e. the MAP estimate).

The following experiment was repeated 30 times and the results averaged: two scans (at least one with all the organs of interest) are randomly selected to form the initial training set. The remaining 194 are randomly split into unlabeled pool and test data. Then, unlabeled scans are iteratively added to the system according to the five algorithms, and the performance recorded using Dice's coefficient $D(A,B) = 2\frac{|A\cap B|}{|A|+|B|}$ and Haussdorff distance (i.e. maximal surface-to-surface distance). These two metrics provide complementary perspectives: gross overlap and robustness, respectively. In case of total miss of an organ, the Haussdorff distance for an algorithm is assumed to be equal to the maximal Haussdorff distance of the other tested methods for that organ in that subject.

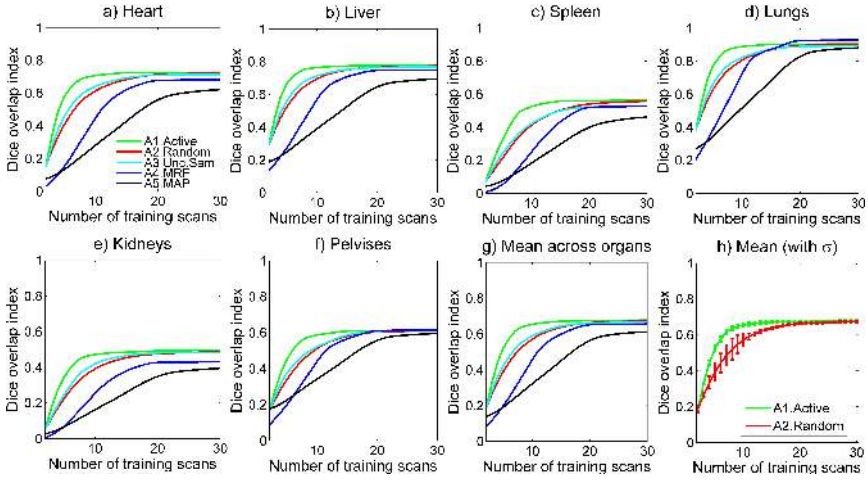

Fig. 4. Segmentation accuracy (Dice) vs. training set size for different database construction approaches. Plots a-f) are organ specific. Plot g) displays the mean for all organs. Plot h) is a zoom-in of f) that displays standard deviations (but only shows random selection and active learning for clarity).

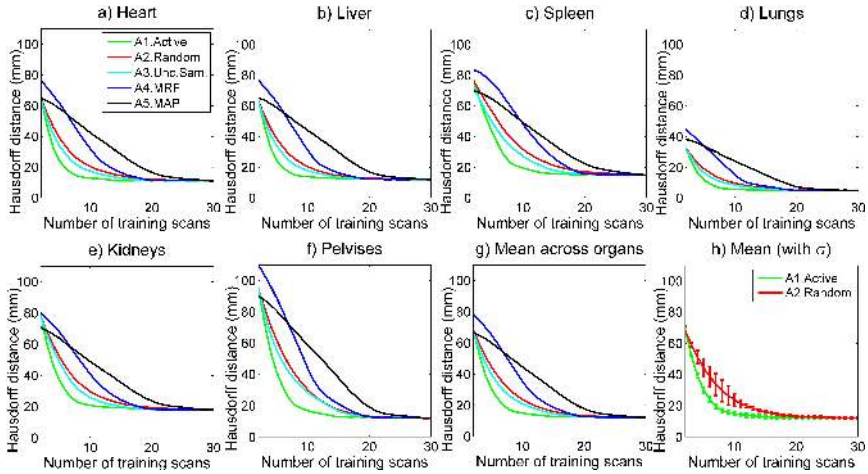

Fig. 5. Haussdorff distance vs. training set size for different database construction approaches. See caption of Figure 4 for the explanation of the plots.

The system parameters were set to the following values: $T = 14$ trees, max. tree depth $= 16$, features tested per node $= 200$, training voxels $= 4 \cdot 10^5$, min. info. gain $= 0.01$, width of kernel to blur prob. atlases $= 20 \cdot n_{atl}^{-3/2}$, MRF smoothness $\gamma = 1/3$, LOF threshold $= 2.0$, LOF nearest neighbors $= \lceil \frac{n_{unl}}{20} \rceil$. The random forest parameters were partially motivated by computational limitations. The rest of the parameters were tuned by pilot experiments.

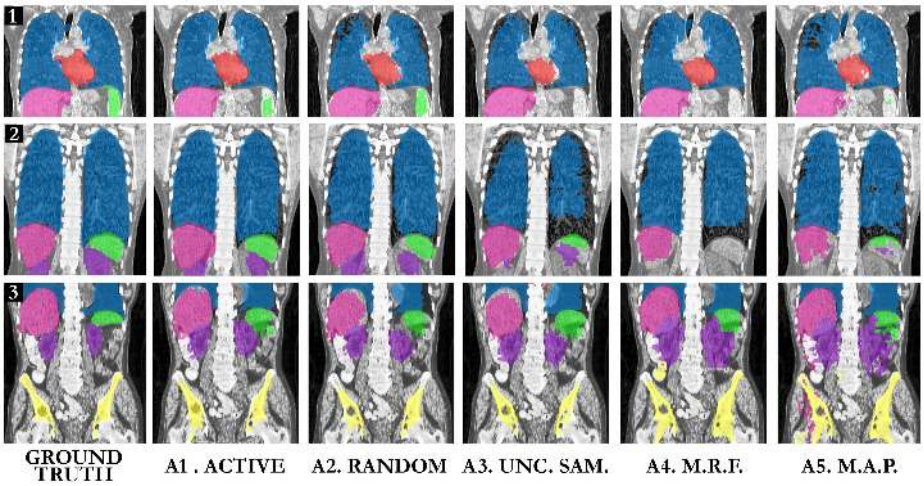

Fig. 6. Comparing segmentation results for different algorithms. Coronal slices of CT segmentations of three test scans using all compared approaches. Different colors indicate different segmented organs. Only ten labeled scans where used for training. Note how our algorithm (A1. Active) produces better segmentation alignment than other techniques, for the same training set size.

4.1 Bayesian segmentation results

Figures 4 and 5 display the Dice's coefficient and the Haussdorff distance vs. the number of training examples for the five algorithms, whereas Fig. 6 shows the segmentations of three test scans. The accuracy is not high in absolute terms due to the difficulty of the dataset (pathologies, large variations). However, the differences between active learning and random selection are still illustrated. Comparing curves A.2 and A.4, we see that our model outperforms the MRF for almost every organ, attributed to having been specifically designed for this task. For example, the MRF misses the spleen and kidneys in scans 1 and 2 in Fig. 6, whereas our method does not. Our model is also useful at the interface of neighboring organs, such as the liver and right kidney in scan 3 in the figure. The MRF only offers good results for the lung and pelvis, whose high image contrast makes discrimination possible with little contextual information. Finally, the MRF produces better segmentations than the direct MAP estimates, since the former is able to clean $\boldsymbol{p}_d(\boldsymbol{r})$ to some extent (e.g. the liver in scan 3 in Fig.6).

4.2 Training database construction results

Figures 4 and 5 also show that our framework reduces the number of labeled scans that are necessary to achieve a desired accuracy (Table 3). Active selection is consistently better than random for every organ. At higher accuracies (Dice $\geq$ 0.6), the decrease in labeling effort is 40-45%. The improvement is approximately

Target Dice's index	0.40	0.50	0.60	0.65	0.66	0.67
Number of training scans for A1. Active	3.6	4.7	8.2	9.4	11.1	14.3
Number of training scans for A2. Random	5.1	7.4	15.2	16.8	18.9	25.8
Number of training scans for A3. Unc.sam.	4.6	6.6	10.7	16.1	19.4	>30

Table 3. Required number of scans to achieve different levels of accuracy using our active learning framework, random selection and uncertainty sampling.

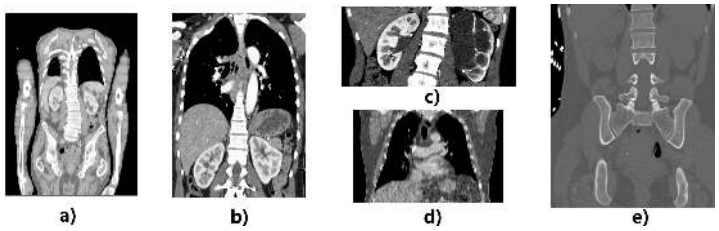

Fig. 7. Active selection of training scans. a-b) initial training set. c) Scan with minimal weighted JS score at the first iteration, which displays a kidney with a large cyst and is rejected by LOF. d-e) Scans actually selected in the first two iterations.

the same with respect to uncertainty sampling, which is marginally better than random selection when Dice ≤ 0.65 and marginally worse above 0.65.

Finally, Fig. 7 illustrates the scan selection process in our framework. The initial training set consists of the scans in Fig.7a-b. Fig. 7c shows the scan with the highest disagreement (because of the cyst in the left kidney and the lack of context due to reduced field of view). Adding it to the training set could negatively affect the performance of the system, therefore the importance of the outlier rejection stage. Fig. 7d-e displays the two first scans from the unlabeled pool actually selected by our algorithm. Next to the initial two, they represent the main types of scan present in the dataset (full body, chest, abdominal, pelvic).

5 Discussion and conclusion

A new framework for the automatic segmentation of anatomy within CT scans has been presented. A joint discriminative-generative model has been introduced to address two tasks: estimating voxel segmentation posteriors and constructing a minimal training dataset of manually segmented CT scans. Quantitative experiments demonstrated that the joint model considerably helps in: 1. attaining higher segmentation accuracy than generic smoothness priors (e.g. MRF); and 2. reducing the size of the training dataset by $\sim 45\%$ compared to alternatives.

In our Matlab / C# implementation, forest training takes $3-10$ min. on a desktop. Segmenting a new CT scan takes ≈ 4 s. (≈ 1 s for forest testing and ≈ 3 s for model alignment). In active learning, ranking all unlabeled scans takes $\approx 4N_{unlabeled}$ s. Next we wish to exploit the algorithm's parallelism to reduce

execution times further, combine the method with complementary work [11] to improve segmentation accuracy, and apply it to other image modalities, e.g. MR.

Acknoledgements

The authors would like to acknowledge Dr. S. Pathak and Dr. S. White for the data, and Dr. C. Rother and Dr. J. Wortman Vaughan for useful suggestion. J.E. Iglesias and Dr. Tu were partially supported by NSF career award IIS-0844566.

References

1. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
2. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Proc. ACM SIGIR Conf. Res. and Dev. in Inf. (1994) 3–12
3. Freund, Y., Seung, H., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Mach. Learn. **28**(2) (1997) 133–168
4. Muslea, I., Minton, S., Knoblock, C.: Active learning with multiple views. J. Artif. Intell. Res. **27**(1) (2006) 203–233
5. Linguraru, M., Sandberg, J., Li, Z., Pura, J., Summers, R.: Atlas-based automated segmentation of spleen and liver using adaptive enhancement estimation. In: Proc. of MICCAI. (2009) 1001–1008
6. Shimizu, A., Ohno, R., Ikegami, T., Kobatake, H., Nawano, S., Smutek, D.: Segmentation of multiple organs in non-contrast 3D abdominal CT images. Int. J. Comput. Assisted Radiol. and Surg. **2**(3) (2007) 135–142
7. Park, H., Bland, P., Meyer, C.: Construction of an abdominal probabilistic atlas and its application in segmentation. IEEE Trans. Med. Im. **22**(4) (2003) 483–492
8. Seifert, S., Barbu, A., Zhou, S., Liu, D., Feulner, J. Huber, M.S.M., Cavallaro, A., Comaniciu, D.: Hierarchical parsing and semantic navigation of full body CT data. In: Proc. of SPIE. Volume 7258. (2009) 725902–725909
9. Criminisi, A., Sharp, T., Blake, A.: Geos: Geodesic image segmentation. In: Proc. of ECCV. (2008) 99–112
10. Geremia, E., Menze, B., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N.: Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: Proc. of MICCAI. (2010) 111–118
11. Montillo, A., Shotton, J., Winn, J., Iglesias, J., Metaxas, D., Criminisi, A.: Entangled decision forests and their application for semantic segmentation of CT images. IPMI 2011, accepted for publication
12. Breiman, L.: Random forests. Mach. Learn. **45**(1) (2001) 5–32
13. Ten Berge, J.: Orthogonal Procrustes rotation for two or more matrices. Psychometrika **42**(2) (1977) 267–276
14. Tipping, M., Bishop, C.: Probabilistic principal component analysis. J. R. Stat. Soc.: Series B **61**(3) (1999) 611–622
15. Belis, M., Guiasu, S.: A quantitative-qualitative measure of information in cybernetic systems. IEEE Trans. Inf. Theory **14**(4) (1968) 593–594
16. Breunig, M., Kriegel, H., Ng, R., Sander, J., et al.: LOF: identifying density-based local outliers. Sigmod Rec. **29**(2) (2000) 93–104
17. Boykov, Y., Veksler, O., Zabih, R.: Efficient approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12) (2001) 1222–1239