Medical Image Analysis 54 (2019) 45–62

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

AAR-RT – A system for auto-contouring organs at risk on CT images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases ☆

Xingyu Wu [a], Jayaram K. Udupa [a,1,*], Yubing Tong [a], Dewey Odhner [a], Gargi V. Pednekar [b], Charles B. Simone II [c], David McLaughlin [b], Chavanon Apinorasethkul [d], Ontida Apinorasethkul [d], John Lukens [d], Dimitris Mihailidis [d], Geraldine Shammo [d], Paul James [d], Akhil Tiwari [d], Lisa Wojtowicz [d], Joseph Camaratta [b], Drew A. Torigian [a]

[a] *Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard building, 3710 Hamilton Walk, 6th Floor, Rm 602W, Philadelphia, PA 19104, United States*
[b] *Quantitative Radiology Solutions, 3624 Market Street, Suite 5E, Philadelphia, PA 19104, United States*
[c] *Department of Radiation Oncology, Maryland Proton Treatment Center, School of Medicine, University of Maryland 850W, Baltimore, MD 21201, United States*
[d] *Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA 19104, United States*

ARTICLE INFO

Article history:
Received 4 August 2018
Revised 4 December 2018
Accepted 26 January 2019
Available online 29 January 2019

ABSTRACT

Contouring (segmentation) of *Organs at Risk* (OARs) in medical images is required for accurate radiation therapy (RT) planning. In current clinical practice, OAR contouring is performed with low levels of automation. Although several approaches have been proposed in the literature for improving automation, it is difficult to gain an understanding of how well these methods would perform in a realistic clinical setting. This is chiefly due to three key factors – small number of patient studies used for evaluation, lack of performance evaluation as a function of input image quality, and lack of precise anatomic definitions of OARs. In this paper, extending our previous body-wide Automatic Anatomy Recognition (AAR) framework to RT planning of OARs in the head and neck (H&N) and thoracic body regions, we present a methodology called AAR-RT to overcome some of these hurdles.

AAR-RT follows AAR's 3-stage paradigm of model-building, object-recognition, and object-delineation. *Model-building*: Three key advances were made over AAR. (i) AAR-RT (like AAR) starts off with a computationally precise definition of the two body regions and all of their OARs. Ground truth delineations of OARs are then generated following these definitions strictly. We retrospectively gathered patient data sets and the associated contour data sets that have been created previously in routine clinical RT planning from our Radiation Oncology department and mended the contours to conform to these definitions. We then derived an Object Quality Score (OQS) for each OAR sample and an Image Quality Score (IQS) for each study, both on a 1-to-10 scale, based on quality grades assigned to each OAR sample following 9 key quality criteria. Only studies with high IQS and high OQS for all of their OARs were selected for model building. IQS and OQS were employed for evaluating AAR-RT's performance as a function of image/object quality. (ii) In place of the previous hand-crafted hierarchy for organizing OARs in AAR, we devised a method to find an optimal hierarchy for each body region. Optimality was based on minimizing object recognition error. (iii) In addition to the parent-to-child relationship encoded in the hierarchy in previous AAR, we developed a directed probability graph technique to further improve recognition accuracy by

☆ **Conflict of interest:** Udupa and Torigian are co-founders of Quantitative Radiology Solutions, LLC. Pednekar, McLaughlin, and Camaratta were employees of Quantitative Radiology Solutions. This is the solo submission to Medical Image Analysis.

* Corresponding author.
E-mail address: jay@mail.med.upenn.edu (J.K. Udupa).

[1] Chief contributor to this entire study – for the underlying ideas, overall methodology, design of algorithms and evaluation strategies, and manuscript preparation.

https://doi.org/10.1016/j.media.2019.01.008

learning and encoding in the model "steady" relationships that may exist among OAR boundaries in the three orthogonal planes. *Object-recognition*: The two key improvements over the previous approach are (i) use of the optimal hierarchy for actual recognition of OARs in a given image, and (ii) refined recognition by making use of the trained probability graph. *Object-delineation*: We use a kNN classifier confined to the fuzzy object mask localized by the recognition step and then fit optimally the fuzzy mask to the kNN-derived voxel cluster to bring back shape constraint on the object.

We evaluated AAR-RT on 205 thoracic and 298 H&N (total 503) studies, involving both planning and re-planning scans and a total of 21 organs (9 – thorax, 12 – H&N). The studies were gathered from two patient age groups for each gender – 40–59 years and 60–79 years. The number of 3D OAR samples analyzed from the two body regions was 4301. IQS and OQS tended to cluster at the two ends of the score scale. Accordingly, we considered two quality groups for each gender – good and poor. Good quality data sets typically had OQS $\geq$ 6 and had distortions, artifacts, pathology etc. in not more than 3 slices through the object. The number of model-worthy data sets used for training were 38 for thorax and 36 for H&N, and the remaining 479 studies were used for testing AAR-RT. Accordingly, we created 4 anatomy models, one each for: Thorax male (20 model-worthy data sets), Thorax female (18 model-worthy data sets), H&N male (20 model-worthy data sets), and H&N female (16 model-worthy data sets). On "good" cases, AAR-RT's recognition accuracy was within 2 voxels and delineation boundary distance was within ~1 voxel. This was similar to the variability observed between two dosimetrists in manually contouring 5–6 OARs in each of 169 studies. On "poor" cases, AAR-RT's errors hovered around 5 voxels for recognition and 2 voxels for boundary distance. The performance was similar on planning and replanning cases, and there was no gender difference in performance.

AAR-RT's recognition operation is much more robust than delineation. Understanding object and image quality and how they influence performance is crucial for devising effective object recognition and delineation algorithms. OQS seems to be more important than IQS in determining accuracy. Streak artifacts arising from dental implants and fillings and beam hardening from bone pose the greatest challenge to auto-contouring methods.

1. Introduction

1.1. Background and rationale

Cancer is a major public health problem worldwide and is the 2nd most common cause of death in the US, with ~1.7 million new cancer cases expected to be diagnosed in the US in 2018, and with an estimated 609,640 American deaths to occur in 2018 (Siegel et al., 2018). Among several therapeutic options, nearly two thirds of cancer patients will have treatment that will involve radiation therapy (RT) (ASTRO website, 2018). Contouring of critical organs, called *Organs at Risk* (OARs), and target tumor in medical images taken for the purpose of RT planning (referred to as *planning* images) is required for accurate RT planning to ensure that a proper dose of radiation is delivered to the tumor while minimizing the radiation dose to healthy organs. In current clinical practice, OAR contouring is still performed with low levels of automation due to lack of highly automated commercial contouring software. This deteriorates RT planning. There are two major issues with the current clinical practice of OAR contouring: (1) Poor accuracy. (2) Poor efficiency, throughput, and reproducibility.

Poor accuracy, and consequently poor efficiency/acceptability, of OAR contours produced by existing software platforms on planning images is the main hurdle in auto-contouring for RT planning. The problem is well summarized in Whitfield et al. (2013): "Rapid and accurate delineation of target volumes and multiple organs at risk,… is now hugely important in radiotherapy, owing to the rapid proliferation of intensity-modulated radiotherapy … Nevertheless, delineation is still clinically performed with little if any machine assistance, even though it is both time consuming and prone to inter-observer variation." Many commercial auto-contouring systems are currently available (Thomson et al., 2014; Lustberg et al., 2017), but their poor accuracy leads to poor clinical acceptability of the contours and hence poor efficiency. As we demonstrate in Section 5 involving a large realistic study, in the clinical setting, OAR contouring can take anywhere from 40 min to 2 h depending on the number of OARs to be contoured.

The efficiency problem is exacerbated in advanced RT methods such as intensity modulated radiotherapy (IMRT) and proton beam radiation therapy (PBRT) (McGowan et al., 2013). Adaptive RT can allow for modifying the treatment plan to account for anatomic changes occurring during a 5–8-week course of treatment due to weight loss or deformation of tumor and normal tissues. Such changes are particularly common during head and neck (Simone et al., 2011) and thoracic (Veiga et al., 2016) radiation and can significantly affect the total dose delivered to the tumor and normal surrounding organs and are particularly important when treating most thoracic malignancies (Veresezan et al., 2017). PBRT can allow for ultra-precise delivery of treatment due to the physical characteristics of the proton beam, eliminate exit dose, maximize dose delivered to the tumor, and minimize radiation dose to adjacent OARs, reducing toxicity and patient morbidity (Roelofs et al., 2012), and improving clinical outcomes like overall survival (Leeman et al., 2017). Yet, because of the poor accuracy, and hence efficiency of current software products, re-contouring on images taken during treatment (referred to as *evaluation* or *re-planning* images) is rarely done. While the impact of this issue on patient outcome has sparsely been studied (Dolz et al., 2016), with accurate automated contouring, advanced IMRT and PBRT methods can be employed more extensively and may allow for these advanced radiotherapy modalities to achieve toxicity reductions or outcomes benefits to a large subset of patients.

The current gaps/challenges in auto-contouring for the RT application, which motivated the development of AAR-RT[1], may be summarized as follows. (1) Evaluation: Testing on a large number of independent data sets versus on the same data sets in a multi-fold cross validation manner is vital to get a real understanding of the behavior of the method independent of the data sets. This is currently lacking. Generally, performance evaluation is done only on planning and not evaluation images. In our study cohort, we found the quality of the images to be lower in evaluation scans than in planning scans. (2) Data quality: The quality of the image

[1] AAR: Automatic Anatomy Recognition. RT: Radiation Therapy.

data sets used, presence and severity of the artifacts/deviations from normality in these data sets, and how they might influence results are not usually discussed in published methods. No examples of performance on scans with artifacts are given and there is no discussion of how the training and testing data sets are selected with regard to artifacts and other distortions. (3) *OAR definition:* Although some contouring guidelines are followed by dosimetrists and oncologists (Brouwer et al., 2015a, 2015b; Kong et al., 2011), the flexibility allowed, site-to-site variations, and the looseness of the definitions make the resulting contours unsuitable for building precise computational population object models/schemas.

In an attempt to address some of these challenges, we adopted our previous body-wide Automatic Anatomy Recognition (AAR) framework (Udupa et al., 2014) and refined its three main steps, namely, fuzzy anatomy model building for a body region, object recognition/localization, and object delineation, with further advances in each step. Key innovations and improvements over the previous AAR framework are as follows. (1) OAR definition: To overcome the non-standardness hurdle, following published guidelines for head and neck (H&N) (Brouwer et al., 2015a, 2015b; Hall et al., 2008) and thoracic (Kong et al., 2018; Kong et al., 2011; Hall et al., 2008) anatomic OAR definitions, we formulated detailed and precise operational definitions and a reference document for specifying and delineating each of the 21 OARs considered in this work on axial CT slices, as explained in Section 2. (2) Optimal hierarchy: The AAR approach arranges OARs in a hierarchy by learning object relationships. Previously, we used an anatomically motivated hierarchy for OARs. In this work, we find an optimal hierarchy that actually minimizes OAR recognition error, as described in Section 3. (3) Image texture: The best OAR-specific image texture property is found and used for both object recognition and delineation, as outlined in Section 3. (4) Recognition refinement using Directed Probability Graph (Section 3): In the previous approach, object localization accuracy was inferior in the *z*- (cranio-caudal) direction to that in the *xy* (axial) plane. We train and employ a Directed Probability Graph to improve this accuracy. (5) Delineation via voxel classification and fuzzy model fitting: The previous approach used fuzzy connectedness which had issues with automatically finding seeds required for its delineation engine. We replace that strategy by a fuzzy classification and fuzzy model fitting step to improve accuracy (Section 3). (6) Large-scale evaluation of recognition and delineation: We evaluate both recognition and delineation performance of AAR-RT on clinical CT scans of over 500 cancer patients randomly selected from our hospital database for the two body regions involving both planning and evaluation scans (Sections 2 and 4). (7) Evaluation as a function of image/object quality: To understand dependence of performance on image/object quality, we define image/object quality metrics, build models using highest quality data sets, and evaluate recognition/delineation accuracy on all data sets as a function of quality (Sections 2–4).

1.2. Related work: approaches to segmentation of OARs

There is a large body of literature on segmentation of individual objects/OARs on images from different modalities. However, not all of them are applicable to the problem of body-region-wide OAR segmentation. It takes a lot of effort to understand the application-specific issues, solve each of them satisfactorily, and evaluate them in a realistic manner to gain confidence on the behavior of the method on real clinical data sets. We shall therefore review works specifically related to body-region-wide OAR segmentation for the RT application on CT images of cases involving H&N and thoracic malignancies. We will perform a comparative analysis of AAR-RT and key published works from in Section 4.

Atlas-based methods are quite popular in RT application due to their robustness and requirement for a small number of training samples. These methods register the training images to the test image and correspondingly propagate the training OAR contours to the test image. The anatomy information in the training set is described by one or a group of images called *atlas*. Reported atlas generation methods include a single training image (Han et al., 2008; Voet et al., 2011), averaging multiple images (Sims et al., 2009), and simulated images with standard anatomy (Isambert et al., 2008). More recently, multi-atlas methods have shown better accuracy with a more elaborate training step which groups patients first for atlas generation (Saito et al., 2016; Schreibmann et al., 2014; Teguh et al., 2011), and then selects the most similar group to the test image subsequently for object segmentation. One disadvantage of the atlas-based methods is that they require accurate registration to align the patient and target image, which is hard to make robust to shape variations, anatomy changes, and image quality variations. More importantly, it is hard to handle non-smooth geometric relationships that exist among objects in their geographic layout, size, and pose (Matsumoto et al., 2016) via smooth registration operations, although grouping helps to circumvent this issue to some extent.

Besides atlas-based methods, the approach of using landmarks on each object to handle local variations (Ghesu et al., 2017; Ibragimov et al., 2014; Zheng et al., 2015) received considerable attention in recent years due to the better local adaptability of such approaches. These methods can be categorized as global approaches because they start from the entire patient image rather than a local region of interest (ROI), so a registration step becomes necessary. However, the orientation and position variations between H&N and thoracic regions and curvature variations of the spine often pose extra difficulties for registration (Daisne and Blumhofer, 2013) which are addressed via the use of landmarks. As an alternative, our previous AAR works (Udupa et al., 2014; Phellan et al., 2016) build fuzzy models for each object and encode object relationships pairwise explicitly in a hierarchical arrangement of objects for facilitating recognition, which eliminates the registration step and can also handle non-smooth object relationships.

More recent approaches tend to explore local methods that start from an ROI for each object. The ROI may be determined either manually or by global methods. This kind of global-to-local strategy has lower requirements on the precision of registration and can become more robust under anatomy variations and image quality vagaries. Some studies cascade atlas-based methods for ROI initialization followed by a local boundary extraction approach, such as geodesic active contours (Fritscher et al., 2014), graph-cut (Fortunati et al., 2015), and appearance models (Wang et al., 2018). In recent years, delineation methods using convolutional neural networks (CNNs) (de Vos et al., 2017; Ibragimov and Xing, 2017a) and fully convolutional networks (FCNs) (Çiçek et al., 2016; Dou et al., 2017; Trullo et al., 2017a; Zhou et al., 2017a) have started showing improved results under the prerequisite of correct local ROI selection. Deep learning approaches seem to outperform other methods in learning local anatomy patterns, but challenges still exist in localizing OARs in the whole given image (object *recognition* problem), especially for sparse and small objects. It is worth investigating, therefore, how to incorporate the anatomy prior information to reduce the amount of total input information to these networks to make them more effective and specific. Recent research shows the benefit of incorporating shape prior as a constraint for neural network strategies (Oktay et al., 2018), but this is only prior information on each individual OAR. The problem of determining the manner in which to utilize global information, especially the relationship among OARs for localization before delineation, is still unsolved in these approaches.

The progress in research over the years in multi-object segmentation suggests a dual paradigm for segmentation: (1) object *recognition* (or localization), which uses prior information to

Table 1
Thoracic and H&N OARs included in our study and some study statistics.

Abbr	OAR	Abbr	OAR	Abbr	OAR	Study statistics	Thorax	H&N
tSB	Thoracic skin outer boundary	LBP	Left brachial plexus	LPG	Left parotid gland	**#Planning scans**	118	216
Hrt	Heart	RBP	Right brachial plexus	RPG	Right parotid gland	#OARs	9	12
LLg	Left lung	hSB	H&N skin outer boundary	LSG	Left submandibular gland	#OAR samples	1175	2199
RLg	Right lung	SBi	hSB inferior part	RSG	Right submandibular gland	#Good Quality samples	718	905
						#Poor Quality samples	457	1294
TB	Trachea & proximal bronchi	SBs	hSB superior part	MD	Mandible	#Model worthy scans	38	36
tSC	Thoracic spinal cord	cSC	Cervical spinal cord	OHP	Orohypopharynx constrictor muscle	**#Replanning scans**	87	82
tES	Thoracic esophagus	LX	Larynx	cES	Cervical esophagus	#OAR samples	516	411

define the whereabouts of the object, and (2) object *delineation*, which employs local information to precisely define the object's spatial extent in the image. This dichotomous strategy for image segmentation was first suggested in the live wire method (Falcao et al., 1998) where recognition is done manually but delineation is automatic and occurs in real time, and the two processes are tightly coupled. Our entire AAR framework operates on this dual recognition-delineation premise and we try to advance recognition and delineation methods separately and synergistically. This is the key idea behind our AAR-RT framework.

A very preliminary report on this investigation appeared in the proceedings of the 2018 SPIE Medical Imaging Conference (Wu et al., 2018). The present paper includes the following significant enhancements over the conference paper: (i) A comprehensive literature review. (ii) Full description of the methods and the underlying algorithms. None of the object recognition and delineation algorithms were described in the conference paper. (iii) Comprehensive evaluation. The conference paper preliminarily tested and presented results for 6 H&N OARs and none from the thorax. This paper analyzes results for recognition and delineation for all 21 OARs from both H&N and thoracic regions and their dependence on image/object quality. (iv) Evaluation on both planning and evaluation scans. The conference paper considered only a subset of the planning data sets used in this paper and no evaluation scans. (v) A detailed comparison of AAR-RT with key auto-contouring methods from the literature for the two body regions which was not undertaken in the conference paper.

2. Materials

2.1. Image and contour data

This retrospective study was conducted following approval from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act waiver. We collected planning CT image and contour data sets from existing patient databases from the Department of Radiation Oncology, University of Pennsylvania, under four patient groups: 40–59-year-old males and females (denoted G_{M1} and G_{F1}, respectively), 60–79-year-old males and females (denoted G_{M2} and G_{F2}, respectively). For thorax and H&N, data sets respectively from 210 to 216 cancer patients (with different types of cancer) were gathered, with at least 50 data sets per group; pixel size: 1–1.6 mm, slice spacing: 1.5–3 mm. Similarly, we gathered *replanning* (evaluation) scans from 30 patients (for each body region) who underwent PBRT fractionated treatment serially. For each patient, we selected image data at 2 or more, commonly 3, serial time points, accounting for a total of 87 scans for thorax and 82 scans for H&N. The OARs considered for the two body regions (9 for thorax and 12 for H&N for planning cases and 6 for thorax and 5 for H&N for replanning cases), their abbreviations used, and their total number are listed in Table 1. The total number of 3D OAR samples considered in this study from planning and replanning scans was 4301 (1691 for thorax and 2610 for H&N) from a total of 595 patient scans.

OAR contours for the planning cases were previously drawn by the dosimetrists (and approved by attending physicians) in the process of routine clinical RT planning of these patients. Note that not all OARs were delineated in each planning scan. The number of OARs for which dosimetrist-drawn contours were available in each scan was 5–9 for thorax and 5–12 for H&N. Since manual object contouring is impractical to perform and hence not done clinically for every replanning scan associated with treatment fractions, we do not have ground truth OAR contours for the corresponding data sets. Therefore, to generate ground truth data for replanning scans and to gain insight into how contouring is done in practice, we recruited four dosimetrists (two for each body region) from the Penn Radiation Oncology department to perform manual contouring on *all* 169 replanning studies from the two body regions. The following OARs were considered for the replanning studies (see Table 1). Thorax: RLg, LLg, Hrt, tES, tSC, and TB. H&N: hES, hSC, MD, OHP, and LX. The dosimetrists were asked to record the start time and end time for each contouring session for each object. Also, we noted down other preparatory time and time for ancillary efforts during the contouring process.

2.2. Standardizing OAR definition and ground truth contouring

Although some object contouring guidelines are followed by dosimetrists and oncologists (Brouwer et al., 2015a, 2015b), the flexibility allowed and the looseness of the definitions make ground truth contouring less precise and the resulting contours unsuitable for building precise computational population object models. To overcome this hurdle, following the above guidelines for anatomic object definitions, we formulated detailed and precise operational definitions and a document (Wu et al., 2017a, 2017b) for specifying each object and for delineating its boundaries on axial CT slices. For illustrating the level of detail involved in our specification, we show in Fig. 1 the mandible in the H&N region as an example. Two software engineers (co-authors GVP and DM) were thoroughly trained on these definitions who then mended dosimetrist-drawn contours of all 21 OARs on all 426 planning scans by strictly following this document under the supervision of a radiologist with 22 years of experience (co-author DAT). The resulting contours were used as ground truth object delineations for building models and for evaluating AAR-RT. The two dosimetrists followed these documents as well for contouring the 11 OARs on the 169 replanning scans.

2.3. Image/object quality consideration for model building and evaluation

Algorithms for image segmentation are influenced by the quality of appearance of each object in the image and overall image quality. For holistic evaluation, it is important to define object and

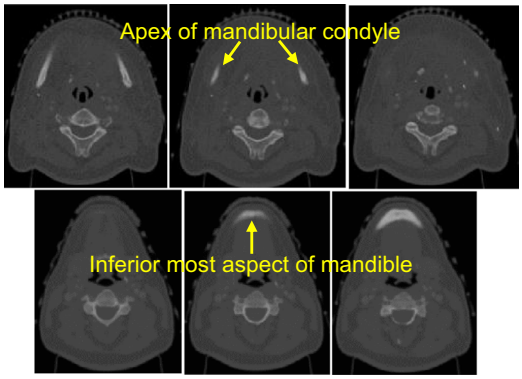

Fig. 1. Specification of Mandible. Top row: Superior boundary is the superior-most aspect of the mandible (typically the apex of the condyle) as shown in axial slice in the middle. The slices on the left and right are immediately inferior and superior to the slice in the middle, respectively. Bottom row: Inferior boundary of the mandible is the inferior-most aspect of the mandible as shown in slice in the middle. The slices on the left and right are immediately inferior and superior to the slice in the middle, respectively. The slices are displayed at bone window.

image quality metrics and perform segmentation evaluation as a function of these quality metrics. No such efforts seem to have been undertaken to date in segmentation challenges and other quantitative medical imaging application efforts. We developed a method (Pednekar et al., 2018) to assign a quality grade to the image appearance of each object (OAR) in each image based on a set of 9 criteria: neck posture deviation, mouth position, other types of body posture deviations, image noise, beam hardening artifacts (streak artifacts), shape distortion, presence of pathology, object intensity deviation, and object contrast. Fig. 2 displays patient cases illustrating some of these criteria. We converted these criterion grades into an object quality score (OQS) on a 1 to 10 scale using logical predicates (Pednekar et al., 2018). The OQSs were also used to determine an integrated image quality score (IQS), also on a 1 to 10 scale. OQS and IQS served two purposes: (i) for determining patient scans in our cohort that can be utilized for model building, which we refer to as *model-worthy* data sets; and (ii) for segmentation evaluation.

The number of scans in our cohort that were completely free of deviations on the basis of the above 9 factors was 0 for thorax and 1 for H&N. Generally, younger patients had better quality than older patients. We observed that OQS and IQS mostly clustered at the low and high end of the score scale (see Fig. 3). We therefore defined an OAR sample (i.e., an OAR as a 3D object in a given patient image data set) as of *good* quality if it did not carry deviations in more than 3 slices (this corresponded roughly to OQS>6); otherwise the sample was considered as of *poor* quality. A scan (image data set) was considered *model-worthy* if all of its OARs were good-quality samples. Following the basic principle of the AAR framework (Udupa et al., 2014) of using near-normal data sets for building anatomy models of a body region, only model-worthy data sets were used for model building: Thorax: 20 males, 18 females; H&N: 20 males, 16 females. Table 1 (last column) lists statistics related to good and poor OAR samples and model-worthy data sets for the two body regions among our planning/evaluation scans. Since the number of model-worthy data sets in each of the 4 patient groups was not large enough, we built only 2 models, called *fuzzy anatomy models*, one for males and one for females for each body region *B*: $FAM(B, G_M)$ by combining groups G_{M1} and G_{M2}, and $FAM(B, G_F)$ by combining groups G_{F1} and G_{F2}. These model-worthy data sets did not participate in testing recognition and delineation algorithms. We performed evaluation of OAR recognition and delineation separately for the four categories: male-good, male-poor, female-good, and female-poor.

3. Methods

3.1. Overview

Our previous AAR approach (Udupa et al., 2014) consists of three stages – model building, object recognition, and object delineation. Model building involves creating a *Fuzzy Anatomy Model*, $FAM(B, G) = (H, M, \rho, \lambda, \eta)$, of the body region B of interest for a group G of subjects. In this expression, H denotes a hierarchical arrangement (tree structure) of the objects (OARs); M is a set of fuzzy models with one model for each object; ρ represents the parent-to-child relationship in G in the hierarchy; λ is a set of scale ranges, one for each object; η includes a host of parameters representing object properties such as the range of variation of size, image intensity and texture properties, etc., of each object. $FAM(B, G)$ is built from a set of *good* quality (model-worthy) CT images of B and the binary images representing a set of OARs in B for each of these images.[2] After $FAM(B, G)$ is built, it is used to recognize and delineate any OAR in any patient image of B. Recognition and delineation proceed hierarchically in H, starting from the root OAR, then proceeding to the child.

AAR-RT incorporates several advances made in AAR in each of the three stages. *Model building*: (i) In place of the handcrafted hierarchy H that was employed in the previous approach to build $FAM(B, G)$, we use an algorithm to construct a hierarchy that yields close to the least recognition error among all possible hierarchies. (ii) Previously, the parent-child relationship ρ was expressed by just the vector connecting the geometric centers of the parent and child and its statistics over G. Now, based on experience with the previous approach, this is further refined by including the relationship among inferior-to-superior (z direction), lateral-to lateral (x direction), and anterior-to-posterior (y direction) boundaries of the OARs using a Directed Probability Graph. *Object recognition*: (i) The order specified by the optimal hierarchy found in the model building stage is followed for localizing OARs in a given patient image using the previous optimal threshold approach. (ii) This recognition result is refined using the Directed Probability Graph constructed in the model building stage. *Object delineation*: (i) To overcome seed specification issues, in place of the previous fuzzy connectedness engine, a kNN scheme is used. (ii) The final refined fuzzy model resulting at the recognition stage is fitted optimally to the kNN delineation result to produce the final OAR delineation.

The flow diagram of the overall approach underlying AAR-RT is depicted in Fig. 4. The three stages are described separately below in detail.

3.2. Building fuzzy anatomy model

Given a set of images $\mathcal{I} = \{I_1, \ldots, I_N\}$ of B for group G and the associated binary images (the next $\mathcal{I}^b = \{I_{n,l}: 1 \le n \le N \ \& \ 1 \le l \le L\}$ representing the L OARs $\mathcal{O} = \{O_1, \ldots, O_L\}$ in B, building $FAM(B, G) = (H, M, \rho, \lambda, \eta)$ involves determining each of the 5 parameters in this quintuple. Hierarchy H and object relationships ρ in H are found as described below. Other parameters are found as described in the original AAR framework (Udupa et al., 2014). Briefly, $M = \{FM(O_l): 1 \le l \le L\}$ is a set of fuzzy models, one fuzzy model for each OAR. The fuzzy model $FM(O_l)$ of an OAR O_l is created by scaling all binary samples of O_l to a mean size, repositioning all samples to a mean location, and averaging the result (see Udupa et al., 2014, for details). Parameter λ is a set of scale

[2] Binary images of all objects considered in B are expected to be available for each image that is selected for building the model. Only these objects can then be recognized and delineated in any given patient image. In other words, the set of OARs to be segmented in a given patient image should always be a subset of the set of OARs considered for building $FAM(B, G)$.

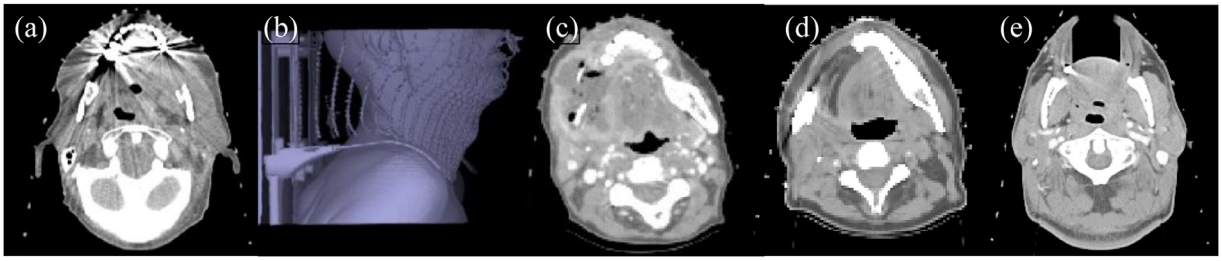

Fig. 2. Examples of factors that can downgrade the image quality of CT scans. (a) Streak artifacts due to dental fillings and implants. (b) Body posture deviation (neck rotation). (c) Pathology (centrally necrotic lesion predominantly in right masticator space). (d) Shape distortion (post-surgical change). (e) Body posture deviation (mouth open).

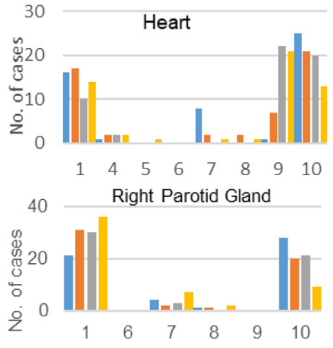

Fig. 3. OQS distribution in our planning scans for Hrt and RPG. Colors denote different groups: Blue: G_{M1}. Red: G_{F1}. Gray: G_{M2}. Orange: G_{F2}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ranges in which each element of the set indicates the size variation of each OAR. This parameter is utilized in confining recognition search in the pose space to previously known ranges in the population G. Parameter η stores population statistics over G pertaining to OARS such as their intensity and texture properties etc., which are used in recognition and delineation.

3.2.1. Finding optimal hierarchy of OARs

There are several reasons for a hierarchical arrangement of objects. Objects have steady geometric relationships (Matsumoto et al., 2016) and they generally do not depend on image/object quality. This implies that if the relationships can be learned, then OAR recognition can be made quite robust with respect to image/object quality. Furthermore, the relationships are non-smooth and non-linear (Matsumoto et al., 2016), implying that some relationships are much less variable than others. Therefore, we contend that for any object O_1, there is a best (most optimal) object O_2 to be paired as its child. Since our goal is achieving accurate recognition of objects, the optimality criterion here should be the accuracy of recognition of the child given the parent. This naturally leads to the following formulation for optimal hierarchy: Given image sets $\mathcal{I}$ and $\mathcal{I}^b$ and the set $\mathcal{O}$ of OARs for B, find that hierarchy H over which the total recognition error is minimized. To solve this problem, we may form a complete graph $\mathcal{G} = (\mathcal{O}, E)$, $E = \{(O_i, O_j) : O_i, O_j \in \mathcal{O} \& O_i \neq O_j\}$, in which the nodes are the OARs and every pair of OARs is connected by two directed arcs; then determine all possible L^{L-2} trees that span G, and find among them the tree that yields the least recognition error. Given that each recognition experiment requires about 30 s, when $L = 12$ (H&N body region, for example) and assuming the number of images in $\mathcal{I}$ to be $N = 50$, finding a globally optimal tree following a brute-force approach would take about 17.5 million days! We take a greedy approach to find optimal H.

We convert the above graph into a weighted graph $\mathcal{G} = (\mathcal{O}, E, \omega)$, where $\omega(O_i, O_j)$ is the weight assigned to directed arc (O_i, O_j). Our idea is to make $\omega(O_i, O_j)$ small when a mini hierarchy, where O_j is a child of O_i, yields small error. Subsequently, we can find an optimum spanning tree $OST(\mathcal{G}, O_r)$ in $\mathcal{G}$ that is rooted at O_r using a minimum spanning tree algorithm (Cormen et al., 2009). In our approach, we fix O_r to be the skin object (tSB for thorax and hSB for H&N). We take a greedy approach that is computationally feasible although it cannot guarantee that $OST(\mathcal{G}, O_r)$ is a hierarchy that yields globally the best possible recognition results for the objects in $\mathcal{O}$, to yield minimum total error in recognition of all objects over the images in $\mathcal{I}$. To implement the approach, we form all possible mini hierarchies of the form shown in Fig. 5, where O_r is the root object and O_i and O_j are other (non-root) objects. Then, for all arcs of the form (O_r, O_j), we set $\omega(O_r, O_j)$ to the mean of the recognition error of O_j over all images in $\mathcal{I}$ resulting by using the mini hierarchy of Fig. 5(a). For all arcs (O_i, O_j) of the form shown in Fig. 5(b), the arc weight $\omega(O_i, O_j)$ assigned is the mean over all images of $\mathcal{I}$ of the recognition errors of O_j resulting by using the mini hierarchy of Fig. 5(b). The idea here is that, in this basic hierarchical form, which is different from that in Fig. 5(a), the recogni-

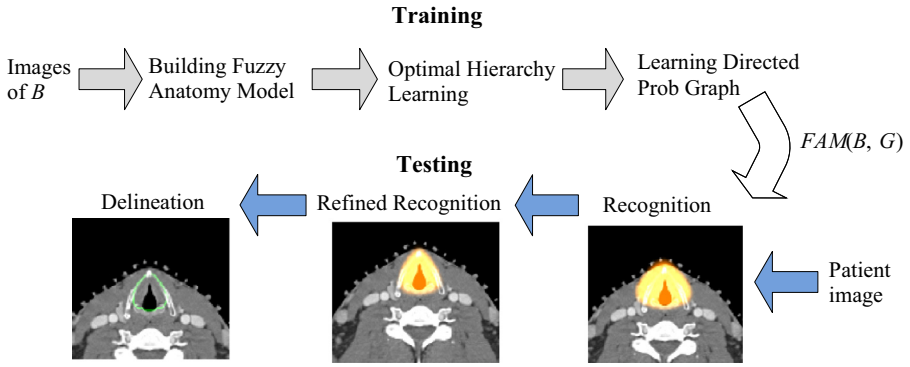

Fig. 4. Flow diagram illustrating the overall approach underlying AAR-RT.

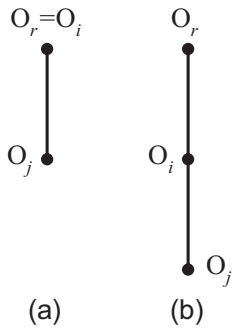

Fig. 5. Mini hierarchies considered in the greedy algorithm for estimating arc weight based on recognition error. In (a), all mini hierarchies that include the root object O_r and any other object O_j are considered. In (b), all mini hierarchies that include arcs (O_i, O_j) where O_i and O_j are different from O_r are considered.

tion accuracy of both O_i and O_j should influence the cost assigned to O_j being the child of O_i.

In the AAR approach, recognition error for an object O is expressed via its scale error *SE*(O), location error *LE*(O), and false positive and false negative volumes, all with respect to the known ground truth object. In our implementation, we set $\omega(O_i, O_j) = LE(O_j)$ in the situations shown in Fig. 5(a) and (b). In finding $OST(\mathcal{G}, O_r)$, we set a limit of 4 for the depth of the tree to generate more balanced trees and to avoid long paths in resulting hierarchies. We developed an algorithm that finds a hierarchy seeking to minimize the sum of the arc weights while keeping the depth limited. This hierarchy has an arc weight cost close to that of the tree found by the minimum spanning tree algorithm, but a smaller recognition error when the whole tree is used for recognition.

3.2.2. Refining object relationships

Object hierarchical relationships learned in the previous step allow overall placement of object models in a test image. Based on this placement, model boundary extents in the three anatomic planes are refined by exploiting the relationship that may exist among these boundary planes. Learning and refining this relationship are done independently in the three directions (left-to-right or $\pm x$ direction, antero-posterior or $\pm y$ direction, and cranio-caudal or $\pm z$ direction). We will use the notation $x_l(O)$ and $x_h(O)$ to denote the boundary extent of an object O in the $-x$ and $+x$ directions, respectively. Similarly, $y_l(O)$, $y_h(O)$, $z_l(O)$, and $z_h(O)$ are defined. An example to illustrate the idea is shown in Fig. 6 involving 3 OARs: MD, LPG, and RPG. Since the parotid glands are situated close to the mandibular condyle laterally (Fig. 6(a)), we expect x_l(RPG) and x_h(RPG) to have a steady relationship with respect to x_l(MD) (Fig. 6(b)) due to anatomic constraints. A similar remark applies to x_l(LPG) and x_h(LPG) with respect to x_h(MD). This implies that, if we localize (recognize) MD, and if we learn the above relationships in the $\pm x$ direction between MD and the parotid glands, we may be able to refine (in the Bayesian sense) the extents of the localized models of RPG and LPG in the $\pm x$ direction. In the model building stage, we learn such relationships and incorporate them into *FAM*(*B*, *G*) (in the ρ component), and in the recognition stage, this information is exploited to predict locations $x_l(O)$ and $x_h(O)$ for RPG and LPG. Only those relationships that are "steady" are utilized for this learning (modeling) and prediction processes. We will describe in this section the modeling part. The prediction process will be explained in Section 3.3 on object recognition. We assume that all locations (*x*, *y*, and *z*) are specified with respect to the scanner coordinate system and *after* binary object samples are scaled and aligned in the process of creating fuzzy models $FM(O_l)$. Since the three coordinate directions are handled in exactly the same manner, we present the details for the *z* direction only.

We employ the mechanism of a Directed Probability Graph, which is a directed acyclic graph, to model the above location relationships. In our case, the graph is expressed as $DG_z = (V_z, \mathcal{E}_z)$, whose set of nodes is $V_z = \{z_l(O_k), z_h(O_k): 1 \le k \le L\}$ and set $\mathcal{E}_z$ of directed arcs is a special subset of the set of all possible directed arcs $A_z = \{(v_i, v_j) : v_i, v_j \in v_z \& V_i \neq v_j \& (v_i, v_j) \neq (z_l(O_k), z_h(O_k))$ for any object $O_k\}$. The subset $\mathcal{E}_z$ is chosen from A_z as described below. Note that V_z has 2L nodes and each node represents a random variable (*z* location). Some elements of V_z are special which we refer to as *anchor nodes*. They represent *z*-locations (superior and inferior boundaries) of OARs which coincide with the *z*-location (superior and inferior boundaries) of the body region *B*. For example, for *B* = thorax, the superior boundary of *B* is defined to be 15 mm above the apex of the lungs and the inferior boundary is 5 mm below the base of the lungs (Udupa et al., 2014; Wu et al., 2017a, 2017b), and so, $z_l(tSC) = z_l(tSB) = z_l(tES) = z_l(TB) = z_l(B)$, and $z_h(tSC) = z_h(tSB) = z_h(B)$. That is, for these 4 OARs, one (in the superior direction for tES and TB) or both (both superiorly and inferiorly for tSC and tSB) of their *z*-location boundaries coincides with the corresponding boundaries of *B*. Since these anchor boundary locations are known precisely in the AAR approach due to the definition of *B*, we can exploit this prior knowledge to refine the automatically-identified boundary locations of all OARs. The directed arcs in $\mathcal{E}_z$ represent conditional dependencies between

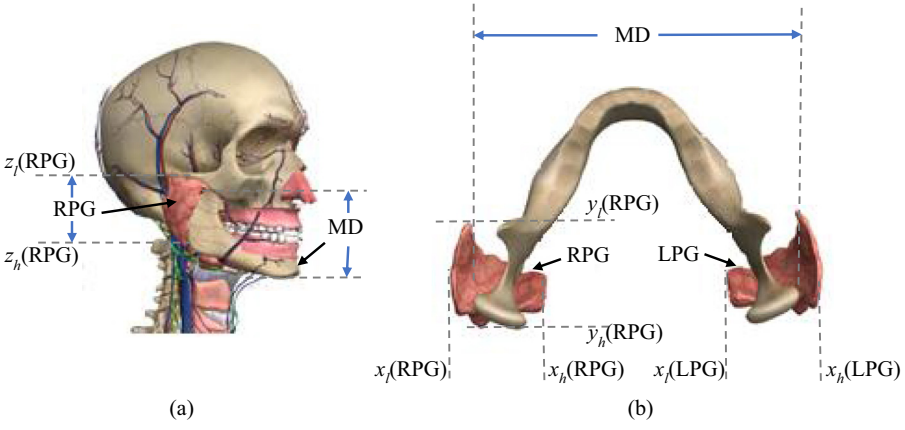

Fig. 6. Illustration of boundary relationship among Mandible (MD) and Right and Left Parotid Glands (RPG and LPG) in (a) sagittal view, (b) axial view. Boundary plane locations of RPG in the three coordinate directions are shown as x_l(RPG), x_h(RPG), y_l(RPG), y_h(RPG), z_l(RPG), and z_h(RPG). Reproduced with permission from https://zygotebody.com.

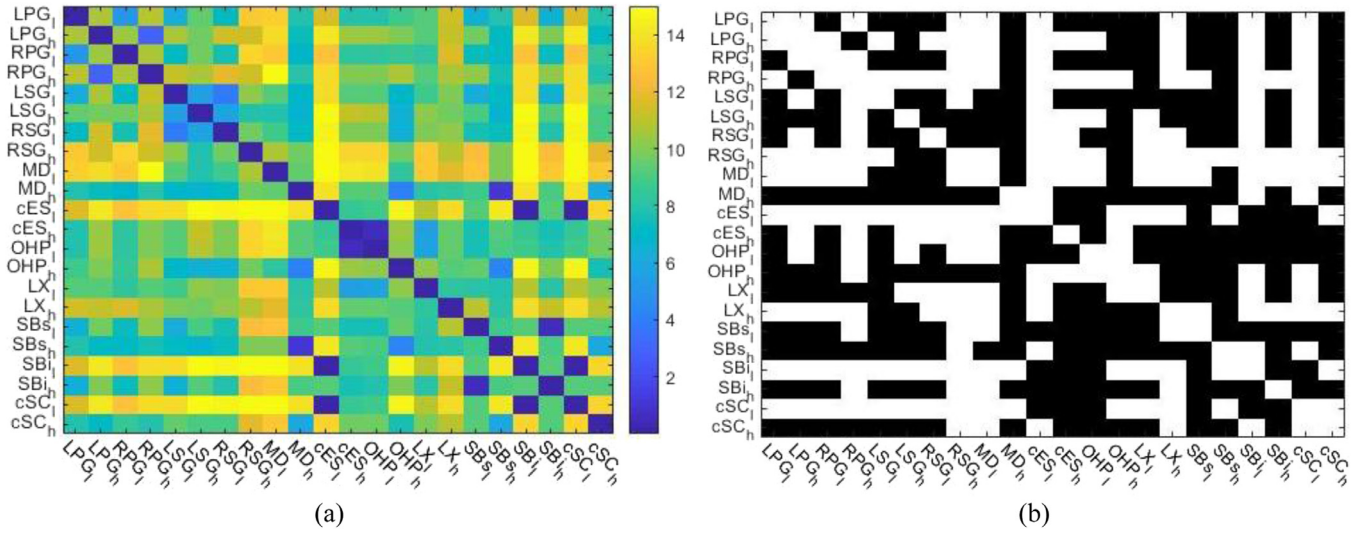

Fig. 7. (a) Matrix of the standard deviation $\sigma_{d(e)}$ values of the distances for all edges over our training data set. Edges e here denote arcs connecting boundary locations in the z-direction. The color scale is shown on the right. (b) A binary matrix obtained from (a) where cells with values $\sigma_{d(e)} \leq x$ mm are shown black. These cells suggest that the associated objects have a "steady" relationship between their z-boundary locations. For each object O, O_l and O_h denote, respectively, $z_l(O)$ and $z_h(O)$.

nodes. Nodes that are not connected represent variables that are conditionally independent of each other. Each node has a probability function associated with it which takes as input a particular set of values of the node's parent variables and gives as output the probability of the variable represented by the node.

If (v_i, v_j) is a directed arc selected from A_z to be included in $\mathcal{E}_z$, our desire is to assign a conditional probability to (v_i, v_j) such that we can reliably estimate the contribution from parent v_i to the probability of the random variable associated with v_j. Once we specify how the arcs are selected and the conditional probability $P_z(v_j/v_i)$ associated with these arcs (v_i, v_j) are determined, the Directed Probability Graph is fully specified.

We determine $\mathcal{E}_z$ in two stages. In the first stage, we determine a subset U_z of A_z of edges that show a "steady" relationship. Consider any edge $e = (v_i, v_j) \in A_z$. Let d(e) be the distance between locations denoted by v_i and v_j and $\sigma_{d(e)}$ be the standard deviation of this distance over all samples of O_i and O_j in our training set $\mathcal{I}^b$. First, we find a subset U_z of V_z by

$$U_z = \left\{ e = (v_i, v_j) : \sigma_{d(e)} \leq \tau \right\}, \tag{1}$$

where τ is a fixed threshold. The idea behind U_z is to include only those pairs of nodes which have a "steady" relationship. Note that since this distance is symmetric, if (u, w) is in U_z, so will be (w, u). $\sigma_{d(e)}$ values obtained for all edges in A_z are shown as a color matrix in Fig. 7(a) for the H&N body region for the training data cohort used for model building. Fig. 7(b) shows the result of thresholding the matrix in (a) at $\tau = 10$ mm.

In the second stage, we decide which directed edges in U_z are to be retained for inclusion in $\mathcal{E}_z$. Consider an edge $e = (v_i, v_j) \in U_z$. We include e in $\mathcal{E}_z$ iff one of the following conditions holds.

(1) v_i is an anchor node but not v_j.
(2) Both v_i and v_j are not anchor nodes and object O′ associated with v_j appears *after* object O associated with v_i in the breadth first order in the optimal hierarchy.

The rationale for condition (i) is obvious – we would like the known location represented by v_i to be utilized to predict location v_j. Note that if (u, w) is in U_z and if only u is an anchor node but not w, then although (w, u) is in U_z it will not be included in $\mathcal{E}_z$. If (u, w) is in U_z and both u and w are anchor nodes, then both edges will not be included in $\mathcal{E}_z$ since this will not be useful as both (anchor) nodes are known and there is no need to predict either node. The reason for condition (ii) is that, in the hierarchical order of recognition, object O will already have been recognized before dealing with object O′.

Finally, we assume that the conditional probabilities $P_z(v_j/v_i)$ associated with arcs $e = (v_i, v_j)$ follow a Gaussian distribution $p_z(e)$ with mean $\mu_{d(e)}$ (which is the mean of the d(e) values over the training samples) and standard deviation $\sigma_{d(e)}$. This completes the specification of the network DG_z. Similarly, networks DG_x and DG_y are constructed, except that in these cases, there are no anchor nodes since boundaries of body region B are defined only in the z-direction. In our fuzzy anatomy model $FAM(B, G)$, the ρ component is thought of as consisting of two parts, $\rho = (\rho_R, \rho_{DG})$, where ρ_R denotes parent-to-child relationship in hierarchy H (as in Udupa et al., 2014), and ρ_{DG} represents the triplet of learned Directed Probability Graphs (DG_x, DG_y, DG_z).

3.3. Object recognition

The recognition process proceeds in two steps. Initially, the optimal hierarchy H found during model building is used for locating all objects in the hierarchical order. Then, after all objects are recognized in this manner, object localization is refined using the previously-built Directed Probability Graph by again going through the hierarchical order in H.

3.3.1. Recognition via optimal hierarchy

The fuzzy anatomy model $FAM(B, G)$ built for B and G is utilized for recognizing objects in an image I of B of any patient belonging to group G. Recall that the purpose of recognition is to determine the whereabouts of the objects in I and not their precisely delineated boundaries. The AAR-RT recognition process takes, as input, image I, $FAM(B, G)$, and the names of OARs that need to be contoured among the OARs in $\mathcal{O}$, and outputs the recognized (localized) fuzzy model $FM^T(O)$ of each O that is optimally transformed to image I starting from the version of the fuzzy model $FM(O)$ in $FAM(B, G)$. This process takes place in several steps. AAR-RT first recognizes the skin object (tSB in thorax and hSB in H&N) following the original AAR approach (Udupa et al., 2014). This initializes the hierarchical recognition process. Subsequently, following

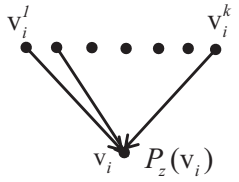

Fig. 8. A portion of DG_z is shown to illustrate the estimation of a refined z-location v_i of the fuzzy model of object O.

the optimal hierarchy H, for any object O, since the parent is already recognized and hence parent-to-child relationship ρ_R stored in the model is known, it first scales and places $FM(O)$ in I based on just ρ_R. This is called *one-shot* recognition in the original AAR approach. This placement (pose) is further refined by using the *optimal thresholded-search* strategy of the previous AAR approach. Briefly, an object-specific optimal threshold, previously learned in the model building stage and stored in the 5th parameter η of $FAM(B, G)$, is applied to I, and the pose parameters of the fuzzy model $FM(O)$ are adjusted for best fit between the thresholded image and the fuzzy model. For our discussion in the next step, we will refer to this resulting pose-adjusted model of O by $FM^t(O)$. At the end of this first step, we have $FM^t(O)$ for all $O \in \mathcal{O}$ in I.

3.3.2. Recognition refinement via directed probability graphs

In this step, we will refine the model $FM^t(O)$ obtained in the previous step to its final recognized form $FM^T(O)$ in image I by using the previously trained Directed Probability Graphs. In this process, we will refine locations $x_l(O)$, $x_h(O)$, $y_l(O)$, $y_h(O)$, $z_l(O)$, and $z_h(O)$ of object O as represented in $FM^t(O)$ by using the information stored in the ρ_{DG} component of ρ. Obviously, these refinements are made for all objects except the root object. As in Section 3.2(ii), we will take the z-direction to describe the refinement process. The x- and y-directions follow the same procedure.

For any node v_i of V_z, let its parent nodes in $DG_z = (V_z, \mathcal{E}_z)$ be denoted by $\{v_i^1, \ldots, v_i^k\}$.[3] Examine the situation shown in Fig. 8. By the Markov property of DG_z, we can write

$$P_z(v_i | v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{2L}) = P_z\left(v_i \middle| v_i^1, \ldots, v_i^k\right). \quad (2)$$

Assuming that the parents are conditionally independent, we can write

$$P_z(v_i) = P_z\left(v_i \middle| v_i^1\right) P_z\left(v_i^1\right) + \cdots + P_z\left(v_i \middle| v_i^k\right) P_z\left(v_i^k\right). \quad (3)$$

Since our recognition process proceeds hierarchically, when we are dealing with object O whose z-location node v_i is being refined, the z-locations of its parents $\{v_i^1, \ldots, v_i^k\}$ have all been refined already and hence known. Let these refined actual locations for O in I be $\{u_i^1, \ldots, u_i^k\}$. Based on these known parent locations and the priors $p_z(e)$ (Gaussian distributions associated with each edge $e = (v_i^j, v_i)$ with parameters $\mu_{d(e)}$ and $\sigma_{d(e)}$ as described previously in Section 3.2(ii)), we approximate the probability in Eq. (3) of node (location) v_i as follows.

$$P_z(v_i) \approx \max\left[g_{ij}\left(v_i \middle| v_i^1, \ldots, v_i^k\right)\right], \text{ where}$$

$$g_{ij}\left(v_i \middle| v_i^1, \ldots, v_i^k\right) = \sum_{j=1}^{k} \left[u_i^j + p_z(e)\right]. \quad (4)$$

That is, the individual priors associated with each edge are shifted by the known location of the parent and then added. The result is a mixture of Gaussians whose maximum is taken to be the predicted probability of v_i. We denote the predicted location where this maximum occurs by w_i^p.

[3] Note that although each node has exactly one parent in H, a node may have several parents in DG_z (and DG_x and DG_y).

From the known location of model $FM^t(O)$ before performing refinement, we know a location for node v_i coming from the recognition process. Let this location be denoted by w_i^r. We will make use of both these locations w_i^r and w_i^p and their associated probabilities $P_z(w_i^r)$ and $P_z(w_i^p)$ to fuse them to estimate the final refined location w_i.

$$w_i = w_i^p + \left(w_i^r - w_i^p\right) \frac{P_z\left(w_i^r\right)}{P_z\left(w_i^r\right) + P_z\left(w_i^p\right)}. \quad (5)$$

The refinement process proceeds in this manner in the hierarchical order. The fuzzy model $FM^t(O)$ found before refinement is finally rescaled to fit the refined boundary locations, which yields the refined model $FM^T(O)$ for all $O \in \mathcal{O}$ in I.

3.4. Object delineation

The delineation process proceeds in two steps. First, the localized fuzzy model $FM^T(O)$ output in the recognition step for each $O \in \mathcal{O}$ in I is utilized to identify all voxels that are within the object region in I via a kNN classifier. In the second step, to this cluster of identified voxels the fuzzy model $FM^T(O)$ is optimally fit to produce the final delineation.

3.4.1. Delineation via fuzzy connectedness/kNN voxel classification

For skin objects, we use Iterative Relative Fuzzy Connectedness (IRFC) algorithm (Ciesielski et al., 2007) as elaborated in (Udupa et al., 2014). IRFC is an image-based delineation engine that requires the specification of a set of seed voxels and a local affinity function. It is suitable for large objects with sufficient intensity contrast wherein automatic seed selection and object-specific affinity specification work well. For small and sparse objects, automatically selecting seeds often fails. Therefore, for all objects other than skin objects, we employ a trained k-nearest-neighbor (kNN) voxel-wise classifier to find the object voxels within the fuzzy mask specified by $FM^T(O)$. For this purpose, we use a 3-dimensional feature vector $[f_{FM}, f_I, f_T]^t$ associated with each voxel, where f_{FM} denotes the fuzzy membership value of O as expressed in $FM^T(O)$, f_I denotes voxel intensity in I, and f_T represents a texture property value assessed at the voxel. All texture properties are from among those derived from gray-level co-occurrence matrix (Sonka et al., 2007). The texture property that is optimal for each OAR is found at the model building stage. For example, the glands (LSG, RSG, LPG, and RPG) have a similar textural characteristic among themselves but different from other objects. kNN training and estimation of all required parameters including the determination of optimal texture properties are performed automatically from the training data sets at the model building stage.

3.4.2. Optimal fuzzy model fitting

The result of kNN classification for object O is a cluster of voxels $C(O)$ in I which is typically a scatter of voxels without proper boundaries and potentially with holes (false negatives) and extraneous voxels (false positives) although all within the generous fuzzy mask defined by $FM^T(O)$. To minimize these issues, we transform the fuzzy model $FM^T(O)$ optimally to $C(O)$ by minimizing the sum of squared difference between $C(O)$ as a binary mask and $FM^T(O)$ as a fuzzy mask. The transformation involves x, y, z translations, uniform scaling, and a threshold. We will denote these 5 parameters by a vector $\boldsymbol{p}$, the binary mask resulting from $FM^T(O)$ after transforming by a given $\boldsymbol{p}$ by $BM(O, \boldsymbol{p})$, and the sum of squared difference between $C(O)$ and $BM(O, \boldsymbol{p})$ by $||C(O) - BM(O, \boldsymbol{p})||$. The final delineation of object O is found as $BM(O, \boldsymbol{p}^*)$, where $\boldsymbol{p}^*$ is the optimal transformation parameter

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \{||C(O) - BM(O, \boldsymbol{p})||\}. \quad (6)$$

Table 2
Age statistics of patients in planning and replanning studies.

Group	Thorax			H&N		
	# Patients	Mean	SD	# Patients	Mean	SD
Planning data						
Male 40–59	50	50	5	54	52.7	5.1
Female 40–59	52	51	5	54	52.4	5.0
Male 60–79	54	72	3	54	64.6	2.6
Female 60–79	54	71	3	54	67.6	4.2
Replanning data						
Male	18	66	9	22	56.9	19.2
Female	12	70	5	8	57.9	16.8

4. Experiments, results, discussion

4.1. Data-related

As mentioned previously, we created 4 anatomy models, one each for: Thorax male (20 model-worthy data sets), Thorax female (18 model-worthy data sets), H&N male (20 model-worthy data sets), and H&N female (16 model-worthy data sets). These models were used in a gender-specific manner to test recognition and delineation performance on all test data sets. The model-worthy data sets (Table 1) did not participate in any experiments involving the testing of recognition and delineation algorithms. We performed evaluation of OAR recognition and delineation separately under four categories based on object quality: male-good, male-poor, female-good, and female-poor. We conducted two groups of experiments – the first on planning data sets and the second on replanning scans.

Table 2 lists statistics related to the age distribution of our study patients. For both body regions and for planning data sets, there is no statistically significant difference ($P > 0.05$) in age distribution between the male and female cohorts for the younger age group (G_{M1} and G_{F1}), although the difference is statistically significant ($P < 0.05$) for the older age group (G_{M2} and G_{F2}), the female group being older than the male group. The replanning CT data sets were selected randomly and not by age group. There is no statistically significant difference ($P > 0.05$) in age distribution between the male and female groups for these data sets for both body regions.

4.2. Models

Fig. 9 displays each OAR (except skin objects) selected from several model-worthy studies for male and female subjects for H&N and thorax body regions as well as the models generated for the two body regions. The optimal hierarchies found from model-worthy male data sets for the OARs in the two body regions are also included in the figure. Note that here the skin boundary was specified explicitly as the root object since it is easy to locate and delineate in CT images compared to other objects. This hierarchy was used for building all models.

All parameters involved in AAR-RT are estimated automatically from model-worthy data sets during the model building stage. There are only two additional parameters: τ (Eq. (1), Fig. 7), k in the kNN method. The values of these parameters are experimentally determined and fixed once for all at $\tau = 10$ mm and k = 50 for both thorax and H&N.

4.3. Object recognition and delineation in planning scans

We will present accuracy results from the following four experiments.

(i) E1: Auto-contouring on high-OQS objects from the male group G_M. The objects involved in this evaluation generally have streak artifacts and pathologies in not more than 3 slices and may have come from any data sets in G_M with any IQS value. Although the objects in this group had minimal artifacts, they may be still affected by pathology. OQS for these objects were in the upper end of the score scale.
(ii) E2: Similar to E1 but on the female group G_F.
(iii) E3: Auto-contouring on low-OQS objects from the male group G_M. The data sets involved in this experiment were the complement of the subset of G_M used in E1.
(iv) E4: Similar to E3 but on the female group G_F.

We express recognition accuracy in terms of location error and scale error. Location error (LE) is the distance (in millimeter) of the geometric center of the object model at final recognition to the known true geometric center of the object, ideally 0 mm. Scale error (SE) is the ratio of the estimated object size to its true size, with the ideal value of 1. We describe delineation accuracy/error via Dice Coefficient (DC) and Hausdorff boundary distance (HD). The ideal values for these parameters are 1 and 0 mm, respectively. Fig. 10 displays recognition and delineation accuracies graphically for experiments E1 through E4 on different OARs in the two body regions. Sample recognition and delineation results for the two body regions (for both good- and poor-quality cases) are displayed in Figs. 11 and 12, respectively, with a slice of the recognized model and the delineated contour overlaid on the original slice.

We make the following observations from the quantitative results depicted in Fig. 10:

(i) E1 and E2 (good object quality): There were 461 object samples (360 H&N, 101 thorax) involved in E1 described above in group G_M out of a total of 1479 samples (32%). The corresponding numbers for group G_F in E2 were 658 object samples (545 H&N, 113 thorax) from a total of 1391 (57%). The voxel size in our data sets varied from $0.93 \times 0.93 \times 1.5\,mm^3$ to $1.6 \times 1.6 \times 3\,mm^3$, most with a slice spacing of 2 mm or more. Overall, the location error of recognition (object localization) in these experiments, as seen from the last bar in Fig. 10 labeled "All", is 4.28 mm for the male group and 4.01 mm for the female group, which is about 2 voxels, and the overall accuracy in delineation is close to 0.7 for DC and within about a voxel from true boundary for HD. The difference in recognition and delineation accuracy between the male and female groups is not statistically significant ($P > 0.05$). Note a similar trend in accuracy for the different objects between the male and female groups. Some OARs like OHP, cES, tES, LBP, and RBP are more challenging than others, but given images with minimal artifacts and generally nominal pathology, they can all be located and delineated quite accurately via AAR-RT.

In understanding these results, two points should be noted: (1) DC is known to be very sensitive to errors in small and sparse objects where small errors of the order of a voxel at different parts of the boundary can deteriorate DC drastically. In this sense, HD may be a more robust measure. (2) There is considerable variation in the ground truth delineations themselves. As will be shown under results from the second experiment (Section 4.4), DC depicting the variability between two dosimetrists is significant. Considering these two points, our results from E1 and E2 are excellent: HD is about 1 voxel and DC is comparable to DC between dosimetrists. If we exclude the five most challenging objects tES, cES, OHP, LBP and RBP,[4] then

[4] These objects are rarely considered in most published papers but need to be contoured frequently in RT planning. LBP and RBP are hard to even visually locate on slices and so also esophagus on some slices in the H&N and thoracic regions.

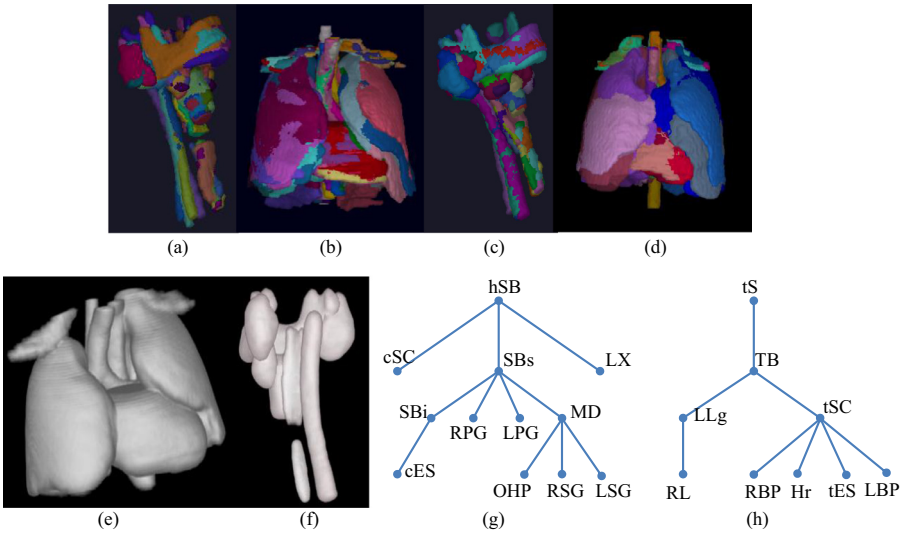

Fig. 9. Object samples from model-worthy data sets shown as surface renditions, models built from model-worthy data sets shown as volume renditions, and optimal hierarchies for H&N and thorax. (a) Male H&N, 9 OARs each from 5 studies. (b) Male thorax, 8 OARs each from 5 studies. (c) Female H&N, 9 OARs each from 5 studies. (d) Female thorax, 8 OARs each from 3 studies. Object samples represent binary objects after they are aligned in the scanner coordinate system during the model building process. Models for (e) thorax (right anterior oblique view), and (f) H&N (left posterior oblique view). Optimal hierarchies found for (g) H&N, and (h) thorax. See Table 1 for object names.

the overall DC for our results for the remaining 16 OARs over the test object samples becomes 0.80 compared to DC of 0.81 (see below) between the dosimetrists.

(ii) E3 and E4 (poor quality): When the objects have significant artifacts, the results are much worse, hovering around 5 voxels for recognition and 5 mm for HD. The streak artifacts pose serious challenges to object recognition and delineation, especially in the H&N images due to dental implants and tooth fillings. Among our planning data in H&N, only 2 out of 216 scans were completely free of any streak artifacts (although they contained other deviations), 188 (87%) had streak artifacts arising from beam hardening from metal (Fig. 2), and 26 (12%) had beam hardening effects from bone. The other factors encoded in OQS, such as existence of pathology and body posture deviation also lead to much worse results compared with E1 and E2. Again, there is no statistically significant difference between results for the male and female subjects in these experiments.

The previous AAR method (Udupa et al., 2014) was tested on near-normal studies. Since all scans considered in the present work contained deviations from normalcy, in almost all cases, AAR-RT yielded better results than the previous approach. To illustrate the improvements brought about by the innovations incorporated in AAR-RT, we compare in Fig. 13 final delineation accuracies achieved for some sample OARs before and after the proposed improvements.

4.4. Object recognition and delineation in replanning scans

We have gathered retrospectively image data from 60 patients who underwent PBRT fractionated treatment serially (see Table 1). For each patient, we selected image data at 2 to 3 serial time points, accounting for a total of 82 studies in H&N and 87 in thorax. Like Fig. 10, we show in Fig. 14 recognition and delineation accuracy for replanning data sets for good- and poor-quality object cases. Generally, replanning data sets had a much larger percentage of cases with low scores (poor quality) for both OQS and IQS.

Unlike testing on planning scans, we used two methods for testing replanning studies. *Method* 1: It uses the contours drawn by the dosimetrists for a patient case at an earlier time point t_0 in the serial study as a patient-specific model to recognize and delineate object contours at later time points for the same patient. *Method* 2: At each time point, we perform AAR-RT recognition and delineation afresh by selecting the gender-specific model for the test scan. Note that no registration step is required in either method. We also assess inter-dosimetrist variation for each of the 11 tested OARs (5 in H&N and 6 in thorax, see Section 2.1) based on the contours drawn by four dosimetrists (two for each body region, drawing on the same data sets) on these 82 H&N studies and 87 thoracic studies.

From the results in Fig. 14, we make the following observations.

(1) Method 1 achieves better delineation accuracy (statistically significant for All, $P < 0.001$) overall than Method 2 on both high OQS and low OQS groups. This is because using manual contour at t_0 as the model for t_1 and t_2 preserves better patient-specific object information than the fuzzy model over population as in Method 2. Meanwhile, on most OARs, results from Method 1 are better than results in Fig. 10 for planning scans. This shows that when introducing manual assistance in auto-contouring, the performance could be further improved over fully automatic methods.
(2) Analogous to Fig. 10, when objects have high OQS (artifacts and pathologies on less than 4 slices), overall DC is 0.77 and HD is 1 voxel or less for Method 1. The corresponding values for Method 2 are 0.71 and around 1 voxel. For the low OQS data, overall DC was ~0.6 and HD was 2–3 voxels for both methods. We observe that the H&N OARs are more influenced by image quality than thoracic OARs; this is mainly due to strong streak artifacts in the H&N region on all replanning data sets.
(3) The overall inter-observer variability expressed in DC and HD between the two dosimetrists (per body region) is slightly above the DC and HD obtained via AAR-RT on high OQS data sets as shown by the hatched bar in Fig. 14. We noticed a dichotomy between DC and HD results as compared to dosimetrists' variation. Considering DC and HD, AAR-RT performance on high-OQS large/non-sparse objects such as MD, LLg, RLg, and Hrt is comparable to dosimetrists' variation, which implies that major editing effort can be

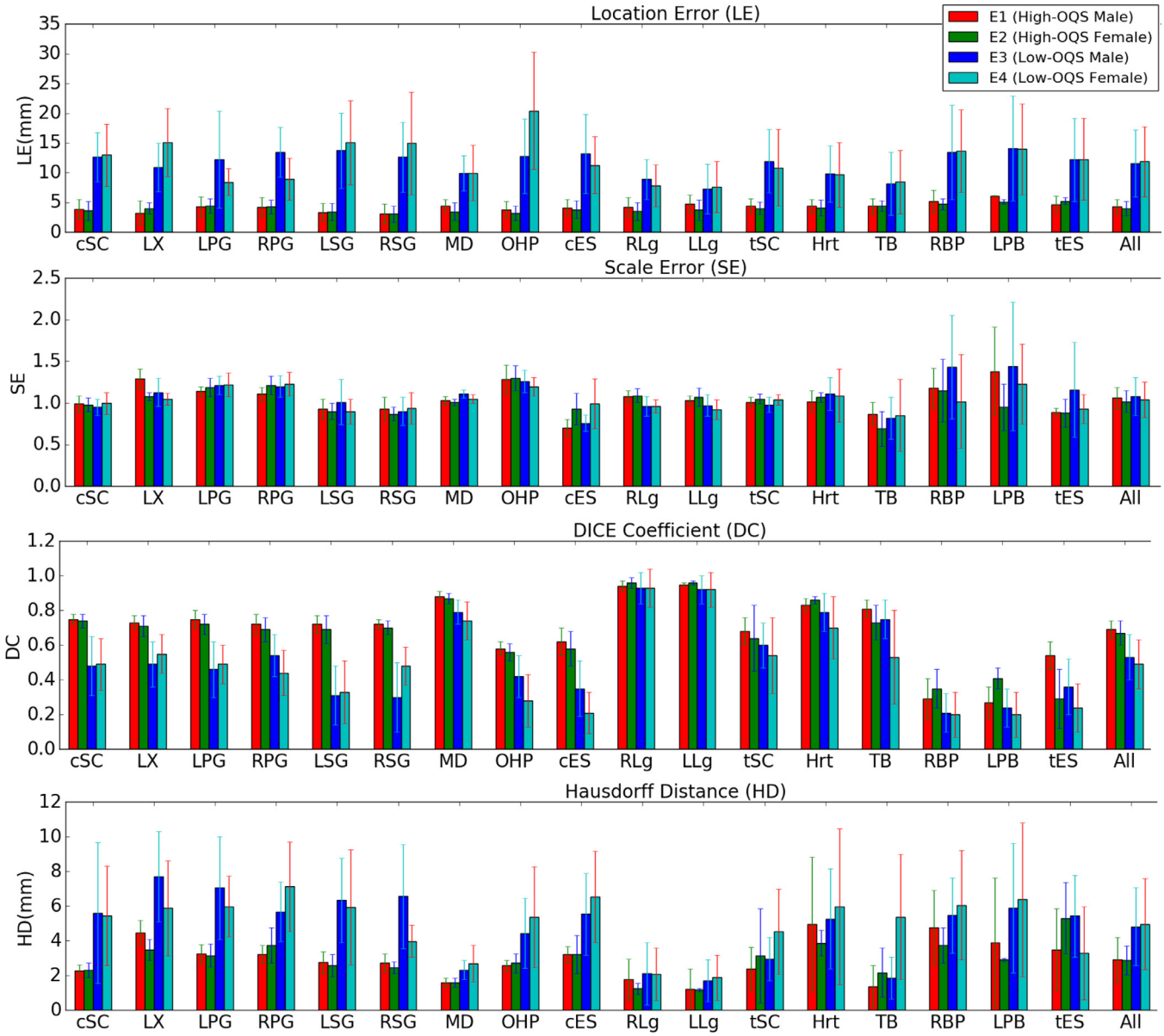

Fig. 10. Recognition errors LE and SE (Rows 1, 2) and delineation accuracies/errors DC and HD (Rows 3, 4) for the OARs in the two body regions in experiments E1–E4. See Table 1 for OAR abbreviations. The bars represent mean value over the tested object samples and the whiskers denote standard deviation.

saved when image quality is sufficient on such objects. On smaller/sparse objects, the behavior is different. Some sparse objects like cSC, cES, tSC, and TB were comparable in HD between AAR-RT and dosimetrists although their DC values for dosimetrists seem better than those from AAR-RT. This suggests a non-uniformity in the meaning of these metrics between large globular objects versus sparse objects. This is a drawback of these metrics, especially DC. Even small deviations from reference segmentations can cause drastic lowering of DC for sparse objects. Notably, for high-OQS samples, there seems to be a significant difference in AAR-RT performance between the two genders (for example, tSC, LX, cES, and tES).

(4) We are not aware of any studies in the literature that evaluated performance of algorithms on replanning data sets directly. However, earlier methods have been reported (La Macchia et al., 2012; Tsuji et al., 2010) to propagate segmentations from planning CT to replanning studies by applying image registration techniques. These techniques assume, like our Method 1, that contours are already available on planning images, mostly by manual drawing.

4.5. Computational considerations

All experiments were conducted on a PC with an Intel i7-6670 processor and 16 GB RAM. In our current implementation of AAR-RT, once the anatomy model is built, auto-contouring of all OARs for each patient study in each body region can be completed in 5–6 min, which translates roughly to 30 s/OAR. Model building itself, however, takes about 6 h for each body region, out of which the most time-consuming step is finding the optimal hierarchy which consumes about 5 h. In a clinical RT set up, this does not matter since there is no need to repeat this step very frequently.

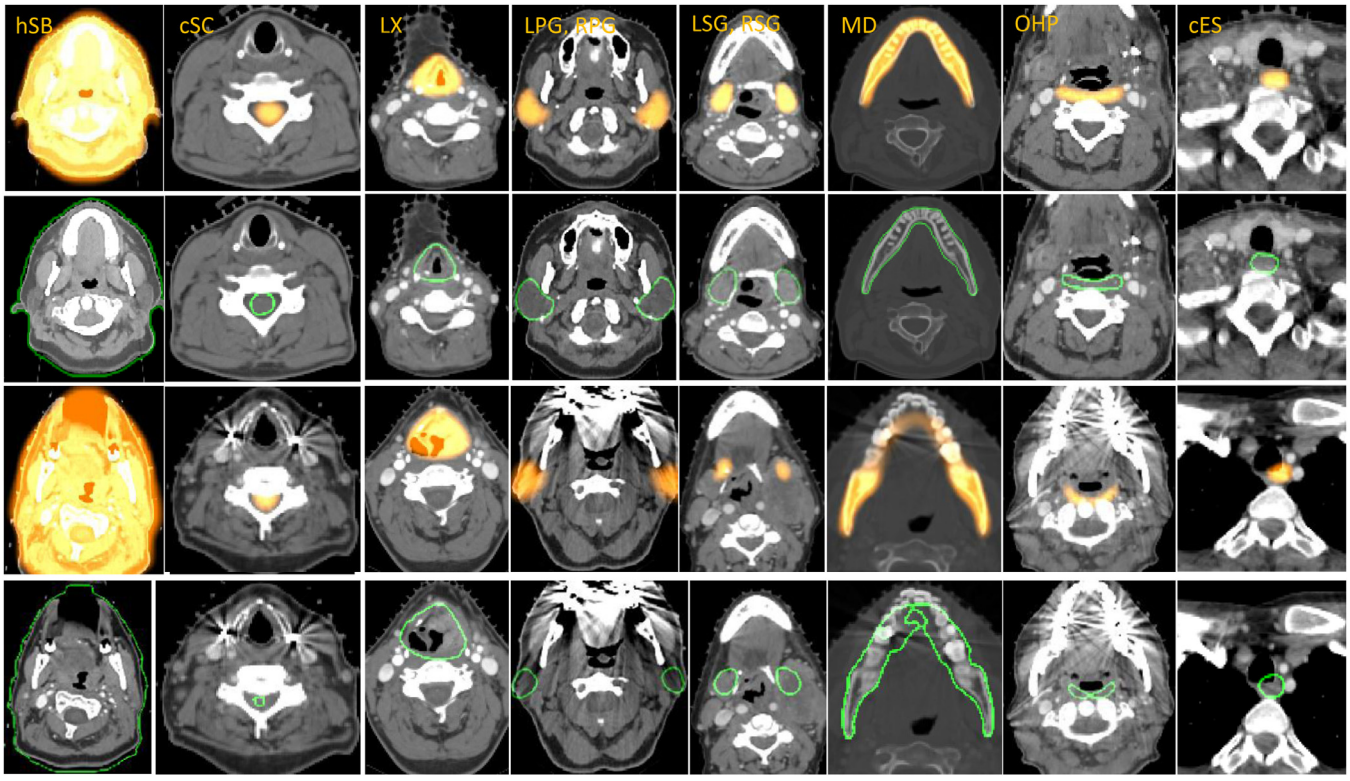

Fig. 11. Sample results for different OARs on representative H&N CT images. Rows 1 & 2: Recognition and delineation from good-quality data sets. Rows 3 & 4: Recognition and delineation from poor-quality data sets.

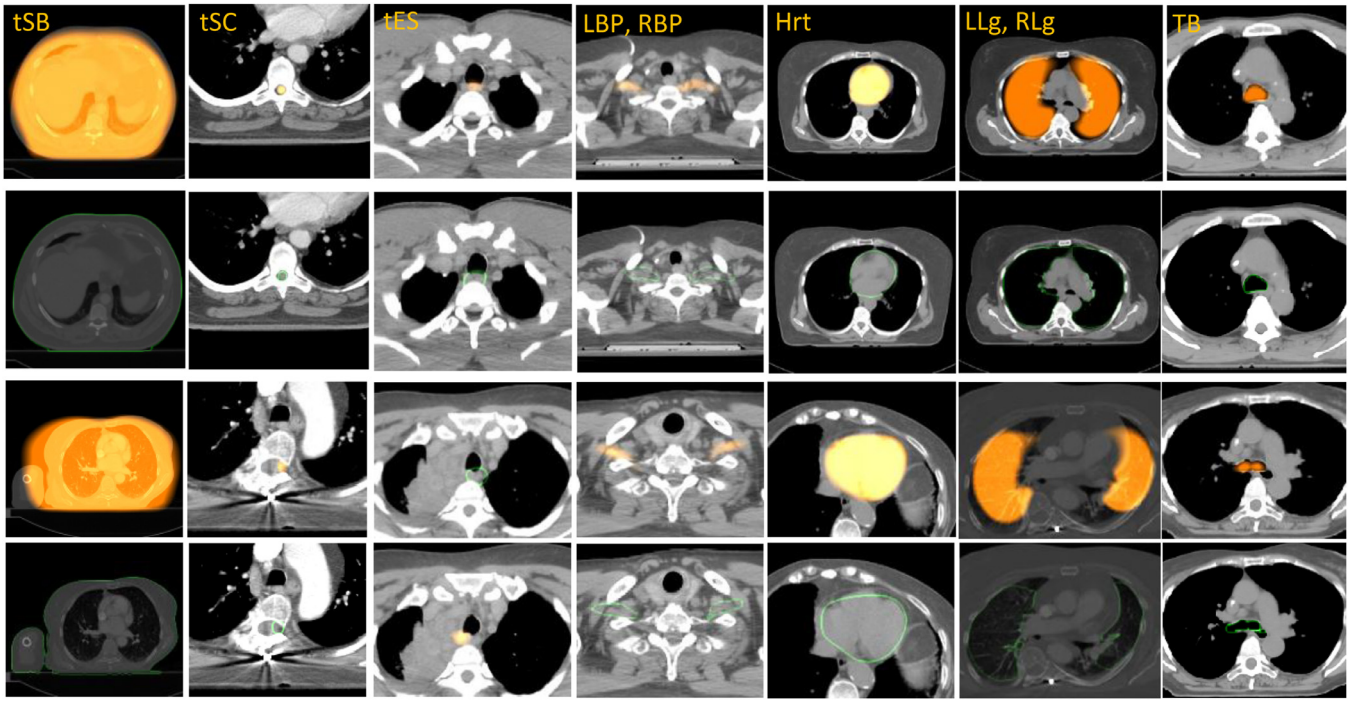

Fig. 12. Sample results for different OARs on representative thoracic CT images. Rows 1 & 2: Recognition and delineation from good-quality data sets. Rows 3 & 4: Recognition and delineation from poor-quality data sets.

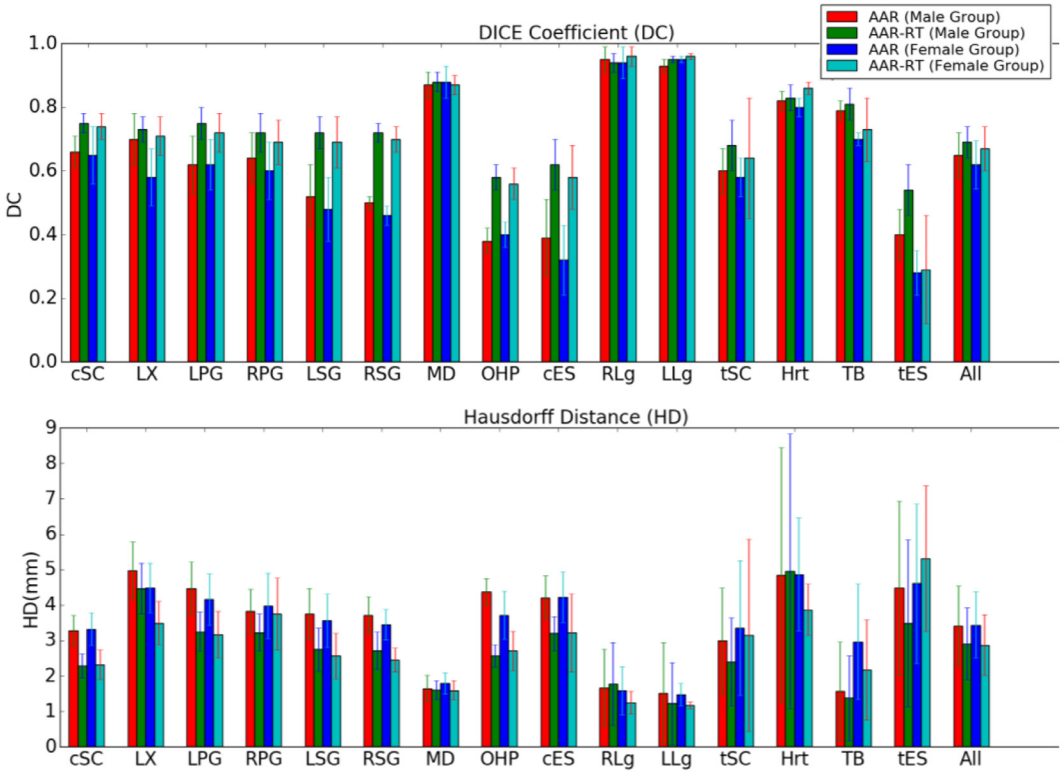

Fig. 13. Delineation accuracies/errors DC and HD for some OARs in the two body regions in experiments E1 and E2 for the previous AAR approach (Udupa et al., 2014) and AAR-RT. See Table 1 for OAR abbreviations. The bars represent mean values over the tested object samples and the whiskers denote standard deviations.

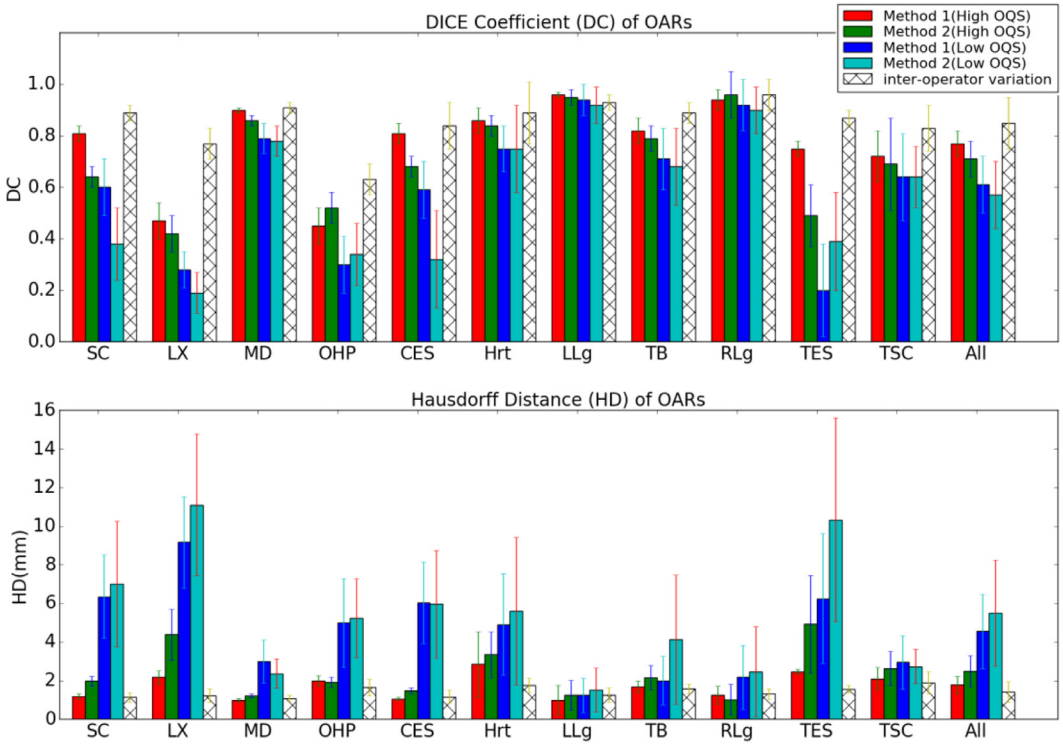

Fig. 14. Delineation accuracies/errors DC and HD for the OARs in the two body regions in experiments on replanning data sets. See Table 1 for OAR abbreviations. The hatched bar denotes metric values between the dosimetrists. The bars represent mean value over the tested object samples and the whiskers denote standard deviation.

Table 3
Comparison with methods from the literature for H&N OARs. Results are illustrated by either mean value or range. Unclear content from the literature is indicated as n/a. AAR-RT results shown are for the good-quality cases including both genders.

Approach	Number of cases Train/Test	Number of test objects	Image/object quality	DC (mean or range)
(Chen and Dawant, 2015)	25/10	60	Not mentioned	MD: 0.86–0.94 LPG, RPG: 0.74–0.87 LSG, RSG: 0.55–0.8
(Jung et al., 2015)	25/10	n/a	Not mentioned	MD: 0.77–0.86 LPG, RPG: 0.56–0.79 LSG, RSG: 0.29–0.59
(Albrecht, 2015)	25/10	n/a	Not mentioned	MD: 0.75–0.93 LPG, RPG: 0.73–0.88 LSG, RSG: 0.56–0.81
(Mannion-Haworth, 2015)	25/10	n/a	Not mentioned	MD: 0.92–0.94 LPG, RPG: 0.74–0.89 LSG, RSG: 0.65–0.87
(Orbes Arteaga et al., 2015)	25/10	n/a	Not mentioned	MD: 0.9–0.96 LPG, RPG: 0.68–0.85
(Ibragimov and Xing, 2017b)	40/10	112	Mentioned cases with streak artifacts	MD: 0.89 LPG: 0.77 RPG: 0.77 LSG: 0.7 RSG: 0.73 LX:0.86 cSC: 0.87
(Thomson et al., 2014)	n/a / 10	70	Cases not distorted by tumor or artifacts are selected	LPG, RPG: 0.74–0.83 LSG, RSG: 0.7–0.85 LX: 0.5–0.62 OHP: 0.4–0.6
(Tao et al., 2015)	n/a /16	16	Not mentioned	LX: 0.73, OHP: 0.64
(Duc et al., 2015)	100/100	600	Not mentioned	LPG: 0.65, RPG: 0.65, cSC 0.75
AAR-RT	**36/262**	**2200**	**Quality as encountered in clinical practice**	**MD: 0.89, LPG: 0.74, RPG: 0.75, LSG: 0.73 RSG: 0.73, LX: 0.74, OHP: 0.58, cES: 0.62, cSC: 0.75**

Table 4
Comparison with methods from the literature for thoracic OARs. Results are illustrated by either mean value or range. Unclear content from the literature is indicated as n/a. AAR-RT results shown are for the good-quality cases including both genders.

Approach	Number of cases Train/Test	Number of test objects	Image/object quality/artifacts	DC (mean or range)
(Zhu et al., 2013)	n/a /40	160	Not mentioned	LLg 0.95; RLg 0.95; Hrt 0.90; tSC 0.52
(Velker et al., 2013)	n/a /50	150	Not mentioned	LLg 0.95–0.98; RLg 0.95–0.98; Hrt 0.81–0.95
(Lustberg et al., 2017)	20/20	n/a	Not mentioned	LLg 0.96–0.98; RLg 0.97–0.98; tES 0.35–0.57; TSC 0.83–0.87; Hrt 0.87–0.93
(Lustberg et al., 2017)	450/20	n/a	Not mentioned	LLg 0.97–0.98; RLg 0.97–0.98; tES 0.65–0.76; tSC 0.80–0.88; Hrt 0.83–0.93
(Schreibmann et al., 2014)	n/a /46	70	Not mentioned	LLg 0.92–0.98; RLg 0.88–0.98; tES 0.01–0.54; tSC 0.52–0.87; Hrt 0.83–0.93; TB 0.81–0.95
(Trullo et al., 2017b)	30/30 (6-fold)	120	Not mentioned	tES 0.67; Hrt 0.90; TB 0.82
AAR-RT	**38/167**	**1187**	**Quality as encountered in clinical practice**	**LLg 0.95; RLg 0.96; tES 0.68; tSC 0.68; Hrt 0.86; TB 0.81; LBP 0.38; RBP 0.40**

4.6. Comparison with results from literature

We summarize some key studies from the recent literature that are related to our work in H&N (Table 3) and thorax (Table 4). Our work differs from these studies in several import ways:

(1) *Size*: Our study is much larger than any from the literature and deals with data sets that constitute the real heterogeneity that exists in clinical cases. The largest previous study from the literature considered 40 cases in H&N (Ibragimov and Xing, 2017b) and 240 cases in thorax (Zhou et al., 2017b) where most studies are used for training and only 12 cases used for testing. Including both planning and replanning scans our evaluation involved: 503 studies with 74 used for training and 429 for testing; a total of 4301 object samples where 774 were involved in training and 3527 in testing. The largest number of object samples tested in prior works is 112 for H&N (Ibragimov and Xing, 2017b) and 413 (Zhou et al., 2017b) for thorax. Testing on a large number of independent data sets (as opposed to on the same data sets in a multifold cross validation manner) is vital to get a real understanding and develop confidence for the behavior of the method in the long run independent of the data sets on which the method is tested. Testing on a large number of object samples from different body regions separately is important for similar reasons since the performance behavior of methods can be different on different objects.

(2) *Scope*: We investigated both planning and replanning studies by using the *same* approach. We did not come across any work in the literature that performed such an analysis; all reported studies tested the planning and replanning cases separately – the former by using an auto-contouring method, the latter by propagating expert-drawn contours from planning studies to replanning studies via deformable image registration. In our study cohort, as described pre-

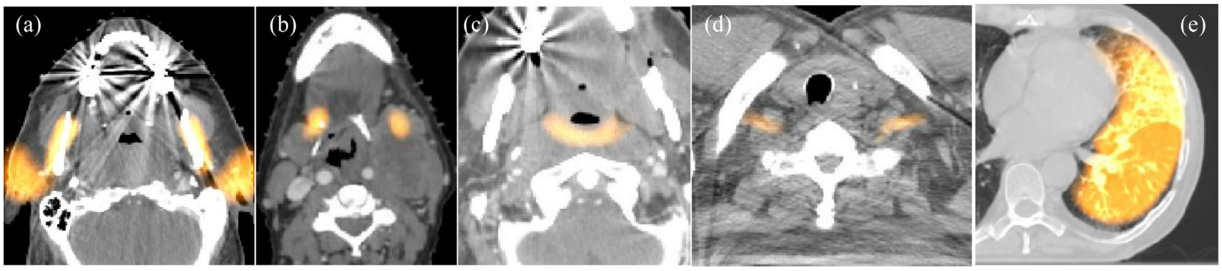

Fig. 15. Sample recognition results to illustrate the robustness of the AAR-RT recognition process. Note particularly how the model is positioned correctly/closely for (a) MD, RPG, and LPG in spite of severe streak artifacts, for (b) LSG and RSG in spite of distortions due to pathology, for (c) OHP and (d) LBP and RBP in spite of absence of sufficient appearance information, and for (e) LLg in spite of presence of large pathology.

viously, we found the quality of the images to be generally slightly lower in replanning studies than in planning cases although the performance was similar. Therefore, testing auto-contouring methods should be done separately on these data sets.

(3) *Data quality*: None of the studies from the literature discussed the quality of the data sets used, presence and severity of the artifacts in their data, and how they might influence their results. No examples of performance on cases with artifacts are given and there is no discussion of how the training and testing data sets are selected with regard to artifacts and other deviations, particularly streak artifacts.

(4) *Gender*: We analyzed gender and age dependence of image and object quality and their influence on results. Such information may be useful in the future for developing effective model building/training strategies and creating standardized and generalizable databases for evaluation.

The results listed in Tables 3 and 4 are influenced by multiple factors, such as patient gender and group, image/object quality, image resolution, definitions used (if any) for body region and OARs, manual ground-truth quality, etc. In view of the reasons listed above, a fair comparison with AAR-RT is hard to garner from these tables. Note that although the results listed for AAR-RT are for the good-quality cases, by definition, these cases include objects with artifacts and other deviations in multiple but not exceeding 3 slices. Keeping these factors in mind, our fully-automated method not only covers the largest number of OARs but also achieves very competitive performance. There is no mentioned result for cES, LBP, and RBP for comparison in the literature. One advantage of our method is its steadiness. Indeed, on every object the performance is above average standard from references, which implies the robustness of the proposed method to OAR variations in size, shape, and appearance. This is mainly because of the recognition step which is capable of overlaying the fuzzy model on the image within a location error of ∼1.5 voxels (∼3 mm) with respect to the actual location of the object. To illustrate the robustness of AAR-RT recognition process, we show in Fig. 15 several examples of severe artifacts and absence of adequate image information for locating objects even by experts, where AAR-RT successfully places the object model close to the true location on the image because of the rich prior information encoded in its anatomy model *FAM*(*B, G*).

5. Concluding remarks

In this paper, we significantly extended our previous body-wide AAR framework through several innovations and evaluated its performance comprehensively from the perspective of the RT application. Some key and unique elements of the new AAR-RT framework are as follows. (i) It uses computationally directed precise definitions of the body regions and the OARs. This becomes essential for encoding prior information consistently and faithfully and for bringing about maximum impact from prior information on object recognition. (ii) It employs a strategy to find a hierarchy for arranging OARs in each body region that seeks to minimize the error in recognition in place of a hand-crafted hierarchy in the previous AAR approach. (iii) It uses Directed Probability Graphs to encode OAR boundary relationships and to predict them at the recognition stage. (iv) Its recognition process follows the found optimal hierarchy and the trained probability graph to localize objects in a robust manner even in the presence of significant image artifacts and deviations. (v) Its delineation process uses the localized fuzzy model of the object and object-specific intensity and texture properties to identify voxels indicating strong membership within the object and to fit the model optimally to the identified voxels. (vi) It uses an image/object quality-based evaluation of both recognition and delineation processes utilizing over 500 CT scans of cancer patients undergoing RT and over 4000 object samples in these scans involving both planning and replanning studies. Our conclusions and remarks based on this study are as follows.

1. On data sets with artifacts and deviations in not more than 3 slices, AAR-RT yields recognition accuracy within 2 voxels and delineation HD within about 1 voxel. This is close to the variability observed among dosimetrists in manual contouring. When artifacts and deviations are more severe, the results are much worse, hovering around 5 voxels for recognition and 5 mm for HD. AAR-RT's performance is similar on planning and replanning cases (when using Method 2) although we observed a slightly lower object and image quality for the latter.
2. Understanding object and image quality and how they influence performance is crucial for devising effective object recognition and delineation algorithms. At present, it is very difficult to gain an understanding of the behavior of segmentation methods as a function of image/object quality in spite of the availability of large databases and many segmentation challenges. Streak artifacts arising from dental implants and fillings and beam hardening from bone pose the greatest challenge to auto-contouring methods. They cast streaks that are much brighter or darker than the actual tissue intensity and affect almost all H&N structures in almost all studies.
3. AAR's dichotomous treatment of the segmentation methodology as dual recognition and delineation processes is helpful in understanding and addressing challenges due to image artifacts and deviations. AAR's recognition operation is much more robust than delineation. We observed that often even when the models were placed very close (within 2 voxels) to the actual object with strong streak artifacts and/or deviations, delineation failed to retain that accuracy since the object intensity patterns were greatly distorted. We are studying ways to combine AAR-RT with deep learning methods to improve delineation robustness. A price to be paid for recognition robustness is the expensive computational time at the model building stage of

finding optimal hierarchies, although this step needs to be executed very infrequently.

4. Individual object quality expressed by OQS seems to be much more important than the overall image quality expressed by IQS in determining accuracy. There is an interesting phenomenon underlying OQS, object hierarchy, and accuracy of recognition. Not all ancestors influence accuracy in the same manner. A study of the relationship among OQS, IQS, recognition accuracy, delineation accuracy, and object hierarchy may help to improve robustness of recognition and delineation strategies.

Acknowledgment

This work was supported by grants from the National Science Foundation [IIP1549509] and National Cancer Institute [R41CA199735-01A1]. The auto-contouring problem was suggested to Udupa by Dr. Peter Bloch, Emeritus Professor, Department of Radiation Oncology, University of Pennsylvania, during an MIPG seminar presented by Udupa on the AAR framework in 2012.

References

ASTRO Website, https://www.astro.org/News-and-Publications/News-and-Media-Center/Media-Resources/Frequently-Asked-Questions/, Accessed June 2018.

Albrecht, T., Gass, T., Langguth, C., Lüthi, M., 2015. Multi atlas segmentation with active shape model refinement for multi-organ segmentation in head and neck cancer radiotherapy planning. In: Presented in head and neck auto-segmentation challenge 2015. MICCAI, Munich. [Online] Available at: http://midasjournal.org/browse/publication/968.

Brouwer, C.L., Steenbakkers, R.J., Bourhis, J., Budach, W., Grau, C., Grégoire, V., van Herk, M., Lee, A., Maingon, P., Nutting, C., 2015a. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiother. Oncol 117, 83–90.

Brouwer, C.L., Steenbakkers, R.J., Bourhis, J., Budach, W., Grau, C., Grégoire, V., van Herk, M., Lee, A., Maingon, P., Nutting, C., 2015b. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG oncology and TROG consensus guidelines. Radiother. Oncol. 117, 83–90 Supplement Material.

Chen, A., Dawant, B., 2015. A multi-atlas approach for the automatic segmentation of multiple structures in head and neck CT images. Presented in Head and Neck Auto-Segmentation Challenge 2015 (MICCAI). Munich.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Ciesielski, K.C., Udupa, J.K., Saha, P.K., Zhuge, Y., 2007. Iterative relative fuzzy connectedness for multiple objects with multiple seeds. Comput. Vis. Image Underst. 107, 160–182.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms, 3rd ed. MIT Press.

Daisne, J.-F., Blumhofer, A., 2013. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiat. Oncol. 8, 154.

de Vos, B.D., Wolterink, J.M., de Jong, P.A., Leiner, T., Viergever, M.A., Išgum, I., 2017. ConvNet-based localization of anatomical structures in 3-D medical images. IEEE Trans. Med. Imaging 36, 1470–1481.

Dolz, J., Kirişli, H.A., Fechter, T., Karnitzki, S., Oehlke, O., Nestle, U., Vermandel, M., Massoptier, L., 2016. Interactive contour delineation of organs at risk in radiotherapy: clinical evaluation on NSCLC patients. Med. Phys. 43, 2569–2580.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54.

Duc, H., Albert, K., Eminowicz, G., Mendes, R., Wong, S.L., McClelland, J., Modat, M., Cardoso, M.J., Mendelson, A.F., Veiga, C., 2015. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Med. Phys. 42, 5027–5034.

Falcao, A., Udupa, J.K., Samarasekera, S., Sharma, S., Hirsch, B.E., Lotufo, R., 1998. User-steered image segmentation paradigms: Live wire and live lane. Graphical Models and Image Process. 60 (4), 233–260.

Fortunati, V., Verhaart, R.F., Niessen, W.J., Veenland, J.F., Paulides, M.M., van Walsum, T., 2015. Automatic tissue segmentation of head and neck MR images for hyperthermia treatment planning. Phys. Med. Biol. 60, 6547.

Fritscher, K.D., Peroni, M., Zaffino, P., Spadea, M.F., Schubert, R., Sharp, G., 2014. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Med. Phys. 41, 051910-1–051910-11.

Ghesu, F.C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2017. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. IEEE Trans. Pattern Anal and Mech Intel. 41, 176–189.

Hall, W.H., et al., 2008. Development and validation of a standardized method for contouring the brachial plexus: preliminary dosimetric analysis among patients treated with IMRT for head-and-neck cancer. Int. J. Radiat. Oncol. Biol. Phys. 72 (5), 1362–1367.

Han, X., Hoogeman, M.S., Levendag, P.C., Hibbard, L.S., Teguh, D.N., Voet, P., Cowen, A.C., Wolf, T.K., 2008. Atlas-based auto-segmentation of head and neck CT images. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 434–441.

Ibragimov, B., Likar, B., Pernus, F., Vrtovec, T., 2014. Shape representation for efficient landmark-based segmentation in 3-D. IEEE Trans. Med. Imaging 33, 861–874.

Ibragimov, B., Xing, L., 2017a. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med. Phys. 44, 547–557.

Isambert, A., Dhermain, F., Bidault, F., Commowick, O., Bondiau, P.-Y., Malandain, G., Lefkopoulos, D., 2008. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. Radiother. Oncol. 87, 93–99.

Jung, F., Knapp, O., Wesarg, S., 2015. CoSMo - coupled shape model segmentation. Presented in Head and Neck Auto-Segmentation Challenge 2015 (MICCAI). Munich.

Kong F.M., Quint L., Machtay M., Bradley J. Atlas for organs at risk (OARs) in thoracic radiation therapy. RTOG website. https://www.rtog.org/LinkClick.aspx?fileticket=qlz0qMZXfQs%3d&;tabid=361 (Accessed 20, July 2018).

Kong, F., Ritter, T., Quint, D., Senan, S., Gaspar, L., Komaki, R., Hurkmans, C., Timmerman, R., Bezjak, A., Bradley, J., Movsas, B., Marsh, L., Okunieff, P., Choy, H, Curran, W., 2011. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. Int. J. Radiat. Oncol. Biol. Phys. 81 (5), 1442–1457.

La Macchia, M., Fellin, F., Amichetti, M., Cianchetti, M., Gianolini, S., Paola, V., Lomax, A.J., Widesott, L., 2012. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. Radiat. Oncol. 7, 160.

Leeman, J.E., Romesser, P.B., Zhou, Y., McBride, S., Riaz, N., Sherman, E., Cohen, M.A., Cahlon, O., Lee, N., 2017. Proton therapy for head and neck cancer: expanding the therapeutic window. Lancet Oncol. 18 (5), e254–e265.

Lustberg, T., van Soest, J., Gooding, M., Peressutti, D., Aljabar, P., van der Stoep, J., van Elmpt, W., Dekker, A., 2018. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother. Oncol. 126, 312–317.

Mannion-Haworth, R., Bowes, M., Ashman, A., Guillard, G., Brett, A., Vincent, G., 2015. Fully automatic segmentation of head and neck organs using active appearance models. In: Presented in head and neck auto-segmentation challenge 2015. MICCAI, Munich. [Online] Available at: http://midasjournal.org/browse/publication/967.

Matsumoto, M.S.M., Udupa, J.K., Tong, Y., Saboury, B., Torigian, D.A., 2016. Quantitative normal thoracic anatomy at CT. Comput. Med. Imaging Gr. 51, 1–10.

McGowan, S.E., Burnet, N.G., Lomax, A.J., 2013. Treatment planning optimization in proton therapy. Br. J. Radiol. 86 (1021), 20120288.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., 2018. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging 37, 384–395.

Orbes Arteaga, M., Cardenas Peña, D., C.D., G., 2015. Head and neck auto segmentation challenge based on non-local generative models. Presented in Head and Neck Auto-Segmentation Challenge 2015 (MICCAI). Munich.

Roelofs, E., Engelsman, M., Rasch, et al., 2012. Results of a multicentric in silico clinical trial (ROCOCO): comparing radiotherapy with photons and protons for non-small cell lung cancer. J. Thorac. Oncol. 7 (1), 165–176.

Pednekar, G.V., Udupa, J.K., McLaughlin, D.J., Wu, X., Tong, Y.T., Simone, C.B.I., Camaratta, J., Torigian, D.A., 2018. Image quality and segmentation. In: Proceedings of the SPIE Medical Imaging.

Phellan, R., Falcão, A.X., Udupa, J.K., 2016. Medical image segmentation via atlases and fuzzy object models: improving efficacy through optimum object search and fewer models. Med. Phys. 43, 401–410.

Saito, A., Nawano, S., Shimizu, A., 2016. Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs. Med. Image Anal. 28, 46–65.

Schreibmann, E., Marcus, D.M., Fox, T., 2014. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. J. Appl. Clin. Med. Phys. 15, 22–38.

Siegel, R.L., Miller, K.D., Jemal, A., 2018. Cancer statistics, 2018 CA: a cancer. J. Clin. 68, 7–30.

Simone 2nd, C.B., Ly, D., Dan, T.D., Ondos, J., Ning, H., Belard, A., O'Connell, J., Miller, R.W., Simone, N.L., 2011. Comparison of intensity-modulated radiotherapy, adaptive radiotherapy, proton radiotherapy, and adaptive proton radiotherapy for treatment of locally advanced head and neck cancer. Radiother. Oncol. 101, 376–382.

Sims, R., Isambert, A., Grégoire, V., Bidault, F., Fresco, L., Sage, J., Mills, J., Bourhis, J., Lefkopoulos, D., Commowick, O., 2009. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiother. Oncol. 93, 474–478.

Sonka, M., Hlavac, V., Boyle, R., 2007. Image processing, analysis, and Machine Vision, 4th ed. Thomson-Engineering Chapter 132007.

Tao, C.-J., Yi, J.-L., Chen, N.-Y., Ren, W., Cheng, J., Tung, S., Kong, L., Lin, S.-J., Pan, J.-J., Zhang, G.-S., 2015. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. Radiother. Oncol. 115, 407–411.

Teguh, D.N., Levendag, P.C., Voet, P.W., Al-Mamgani, A., Han, X., Wolf, T.K., Hibbard, L.S., Nowak, P., Akhiat, H., Dirkx, M.L., 2011. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int. J. Radiat. Oncol. Biol. Phys. 81, 950–957.

Thomson, D., Boylan, C., Liptrot, T., Aitkenhead, A., Lee, L., Yap, B., Sykes, A., Rowbottom, C., Slevin, N., 2014. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiat. Oncol. 9, 173.

Trullo, R., Petitjean, C., Nie, D., Shen, D., Ruan, S., 2017a. Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures. In: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R.S., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (Eds.), Medical Image Computing and Computer-Assisted Intervention (MICCAI) workshop 2017.

Trullo, R., Petitjean, C., Ruan, S., Dubray, B., Nie, D., Shen, D., 2017b. Segmentation of organs at risk in thoracic CT images using a sharpmask architecture and conditional random fields. In: Proceedings of the IEEE Fourteenth International Symposium on Biomedical Imaging (ISBI 2017), pp. 1003–1006.

Tsuji, S.Y., Hwang, A., Weinberg, V., Yom, S.S., Quivey, J.M., Xia, P., 2010. Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. Int. J. Radiat. Oncol. Biol. Phys. 77, 707–714.

Udupa, J.K., Odhner, D., Zhao, L., Tong, Y., Matsumoto, M.M., Ciesielski, K.C., Falcao, A.X., Vaideeswaran, P., Ciesielski, V., Saboury, B., Torigian, D.A., 2014. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Med. Image Anal. 18, 752–771.

Veiga, C., Janssens, G., Teng, C.L., Baudier, T., Hotoiu, L., McClelland, J.R., Royle, G., Lin, L., Yin, L., Metz, J., Solberg, T.D., Tochner, Z., Simone 2nd, C.B., McDonough, J., Teo, B.K., 2016. First clinical investigation of cone beam computed tomography and deformable registration for adaptive proton therapy for lung cancer. Int. J. Radiat. Oncol. Biol. Phys. 95, 549–559.

Velker, V.M., Rodrigues, G.B., Dinniwell, R., Hwee, J., Louie, A.V., 2013. Creation of RTOG compliant patient CT-atlases for automated atlas based contouring of local regional breast and high-risk prostate cancers. Radiat. Oncol. 8, 188.

Veresezan, O., Troussier, I., Lacout, A., Kreps, S., Maillard, S., Toulemonde, A, Marcy, P.Y, Huguet, F., Thariat, J., 2017. Adaptive radiation therapy in head and neck cancer for clinical practice: state of the art and practical challenges. Jpn. J. Radiol. 35 (2), 43–52.

Voet, P.W., Dirkx, M.L., Teguh, D.N., Hoogeman, M.S., Levendag, P.C., Heijmen, B.J., 2011. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? Dosim. Anal. Radiother. Oncol. 98, 373–377.

Wang, Z., Wei, L., Wang, L., Gao, Y., Chen, W., Shen, D., 2018. Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning. IEEE Trans. Image Process. 27, 923–937.

Whitfield, G.A., Price, P., Price, G.J., Moore, C.J., 2013. Automated delineation of radiotherapy volumes: are we going in the right direction? Br. J. Radiol. 86 (1021), 20110718. doi:10.1259/bjr.20110718.

Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone, C.B., ..., Shammo, G., 2018. Auto-Contouring Via Automatic Anatomy Recognition of Organs at Risk in Head and Neck Cancer on CT images. 2018. SPIE Medical Imaging.

Wu, X., Udupa, J.K., Torigian, D.A.. Thoracic object definition document.

Wu, X., Udupa, J.K., Torigian, D.A.. H&N object definition document.

Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D., 2015. 3D deep learning for efficient and robust landmark detection in volumetric data. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015.

Zhou, X., Takayama, R., Wang, S., Hara, T., Fujita, H., 2017a. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. Med. Phys. 44, 5221–5233.

Zhu, M., Bzdusek, K., Brink, C., Eriksen, J.G., Hansen, O., Jensen, H.A., Gay, H.A., Thorstad, W., Widder, J., Brouwer, C.L., 2013. Multi-institutional quantitative evaluation and clinical validation of smart probabilistic image contouring engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography images in the head and neck, thorax, liver, and male pelvis areas. Int. J. Radiat. Oncol. Biol. Phys. 87, 809–816.