An Evaluation of Atlas Selection Methods for Atlas-Based Automatic Segmentation in Radiotherapy Treatment Planning

Bas Schipaanboord, Djamal Boukerroui, Devis Peressutti, Johan van Soest, Tim Lustberg, Andre Dekker, Wouter van Elmpt, and Mark J. Gooding

***Abstract*— Atlas-based automatic segmentation is used in radiotherapy planning to accelerate the delineation of organs at risk (OARs). Atlas selection has been proposed as a way to improve the accuracy and execution time of segmentation, assuming that, the more similar the atlas is to the patient, the better the results will be. This paper presents an analysis of atlas selection methods in the context of radiotherapy treatment planning. For a range of commonly contoured OARs, a thorough comparison of a large class of typical atlas selection methods has been performed. For this evaluation, clinically contoured CT images of the head and neck ($N = 316$) and thorax ($N = 280$) were used. The state-of-the-art intensity and deformation similarity-based atlas selection methods were found to compare poorly to perfect atlas selection. Counter-intuitively, atlas selection methods based on a fixed set of representative atlases outperformed atlas selection methods based on the patient image. This study suggests that atlas-based segmentation with currently available selection methods compares poorly to the potential best performance, hampering the clinical utility of atlas-based segmentation. Effective atlas selection remains an open challenge in atlas-based segmentation for radiotherapy planning.**

***Index Terms*— Multi-atlas segmentation, atlas selection, radiotherapy.**

I. Introduction

CONTOURING of organs-at-risk (OARs) in radiotherapy (RT) is a necessary step prior to treatment planning. Accurate delineation of OARs to be avoided is required to estimate and reduce the delivered dose in order to control side effects of treatment [1]. Delineation of OARs and targets (tumors) is typically performed by clinical experts on a 3D anatomical image, normally Computed Tomography (CT). However, manual segmentation of structures is prone to inter- and intra-observer variability and can be very time consuming. For instance, mean manual segmentation times for Head & Neck cases ranging from 28.5 minutes [2] up to 3 hours [3] are reported in literature, depending on which structures are delineated. As a consequence, automatic techniques have gained popularity in the last decade [1], [4], with atlas-based segmentation being favored in commercial systems [1], [5].

A. Background on Atlas-Based Auto-Segmentation

Atlas-based automatic segmentation (auto-segmentation) uses the prior knowledge provided by previously contoured images (i.e. atlases) to automatically delineate OARs on an image [5], [6]. First, a spatial transformation is calculated using deformable image registration (DIR), which maps the atlas image coordinate system to that of the patient image. Then, the atlas contours are warped to the patient coordinate system using the estimated transformation. When only one atlas is used, the process is referred to as *single-atlas* segmentation (SAS). In multi-atlas segmentation (MAS), the registration and contour warping is repeated for several atlases, and the warped contours are fused into a consensus contour. By averaging out random registration errors, MAS has been shown to outperform SAS [5]–[7]. It is important to highlight that the quality of the obtained contours depends not only on the DIR and fusion algorithms but also on the quality of the atlases themselves.

The literature on medical image registration is enormous [8] and popular DIR algorithms such as optical flow [9], Demons [10], and free-form-deformations [11] have been used in atlas-based segmentation solutions. Note that the computational bottleneck of atlas-segmentation techniques is

Manuscript received January 30, 2019; revised March 16, 2019; accepted March 16, 2019. Date of publication April 9, 2019; date of current version October 25, 2019. This work was supported by InnovateUK, Grant 600277, as part of Eureka Eurostars project "Cloud Atlas" Grant E!9297. *(Corresponding author: Djamal Boukerroui.)*

B. Schipaanboord is with the Department of Radiotherapy, Erasmus MC Cancer Institute, 3015 CE Rotterdam, The Netherlands.

D. Boukerroui, D. Peressutti, and M. J. Gooding are with Mirada Medical Ltd., Oxford OX1 1BY, U.K. (e-mail: djamal.boukerroui@mirada-medical.com).

J. van Soest, T. Lustberg, A. Dekker, and W. van Elmpt are with the Department of Radiation Oncology (MAASTRO), GROW-School for Oncology and Developmental Biology, Maastricht University Medical Centre, 6229 ET Maastricht, The Netherlands.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2907072

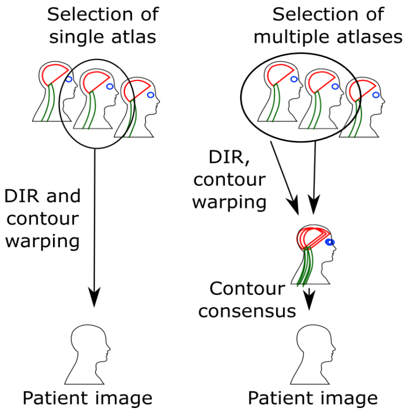

Fig. 1. Overview of atlas-based segmentation with selection (from [22]). Left: single-atlas selection and segmentation, Right: multi-atlas selection and consensus segmentation.

the sregistration step. The estimation of a consensus has also been an active field of research. Popular fusion and label propagation strategies include majority voting [12], STAPLE [13], joint label fusion [14] and other alternatives and variants [15]–[21]. An extensive review of the development of atlas-segmentation techniques over the last decade can be found in [6].

Given a large database of atlases, the selection of a subset of atlases more suitable for the patient has been proposed as a way to improve accuracy [23] (Fig. 1). Theoretically, in the extreme case using perfect atlas selection on a very large database, this may lead to performance equivalent to expert clinical contouring [24]. Selection may also improve computational speed by only performing DIR for the subset instead of the entire database. For these reasons, atlas selection methods have been proposed within auto-segmentation solutions.

B. Background on Atlas Selection

Current atlas selection methods can be broadly divided into offline and online methods, depending on their use of the patient image [6]. Offline methods select a fixed set of atlases in a training phase without reference to the patient image. Conversely, online methods make use of the patient image to search for the best atlas(es) for that specific patient. The latter assumes that the more similar the atlas is to the patient, the better the segmentation results will be. This is the motivation for contour propagation in adaptive re-planning and re-treatment clinical workflows, where the patient's previous image and contours are expected to provide better results compared to an atlas derived from a different patient. However, successful atlas selection relies on finding a good and efficient metric or similarity measure as a surrogate for contouring performance with which to rank the atlases. While this problem has been addressed in a number of works [6], [52], it still appears as a largely overlooked research topic, particularly in RT. A summary of published contributions on atlas selection is given in Table I. Note that the majority of these works focus on MR rather than the predominant modality for RT treatment planning, which is CT.

Perhaps, the simplest atlas selection approach is to stratify atlases according to non image-based information, such as age [28], [53], sex [53] or bladder filling [54]. However, it is natural to assume that the "effectiveness and efficiency of atlas selection is closely related to the registration step" [6]. Therefore, a stronger assumption is normally made: that the more similar the atlas *image* is to the patient *image*, the better the segmentation results will be. Thus, early selection methods are inspired by intensity-based image similarity measures used in registration methods [52]. Proposed measures include mean absolute difference [33], sum-of-squared difference [30] (SSD), cross-correlation [35], [38], [50], [51] (CC) and normalized mutual information [23], [25], [27], [28], [31], [34], [37] (NMI). Following the same line of argument, it may be assumed that mapping a 'similar' atlas image to the patient image will require less deformation than mapping a less similar one. Thus, similarity descriptors based on the registration deformation vector field (DVF) have been proposed. For instance, direct statistical measures, such as the mean or max magnitude of the DVF [23], or indirect geometric properties, such as the determinant of the Jacobian [32], have been used as surrogate similarity measures for atlas selection.

Both intensity and deformation similarity descriptors have been computed over the whole overlapping image region [20], [23], [27], [30], [33] and over local regions surrounding the organ of interest [25], [26], [28], [31]–[35]. Similarity descriptors have been computed following linear (i.e. rigid or affine) registration [23], [28], [30], [31], [33] or DIR [23], [25]–[27], [31], [32], [37].

Atlas selection methods can further differ in the atlas search strategy. Two strategies are commonly used; exhaustive search [20], [23], [25], [27], [33], [36], [38], [43], [50], where the similarity of each atlas to the patient image is computed, and template-based search [26], [28], [31], [32], [34], [39], [42], [44], where one or more template images are employed to reduce the number of registrations to the patient image. In the latter, the similarities of atlases to templates are computed in an offline training phase, and only the similarities between the templates and patient image are evaluated online. In this approach, atlases with an atlas-template similarity closer to the patient-template similarity are then selected. Other works explored the use of manifolds to define a similarity metric in a lower dimensional space. This is performed using linear [35], [41], [44] or non-linear [34], [36], [39], [47] projection of the input intensity or deformation field measures into a lower dimensional feature space. More recently, machine learning tools have been applied to address the selection problem, see e.g. [36], [40], [42]–[44], [48].

The use of offline selection has also been proposed as a preprocessing step to reject atlases that are not descriptive of the population sample. Most commonly, offline methods employ contouring performance indices such as the Dice Similarity Coefficient (DSC) or the Jaccard metric to evaluate the atlas contouring performance in the training phase [38], [40].

C. Study Objective

Despite the large number of selection methods, either proposed in the literature or commercially available, very

TABLE I

OVERVIEW OF ATLAS-BASED SEGMENTATION TECHNIQUES THAT USE ATLAS SELECTION. TYPE INDICATES ONLINE/OFFLINE SELECTION. # REPORTS THE TOTAL NUMBER OF USED IMAGES. ABBREVIATIONS USED ARE: NORMALIZED MUTUAL INFORMATION (NMI), MEAN ABSOLUTE DIFFERENCE (MAD), MEAN SQUARE ERROR (MSE), MANIFOLD DISTANCE (MD), CROSS-CORRELATION (CC), DEFORMATION VECTOR FIELD (DVF), DICE SIMILARITY COEFFICIENT (DSC), HISTOGRAM OF ORIENTED GRADIENTS (HOG), EXPECTATION-MAXIMIZATION (EM), GAUSSIAN MIXTURE MODEL (GMM), MAXIMUM LIKELIHOOD (ML), DEFORMABLE IMAGE REGISTRATION (DIR), CONFOCAL MICROSCOPY (CM), MAGNETIC RESONANCE (MR), COMPUTED TOMOGRAPHY (CT), CARDIAC CT ANGIOGRAPHY (CCTA) AND CONTRAST-ENHANCED CT (CECT)

Reference	Year	Type	Method	Organs	Modality	#
Rohlfing *et al.* [23]	2004	on	Global NMI, DVF after affine, DIR	Bee brain	CM	20
Wu *et al.* [25]	2007	on	Local NMI after DIR	Brain	MR	22
Commowick *et al.* [26]	2007	on	Local DVF wrt template after DIR	Head and neck	CT	45
Klein *et al.* [27]	2008	on	Global NMI after DIR	Prostate	MR	88
Aljabar *et al.* [28]	2009	on	Local NMI wrt template after affine	Brain	MR	275
Sabuncu *et al.* [29]	2009	on/off	GMM of templates using ML+Generalized EM	Brain	MR	50
Išgum *et al.* [17]	2009	off	Sequential forward selection on Jaccard index	Heart and Aorta	CT	29
Gorthi *et al.* [30]	2010	on	Global MSE after affine	Parotids	CT	18
Langerak *et al.* [18],	2010	on	Selection based on an estimated DSC	Prostate	MR	100
Lötjönen *et al.* [31]	2010	on	Local NMI wrt template(s) after: affine, DIR	Brain	MR	78
Ramus *et al.* [32]	2010	on	Local DVF wrt template after DIR	Head and neck	CT	105
van Rikxoort *et al.* [33]	2010	on	Global and local MAD after affine	Heart/caudate nuc.	CT/MR	29/39
Wolz *et al.* [34]	2010	on	MD of local NMI wrt template after DIR	Hippocampus	MR	826
Yang *et al.* [35]	2010	off	CC features from local PCA features	Parotids	CT	18
Cao *et al.* [36]	2011	on	MD of local intensities after DIR	Prostate	MR	40
Dowling *et al.* [37]	2011	on	NMI after DIR	Prostate	MR	50
Akinyemi *et al.* [38]	2012	off/on	Offline selection and local CC	Heart/kidney	CCTA/CECT	21/24
Hoang-Duc *et al.* [39]	2013	on	MD of global DVF wrt template	Hippocampus	MR	826
Langerack *et al.* [40]	2013	off/on	Clustering and selection based on DSC	Prostate	MR	200
Wolz *et al.* [20]	2013	on	Intensity MSE based pre-selection	liver, spleen, pancreas, kidney	CT	150
Asman *et al.* [41]	2014	off/on	PCA appearance after Group-wise rigid	Spinal cord structures	MR	67
Sanroma *et al.* [42]	2014	off/on	DSC regression based HOG after affine wrt template	Brain	MR	183
Sanroma *et al.* [42]	2014	on	HOG after affine wrt template	Brain	MR	183
Wang *et al.* [43]	2014	on	Graph based clustering after DIR and selection	Brain	MR	35
Asman *et al.* [44]	2015	on	AdaBoost on PCA appearance after affine wrt template	Brain	MR	3464
Langerack *et al.* [45]	2015	off	Local selection based on estimated DSC	Prostate	MR	125
Xu *et al.* [46]	2015	on	EM using GMM learned tissue prior	12 abdominal organs	CT	100
Yan *et al.* [47]	2015	on	MD on Intensities + label after DIR	Prostate	MR	60
Zhao *et al.* [48]	2016	off/on	Metric distance learning on intensities in ROI	Corpus callosum	MR	200
Karasawa *et al.* [49]	2017	on	Jaccard index or CC on vessel structures after DIR	Pancreas	CT	150
Yang *et al.* [50]	2017	on	Intensity CC after rigid + KL divergence after DIR	Esophagus	CT	66
Antonelli *et al.* [51]	2019	on/off	Local intensity CC after DIR + DSC overlap	prostate/Left ventricule	T2W/cine MR	190/45

little effort has been dedicated to investigating the underlying assumptions of atlas selection: namely that the similarity descriptors used are actually relevant and find the best atlases for segmentation, particularly when applied in RT. An earlier investigation by Ramus and Malandain [52], on 105 Head and Neck CTs, showed poor correlation between the similarity descriptors and segmentation performance. In this work,[1] we wish to answer the question "how does the best atlas selection method compare to optimal performance in the context of RT?." A large number of atlas selection methods are investigated, using a wide range of similarity descriptors. In this way, an extensive analysis of atlas selection is made on a relatively large dataset, comparing the performance of existing atlas selection methods to the best possible segmentation performance for the given set of atlases when assuming *perfect* atlas selection.

II. MATERIALS AND METHODS

A. Data and Evaluation Metrics

Two databases of clinically contoured cases were created. The head and neck (HN) database comprised 316 patient cases, and the thoracic (LN) database consisted of 280 patient cases. Both databases were acquired at the Department of Radiation Oncology, MAASTRO (Maastricht, Netherlands), and consisted of planning CTs with clinical contours created according to institutional guidelines and reviewed by the treating radiation oncologists. The study was approved by the local Institutional Review Board. The ranges of the isotropic in-plane and through-plane CT image resolution were 0.803-1.602mm and 1-4mm, respectively. Table II reports the OARs considered in this study for both HN and LN databases. The number of occurrences for each OAR considered is reported in the table, since only the clinically necessary OARs had been delineated for each case. Investigations were

TABLE II

LIST OF OARS CONSIDERED, AND THE NUMBER OF OCCURRENCES OF EACH IN THE *HN* AND *LN* DATABASES

HN		LN	
OAR	#	OAR	#
Brain	265	Esophagus	241
Brainstem	236	Heart	240
Cochlea L	262	Lung L	256
Cochlea R	262	Lung R	254
Oral Cavity	271	Mediastinum Env.	221
Parotid L	257	Spinal Cord	262
Parotid R	261		
Spinal Cord	310		
Submand. Gland L	262		
Submand. Gland R	265		

[1] This work builds on preliminary results presented at the AAPM [55] 2016 conference.

performed using only the listed OARs; target volumes and less frequently contoured OARs were not considered. Ideally, all atlases would either be created by a single expert or by merging contours from multiple observers on every case to ensure consistency. However, such an ideal data set would be very time-consuming to create, and it would not be practical to build a large database in this way in a clinical context. Therefore, clinical data, contoured by various experts, was used in this study. Consequently, the results represent what would be achievable in clinical practice, bearing in mind that the contouring was deemed of sufficient quality to deliver radiotherapy treatment.

Segmentation performance was assessed by one volume overlap measure and three metric distances computed between the estimated atlas-based contours and the original clinical contours. Specifically, we used the measures that were used for the liver segmentation challenge [56]: the Dice Similarity Coefficient (DSC), the Hausdorff contour Distance (HD), the Average symmetric surface Distance (AD) and the Root Mean Square symmetric surface Distance (RMSD).

B. Registration and Consensus Contour Generation

In the subsequent experiments, three registration methods were used;

RIGID - Normalized Mutual Information (NMI) intensity-based rigid registration that estimates 3D translations and rotations between the two CT images being registered.

ICP$_b$ - An Iterative Closest Point [57] rigid registration was estimated using the voxels having Hounsfield Unit (HU) values in the range [400, 1000]. Thus, this method computes the rigid transformation that best aligns the bony anatomy of the two images being registered.

DIR - Deformable Image Registration based on Lucas-Kanade optical flow [9]. The RIGID registration was employed as initialization for the DIR algorithm.

The image-based registrations (RIGID, DIR) were performed using methods available in clinical software (Mirada RTx 1.6, Mirada Medical Ltd., Oxford, UK), while the bone-based registration (ICP$_b$) was implemented in Matlab (MathWorks Inc, Natick, MA, USA). Unless otherwise stated, DIR was used for warping the atlas contours to the patient frame of reference. Where required, all consensus contours were generated using a clinically available form of majority voting (Workflow Box 1.4, Mirada Medical Ltd., Oxford, UK).

The choice of a clinically available method for DIR gives an indication of performance achievable in the clinic, with only minor differences being observed between different commercial systems [58]. While the choice of algorithms may impact the absolute contouring performance, and more recent algorithms may result in better contours, it will do so for both selected atlases and perfect atlases. Thus, a fixed choice enables the relative assessment of the selection methods.

C. Exhaustive Pairwise Registration

Using each of the registration methods, exhaustive SAS was carried out between each pair of case combinations of the same anatomical region. The experiment corresponds to a leave-one-out cross-validation, with each image in turn considered to be the patient image and the remaining cases regarded as candidate atlases. The clinical contours of the "patient" were treated as reference contours and were then used to assess the accuracy of the resulting segmentation. For each patient, the atlases resulting in the best segmentation following DIR can be regarded the "perfect" selection set.

D. Online Atlas Selection Methods

In addition to calculating the segmentation performance for all case-to-case combinations, surrogate similarity descriptors were computed. The following classes of similarity measure were used:

Intensity - Three surrogate similarity measures based on image intensity were considered: SSD, NMI and CC.

Deformation - Two measures based on deformation field were considered: Median Harmonic Energy of the DVF and the Median Determinant of Jacobian of the DVF

Tissue - Two tissue overlap measures were considered: DSC overlap of bone and DSC overlap of soft tissue. Regions were defined based on CT Hounsfield Units ranges, with soft tissue defined as 40 to 80 HU and bone as 400 to 1000 HU.

Image and tissue similarity measures were evaluated between the patient image and the transformed atlas image, using each of the three registration algorithms described in Section II-B. DVF-based similarity measures were evaluated only when a DIR is applied. All measures were computed for two types of region of interest:

Global - The similarity measure was computed over the extent of the overlap between the atlas and patient images.

Local - The similarity measure was computed over a bounding box surrounding each OAR. The size of the bounding box was determined in a pre-processing step for each OAR as the smallest box in real-world units that would contain all bounding boxes of the OAR from all atlases, once their centers had been aligned. During similarity computation, the bounding box was centered at the center of bounds of the estimated atlas contour. The use of a fixed size bounding box ensured that all similarity descriptors were computed on the same field of view, enabling their direct comparison.

Consequently, we tested 26 online atlas selection methods for every organ resulting from combinations of a similarity measure, registration method, and ROI, as listed in Table III. While the range of methods used cannot be considered comprehensive due to the unlimited possibilities that could be developed, it represents a broad range of most common methods considered and/or available in radiotherapy applications. Section IV-C gives further discussion on not considered approaches.

E. Offline Atlas Selection

Offline selection aims to select the atlas(es) that outperform other candidate atlases on average in an offline training

TABLE III
SUMMARY OF ONLINE ATLAS SELECTION METHODS

ROI	Similarity	Registration
Local	Intensity (SSD)	Rigid
Global	Intensity (SSD)	Rigid
Local	Intensity (SSD)	ICP_b
Global	Intensity (SSD)	ICP_b
Local	Intensity (SSD)	DIR
Global	Intensity (SSD)	DIR
Local	Intensity (NMI)	Rigid
Global	Intensity (NMI)	Rigid
Local	Intensity (NMI)	ICP_b
Global	Intensity (NMI)	ICP_b
Local	Intensity (NMI)	DIR
Global	Intensity (NMI)	DIR
Local	Intensity (CC)	Rigid
Global	Intensity (CC)	Rigid
Local	Intensity (CC)	ICP_b
Global	Intensity (CC)	ICP_b
Local	Intensity (CC)	DIR
Global	Intensity (CC)	DIR
Local	Deformation (Median Harmonic Energy)	DIR
Local	Deformation (Median Determinant of Jacobian)	DIR
Global	DSC of bone	Rigid
Global	DSC of bone	ICP_b
Global	DSC of bone	DIR
Global	DSC of soft tissue	Rigid
Global	DSC of soft tissue	ICP_b
Global	DSC of soft tissue	DIR

phase [38], [40], [45]. Specifically, the process selects atlases that represent the mode of the segmentation performance distribution from the database, and can be estimated empirically offline since the reference contours of all atlases are available. Such offline selection approaches carry the advantage of vastly reducing the number of registrations required at the time of application, enabling the use of a larger database while not increasing the time required to generate contours for any specific case [40].

In this study, an offline atlas ranking is calculated based on how a given atlas performs when segmenting the remaining cases in the database. The N atlases with the best average rank are then selected as the representative atlas set. Each OAR was considered independently, such that a different set of atlases may be selected for each OAR. Extending this for the leave-one-out evaluation approach, each atlas was in turn regarded as the test patient and removed from the database. A set of atlases to be used on the test patient was then computed from the residual database without reference to the test patient. This offline selection process was performed with each of the four segmentation performance measures, detailed in Section II-A, leading to four sets of atlas selections for each OAR.

F. Evaluation of Selection for Single-Atlas Segmentation

The best ten atlases ranked according to a given selection method implementation were used to evaluate its performance compared to two benchmark selection methods:

Oracle - Selection based on evaluation of the contouring performance of each atlas against the patient reference contour. This method represents the upper bound in atlas selection performance.

Random - Selection of atlases at random for each patient image being contoured from those in the remaining set. This method represents the most ineffective atlas selection.

Given one of the quantitative segmentation performance measures (i.e. DSC, HD, AD, RMSD), the effectiveness of selection methods was assessed using two quantitative criteria:

1) The mean value of the segmentation performance measure is computed over the best ten selected atlases for each organ and for each patient case.
2) The mean rank of the best ten selected atlases, when ranked according to the Oracle for each performance measure. Ranks were normalized between [0 1], where 0 and 1 correspond to the best and worst atlas as ranked by the Oracle respectively. This was computed for each organ and for each patient case independently.

The evaluation of single-atlas performance was made for all online, offline and benchmark selection methods. Average SAS performance of the best ten selected atlases was computed for each selection method to give a more robust estimate of the selection method's performance, and to enable the variance of the performance of each selection method to be calculated.

G. Evaluation of the Impact of Consensus Generation Following Multi-Atlas Selection

The evaluation methods detailed above represent the average performance of the selected atlases. In practice, a consensus contour would be generated from the contours resulting from these atlases to generate a final segmentation to be evaluated. However, given the large number of selection methods employed and the additional computational burden of performing selection and consensus generation per OAR, this was not reasonably tractable. Therefore, to get an indication of the impact of consensus generation compared to the impact of atlas selection, consensus generation was performed for a single OAR (the right parotid) on a small set of selection methods under the assumption that the impact of consensus generation is independent of the selection method employed.

The right parotid was selected as it is a structure that is highly variable, both in anatomical appearance and in atlas segmentation performance. Consensus generation was applied following selection using the DSC Oracle, Offline (DSC), Random and online selection based on the overlap of bony and soft tissue. As in Section II-E, the 10 best performing atlases were selected using greedy single atlas selection rather than evaluating segmentation performance combinatorially (see eg. [59]), as it is the most practical approach to implement.

In addition, MAS results were evaluated using groups of 10 atlases at ranking intervals sampled across the performance range, as ranked according to the DSC Oracle. This was done to assess the relative impact of fusion across the full range of potential segmentation performance.

III. RESULTS

A. Results of Single-Atlas Selection Method Evaluation

Given the large number of selection methods evaluated and segmentation performance measures examined, it is only

TABLE IV

AVERAGE DISTANCE MEASUREMENTS (mm), AVERAGED OVER ALL PATIENTS, FOR THE BEST PERFORMING SELECTION ACCORDING TO EACH SIMILARITY MEASURE TYPE (I.E. TAKING THE LOWEST VALUE REGARDLESS OF REGISTRATION TYPE OR ROI) FOR THE H&N STRUCTURES

Similarity type	Brain	Brainstem	Oral Cavity	Cord	Parotid L	Parotid R	Submand L	Submand R	Cochlea L	Cochlea R
Oracle	0.78	1.60	2.96	1.14	2.20	2.37	1.83	2.09	0.70	1.11
Random	1.76	3.15	8.17	5.00	4.67	5.04	4.64	5.34	2.55	4.51
Intensity	1.04	2.54	6.84	4.40	3.72	4.02	3.77	4.18	1.48	2.13
Deformation	1.20	2.49	7.04	6.41	3.60	4.18	3.25	4.11	1.54	2.08
Tissue	1.07	2.54	7.36	4.98	3.81	4.18	3.71	4.14	1.82	2.88
Offline	0.93	2.16	5.92	3.80	3.17	3.53	2.96	3.44	1.40	2.14

TABLE V

MEAN NORMALIZED RANK ACCORDING TO DSC, AVERAGED OVER ALL PATIENTS, FOR THE BEST SELECTION METHOD USING EACH SIMILARITY MEASURE TYPE (I.E. TAKING THE LOWEST VALUE REGARDLESS OF REGISTRATION TYPE OR ROI) FOR THE THORACIC STRUCTURES

Similarity type	Lung L / R	Med Env	Heart	Cord	Esoph.
Oracle	0.02 / 0.02	0.03	0.02	0.02	0.02
Random	0.51 / 0.50	0.51	0.50	0.51	0.50
Intensity	0.17 / 0.20	0.26	0.26	0.32	0.27
Deformation	0.23 / 0.25	0.28	0.28	0.40	0.28
Tissue	0.44 / 0.37	0.45	0.40	0.42	0.41
Offline	0.24 / 0.19	0.23	0.22	0.27	0.21

possible to present a broad summary of the findings. Table IV shows the best mean AD measurement, over the 10 selected atlases, for all patients for each similarity type, regardless of registration type or region of interest for the head and neck structures. It can be observed that the performance of the best online selection methods is generally closer to random than the best offline selection. Table V shows the mean normalized rank (according to the DSC Oracle) for all patients for each similarity type, regardless of registration type or region of interest for the thoracic structures. Again, it can be observed that the best performing online selection methods perform similarly, or worse than, the offline selection. Interestingly Table IV shows an asymmetry of performance of selection methods for most paired (left/right) organs. In some instances, this asymmetry appears to be statistically significant; further investigation would be required to understand this asymmetry and whether any left-right bias exists in manual delineations. Further summary results for all evaluation metrics, organs and selection methods are given in the supplementary materials, available in the supplementary files /multimedia tab.

Considering the rank of the selected atlases with respect to the Oracle rank, all selection methods, including offline, select atlases ranging from the 20^{th} quantile to the 90^{th} quantile, with random selection being approximately at the 50^{th} quantile. This indicates that the selection methods are unable consistently to select the very best atlases.

Fig. 2 shows the range of performance for each selection method according to organ, with the organs sorted left to right according to descending rank of the DSC Oracle. The selection methods are grouped according to similarity type, with the ranges indicating variation in performance resulting from choices of registration method or ROI type according to Table III. The top plot shows the mean DSC score over all patients, while the bottom plot shows the mean normalized rank over all patients (with ranks according to the AD Oracle).

The best performing selection method varied according to the OAR. For the online selection methods, local intensity-based similarity measures performed better for the lungs while performing poorly (worse than random) for the cord.

Observe that online selection methods performed marginally better than random for the majority of organs. This can be seen in the bottom plot of Fig. 2, where, for organs to the right of the heart, almost all online selection methods have an average normalized rank between 0.4 and 0.6. In contrast, offline selection performed more consistently with a normalized average rank around 0.3.

The impact of poor selection on the quantitative evaluation as measured with DSC score is small for large organs because small differences in overlap have a minimal impact on DSC values for large volumes. However, the impact is more marked with distance measures because they are not size dependent. For example in the brain – the organ with the highest DSC scores ($\geq$0.94) – the best offline selection has an average AD error of 0.93mm while random selection results in an error nearly twice that.

The plots in Fig. 3 show the mean, calculated over all test cases, of the standard deviation of rank position, calculated over 10 selected atlases for each patient, plotted against the mean, over all test cases, of the average rank position, over the selected atlases for the various selection methods. Thus, the bottom-left corner of each plot indicates selection methods where the result is both good (low average rank of the 10 selected) and consistent (low standard deviation). Moving to the right within the plot represents a decrease in average performance, while moving upward in the plot represents an increase in the inconsistency of the rank of the selected pool of atlases. From this, it can also be observed that the offline selection methods generally outperforms online selection methods both in average rank and in terms of the consistency of the ranks of the selected atlases. Plots for all organs and measures are given in the supplementary materials, available in the supplementary files /multimedia tab.

For the online selection, Fig. 3 shows that methods using local ROIs generally outperformed methods based on global ROIs, while Fig. 2 indicates tissue-based similarity measures performing poorly compared to intensity-based ones. Selection using harmonic energy after DIR generally performed better than most online selection methods, notably in the submandibular glands, but also performed poorly in the cord.

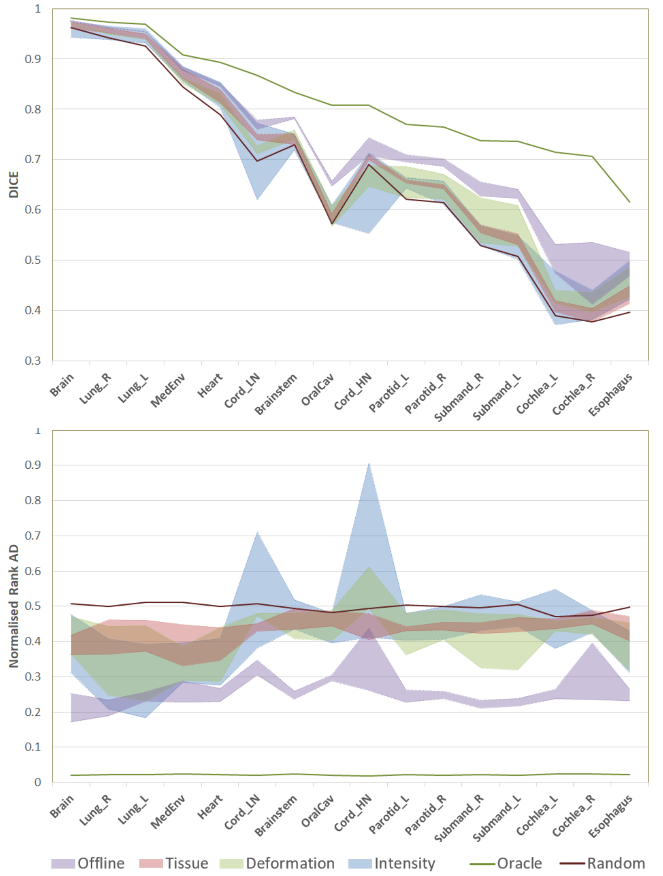

Fig. 2. Mean DSC (top) and normalized AD rank (bottom) of the 10 selected atlases for each of the organs (ordered left to right according to reducing DSC of the Oracle) averaged over all patients. Vertical ranges of the filled regions indicate variation according to changes in the registration method and ROI type used for the various similarity measures. The Oracle selection can be seen to be substantially better than the offline selection, which in turn is generally better than online selection regardless of the similarity measure. Color version and plots for other metrics are available in the online supplementary materials, available in the supplementary files/multimedia tab.

B. Impact of Consensus Generation Following Multi-Atlas Selection

As mentioned in subsection II-G, consensus generation was performed on a subset of selection methods. Results for the right parotid are illustrated in Fig. 4. It shows the SAS performance (DSC) boxplots using atlases ranked according to the Oracle. These are shown as continuous lines as they were calculated for consecutive atlases over the full performance range. The magenta boxplots show the average performance prior to fusion following selection, while the red boxplots show the results following consensus generation. These have been plotted at the mean rank position of the selected atlases over all patients, although variability in this rank exists as indicated in Fig. 3. The green boxplots show the consensus results evaluated using groups of ten consecutive atlases at sampled ranking positions. The figure shows that the impact of consensus generation is consistent throughout the range of single atlas performance, resulting in an improvement of DSC of about 0.1 for the right parotid. This suggests that atlas-based segmentation still benefits from consensus generation regardless of the quality of the selection method.

IV. Discussion

A. Performance of Atlas Selection as Observed in This Study

Comparison of atlas selection methods has shown that typical methods perform poorly compared to a *perfect* selection by an Oracle. Counter-intuitively, offline selection outperformed online selection. However, this does not imply that a generally representative case better represents a particular patient than a selected one, since the Oracle is able to perform systematically better than offline selection. Therefore, this result indicates that the surrogates of similarity used (e.g. image difference etc.) do not reflect the differences that affect segmentation performance. This poor correlation between similarity descriptors and segmentation performance was previously observed in the earlier work of [52].

The lack of correlation might be explained by clinical variation in contouring. Both the Oracle and the offline selection methods consider contours themselves. The Oracle selects based on the known reference contour; therefore, the atlas that generates the best contour result after registration is selected regardless of how accurate the initial contours are or how well the registered images match. For offline selection, this process is used in the training phase to find the atlases that best represent the patient population - i.e. the atlases are clustered according to their contour appearance rather than their image appearance. Such training process may also overcome any shortcomings of the DIR algorithm. However, online selection methods cannot consider the contours. If the images from two cases were identical but the contours showed normal clinical variation, then the result of segmentation would be less than perfect despite a good similarity score.

If this explanation is correct, then better selection performance might be found with very carefully curated data. However, to build a very large database of such carefully curated data would be impractical, both for investigation and subsequent clinical deployment, thus would not represent what could be achieved in clinical practice. Furthermore, atlas-based segmentation is ultimately employed on clinical data, and therefore an evaluation of such techniques on clinical data is of paramount importance.

B. Performance of Atlas Selection in Clinical Practice

Many works have found time saving using atlas-based autosegmentation for contouring in RT. Table VI summarizes a selection of such papers. Two studies used a relatively large number of atlases in conjunction with atlas selection. The time savings reported in these studies are comparable to similar studies without atlas selection, suggesting that the atlas selection methods utilized in these studies have little impact in clinical practice.

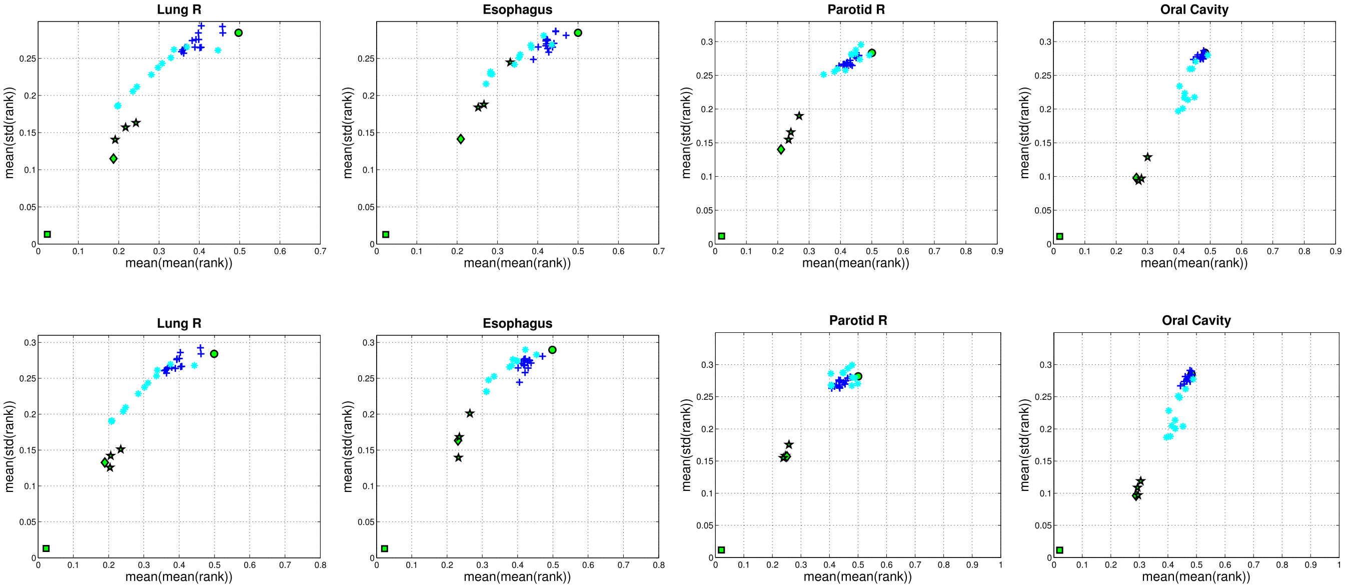

Fig. 3. Scatter plots reporting the distribution of the mean (over the population) of the standard deviation of rank position (over the 10 selected atlases for each patient) against the mean of average rank. Better performing selection methods have low mean average rank (i.e. low x-value) and low mean standard deviation (i.e. low y-value). The Oracle is represented as a square, the offline method with respect to the chosen Oracle as a diamond, the offline method with respect to alternative Oracles as stars, the random selection as a circle, and the remaining online methods as asterisks and pluses respectively for the local and global ROIs. First row: Oracle defined according to DSC. Second row: Oracle defined according to AD.

C. Limitations of This Study

Although this study includes a wide range of selection methods, the approaches evaluated are not exhaustive, even compared to the reviewed literature (see Table I).

A range of approaches could be considered to define a local ROI for similarity computation. The choice to use a simple box within this study was made for the pragmatic reason that it could be calculated once for each organ for all experiments. Furthermore, having a fixed size ROI enables direct comparison without the need for normalization. Alternative approaches, such as selecting atlases iteratively based on an initial segmentation [18], would be too computationally expensive to conduct in the leave-one-out experimental approach adopted in this work. Nevertheless, other definitions of a local ROI, including organ specific definitions, may result in improved selection.

Template based approaches were mainly introduced to reduce the computation burden at run-time and can be regarded as pure optimization techniques for the search rather than alternative selection methods. For example, in [28] comparison of the atlases to the patient using NMI is performed in template reference space so that all atlases can be pre-registered to this template. Only the selected atlases are subsequently registered directly to the patient. Commonly used similarity measures in published works are considered in this study.

It may also be argued that manifold-based selection methods only optimize the search space for the similarity measure considered, and there should not be a fundamental difference in the atlases selected in practice as locally the geometry of the manifold is assumed flat. Common similarity measures used either to build the manifold or to project the test case on the manifold to perform the search were considered in this study. For example, Wolz et al. [34] use NMI, Cao et al. [36] use a weighted similarity of pixel intensities, which can be thought of as a form of correlation, and Hoang Duc et al. [39] use the degree of deformation to build their manifold.

Only deformation measures in the ROI containing the organ were considered in this study, since the deformation of the image away from the organ does not impact the organ contour. In contrast, tissue-based measures were only calculated globally to confirm gross patient alignment. Indeed, local assessment would have resulted in many uninformative tied ranks where the region around an organ is of a single tissue type (i.e. many organs would have DSC of 1 for soft-tissue and 0 for bone.)

It is worth mentioning that this study only considered a greedy selection approach, selecting atlases independently when generating a consensus. The alternative approach which chooses the best combination of atlases is known in theory to perform better [59], but comes with an prohibitive computational cost for the size of the database considered in our study. A recent contribution suggested to reduce the complexity of the combinatorial selection by means of an optimization technique to select a ‘near-optimal’ subset [51]. Similarly, selection methods that learn surrogate measures of segmentation performance as in [42], [48] were also excluded because of the additional computational complexity associated with the need to train the system on an independent dataset. Furthermore, the challenge of accurately reproducing such methods means that such approaches have not been used clinically to date for atlas selection in radiotherapy [5].

Finally, this study has been performed for a specific clinical context using appropriate clinical data. Consequently, it should not be assumed that the findings of this study would generalize to other areas where atlas-based segmentation is used, for

TABLE VI

TIME SAVINGS FOR MULTI-ATLAS SEGMENTATION REPORTED FOR OARS IN RADIATION ONCOLOGY

Reference	Anatomy	Number of Atlases	Time saving (mins)	Time saving (%)	Selection?
Teguh et al. [3]	Head and Neck	10	114	63	Not for editing time evaluation
Young et al. [60]	Endometrial	15	9	26	Online selection
Gambacorta et al. [61]	Rectal	4	13	34	No
Hwee et al. [62]	Prostate	75	3	24	Online selection
Lin et al. [54]	Prostate	97	12	45	Stratification, followed by online selection
Granberg [63]	Prostate	15	10	26	No
Langmack et al. [64]	Prostate	8	10	40	No

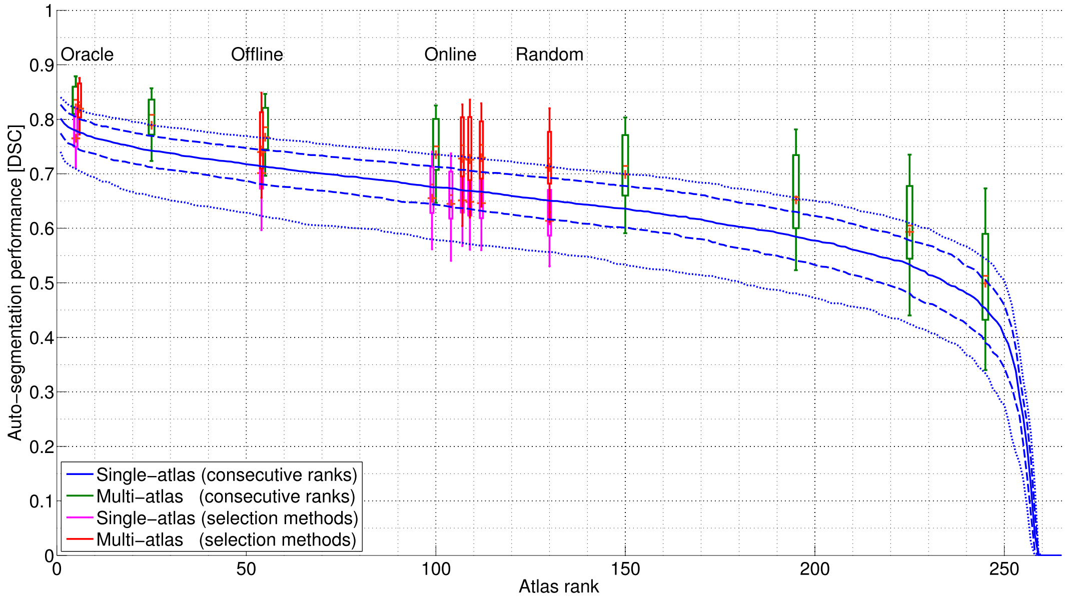

Fig. 4. Impact of consensus generation for the right parotid. Atlas ranks are given according to the DSC Oracle. Boxplots of consecutive ranks of SAS are shown by the dark blue lines, while green boxplots show MAS using consecutive rank atlases at sampled ranking positions. Results of the SAS and MAS when the atlas is selected by: the DSC Oracle, DSC offline, random, bone overlap-based and soft tissue overlap-based are shown by magenta and red boxplots respectively. These are plotted at the mean rank position of the selected atlases, but do not indicate consecutive ranks. The color figure is available online.

example anatomical localization of brain regions highlighted as active by functional imaging methods.

V. CONCLUSION

In this work, an evaluation of atlas selection methods, based on intensity and deformation measures, has been performed in the context of radiotherapy treatment planning. Offline selection of a fixed set of representative atlases was found to be the best of the selection methods evaluated. Commonly-used online selection techniques were found not to perform well, with some online selection methods performing only marginally better than random selection. We believe that atlas selection remains an open challenge in atlas-based auto-segmentation in clinical radiotherapy practice, with clinical studies using larger atlas database sizes not demonstrating improved clinical benefit compared to those using fewer atlases.

REFERENCES

[1] C. Garibaldi *et al.*, "Recent advances in radiation oncology," *Ecancermedicalscience*, vol. 11, p. 785, Nov. 2017.

[2] G. V. Walker *et al.*, "Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer," *Radiotherapy Oncol.*, vol. 112, no. 3, pp. 321–325, 2014.

[3] D. N. Teguh *et al.*, "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.*, vol. 81, no. 4, pp. 950–957, Nov. 2011.

[4] J. Y. Lim and M. Leech, "Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck," *Acta Oncologica*, vol. 55, no. 7, pp. 799–806, 2016.

[5] G. Sharp *et al.*, "Vision 20/20: Perspectives on automated image segmentation for radiotherapy," *Med. Phys.*, vol. 41, no. 5, 2014, Art. no. 050902.

[6] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[7] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, "Quo vadis, atlas-based segmentation?" in *Handbook of Biomedical Image Analysis: Volume III: Registration Models*, J. S. Suri, D. L. Wilson, and S. Laxminarayan, Eds. Boston, MA, USA: Springer, 2005, pp. 435–486. doi: 10.1007/0-306-48608-3_11.

[8] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[9] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, vol. 2. 1981, pp. 674–679.

[10] J.-P. Thirion, "Image matching as a diffusion process: An analogy with Maxwell's demons," *Med. Image Anal.*, vol. 2, no. 3, pp. 243–260, 1998.

[11] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[12] J. Kittler, M. Hater, and R. P. W. Duin, "Combining classifiers," in *Proc. 13th Int. Conf. Pattern Recognit.*, Aug. 1996, pp. 897–901.

[13] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[14] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.

[15] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 983–994, Aug. 2004.

[16] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1266–1277, Aug. 2009.

[17] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1000–1010, Jun. 2009.

[18] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Trans. Med. Imag.*, vol. 29, no. 12, pp. 2000–2008, Dec. 2010.

[19] M. Sdika, "Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote," *Med. Image Anal.*, vol. 14, no. 2, pp. 219–226, Apr. 2010.

[20] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Trans. Med. Imag.*, vol. 32, no. 9, pp. 1723–1730, Sep. 2013.

[21] M. R. Sabuncu, B. T. Yeo, K. V. Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1714–1729, Oct. 2010.

[22] D. Peressutti and B. Schipaanboord. (2018). *Overview of Atlas-Based Segmentation Approaches*. [Online]. Available: https://figshare.com/articles/Overview_of_atlas-based_segmentation_approaches/6809597

[23] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, Apr. 2004.

[24] B. Schipaanboord *et al.*, "Can atlas-based auto-segmentation ever be perfect? Insights from extreme value theory," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 99–106, Jan. 2019.

[25] M. Wu, C. Rosano, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein, "Optimum template selection for atlas-based segmentation," *NeuroImage*, vol. 34, no. 4, pp. 1612–1618, 2007.

[26] O. Commowick and G. Malandain, "Efficient selection of the most similar image in a database for critical structures segmentation," in *Proc. Conf. Med. Image Comput.-Assist. Intervent.*, Oct. 2007, pp. 203–210.

[27] S. Klein, U. A. van der Heide, I. M. Lips, M. van Vulpen, M. Staring, and J. P. Pluim, "Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information," *Med. Phys.*, vol. 35, no. 4, pp. 1407–1417, 2008.

[28] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.

[29] M. R. Sabuncu, S. K. Balci, M. E. Shenton, and P. Golland, "Image-driven population analysis through mixture modeling," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1473–1487, Sep. 2009.

[30] S. Gorthi, M. Bach Cuadra, U. Schick, P.-A. Tercier, A. S. Allal, and J.-P. Thiran, "Multi-atlas based segmentation of head and neck CT images using active contour framework," in *Proc. MICCAI Workshop 3D Segmentation Challenge Clin. Appl.*, Dec. 2010, pp. 20–24.

[31] J. M. Lötjönen *et al.*, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.

[32] L. Ramus, O. Commowick, and G. Malandain, "Construction of patient specific atlases from locally most similar anatomical pieces," in *Proc. Med. Image Comput.-Assist. Intervent.*, Sep. 2010, pp. 155–162.

[33] E. M. van Rikxoort *et al.*, "Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus," *Med. Image Anal.*, vol. 14, no. 1, pp. 39–49, 2010.

[34] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert, "LEAP: Learning embeddings for atlas propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, 2010.

[35] J. Yang, Y. Zhang, L. Zhang, and L. Dong, "Automatic segmentation of parotids from CT scans using multiple atlases," in *Proc. Med. Image Anal. Clinic, A Grand Challenge*, Dec. 2010, pp. 323–330.

[36] Y. Cao, Y. Yuan, X. Li, and P. Yan, "Putting images on a manifold for atlas-based image segmentation," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 289–292.

[37] J. A. Dowling *et al.*, *Fast Automatic Multi-atlas Segmentation of the Prostate from 3D MR Images*. Berlin, Germany: Springer, 2011, pp. 10–21.

[38] A. Akinyemi, C. Plakas, J. Piper, C. Roberts, and I. Poole, "Optimal atlas selection using image similarities in a trained regression model to predict performance," in *Proc. 9th IEEE Int. Symp. Biomed. Imag. (ISBI)*, May 2012, pp. 1264–1267.

[39] A. K. Hoang Duc *et al.*, "Using manifold learning for atlas selection in multi-atlas segmentation," *PLoS ONE*, vol. 8, no. 8, Aug. 2013, Art. no. e70059.

[40] T. R. Langerak, F. F. Berendsen, U. A. Van der Heide, A. N. Kotte, and J. P. Pluim, "Multiatlas-based segmentation with preregistration atlas selection," *Med. Phys.*, vol. 40, no. 9, 2013, Art. no. 091701.

[41] A. J. Asman, F. W. Bryan, S. A. Smith, D. S. Reich, and B. A. Landman, "Groupwise multi-atlas segmentation of the spinal cord's internal structure," *Med. Image Anal.*, vol. 18, no. 3, pp. 460–471, Apr. 2014.

[42] G. Sanroma, G. Wu, Y. Gao, and D. Shen, "Learning to rank atlases for multiple-atlas segmentation," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 1939–1953, Oct. 2014.

[43] J. Wang *et al.*, "Multi-atlas segmentation of subcortical brain structures via the autoseg software pipeline," *Frontiers Neuroinformat.*, vol. 8, p. 7, Feb. 2014.

[44] A. J. Asman, Y. Huo, A. J. Plassard, and B. A. Landman, "Multi-atlas learner fusion: An efficient segmentation approach for large-scale data," *Med. Image Anal.*, vol. 26, no. 1, pp. 82–91, 2015.

[45] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, F. Berendsen, and J. P. W. Pluim, "Improving label fusion in multi-atlas based segmentation by locally combining atlas selection and performance estimation," *Comput. Vis. Image Understand.*, vol. 130, pp. 71–79, Jan. 2015.

[46] Z. Xu *et al.*, "Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning," *Med. Image Anal.*, vol. 24, no. 1, pp. 18–27, 2015.

[47] P. Yan, Y. Cao, Y. Yuan, B. Turkbey, and P. L. Choyke, "Label image constrained multiatlas selection," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1158–1168, Jun. 2015.

[48] T. Zhao and D. Ruan, "Learning image based surrogate relevance criterion for atlas selection in segmentation," *Phys. Med. Biol.*, vol. 61, no. 11, pp. 4223–4234, 2016.

[49] K. Karasawa *et al.*, "Multi-atlas pancreas segmentation: Atlas selection based on vessel structure," *Med. Image. Anal.*, vol. 39, pp. 18–28, Jul. 2017.

[50] J. Yang *et al.*, "Atlas ranking and selection for automatic segmentation of the esophagus from CT scans," *Phys. Med. Biol.*, vol. 62, no. 23, pp. 9140–9158, 2017.

[51] M. Antonelli *et al.*, "GAS: A genetic atlas selection strategy in multi-atlas segmentation framework," *Med. Image Anal.*, vol. 52, pp. 97–108, Feb. 2019.

[52] L. Ramus and G. Malandain, "Assessing selection methods in the context of multi-atlas based segmentation," in *Proc. 7th IEEE Int. Symp. Biomed. Imag.*, Apr. 2010, pp. 1321–1324.

[53] V. Dill, P. C. Klein, A. R. Franco, and M. S. Pinho, "Atlas selection for hippocampus segmentation: Relevance evaluation of three meta-information parameters," *Comput. Biol. Med.*, vol. 95, pp. 90–98, Apr. 2018.

[54] A. Lin, G. Kubicek, J. Piper, A. S. Nelson, A. P. Dicker, and R. K. Valicenti, "Atlas-based segmentation in prostate IMRT: Timesavings in the clinical workflow," *Int. J. Radiat. Oncol., Biol., Phys.*, vol. 72, no. 1, pp. S328–S329, 2008.

[55] D. Peressutti *et al.*, "TU-AB-202-10: How effective are current atlas selection methods for atlas-based auto-contouring in radiotherapy planning?" *Med. Phys.*, vol. 43, no. 6, pp. 3738–3739, 2016.

[56] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.

[57] P. J. Besl and D. N. McKay, “A method for registration of 3-D shapes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[58] G. Delpon *et al.*, “Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy,” *Frontiers Oncol.*, vol. 6, p. 178, Aug. 2016.

[59] P. Zaffino *et al.*, “Multi atlas based segmentation: Should we prefer the best atlas group over the group of best atlases?” *Phys. Med. Biol.*, vol. 63, no. 12, p. 12NT01, Jun. 2018.

[60] A. V. Young, A. Wortham, I. Wernick, A. Evans, and R. D. Ennis, “Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 79, no. 3, pp. 943–947, 2011.

[61] M. A. Gambacorta *et al.*, “Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system,” *Acta Oncologica*, vol. 52, no. 8, pp. 1676–1681, 2013.

[62] J. Hwee *et al.*, “Technology assessment of automated atlas based segmentation in prostate bed contouring,” *Radiat. Oncol.*, vol. 6, no. 1, p. 110, 2011.

[63] C. Granberg, “Clinical evaluation of atlas based segmentation for radiotherapy of prostate tumours,” M.S. thesis, Dept. Phys., Umeå Univ., Umeå, Sweden, 2011.

[64] K. Langmack, C. Perry, C. Sinstead, J. Mills, and D. Saunders, “The utility of atlas-assisted segmentation in the male pelvis is dependent on the interobserver agreement of the structures segmented,” *Brit. J. Radiol.*, vol. 87, nos. 10–43, Nov. 2014, Art. no. 20140299.