HHS Public Access

Author manuscript

Phys Med Biol. Author manuscript; available in PMC 2017 September 07.

Published in final edited form as:

Phys Med Biol. 2016 September 7; 61(17): 6553–6569. doi:10.1088/0031-9155/61/17/6553.

Tissue segmentation of Computed Tomography images using a Random Forest algorithm: a feasibility study

Daniel F. Polan, M.S.E.[a,b], **Samuel L. Brady, Ph.D.**[b,1], and **Robert. A. Kaufman, M.D.**[b]

Daniel F. Polan: polandan@umich.edu; Robert. A. Kaufman: robert.kaufman@stjude.org

[a]Department of Nuclear Engineering and Radiological Sciences, University of Michigan, Ann Arbor, MI, USA

[b]Department of Diagnostic Imaging, St Jude Children's Research Hospital, Memphis TN, USA

Abstract

Current innovation in computed tomography (CT) is focused on radiomics, patient-specific radiation dose calculation, and image quality improvement using iterative reconstruction, all of which require specific knowledge of tissue and organ systems within a CT image. The purpose of this study was to develop a fully automated Random Forest classifier algorithm for segmentation of neck-chest-abdomen-pelvis CT examinations based on pediatric and adult CT protocols. Seven materials were classified: background, lung/internal air or gas, fat, muscle, solid organ parenchyma, blood/contrast enhanced fluid, and bone tissue using Matlab and the Trainable Weka Segmentation (TWS) plugin of FIJI. The following classifier feature filters of TWS were investigated: minimum, maximum, mean, and variance evaluated over a voxel radius of 2^n, (n from 0 to 4), along with noise reduction and edge preserving filters: Gaussian, bilateral, Kuwahara, and anisotropic diffusion. The Random Forest algorithm used 200 trees with 2 features randomly selected per node. The optimized auto-segmentation algorithm resulted in 16 image features including features derived from maximum, mean, variance Gaussian and Kuwahara filters. Dice similarity coefficient (DSC) calculations between manually segmented and Random Forest algorithm segmented images from 21 patient image sections, were analyzed. The automated algorithm produced segmentation of seven material classes with a median DSC of 0.86 ± 0.03 for pediatric patient protocols, and 0.85 ± 0.04 for adult patient protocols. Additionally, 100 randomly selected patient examinations were segmented and analyzed, and a mean sensitivity of 0.91 (range: 0.82–0.98), specificity of 0.89 (range: 0.70–0.98), and accuracy of 0.90 (range: 0.76–0.98) were demonstrated. In this study, we demonstrate that this fully automated segmentation tool was able to produce fast and accurate segmentation of the neck and trunk of the body over a wide range of patient habitus and scan parameters.

Keywords

CT; tissue segmentation; pediatrics

[1]Corresponding author: samuel.brady@stjude.org; 262 Danny Thomas Pl, Memphis, TN 38105; (T) 901-595-3927; (F) 901-595-3978.

I. Introduction

In the current era of technical development in diagnostic computed tomography (CT), three areas of research and innovation are prominent, namely: radiomics(Gillies *et al.*, 2016; Kumar *et al.*, 2012; Parmar *et al.*, 2014), patient-specific radiation dose calculation(Chen *et al.*, 2012b), and image quality assessment and improvement using iterative reconstruction(Solomon and Samei, 2014). Each area requires specific knowledge of tissue and organ systems within a CT image, which also requires an understanding of tissue and organ segmentation. The latter two demand challenging multi organ/tissue segmentation over a whole scan volume (e.g., chest, abdomen, and pelvis). Segmentation of CT images has a history within radiotherapy planning and computer-aided detection/diagnosis (CAD) software, but is generally specific to a localized region in the body (i.e., head and neck) (Chen *et al.*, 2012a) or a specific organ (e.g., lungs)(van Rikxoort and van Ginneken, 2013). Currently, a fully automated algorithm for multiple tissue and organ segmentation applied to the whole body or trunk is not available.

Tissue segmentation in CT is complicated by a variety of factors inherent to CT: similar gray scale levels between soft tissues and organs in the abdomen and pelvis, technique factor dependent image quality, organ shape, partial volume artifacts, quantum noise texture variability from reconstructed CT image-to-image and patient-to-patient, and finally, lack of a consistent method for tissue segmentation performance evaluation (Haas *et al.*, 2008; Memon *et al.*, 2008; Padma and Sukanesch, 2011). A variety of approaches to tissue and organ segmentation in a CT environment have been investigated, for example: shape analysis, atlas based localization, thresholding, edge detection, voxel-based texture analysis, artificial neural networks, region growing, deformable models, Markov random field models, morphological operations, and various deep learning methodologies (Pham *et al.*, 2000). However, each technique used for tissue and organ segmentation is largely specific to type of body part or end-point application for the segmentation algorithm (Sharma and Aggarwal, 2010). Additionally, the variation of body habitus from childhood into adulthood limits the more common approach to automated segmentation using shape analysis and atlas based localization. There is no universally accepted algorithm for segmentation of a whole body multi-organ/tissue system that spans the variation of body habitus from pediatrics to adulthood. In fact, different tissue and organ systems present their own specific limitations for any single tissue segmentation algorithm. Furthermore, most segmentation approaches are semi-automated and require *a priori* information and user manipulation, which may be more appropriate for localized regions of tissue and organ segmentation used in radiotherapy planning or CAD detection (Haas *et al.*, 2008). But with the need to process large CT exam data sets for quantitative diagnostic analysis (i.e., radiomics), semi-automated segmentation is too cumbersome and inefficient; the process of CT segmentation needs to be fully automated, without manual user input to make a meaningful impact in the current era of big data and technical development in diagnostic CT.

The purpose of this study was to investigate and optimize a Random Forest algorithm for fully automated multi-tissue and organ segmentation of pediatric and adult neck-chest-abdomen-pelvis CT examinations. Random Forest is a classification and regression tree (CART) decision analysis methodology, previously described in detail (Breiman, 1996,

2001). For Random Forest, the process is two parts: train then detect. To train using a CT image, data are randomly sampled from a test image and assigned to a class, e.g., voxels are sampled and labeled as bone, etc. Features of from each class are then extracted from the training data, e.g., the mean and variance of the voxel CT numbers sampled from voxels classified as bone and surrounding voxels. To train the data and grow a tree, split functions are identified that evaluate the image features from the training dataset and pass to a left or right branch of the tree. A random set of features are selected to grow a single tree; the random selection of features to evaluate the split functions both reduces training time and prevents over fitting of the data. The split functions are calculated to maximize the information gained per node (or point of branching), i.e., to produce the best split separating the class labels. The data are passed down the tree and the tree grows recursively from the root (starting point) to a terminal branch; the point at the end of the terminal branch is called a leaf. Detection is the result of a probability distribution function calculated at the leaf to classify the test sample. The forest is grown from the development of multiple trees, each created from randomized subsets of classification features.

A Random Forest algorithm is desirable for its training/classification, computational efficiency, probabilistic output, ability to handle a large variety of image input features, and iterative improvement based on error handling. Additionally, a Random Forest algorithm prevents over-fitting of the data by injecting randomness into the training of the trees, and combining the output of multiple random trees into the final classifier. In a head-to-head comparison of machine learning algorithms, Random Forest was shown to have a robust performance when compared using eight evaluation metrics (Caruana and Nicules-Mizil, 2006).

A Trainable Weka Segmentation (TWS) plugin of Fiji (Schindelin *et al.*, 2012), a Java-based image processing package that combined ImageJ (Schneider *et al.*, 2012) with open source plugins, was utilized to implement a Random Forest algorithm for tissue segmentation of CT images. TWS is investigated in this study primarily because it is an open source implementation of Weka that is both fast and free, it utilizes Random Forests for machine learning and data mining, and does not utilize atlas or deformable registration methodologies —which is important since a pediatric population significantly varies in size and internal anatomy (Hall *et al.*, 2009). The TWS plugin was first optimized for tissue segmentation in a CT environment. The following image features were evaluated for optimal tissue segmentation: minimum, maximum, mean, and variance were evaluated over a voxel radius of 2^n, with n ranging from 0 to 4, and the noise reduction and edge preserving filters, Gaussian, bilateral, Kuwahara, and anisotropic diffusion, were evaluated. Second, an estimate of differential tissue segmentation ground truth reproducibility was investigated based on an intra-observer study. Third, the optimized TWS Random Forests algorithm was applied to a sample of CT examinations acquired based on pediatric and adult CT protocols, and compared with the established ground truth.

II. Materials and Methods

In order to develop a fully automated differential tissue segmentation Random Forest algorithm for CT examinations, Fiji was utilized in this study. Fiji included two important

features which were utilized. First, the included MIJ package allowed a user to run ImageJ from MATLAB (ver. 2014b The MathWorks, Inc., Natick, MA), and allowed for the exchange of image volumes between MATLAB and ImageJ. In the development of a high-throughput automated segmentation tool, this link allowed for image pre- and post-processing in MATLAB while utilizing segmentation tools previously developed for ImageJ. The second important package included in the distribution is the TWS plugin (version 2.2.1). The TWS tool utilized the open-source machine learning software Weka (University of Waikato, Hamilton, New Zealand) to train a Random Forest classifier based on user input data selected from an image stack (Hall *et al.*, 2009). TWS uses an implementation of Breiman's Random Forests algorithm to develop a collection of random decision trees based on features of a bootstrap sampled training data set (Breiman, 1996, 2001). The TWS implementation of the Random Forest classifier used 200 trees with 2 features randomly selected per node.

II.A. Random Forests Segmentation

II.A.1. Training Data Set—A graphical user interface within the TWS plugin was used to select regions on an image stack and assign those regions to segmentation classes. Seven materials were classified in this study: background, lung/internal air or gas, fat, muscle, solid organ parenchyma, blood/contrast enhanced fluid, and bone tissue. Prior to training the classifier, image features were selected for training the selected segmentation classes. Available image features investigated in this study included voxel intensity statistics: minimum, maximum, mean, and variance, and different noise reduction and edge preserving filters such as: Gaussian, bilateral, Kuwahara, and anisotropic diffusion were also evaluated. Once the classifier was trained based on the user selected feature data, the classifier could be applied to other images.

The automated tissue segmentation method was developed and trained based on a contrast enhanced chest-abdomen-pelvis CT examination of a 20 year old male (72 kg) subsequently described in section II.B.2, [Fig 1]. Regions of voxels from each of the training data set's seven tissue segmentation classes were selected across the complete image volume. The training image volume consisted of 139 512x512 images. The number of voxels selected per segmentation class was: 21311 for background, 5863 lung/internal air or gas, 2512 for fat, 1874 for muscle, 9118 for solid organ parenchyma, 1547 for blood/contrast enhanced fluid, and 426 for bone.

II.A.2. Segmentation Algorithm Optimization—Optimization of the automated segmentation method was investigated using the following feature inputs: minimum, maximum, mean, and variance voxel values, where the voxel region of interest (ROI) was varied over 2^n radius with n ranging from 0 to 4; additionally, Gaussian, bilateral, Kuwahara, and anisotropic diffusion filters were tested to preserve the boundary edge of each tissue class while reducing the impact of stochastic image noise on the material class analysis and classification.

The TWS plugin did not include direct tools to evaluate and optimize each feature, the varying radii of the feature ROI, or noise and edge preserving filters for the accuracy and

specificity of the automated segmentation method in a CT image environment. Therefore, each of the features (i.e., minimum, maximum, mean, and variance voxel values) and noise and edge preserving filters (i.e., Gaussian, bilateral, Linear Kuwahara, and anisotropic diffusion) were tested quantitatively for Dice similarity coefficient (DSC) and out-of-bag (OOB) error. The DSC was used as a validation metric of spatial overlap (Dice, 1945; Zijdenbos *et al.*, 1994; Brock, 2014) of the manual and automated segmentation methods for each of the seven material classes separately: the DSC metric ranges from 0, indicating no overlap of the manual and automated material class segment, to 1, indicating complete overlap or perfect agreement. OOB error was used as an unbiased estimate of misclassification error and served as a surrogate of cross validation (Breiman, 1996; Wolpert and Macready, 1999). OOB was calculated from the approximately 33% of samples left out during the creation of each tree. Each of the 200 trees was created from different, with-replacement, bootstrap samples of the original data set. The randomly selected left-out-samples for each tree were considered “out-of-bag” relative to the corresponding tree. OOB error was then calculated by classifying the OOB samples using each tree in which that specific sample was not used during the training. Since 33% of samples are left out during the creation of each tree, an OOB classification was calculated for each sample in the data set using approximately 33% of the total number of trees. The percentage of resulting material class segmentations generated from the OOB classifications that did not match the true segmentation class for each sample (e.g., voxel values classified as bone but were actually muscle) gave the OOB error.

The training data set was processed using the intensity values (i.e., the CT number only) from the seven material classes to generate a baseline data set that compared input features from the feature filters. The baseline input was a single voxel intensity value taken from each of the seven material classes. Then the baseline input, using CT numbers only, was combined with each individual input feature filter and processed separately as a two-feature segmentation (e.g., CT number and variance, etc.). Each of the features from feature filters and the Gaussian filter were sampled from the input data at various ROI radii (ROI radii varied from 2^n, where $n = 0$ to 4). The segmented seven material classes from the baseline and two-feature data sets, at all 2^n ROI radii dimensions, were compared with the manually segmented training data set at the three axial locations demonstrated in [Fig. 1(a)].

The bilateral, anisotropic diffusion, and Kuwahara noise reduction, edge preserving filters were evaluated. TWS implementation of the bilateral filter formed an edge preserving mean filter using kernel radii of 5, 10, and 20 voxels, which included voxel value ranges of ±50 and ±100. The anisotropic diffusion filter had several fixed parameters that included the number iterations set at 20, the diffusion limiters along minimal variations set at 0.10 and 0.35, and the diffusion limiter along maximal variations set at 0.9. The number of smoothings per iteration was controlled with numerical user input as kernel radius. Therefore, for the anisotropic diffusion filter, the number of smoothings per iteration was varied from 2^n, where $n = 0$ to 4, and the edge threshold, defining the minimum voxel value difference that was preserved by the filter, was controlled by an independent user input. This input was varied from 10 to 100 in increments 10. The Kuwahara filter provided 3 features per linear voxel-based criteria, namely: variance, variance/mean and variance / mean^2, and

was investigated using a linear kernel length that varied from 3 to 35 voxels in increments of 4, and sampled the image at fixed 30^{deg} angles.

In addition to OOB and DSC calculations, a visual inspection of the quality of the two-feature segmentation and degree of mis-segmentation within each segmentation class was performed, and segmentation time per image slice was noted. Once a set of useful features was determined, the classifier was trained and applied to the remaining six patients.

II.B. Quantitative Evaluation Comparing Manual Segmentation

II.B.1 Patient Examinations—Our institutional review board deemed this quality analysis study to be exempt from informed consent. All data were managed in compliance with the Health Insurance Portability and Accountability Act. Seven anonymized CT examinations were selected at random from the institution's CT database and used in this segmentation study. All CT scans were performed with Lightspeed VCT XTe CT scanner (GE Healthcare, Waukesha, WI), using the appropriate CT protocol selected based on patient weight. All patients received oral and IV contrast enhanced chest-abdomen-pelvis CT scans, with one including the neck. Of the seven patients, four patients, median age 19 yr (range 16–21 yr), median weight 71 kg (range 57–81 kg), were considered adults for the purpose of protocol selection. The remaining three patients, median age yr (range 3–16 yr), median weight 30 kg (range 19–54 kg), were scanned using pediatric protocols.

II.B.2 Manual Segmentation Comparison—Manual segmentation served as a quantitative ground truth for evaluation of the TWS automated tissue segmentation method. Using Eclipse treatment planning software (Varian Medical Systems, Palo Alto, CA), the seven segmentation classes were manually segmented. Three slices were contoured per CT examination in a method directed by an experienced radiologist (RAK, 7 years' experience). For each patient, this included reconstructed image locations at the aortic arch, upper-liver, and the immediately above the first appearance of the iliac crests in axial format. These selected reconstructed image locations allowed for evaluation of three distinct portions of a chest-abdomen scan, and comparison of the seven segmentation classes in different anatomical environments (i.e., patterns of fat are different in the upper thorax vs. the abdomen and pelvis).

To evaluate accuracy of the tissue class segmentation specificity and the reproducibility of the manual segmentation, an intra-observer reproducibility study was performed 15 days apart on two segmented patients. To account for variability in anatomy between adult and pediatric sized patients, one patient was selected from the adult scan protocol group and one from the pediatric scan protocol group. The adult scan protocol patient examination was represented by a 20 year old male (72 kg) who was administered 143 ml of intravenous iodixanol 270 contrast, and 16 ml of 1.5% solution oral iohexol 300 contrast mixed in 355 ml (12 oz) of diluent. Total patient exam dose length product (DLP) was 528.2 mGy cm with chest series $CTDI_{vol}$ of 4.92 mGy (4.18 mGy SSDE) and with abdomen-pelvis series $CTDI_{vol}$ of 7.26 mGy (8.64 mGy SSDE). The pediatric scan protocol patient examination was represented by a 3 year old female (19 kg) who was administered 37 ml of intravenous iodixanol 270 contrast and 4 ml of 1.5% solution oral iohexol 300 contrast mixed in 89 ml (3

oz) of diluent. Total patient exam dose length product (DLP) was 128.78 mGy cm with chest series $CTDI_{vol}$ of 1.29 mGy (2.05 mGy SSDE) and with abdomen-pelvis series $CTDI_{vol}$ of 2.46 mGy (4.53 mGy SSDE).

II.C. Evaluation of Implemented Automated Segmentation Algorithm

II.C.1. Patient Examinations—To determine the accuracy, sensitivity, and specificity of the implemented Random Forest algorithm, 100 randomly selected patient examinations were evaluated. The median age of the patient population evaluated was 13 years (range 0.4–26 years), and median weight was 53 kg (range 5–112 kg). All examinations were performed with the same Lightspeed VCT XTe CT scanner used for manual segmentation comparison. The 100 randomly selected studies represented a combination of neck, chest, abdomen, and pelvis studies. Sixty three patients had oral and IV contrast enhanced CT scans, and 37 were non-contrast enhanced CT studies.

II.C.2. Confusion Matrix Comparison—To evaluate the 100 CT examinations a confusion matrix was derived from a qualitative evaluation of the segmented tissue classes. A comparison tool was developed using MATLAB. The tool imported both the anonymized CT images, and the seven segmented class layers (e.g., lung, fat, muscle, etc.). Any combination of segmented class layers could be overlaid with the grayscale CT images at one time. The transparency of the overlapping layers could be controlled; additionally, an outline of the layer could be used instead of overlapping with a filled transparency layer. To evaluate each segmented image, the observer had full control of image zoom, pan, window width, window level, CT voxel analysis (i.e., the CT number of any voxel could be identified), and image scrolling (i.e., the observer could scroll forward, backwards, or jump to any specific image slice through the image stack). Each segmented class was evaluated based on a 100 point scale ranging from 0 to 1.

To derive the confusion matrix, each segmented class was assigned a qualitative score based on the following definitions of a true positive (TP)—segmented class layer overlapping the correct corresponding grayscale tissue, false negative (FN)—grayscale tissue corresponding to the segmented class layer of tissue but not overlaid with the segmented class, false positive (FP)—segmented class layer overlaying non corresponding grayscale tissue, and true negative (TN)—non segmented class layer correctly not overlapping non-corresponding grayscale tissue. From each confusion matrix derived from each segmentation class the following metrics of analysis were derived, sensitivity:

$$\text{Sensitivity}=\frac{\sum TP}{\sum(TP+FN)}; \quad \text{(Eq. 1)}$$

specificity:

$$\text{Specificity}=\frac{\sum TN}{\sum(TN+FP)}; \quad \text{(Eq. 2)}$$

and accuracy:

$$\text{Accuracy}=\frac{\sum(TP+TN)}{\sum(TP+FN+TN+FP)}. \quad \text{(Eq. 3)}$$

III. Results

III.A. Segmentation Algorithm Optimization

Baseline testing of the automated TWS segmentation tool was established using the CT number alone as an input feature and resulted in an OOB error of 1.5% with DSCs of 1.00, 0.97, 0.80, 0.80, 0.74, 0.29, 0.45, and 0.53 for background, lung/internal air or gas, fat, muscle, solid organ parenchyma, blood/contrast enhanced fluid, bone tissue, and combined high contrast material classes, respectively. OOB error, [Fig. 2(a)], for each of the features and Gaussian blur filter in the two-feature tests showed a strong exponential decrease, eventually becoming asymptotic with respect to increasing kernel size. The main inflection point where increasing kernel radii return diminishing improvement in OOB error was for a kernel radius of 4.

For tested features, the DSC of the background segmentation averaged 0.99; therefore, it was excluded from optimization of the segmentation algorithm. Average DSC of the remaining six material segmentation classes was normalized by the baseline, single CT voxel test, for the two-feature tests and showed that, with exception of the minimum feature, the DSC increased relative to the single CT voxel test when using input features at kernel radii of 1, 2, and 4 voxels, [Fig 2(b)]. Since the minimum feature filter did not follow this trend, it was removed as an input.

Visual evaluation of the resulting segmentations for the maximum, mean, and variance texture input features showed that kernel sizes of radii larger than 4 voxels produced large scale segmentation misclassification. For the maximum feature filter input, radii of 8 [Fig 3(b)] and 16 [Fig 3(c)] produced circular spots in the segmentation corresponding to the kernel size. For the mean texture feature filter input, radii of 8 [Fig 3(d)] and 16 [Fig 3(e)] produced CT number intensity gradients that were classified as multiple different material classes. This was due to the mean values of different material types (i.e., CT number values) calculated across boundary interfaces; e.g., the liver/lung interface was classified as soft tissue, muscle, fat, and lung based on the decreasing CT number averaged at the liver/lung interface. For the variance feature filter input, radii of 8 [Fig 3(f)] and 16 [Fig 3(g)] produced less homogeneous segmentations due to the sensitivity of greater intra-tissue CT number variance.

The two-feature test of the bilateral filter had a low OOB error of 0.2% and an average DSC of 0.83 (range: 0.68–0.96) for the six material segmentation classes, but the resulting segmentation had large scale misclassification errors on some slices not included in the three slice DSC calculation, and was not used for classification and segmentation. The linear Kuwahara filter had an optimum linear kernel length of 19 voxels, resulting in an out-of-bag error of 0.2% and an average DSC of 0.83 (range: 0.65–0.97) for the six material classes.

Visually, this segmentation preserved edges and reduced noise well but led to small linear noise due to the linear kernel. This minimal linear noise was deemed acceptable due to the improved DSC relative to the baseline. The ideal edge threshold parameter of the anisotropic diffusion filter was found to be 50 (OOB of 0.2%); however, the TWS parameter controlling smoothing per iteration is based on the same user input as the selected kernel radii. The anisotropic diffusion filter was therefore fixed at 1, 2, and 4 smoothing per iteration, which did not lead to improvement in the OOB error, but increased segmentation time per CT image substantially. For this reason, the anisotropic diffusion filter was not included in the final segmentation classifier.

The final selected texture input feature filters, in addition to the original CT number, included maximum, mean, and variance, with kernel radii of 1, 2, and 4. Additional selected features included the Gaussian blur filter with kernel radii of 1, 2, and 4, and the Linear Kuwahara filter with a linear kernel of 19 voxels. This resulted in a total of 16 image features. Average segmentation time was 3.4 sec/slice (Intel® Xeon® CPU E5-1650 v3 @ 3.50GHz, 16GB RAM) with 100% CPU utilization and 12.5 MB RAM/slice. The fully classified and segmented training data set is shown in [Fig 4], as compared to the original CT images, [Fig 1]; the data set contained 139 slices and required 7 minutes for total segmentation time. In comparison, average pediatric chest, abdomen, and pelvic CT examinations contained ~80 slices and required ~4–5 minutes for segmentation.

III.B. Manual Segmentation Comparison

The DSCs for the intra-observer manual segmentation comparison of the adult imaging protocol patient was found to be 0.99 for background, 0.98 lung/internal air or gas, 0.90 fat, 0.94 muscle, 0.97 solid organ parenchyma, 0.88 blood/contrast enhanced fluid, 0.91 bone tissue, and 0.91 combined high contrast material classes (blood/contrast enhanced fluid and bone tissue), respectively. The pediatric imaging protocol patient had similar manual segmentation results with DSCs of 0.99, 0.97, 0.85, 0.90, 0.97, 0.87, 0.91, and 0.89, respectively.

A summary of the automated segmentation compared to the manual contouring is shown in Table I. These results are separated into average DSC for each of the segmentation classes across the three manually contoured slices and are stratified into results for the training patient, adult imaging protocol patients, and pediatric imaging protocol patients. The average DSCs over all patients also is included. A combined, not averaged, calculation of the DSC over the three slices is included to represent an estimate of the DSC of each segmentation class over the complete imaging volume. Since misclassification between blood/contrast enhanced fluid and bone was noted, both segmentation classes were combined into an additional combined high contrast region for segmentation evaluation.

III.C. Confusion Matrix Comparison

The confusion matrix results for the seven classes are shown in Table 2. Similar to Table 1, the high contrast regions, blood/contrast enhanced liquid and bone were combined and separately calculated for sensitivity, specificity, and accuracy. Since the background score was 1.0 across all categories, it was removed from further analysis. When considering the

combined high contrast regions as one class, the mean sensitivity, specificity, and accuracy across 100 patients was demonstrated to be 0.91 (range: 0.82–0.98), 0.89 (range: 0.70–0.98), and 0.90 (range: 0.76–0.98), respectively.

IV. Discussion

The purpose of this study was to develop an automated tissue segmentation algorithm optimized for CT using the Random Forest statistical classifying concept; to this end, the Fiji-based TWS plugin, in conjunction with MATLAB, was investigated. Generally, tissue segmentation for CT has been broadly investigated using a variety of segmentation approaches (Fortunati *et al.*, 2013; Fritscher *et al.*, 2014; Gao *et al.*, 1996; Heimann *et al.*, 2009; Hu *et al.*, 2001; Koss *et al.*, 1999; van den Boom *et al.*, 2012). The Random Forest algorithm within the TWS plugin has previously been utilized in a wide-range of imaging modalities including magnetic resonance imaging (MRI) and micro-CT (Chyzhyk *et al.*, 2013; Kulinowski *et al.*, 2011; Macdonald and Shefelbine, 2013). To our knowledge, no machine-learning based automated CT tissue segmentation tool has been developed using TWS, and no previous studies have been reported on patient CT tissue segmentation optimized for TWS.

The results of this feasibility study were derived from a sample of pediatric and adult CT examinations. An initial comparison of three regions within seven automatically segmented patient examinations was compared against the same 21 manually segmented images. Manual segmentation in conjunction with an intraobserver study was deemed to be the most robust method for establishing ground truth for tissue and organ segmentation. The intraobserver study was conducted to provide context for interpretation of the DSC values reported in this study.

Following the feasibility study, the automated segmentation algorithm was implemented into the patient examination workflow at St Jude Children's Research Hospital, namely: at the conclusion of each patient CT examination, all image series are archived within PACs for evaluation by a radiologist, and a separate copy is anonymized by stripping protected health information (PHI) and sent to the segmentation server. To date over 700 patient examinations have been segmented using the methodologies described in this work. As a further estimate of sensitivity, specificity, and accuracy of the Random Forest class segmentation, 100 patients were selected at random and visually analyzed for segmentation agreement with the grayscale anatomy in the original reconstructed images. The results of this study suggest that the Random Forest implementation in TWS can robustly segment seven material classes (background, lung/internal air or gas, fat, muscle, solid organ parenchyma, and bone and blood/contrast enhanced fluid combined) from CT imaging examinations of varying patient sizes (i.e., children and adults from 5 months to 26 years) and scan protocols (i.e., neck, chest, abdomen, and pelvis). Both metrics of segmentation accuracy, DSC and confusion matrix, show that the algorithm performs well for all material classes, excluding individualized high contrast regions. High contrast regions including blood/contrast enhanced fluid and bone overlap in CT number range. Furthermore, contrast enhancement is not uniform throughout the body with vascular contrast dependent on bolus timing and gastrointestinal opacification dependent on transit time and patient compliance.

However, when high contrast regions are combined into one class, the average DSC for all material classes is above 0.86 ± 0.04 (range 0.77–0.99) and the algorithm's average measure of sensitivity is 0.90 and specificity is 0.99.

It is important to note that classification training was performed in the presence of intravenous and oral contrast. Subsequent tissue segmentation of the solid organ parenchyma classification is most accurate when intravenous contrast is administered to the patient. In the future, a separate training classifier needs to be investigated for non-contrast CT examinations to more accurately segment this patient population. Additionally, for solid organ DSC values presented in Table I, reconstructed images 1 and 3, corresponding to locations at the aortic arch and immediately above the iliac crests, respectively, no DSC was calculated. This results from no manual segmentation of solid organ parenchyma on these reconstructed images. Furthermore, the DSC of solid organ parenchyma across the combined three slices shows a decrease in spatial overlap when compared to slice 2 (upper-liver/lower lung). The automated segmentation algorithm classified cartilage, ligament, joint fluid, bone marrow, and the trailing edge of bony ossification around the bony anatomy in the thorax and pelvic regions as solid organ parenchyma, [Fig. 5(a–b)]. Similarly, partial volume artifacts from axial imaging of an oblique surface near air/tissue interfaces are misrepresented as fat tissue due to the lower CT number intensity, which lowers the DSC result when comparing manual and automated segmentations, [Fig. 5(c–d)]. Considering only voxel intensity and not location or voxel pattern/geometry are current limitations of the automated segmentation tool as implemented.

Another known misclassification is water/fluid in the body, [Fig. 5(e–f)]. At the time of training the TWS algorithm, water/fluid was not included in the classification training, and as such is routinely classified as muscle since the CT number intensity of water is between fat and solid organ; as an example: the average CT number for fat in the sampled patient population is −109 ± 39 HU, for muscle is 69 ± 49 HU, for water/fluid is 5 ± 32 HU, and for solid organ parenchyma (as measured with IV contrast enhancement at late portal venous phase) is 117 ± 56 HU. Thus, all CT number values between fat and solid organ are classified as muscle. To limit tissue misclassification, a more granular training of different tissue types needs to be investigated.

Optimization of this algorithm is limited within the version of TWS used in the study (version 2.2.1). In this version, the TWS tool does not include direct methods of estimating feature importance, such as calculating the increase in classifier error for each variable resulting from permutations in OOB observations. The TWS tool in this study used a fixed Random Forest classifier; the number of trees and leaf size of the classifier ensemble were not optimized. Additionally, the creation of voxel features for training and classification in TWS is limited to minimum, maximum, mean, and variance of the region of voxels. Ideally, features for CT images would ignore outliers in the ROI around the voxel in which the feature is being calculated. This would aid in preserving tissue edges while not disrupting the segmentation within a tissue. The bilateral, and anisotropic diffusion noise-reduction filters have fixed parameters and could not be optimized for segmentation of CT images, but were likely optimized for microscopy image segmentation, the original purpose of the TWS tool. For CT segmentation, further optimization may be useful in increasing the efficacy of

these filters. Future improvements of this automated segmentation tool will likely rely on the inclusion of more image based features, such as 3D filters—in addition to current filter optimization, more granular tissue classification, spatially-based feature calculations (Pham *et al.*, 2000; Karssemeijer *et al.*, 1988), and combining additional classification algorithms for intelligent misclassification exclusions.

V. Conclusion

In this study, manual segmentation served as the ground truth for evaluation of automated segmentation. Ground truth reproducibility, determined via an intra-observer study and evaluated using DSCs, was found to be 0.94 (range: 0.90–0.99) for adult imaging protocol patients and 0.92 (range: 0.85–0.99) for pediatric imaging protocol patients across seven segmentation classes, background, lung/internal air or gas, fat, muscle, solid organ parenchyma, blood/contrast enhanced fluid, and bone. The optimized auto-segmentation tool included 16 image features calculated using maximum, mean, variance, and Gaussian blur filters with kernel radii of 1, 2, and 4 voxels, in addition to the original CT number and linear Kuwahara filter with a linear kernel of 19 voxels. Overall, the developed automated segmentation tool was found to produce fast and accurate segmentation of the seven material classes with an average DSC of 0.86 ± 0.04 (range: 0.81–0.99) and mean sensitivity of 0.91 (range: 0.82–0.98), specificity of 0.89 (range: 0.70–0.98), and accuracy of 0.90 (range: 0.76–0.98).

Acknowledgments

The authors would like to acknowledge Wilburn “Gene” Reddick, PhD for his advice and expertise. This work was partially funded by American Lebanese Syrian Associated Charities (ALSAC) and the National Cancer Institute (NCI) R25E Grant 5R25CA23944.

References

Breiman L. Bagging predictors. Machine Learning. 1996; 24:123–140.

Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

Brock, K., editor. Image Processing in Radiation Therapy. Boca Raton, FL: CRC Press; 2014.

Caruana, R.; Nicules-Mizil, A. Proceedings of the 23rd International Conference on Machine Learning; Pittsburgh, PA. 2006. vol. Series

Chen A, Niermann KJ, Deeley MA, Dawant BM. Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. Physics in medicine and biology. 2012a; 57:93–111. [PubMed: 22126838]

Chen W, Kolditz D, Beister M, Bohle R, Kalender WA. Fast on-site Monte Carlo tool for dose calculations in CT applications. Med Phys. 2012b; 39:2985–2996. [PubMed: 22755683]

Chyzhyk D, Ayerdi B, Maiora J. Active Learning with Bootstrapped Dendritic Classifier applied to medical image segmentation. Pattern Recognition Letters. 2013; 34:1602–1608.

Dice L. Measures of the amount of ecologic association between species. Ecology. 1945; 26:297–302.

Fortunati V, Verhaart RF, van der Lijn F, Niessen WJ, Veenland JF, Paulides MM, van Walsum T. Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling. Med Phys. 2013; 40:071905. [PubMed: 23822442]

Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Med Phys. 2014; 41:051910. [PubMed: 24784389]

Gao L, Heath DG, Kuszyk BS, Fishman EK. Automatic liver segmentation technique for three-dimensional visualization of CT data. Radiology. 1996; 201:359–364. [PubMed: 8888223]

Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016; 278:563–577. [PubMed: 26579733]

Haas B, Coradi T, Scholz M, Kunz P, Huber M, Oppitz U, Andre L, Lengkeek V, Huyskens D, van Esch A, Reddick R. Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies. Physics in medicine and biology. 2008; 53:1751–1771. [PubMed: 18367801]

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. 2009; 11:10–18.

Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, Bello F, Binnig G, Bischof H, Bornik A, Cashman PM, Chi Y, Cordova A, Dawant BM, Fidrich M, Furst JD, Furukawa D, Grenacher L, Hornegger J, Kainmuller D, Kitney RI, Kobatake H, Lamecker H, Lange T, Lee J, Lennon B, Li R, Li S, Meinzer HP, Nemeth G, Raicu DS, Rau AM, van Rikxoort EM, Rousson M, Rusko L, Saddi KA, Schmidt G, Seghers D, Shimizu A, Slagmolen P, Sorantin E, Soza G, Susomboon R, Waite JM, Wimmer A, Wolf I. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging. 2009; 28:1251–1265. [PubMed: 19211338]

Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Trans Med Imaging. 2001; 20:490–498. [PubMed: 11437109]

Karssemeijer N, van Erning LJ, Eijkman EG. Recognition of organs in CT-image sequences: a model guided approach. Computers and biomedical research, an international journal. 1988; 21:434–448.

Koss JE, Newman FD, Johnson TK, Kirch DL. Abdominal organ segmentation using texture transforms and a Hopfield neural network. IEEE Trans Med Imaging. 1999; 18:640–648. [PubMed: 10504097]

Kulinowski P, Dorozynski P, Mlynarczyk A, Weglarz WP. Magnetic resonance imaging and image analysis for assessment of HPMC matrix tablets structural evolution in USP Apparatus 4. Pharmaceutical research. 2011; 28:1065–1073. [PubMed: 21181545]

Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ. Radiomics: the process and the challenges. Magnetic resonance imaging. 2012; 30:1234–1248. [PubMed: 22898692]

Macdonald W, Shefelbine SJ. Characterising neovascularisation in fracture healing with laser Doppler and micro-CT scanning. Medical & biological engineering & computing. 2013; 51:1157–1165. [PubMed: 23881721]

Memon NA, Miraz AM, Gilani SAM. Deficiencies of Lung Segmentation Techniques using CT Scan Images for CAD. International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2008; 2:2764–2769.

Padma A, Sukanesch R. Automatic Classification and Segmentation of Brain Tumor in CT Images using Optimal Dominant Gray level Run length Texture Features. International Journal of Advanced Computer Science and Applications. 2011; 2:53–59.

Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, Mitra S, Shankar BU, Kikinis R, Haibe-Kains B, Lambin P, Aerts HJ. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PloS one. 2014; 9:e102107. [PubMed: 25025374]

Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. Annu. Rev. Biomed. Eng. 2000; 02:315–337. [PubMed: 11701515]

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. Fiji: an open-source platform for biological-image analysis. Nature methods. 2012; 9:676–682. [PubMed: 22743772]

Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature methods. 2012; 9:671–675. [PubMed: 22930834]

Sharma N, Aggarwal LM. Automated medical image segmentation techniques. Journal of medical physics / Association of Medical Physicists of India. 2010; 35:3–14. [PubMed: 20177565]

Solomon J, Samei E. Quantum noise properties of CT images with anatomical textured backgrounds across reconstruction algorithms: FBP and SAFIRE. Med Phys. 2014; 41:091908. [PubMed: 25186395]

van den Boom, R.; Oei, MTH.; Lafebre, S.; Oostveen, LJ.; Meijer, FJA.; Steens, SCA.; Prokop, M.; van Ginneken, B.; Manniesing, R. Medical Imaging 2012: Image Processing. San Diego, CA, USA: Proc. SPIE; 2012. Brain tissue segmentation in 4D CT using voxel classification.

van Rikxoort EM, van Ginneken B. Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review. Physics in medicine and biology. 2013; 58:R187–R220. [PubMed: 23956328]

Wolpert D, Macready W. An Efficient Method To Estimate Bagging's Generalization Error. Machine Learning. 1999; 35:41–55.

Zijdenbos A, Dawant B, margolin R, AC P. Morphometric analysis of white matter lesions in MR images. IEEE Trans Med Imaging. 1994; 13:716–724. [PubMed: 18218550]

Novelty & Significance

There is a need for robust, fully automated whole body organ segmentation for diagnostic CT. This study investigates and optimizes a Random Forest algorithm for automated organ segmentation; explores the limitations of a Random Forest algorithm applied to the CT environment; and demonstrates segmentation accuracy in a feasibility study of pediatric and adult patients. To the best of our knowledge, this is the first study to investigate a Trainable Weka Segmentation implementation using Random Forest machine-learning as a means to develop a fully automated tissue segmentation tool developed specifically for pediatric and adult examinations in a diagnostic CT environment.

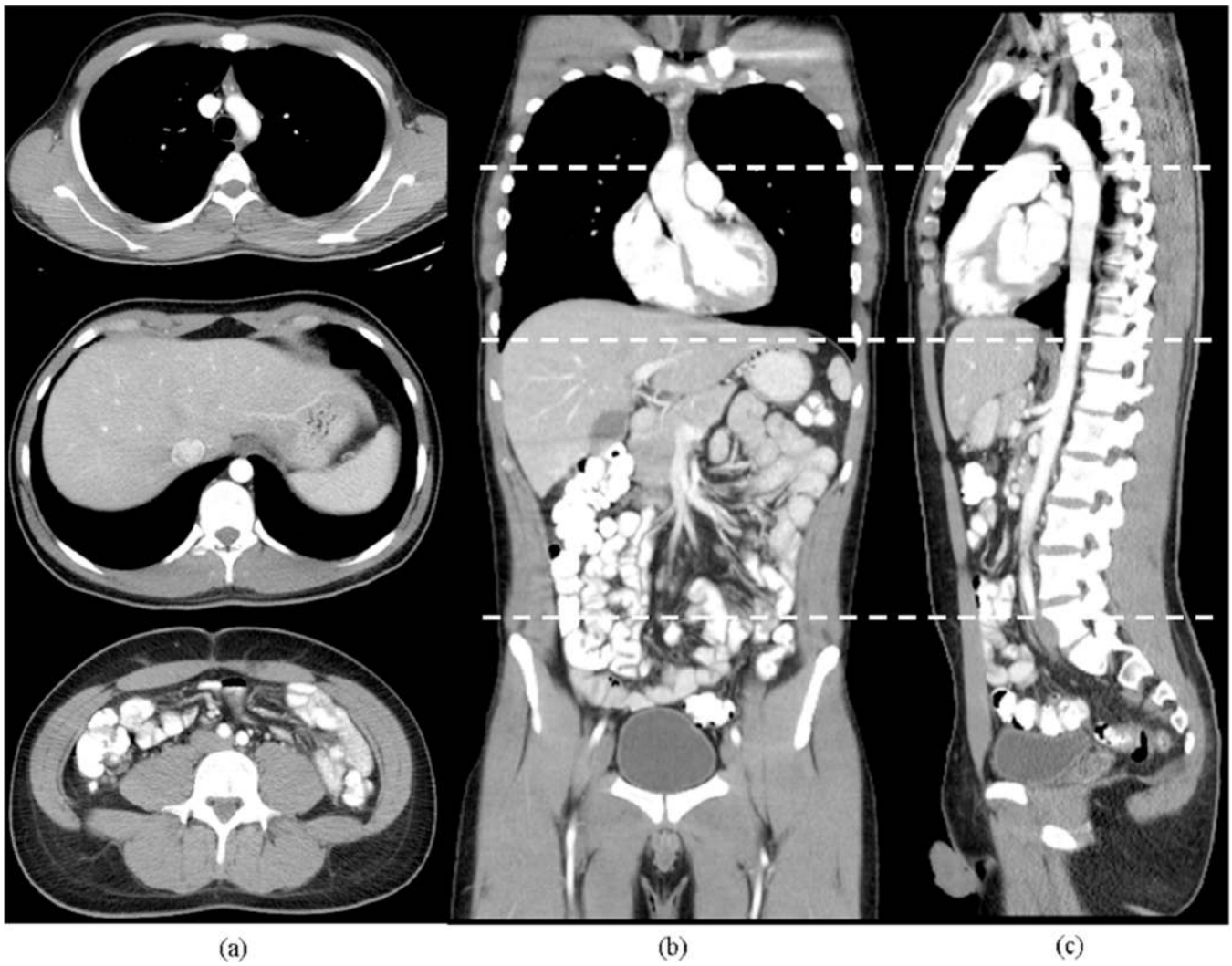

Fig. 1.

Training data set. The training data set was established based on chest-abdomen-pelvis CT examination reconstructed images of a 20 yr male (72 kg). Images from the data set are demonstrated: (a) axial, [sampled at the position of the dashed (– –) lines at the aortic arch, upper-liver, and immediately above the iliac crests], (b) coronal, and (c) sagittal reconstructed images.

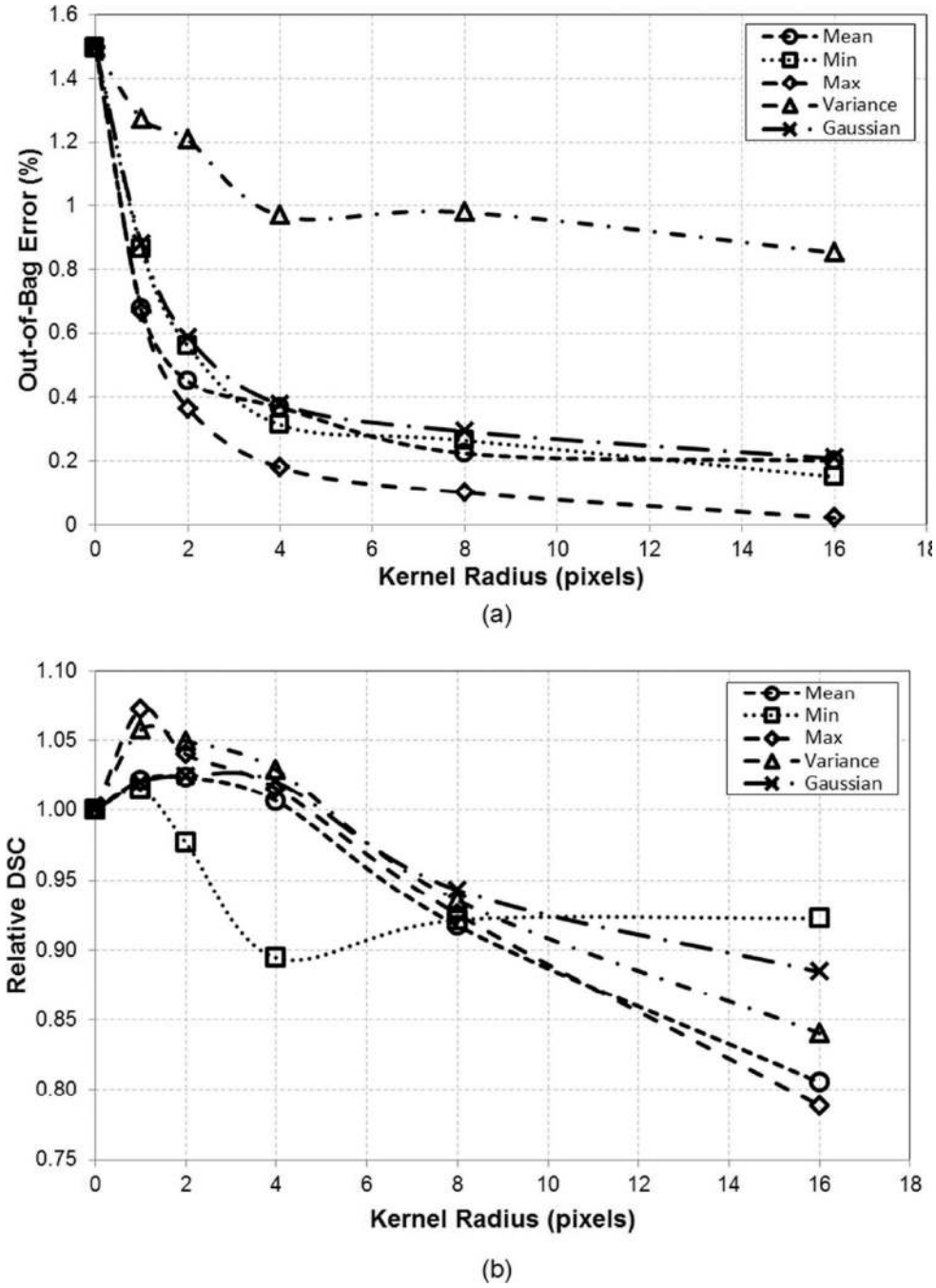

Fig. 2.

Segmentation optimization. (a) Out-of-bag error, and (b) Dice similarity coefficient (DSC) were calculated for material class segmentation optimization. Each plot represents the average value for six material class segmentations: lung/internal air or gas, fat, muscle, solid organ parenchyma, blood/contrast enhanced fluid, and bone tissue (background did not vary so it was removed from optimization process). Four texture feature filter inputs: mean, minimum (min), maximum (max), and variance, along with a noise reduction Gaussian filter, were calculated for the six material classes based on varying kernel radii (radii varied

from 2^n, where n = 0 to 4). Kernel Radius of 0 represented the baseline or individual voxel only.

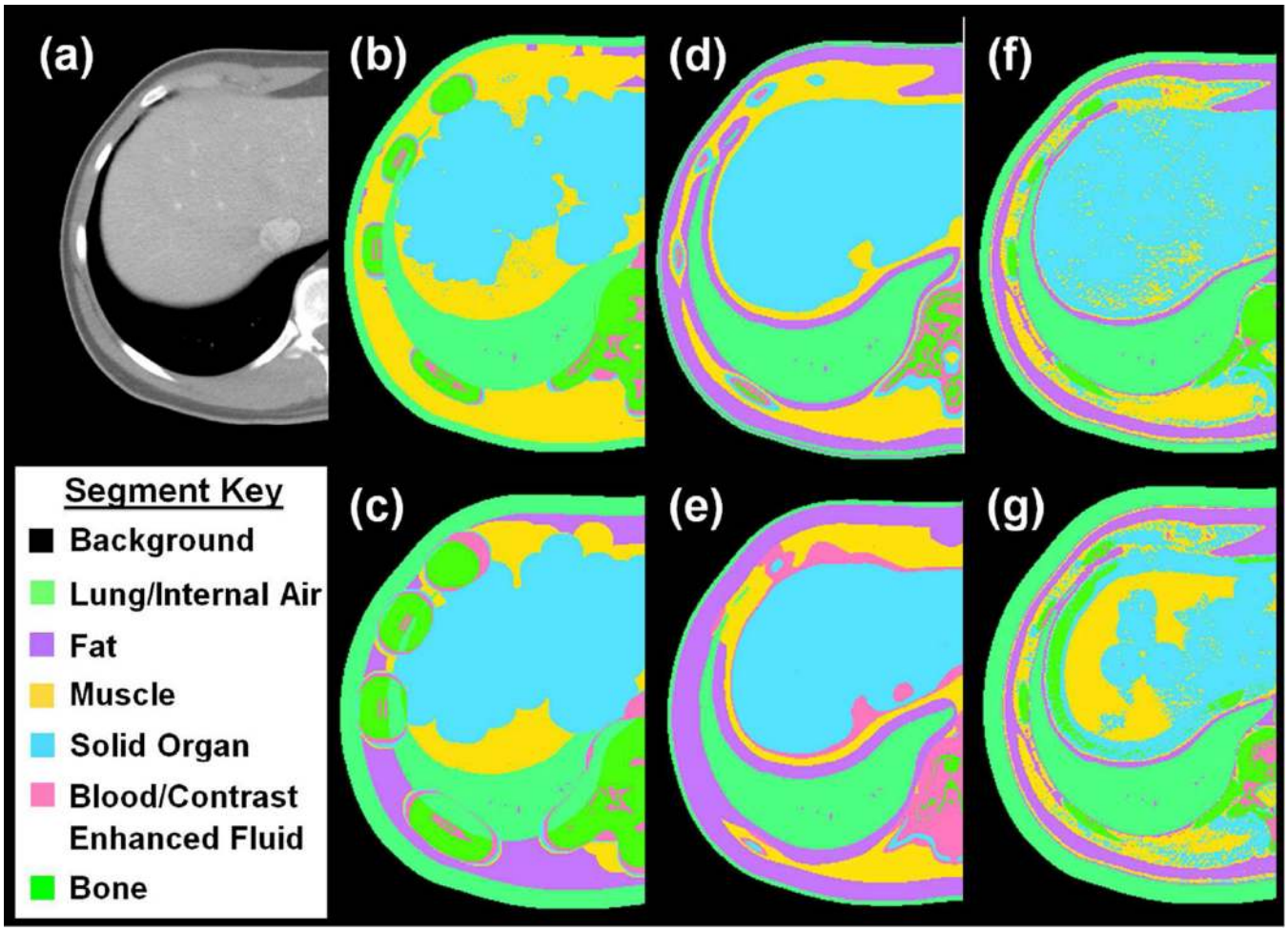

Fig. 3.

Visual evaluation of image features. (a) The upper-liver original CT image from the training data set was classified and seven material classes were segmented. The following feature inputs: (b)–(c) maximum, (d)–(e) mean, and (f)–(g) variance were visually assessed for appropriateness of segmentation. The top row (b), (d), and (f) was segmented using a radius of 8 voxels, and the bottom row (c), (e), and (g) was segmented using a radius of 16 voxels.

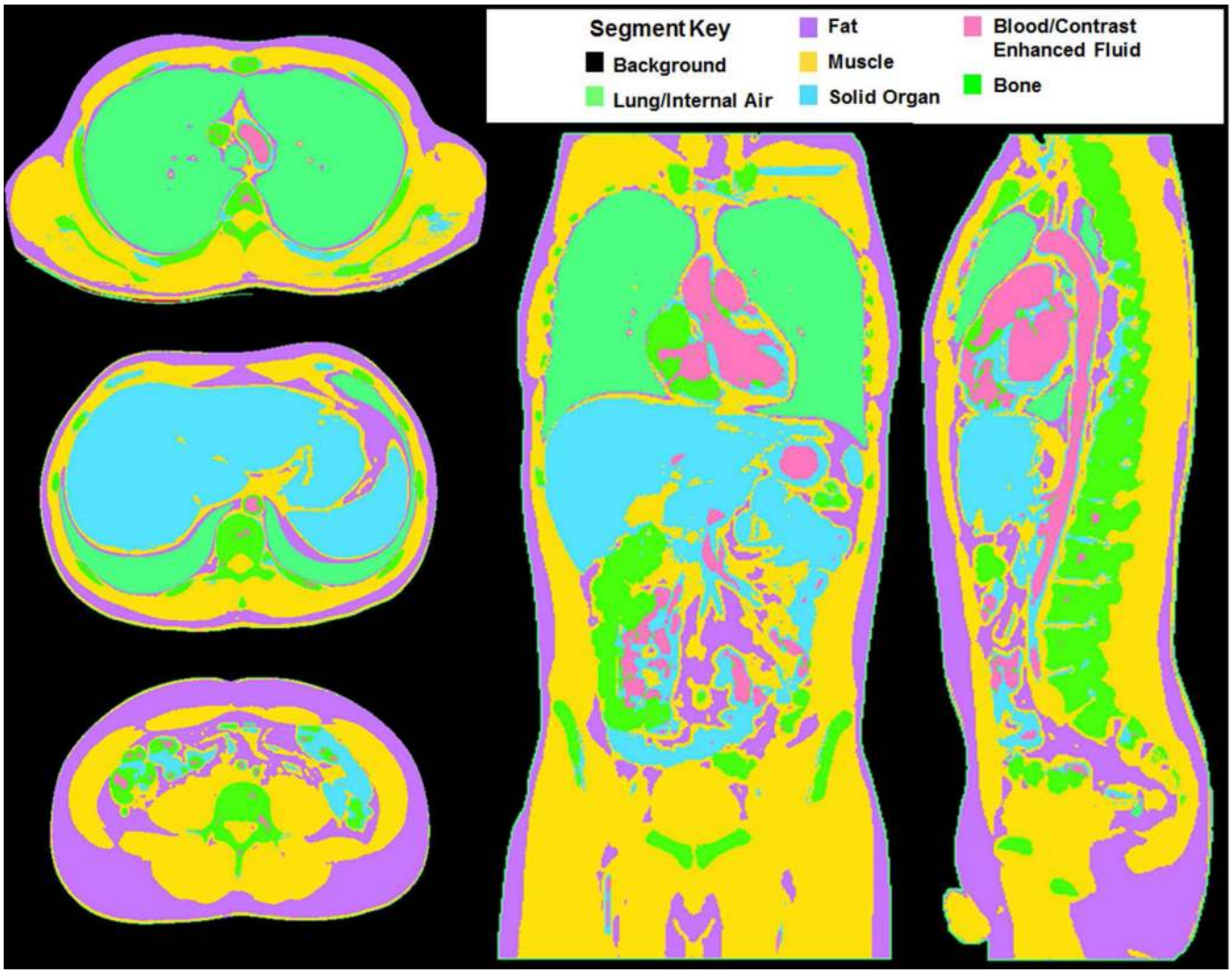

Fig. 4.

Example of segmentation. The training data set was classified and seven material classes segmented. Segmented images from the data set are demonstrated: (a) axial reconstructed images at the aortic arch, upper-liver, and immediately above the iliac crests, (b) coronal, and (c) sagittal. The original CT images are presented in Figure 1.

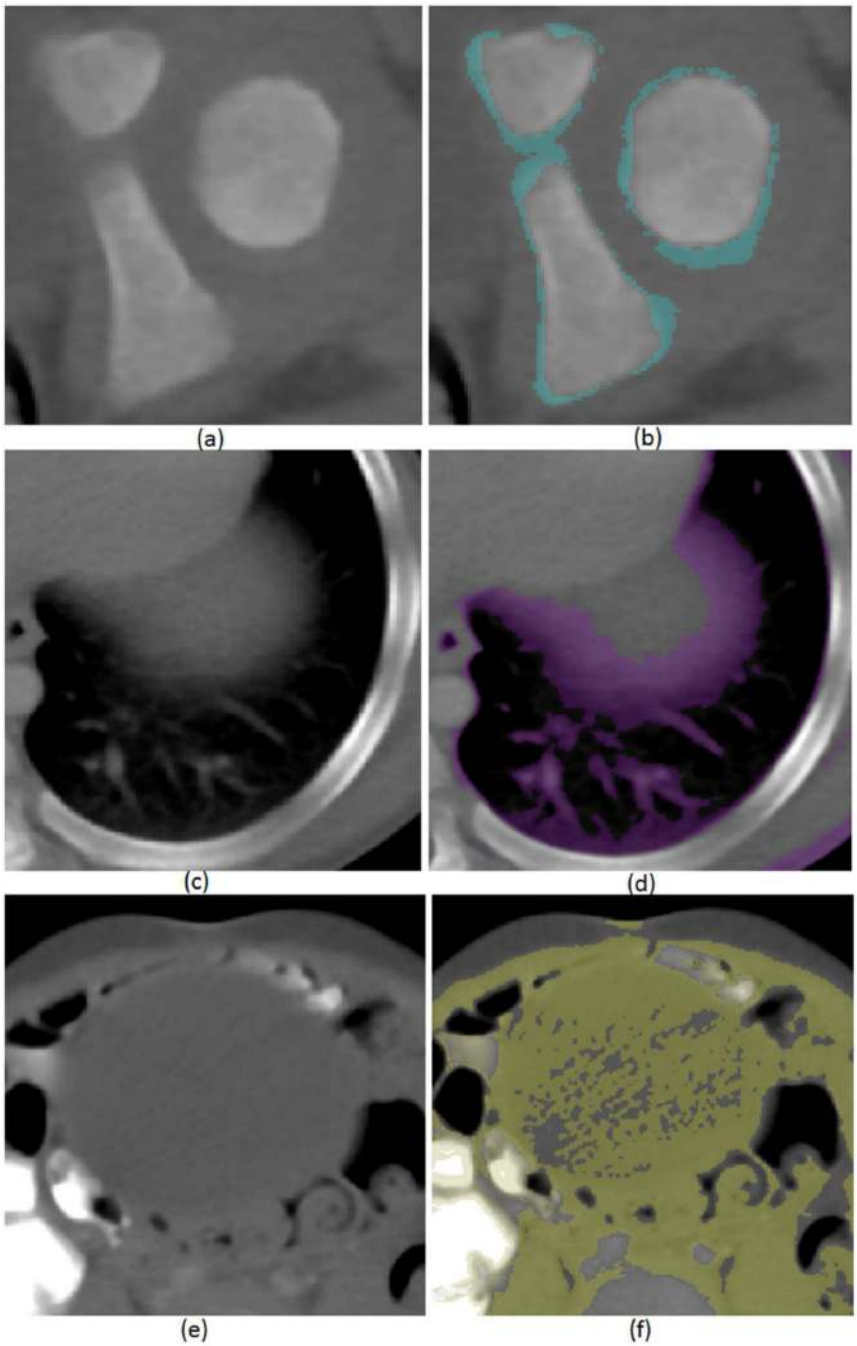

Fig. 5.

Example of segmentation misclassification surrounding. Based on the image intensity of (a) cartilage, ligament, joint fluid, and the trailing edge of bony ossification, (b) the Random Forest algorithm classified these tissues as solid organ parenchyma (blue voxels). Due to partial volume artifact from axial imaging of an oblique surface, (c) air/tissue interface is classified as (d) fat (purple voxels) due to the lower CT number intensity. Since (e) water/

fluid, as shown in the bladder, was not originally classified, (f) the Random Forest algorithm classified these voxels as muscle (yellow voxels).

Table 1

Dice similarity coefficient (DSC) results for manually segmented (ground truth) material classes compared to automated TWS algorithm segmentation. DSC results were compared for the training data set, 3 patients imaged with adult protocols (57–81 kg), 3 patients imaged with pediatric protocols (19–54 kg), and all seven patients averaged together. Comparisons were made at reconstructed image locations (1) aortic arch, (2) upper liver, (3) immediately above iliac crests, and (1–3) where the individual slices were combined into one volume set and compared as a volume. Slice locations 1 and 3 did not include solid organ parenchyma in the manually segmented images and therefore were not calculated for DSC.

Patient Group	Slice Location	Background	Lung/Internal Air	Fat	Muscle	Solid Organ	Blood/Contrast Enhanced Liquid	Bone	Combined High Contrast Regions[a]
Training	1	1.00	0.97	0.66	0.89	-	0.56	0.78	0.97
	2	1.00	0.93	0.76	0.74	0.89	0.30	0.90	0.92
	3	1.00	0.22	0.92	0.88	-	0.11	0.64	0.65
	1–3	1.00	0.95	0.82	0.85	0.81	0.20	0.77	0.81
Adult Protocol	1	1.00	0.96	0.72	0.87	-	0.20	0.82	0.92
	2	0.99	0.89	0.70	0.66	0.80	0.16	0.80	0.96
	3	1.00	0.53	0.90	0.86	-	0.06	0.43	0.75
	1–3	1.00	0.93	0.81	0.81	0.70	0.07	0.68	0.84
Pediatric Protocols	1	0.99	0.92	0.76	0.86	-	0.41	0.76	0.88
	2	0.99	0.78	0.75	0.73	0.90	0.31	0.89	0.92
	3	1.00	0.54	0.85	0.81	-	0.18	0.39	0.85
	1–3	0.99	0.88	0.80	0.81	0.82	0.27	0.69	0.88
All	1	1.00	0.94	0.73	0.87	-	0.34	0.79	0.91
	2	0.99	0.85	0.73	0.70	0.86	0.24	0.85	0.93
	3	1.00	0.49	0.88	0.84	-	0.12	0.44	0.78
	1–3	0.99	0.91	0.81	0.82	0.77	0.18	0.70	0.85

[a]The material class blood/contrast enhanced fluid and bone were compared as a single class entitled combined high contrast regions.

Table 2

Confusion matrices were developed for each of the seven segmented classes. From each matrix, sensitivity (Eq. 1), specificity (Eq. 2), and accuracy (Eq. 3) were calculated.

	Background	Lung/Internal Air	Fat	Muscle	Solid Organ[a]	Blood/Contrast Enhanced Liquid	Bone	Combined High Contrast Regions[b]
Sensitivity	1.0	0.98	0.95	0.88	0.82	0.86	0.22	0.90
Specificity	1.0	0.98	0.86	0.92	0.70	0.81	0.73	0.99
Accuracy	1.0	0.98	0.90	0.91	0.76	0.84	0.47	0.95

[a] Scores calculated for the solid organ class only apply to contrast enhanced CT examinations.

[b] The material class blood/contrast enhanced fluid and bone were compared as a single class entitled combined high contrast regions.