Can Atlas-Based Auto-Segmentation Ever Be Perfect? Insights From Extreme Value Theory

Can Atlas-Based Auto-Segmentation Ever Be Perfect? Insights From Extreme Value Theory

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

***Abstract*— Atlas-based segmentation is used in radiotherapy planning to accelerate the delineation of organs at risk (OARs). Atlas selection has been proposed to improve the performance of segmentation, assuming that the more similar the atlas is to the patient, the better the result. It follows that the larger the database of atlases from which to select, the better the results should be. This paper seeks to estimate a clinically achievable expected performance under this assumption. Assuming a perfect atlas selection, an extreme value theory has been applied to estimate the accuracy of single-atlas and multi-atlas segmentation given a large database of atlases. For this purpose, clinical contours of most common OARs on computed tomography of the head and neck ($N = 316$) and thoracic ($N = 280$) cases were used. This paper found that while for most organs, perfect segmentation cannot be reasonably expected, auto-contouring performance of a level corresponding to clinical quality could be consistently expected given a database of 5000 atlases under the assumption of perfect atlas selection.**

***Index Terms*— Radiotherapy, extreme value theory, atlas-based segmentation, auto-contouring.**

I. Introduction

ACCURATE contours of organs at risk (OARs) are a crucial link in radiotherapy treatment planning. In radiotherapy, ionising radiation is used to treat tumour cells. The aim of radiotherapy is to maximise dose to the tumour, while sparing healthy OARs. Radiotherapy treatment planning involves carefully made decisions to balance conflicting treatment objectives. However, due to the physical nature of irradiation techniques it is unavoidable that dose is also delivered to healthy OARs surrounding the tumour. It is therefore of utmost importance that OAR segmentations are an accurate representation of the in-vivo morphology, such that the actual delivered dose to the patient reflects the intended trade-offs as carefully made during planning. Segmentation of OARs and tumours is typically performed by clinical experts on a 3D anatomical image, normally Computed Tomography (CT). However, manual segmentation of the structures of interest is prone to inter- and intra-observer variability and is time consuming. For instance, mean manual segmentation times for Head & Neck cases ranging from 28.5 minutes [1] up to 3 hours [2] are reported in literature. The large variation in reported segmentation times relates to the number of structures to be segmented, e.g. the number OARs or lymph-node regions. Indeed, it is natural to assume that segmentation times should increase with the increasing number and complexity of structures that have to be segmented. As a consequence, automatic contouring techniques have gained popularity in the last decade to segment OARs, with atlas-based segmentation being favoured in commercial systems [3].

While atlas-based segmentation has been shown to improve consistency and reduce contouring time [3], little is known on the upper bounds in clinically achievable performance by such methods.

In this paper we seek to address this question of what auto-contouring performance could be achieved if a very large, but clinically realistic, size of atlas database was created. To do so, the statistical technique of Extreme Value Theory (EVT) is applied to estimate accuracy performance of single-atlas and multi-atlas segmentation assuming a database of 5000 atlases.[1] It is important to notice the difference between the atlas database size used for atlas selection (i.e. the number of atlases to select from) and the number of atlases used for multi-atlas segmentation (i.e. the number of atlases that will be combined to form one fused contour). The aim of this paper is to investigate the convergence in segmentation performance for increasing the size of the database of atlases to select from, and not the number of atlases used for multi-atlas segmentation.

A. Atlas-Based Auto-Segmentation

Atlas-based segmentation uses prior knowledge provided by previously contoured images (i.e. atlases) to automatically segment OARs of a patient image [3], [5]. This is achieved by calculating a spatial transformation, using deformable image

Manuscript received May 19, 2018; accepted July 9, 2018. Date of publication July 16, 2018; date of current version December 28, 2018. This work was supported in part by Innovate U.K. and in part by the Eurostars Project under Grant ID9297. *(Corresponding author: Djamal Boukerroui.)*

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

D. Peressutti, D. Boukerroui, T. Kadir, and M. 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.

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

Digital Object Identifier 10.1109/TMI.2018.2856464

[1] This work builds on preliminary results presented at the ESTRO [4].

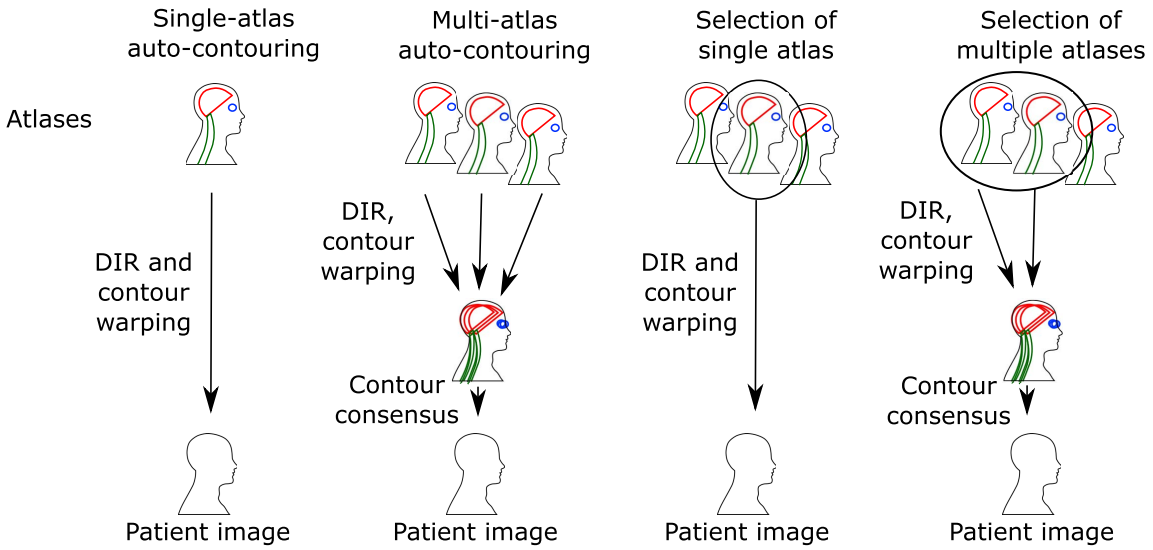


Fig. 1. Overview of approaches to atlas-based segmentation (from [7]). Left: single-atlas only, Center-left: multi-atlas fusion, Center-right: single-atlas selection, Right: multi-atlas selection and fusion.

TABLE I

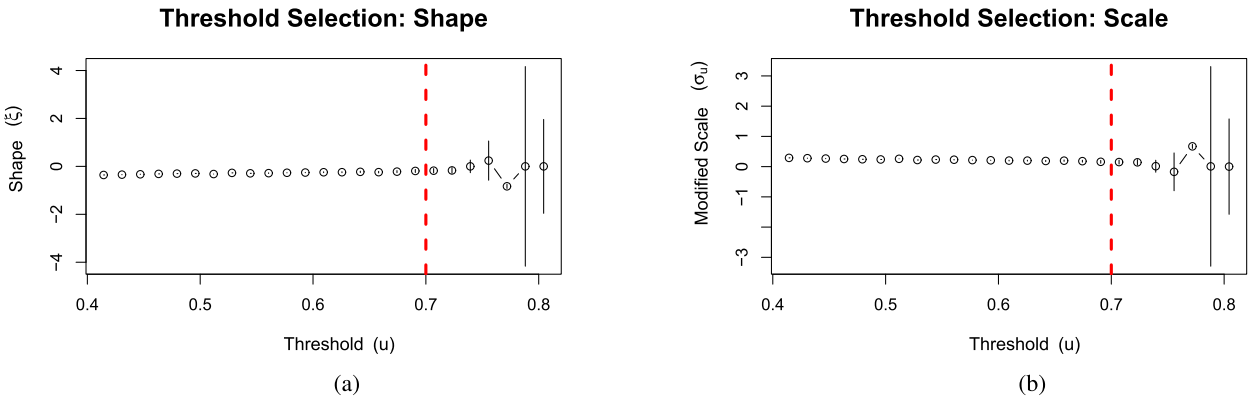
OVERVIEW OF ATLAS SELECTION METHODS PROPOSED IN THE LITERATURE. TYPE INDICATES ONLINE OR OFFLINE SELECTION. ABBREVIATIONS USED ARE: NORMALISED MUTUAL INFORMATION (NMI), MEAN ABSOLUTE DIFFERENCE (MAD), MANIFOLD DISTANCE (MD), CROSS-CORRELATION (CC), DEFORMATION FIELD (DEF), DICE SIMILARITY COEFFICIENT (DSC), HISTOGRAM OF ORIENTED GRADIENTS (HOG), DEFORMABLE IMAGE REGISTRATION (DIR), CONFOCAL MICROSCOPY (CM), MAGNETIC RESONANCE (MR) AND COMPUTED TOMOGRAPHY (CT)

Paper	Type	Method	Organs
Rohlfing, T. *et al.*, 2004 [8]	on	Global NMI, DEF after affine, DIR	Bee brain
Wu, M. *et al.*, 2007 [9]	on	Local NMI after DIR	Brain
Commowick, O. *et al.*, 2007 [10]	on	Local DEF wrt template after DIR	Head and neck
Klein, S. *et al.*, 2008 [11]	on	Global and local NMI after DIR	Prostate
Aljabar, P. *et al.*, 2009 [12]	on	Local NMI wrt template after affine	Brain
Lotjonen, J.M. *et al.*, 2010 [13]	on	Local NMI wrt templates after affine, DIR	Brain
Ramus, L. *et al.*, 2010 [14]	on	Local DEF wrt template after DIR	Head and neck
van Rikxoort, E.M. *et al.*, 2010 [15]	on	Global and local MAD after affine	Heart+caudate nuc.
Wolz, R. *et al.*, 2010 [16]	on	MD of local NMI wrt template after DIR	Hippocampus
Cao, Y. *et al.*, 2011 [17]	on	MD of local intensities	Prostate
Dowling, J.A. *et al.*, 2011 [18]	on	NMI after DIR	Prostate
Akinyemi, A. *et al.*, 2012 [19]	off/on	Offline selection and local CC	Heart and kidney
Hoang Duc, A.K. *et al.*, 2013 [20]	on	MD of global DEF wrt template	Hippocampus
Langerak, T.R. *et al.*, 2013 [21]	off/on	Clustering an selection based on DSC	Prostate
Sanroma, G. *et al.*, 2014 [22]	off/on	Appearance (HOG) wrt template after affine	Brain
Asman, A.J. *et al.*, 2015 [23]	on	MD of intensities wrt template after affine	Brain
Zhao, T. *et al.*, 2016 [24]	off/on	Metric distance learning on intensities	Corpus callosum

registration (DIR), which maps the atlas image coordinate system to the patient image coordinate system. The atlas contours are subsequently warped to the patient coordinate system using the estimated spatial transformation. When only one atlas is used to estimate the contours of the patient image, the process is referred to as single-atlas segmentation (Fig. 1, Left). In multi-atlas segmentation (Fig. 1, Center-left), the registration and contour warping is repeated for several atlases. Then, the warped contours from each atlas are fused into a single contour. By averaging out random registration errors, multi-atlas segmentation has been shown to outperform single-atlas segmentation [3], [5], [6]. Furthermore, the quality of the obtained contours depends not only on the DIR and fusion algorithms used but also on the quality of the atlases themselves.

Given a large database of atlases, the selection of a single atlas or multiple atlases more suitable to segment the patient image has been proposed as a way to improve segmentation accuracy [5], [8] (Fig. 1, Center-right & Right). Selection may also improve computational speed by using only a subset of one or more atlases instead of the entire database. For this reason, methods of atlas selection have been an active research field in the past decade, as demonstrated by the methods summarised in Table I.

Two main types of atlas selection methods exist in the literature [5]: i) offline methods do not use the patient image in the selection process. ii) online methods make use of the current patient image to search for the best atlas(es) for that patient. Selection methods can differ in the type of the employed image similarity measure (e.g. intensity- or deformation-based),



(a) (b)

Fig. 2. Example of threshold selection for the shape (a) and scale (b) parameter for the single-atlas segmentation of the oesophagus. A maximum likelihood estimation of the PDF parameters is performed for a range of threshold values. For each threshold value, the estimated parameter and its Confidence Interval, denoted by the circles and black vertical lines respectively, are plotted. The largest threshold value leading to small variation of the estimate and the associated error of both the shape ξ and modified scale σ_u (modified by subtracting the shape multiplied by the threshold) parameters is selected. The red dashed line represents the selected threshold at 0.7 DSC. The threshold selection was performed using the "tcplot" function from the POT package in R [26].

on the region of interest considered (e.g. global or local) or the utilized atlas strategy (e.g. exhaustive or template-based search). Furthermore, variations in the registration model (e.g. rigid, affine or deformable) or similarity representation (e.g. manifold representation) have been used. A thorough review of atlas selection methods is given in [5]. A selection of the representative methods proposed in the literature is presented in Table I.

The assumption of atlas selection is that finding a more similar atlas to the patient will result in improved auto-segmentation performance. Can it be assumed, therefore, that the larger the pool of atlases from which to choose from, the better the match that can be achieved and correspondingly the better the auto-segmentation results that can be obtained? A priori it is expected that the set of atlases that will perform very well for a given patient and a given OAR is small even when the selection is performed on a very large pool of candidate atlases. In other words, observing a high segmentation performance, specifically higher than a certain given threshold, could be seen as a rare event. Consequently, Extreme Value Theory can be applied to help answer the above question.

B. Extreme Value Theory

Extreme Value Theory (EVT) is a statistical tool used for modeling the probability distribution of rare events. The theory has been successfully applied in finance, engineering and physics, with examples including the estimation of extreme floods in hydrology, environmental loads in mechanical constructions or large insurance losses [25]. A first approach of EVT is the "block maxima" method which leads to the Generalized Extreme Value (GEV) distribution. A popular alternative is Peaks-Over-Threshold (POT) method, aiming at modeling the distribution of observed data exceedances over a sufficiently high threshold. This is achieved by using the Generalized Pareto Distribution (GPD), as a model for the upper tail of the distribution. Both methods do not enforce any assumption on the underlying Probability Density Function (PDF), except that the observations are independent and identically distributed (i.i.d.). Therefore, the theory is particularly suited when no prior knowledge on the PDF of the observed data is available, as is the case for the distribution of segmentation performance measures. POT has the advantage over the GEV of using the available data more efficiently. Note however, the choice of the threshold is crucial, as it defines which part of the data is considered to be extreme.

1) Fundamental POT: Let $X_1, \cdots, X_n$ be a set of i.i.d. random variables with a distribution function F and a right endpoint $x_F = \sup\{x; F(x) < 1\}$. The function $F_u(x) = \mathrm{P}\{X - u \leq x | X > u\}$, $x \geq 0$ is called the distribution function of exceedances above the threshold $u < x_F$. Let $Y_j = X_i - u$ for $X_i > u$, then under certain conditions, and for a large enough threshold u, the distribution of the exceedances $(Y_1, \cdots, Y_{n_u})$, can be approximated by a GPD [25]:

$$G_{\xi,\sigma_u}(y) = 1 - \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-\frac{1}{\xi}}, \quad y \geq 0 \tag{1}$$

where ξ and $\sigma_u > 0$ are respectively the shape and scale parameters and $z_+ = \max(z, 0)$. Given a fitted GPD model (u, ξ, σ_u) to sample data, one can estimate a value, x_M, that is expected to be exceeded once every M observations. M is known as return period. x_M is the return level and is estimated as:

$$x_M = u + \frac{\sigma_u}{\xi}\left[(M \cdot \mathrm{P}\{X > u\})^{\xi} - 1\right], \tag{2}$$

where $\mathrm{P}\{X > u\}$ is approximated by the proportion of observations exceeding u in the observed sample.

As mentioned earlier, a challenge in using the POT method is the determination of a good threshold value u. As u increases, the number of samples exceeding the threshold decreases, impairing the fitting of a robust model. Conversely, as u decreases the validity of assuming a GPD is questionable. To overcome this issue, a maximum likelihood approach was employed to estimate an optimal set of parameters (ξ, σ_u) for a given threshold u. Then, standard errors for the estimated parameters (ξ, σ_u) were computed for varying values of u. The threshold u was then chosen to be the highest value that provided a robust estimate of the model parameters with low standard error (Expected Fisher Information). For example, in Figure 2 the estimated model parameters remain stable

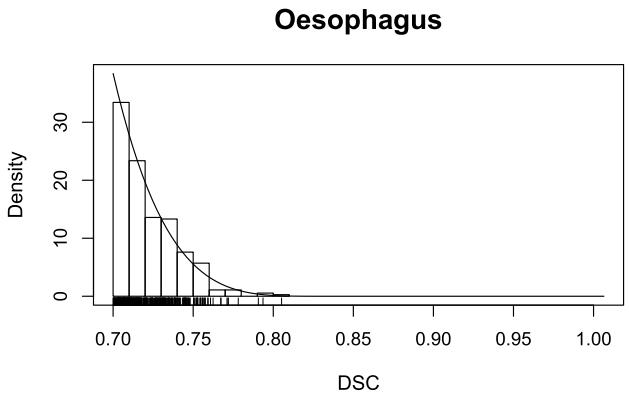


Fig. 3. Example of GPD model fitting for the Dice similarity coefficient exceedances for single-atlas segmentations of the oesophagus.

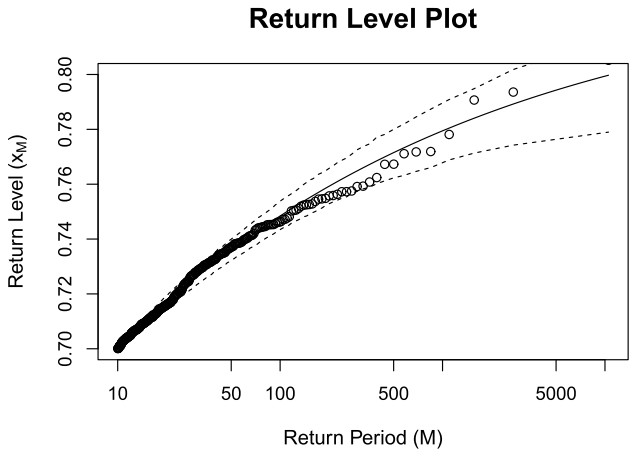


Fig. 4. Example of return level x_M as a function of the return period M. Specifically, the plot shows the expected Dice score performance for the single-atlas segmentation of the oesophagus function of the size of atlas database conditional on having a perfect atlas selection algorithm. The dashed curves show the 95% Confidence Interval (CI). Observe for example that segmentation performance improves by 0.1 (DSC) when the size of the database increases from 10 to 5,000 atlases. Note however that the error on the estimation of the return level increases with the return period. In other words, the larger the return period the bigger the error on the estimation will be. The chosen return period of 5000 atlases allows for a reasonably accurate estimation (95% CI $<$ 0.03 DSC) of the segmentation performance. The asymptotic trend of the curve gives insight on how the segmentation performance may evolve with the increase of the database size.

up to a threshold value of 0.7 DSC. Beyond this point the standard error of the model parameters increases, impairing the fit of a robust model. Therefore, a threshold of 0.7 DSC was selected just before the fitted model becomes unstable. In this way, a robust model is fitted while maintaining validity to use a GPD to model the extremes. The threshold selection was performed using the "tcplot" function from the Points Over Threshold (POT) package in R [26].

2) Practical POT: In the context of atlas-based segmentation, it is reasonable to assume that for a given OAR and a given segmentation performance measure, the observed values are sampled from the same distribution as long as the same atlas based auto-contouring algorithm is used and the contouring guidelines are also the same. Furthermore, the observations are independent as they are obtained from different atlas-patient pairs. Therefore, each segmentation performance measure for a particular OAR can be treated as an observation, x. Fitting a GPD model to the observed data, for every pair of OAR and performance measure, will allow us to investigate the expected extreme performance (i.e. the return level, x_M) for a particular database size (i.e. the return period, M), since this estimates an expected performance once in M observations. The threshold for each OAR and similarity measure is found using the approach described above. Figure 2 illustrates the DSC threshold selection for the single-atlas segmentation of the oesophagus, where the estimates of the shape and scale parameter become unstable beyond a DSC value of 0.7. Figure 3 shows the corresponding GPD model fitting for the values over this threshold for single-atlas segmentation. Following, fitting of the GPD model, the return level is then estimated for particular database size (return period), as illustrated in Fig. 4.

In the subsequent section, we evaluate the potential performance of both single-atlas (Section II-B) and multi-atlas (Section II-C) segmentation using the POT for a clinically achievable but nonetheless large database.

II. Materials and Methods

A. Data

Two databases of clinically contoured cases were created for use in these investigations. A head and neck (HN) database comprised 316 CT patient cases, while the thoracic (LN) database consisted of 280 CT patient cases. Both databases were acquired at the Department of Radiation Oncology MAASTRO (Maastricht, Netherlands), and consisted of the CT with clinical contours created according to institutional guidelines and reviewed by the treating radiation oncologists. The clinical contours are used as a reference for segmentation performance and referred to as reference contours for the remainder of this study. The study was approved by the local Institutional Review Board. The ranges of the isotropic in-plane and through-plane CT image resolution were of 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 also reported in the table, since only the clinically necessary organs had been contoured in each case. Investigations were performed using only the OARs listed in the table, all other less frequently contoured OARs and target volumes were not considered.

TABLE II
List of OAR Considered in This Study for the *HN* and *LN* Databases. The Second and Fourth Columns (#) Report the Number of Occurrences for Each OAR

HN		*LN*	
OAR	#	OAR	#
Brain	265	Oesophagus	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		

B. Performance Expectation for Single-Atlas Segmentation

As input to EVT, samples from distributions of segmentation performance measures for each OAR are required. To generate this, exhaustive single-atlas segmentation was carried out between each possible pair of case combinations within each dataset. This evaluation corresponds to a leave-one-out cross-validation, where each case image is in turn considered to be the patient image and the remaining cases are considered as candidate atlases for segmentation. Deformable image registration between atlases and test patients were performed using methods available in commercial software (Workflow Box 1.4, Mirada Medical Ltd., Oxford, UK).

The reference contours of the case treated as patient are then employed to assess the accuracy of the segmentation. In this study, Dice Similarity Coefficient (DSC) [27], Hausdorff contour distance (HD) [28], average symmetric surface distance (AD) [29] and root-mean-square symmetric surface distance (RMSD) [29] between the estimated atlas-based contour and the clinically delineated contour were used. This process resulted in approximately 100,000 and 78,000 evaluations for the HN and LN databases, respectively for each performance measure. POT was applied to model extreme values of the distribution of each pair of performance measure (i.e. DSC, HD, AD, RMSD) and OAR.

A database of 5,000 atlases is used as representative of a very large, yet clinically achievable, size for which to consider expected performance. Around 600 lung cancer and 900 breast cancer patients are treated with radiotherapy at Maastro Clinic each year. If all previously treated patients from the preceding 5 years were used as thoracic atlases, a databases of 7,500 could be build. However it is expected that not all clinical cases would be suitable to be used as an atlas. Thus 5,000 atlases was chosen as an ambitious, but potentially achievable, size. Additionally, the error on the estimation of the return level was taken into account. The dashed curves in Figure 4 denote the 95% Confidence Interval (CI) of the estimation. Note that the larger the chosen return period the bigger the error on the estimation will be. The chosen return period of 5000 atlases allows for a reasonably accurate estimation (95% CI < 0.03 DSC) of the segmentation performance. For estimation beyond 5000 atlases the 95% CI becomes larger and will yield less meaningful results given the dataset we had available.

C. Performance Expectation for Multi-Atlas Segmentation

Ten atlases were used for multi-atlas segmentation, as evidence shows that this represents a good trade-off between computational efficiency and segmentation accuracy [12], [30], [22]. However, the computational cost of evaluating every possible combination of ten atlases for each case within the databases becomes prohibitive (e.g. assuming 250 atlases, there are $\binom{249}{10} \times 250 = 5.3 \times 10^{19}$ combinations). Therefore, the segmentation results were computed assuming perfect selection of atlases had been performed in a greedy approach from the single-atlases prior to multi-atlas segmentation for the analysis of multi-atlas segmentation performance.

The 11 atlases providing the best single-atlas segmentation performance according to DSC against the reference contours for each OAR and each patient image in the database were considered. Any other segmentation performance measure or a combination of measures could have been utilized. Given the set of best 11 performing atlases for each patient image and OAR, all combinations of 10 atlases were employed to perform contour fusion. Fused contours were generated using a commercially available implementation of majority voting (Workflow Box 1.4, Mirada Medical Ltd., Oxford, UK). For each organ the voxel-wise contour labels are of a binary form: a voxel either belongs to the organ or is located outside of the organ. Considering each of the input contours, the majority voting fusion method predicts binary contour labels (organ/non-organ) on a voxel-based level. Selecting the 11 best atlases and performing multi-atlas segmentation with 11 different combinations of 10 atlases (11 choose 10) provides 11 times more observations than solely performing multi-atlas segmentation with the 10 best atlases. This allows for sufficient observations to provide a performance estimate with a reasonable accuracy. Thus, 11 multi-atlas segmentation contours were generated for each OAR of each case in the database, resulting in approximately 3,400 and 3,000 multi-atlas segmentation results for each organ for the HN and LN databases respectively. Although the results may not be strictly independent because a given atlas may be used in multiple segmentations, we assume that this effect is relatively small and hence POT was applied to the multi-atlas segmentation performance measures for each OAR, in the same manner as for single-atlas segmentation. The impact of this assumption is further discussed in Section IV-A.

III. Results

A. Performance Expectation for Single-Atlas Segmentation

In the head and neck region, the expected single-atlas segmentation performance given 5000 atlases was in excess of a DSC of 0.85 for all organs. For the thoracic region, all organs had a DSC in excess of 0.80. The expected performance for each organ and metric are reported in Table III, including the threshold values (u) used to fit the GPD model.

For some OAR, the best observed performance scores exceeded the estimated return level. It should be noted that the number of observations before applying the threshold is much larger than the return period used of 5000. For example, the HN database consisted of 261 cases including the right parotid contoured. Treating each case in turn as patient image and the remaining cases as candidate atlases for segmentation results in $261 \times 260 = 67860$ observations. For the estimate of the return level, the probability that an observation exceeds the threshold ($\mathrm{P}\{X > u\}$) is approximated based on the complete set of observations, see equation 2. Therefore, it is likely that segmentation performance scores are observed exceeding the calculated expected value for a return period of 5000 atlases. Additionally, the estimated return level for the Points Over

TABLE III

RESULTS OF EVT FOR SINGLE-ATLAS SEGMENTATION. FOR EACH OAR AND FOR EACH SEGMENTATION MEASURE, THE VALUES OF THE THRESHOLD u, THE BEST SCORE ACHIEVED $max\{x_i\}$ AND THE ESTIMATED RETURN LEVEL x_M ARE REPORTED. M = 5,000

	DSC			HD (mm)			AD (mm)			RMSD (mm)		
OAR	u	Best	x_M	u	Best	x_M	u	Best	x_M	u	Best	x_M
Brain	0.980	0.988	0.997	10.00	5.36	5.63	0.73	0.49	0.51	1.20	0.79	0.81
Brainstem	0.850	0.909	0.903	7.00	3.70	3.91	1.50	0.88	0.93	2.00	1.20	1.25
Cochlea L	0.750	1.000	0.998	1.75	0.40	0.48	0.60	0.13	0.15	0.75	0.16	0.19
Cochlea R	0.750	0.957	0.947	1.75	0.51	0.54	0.60	0.17	0.18	0.75	0.21	0.22
Oral Cavity	0.800	0.911	0.904	15.00	6.52	7.11	3.00	1.46	1.56	4.00	1.99	2.10
Parotid L	0.800	0.879	0.874	11.00	4.93	5.49	2.00	1.22	1.27	2.75	1.55	1.64
Parotid R	0.800	0.877	0.870	11.00	5.37	5.83	2.00	1.30	1.30	2.75	1.64	1.70
Spinal Cord	0.830	0.890	0.888	5.00	2.48	2.63	0.90	0.53	0.60	1.30	0.69	0.75
Submand. Gland L	0.750	0.873	0.868	7.00	3.62	3.66	1.75	0.87	0.91	2.20	1.11	1.16
Submand. Gland R	0.750	0.859	0.856	7.00	3.32	3.56	1.75	0.99	1.00	2.20	1.22	1.27
Oesophagus	0.700	0.805	0.809	15.00	6.54	7.08	2.20	1.19	1.23	4.00	1.77	1.86
Heart	0.900	0.944	0.941	20.00	10.27	10.80	3.00	1.72	1.74	4.00	2.36	2.37
Lung L	0.960	0.989	0.990	20.00	9.88	10.79	0.70	0.45	0.47	1.50	0.84	0.86
Lung R	0.960	0.989	0.989	20.00	9.61	11.17	0.70	0.47	0.47	1.50	0.91	0.90
Mediastinum Env.	0.900	0.949	0.946	20.00	13.01	13.08	2.70	1.64	1.65	4.00	2.52	2.52
Spinal Cord	0.880	0.932	0.928	10.00	3.40	3.56	1.00	0.50	0.52	1.40	0.67	0.71

TABLE IV

RESULTS OF EVT FOR MULTI-ATLAS SEGMENTATION. FOR EACH OAR AND FOR EACH SEGMENTATION MEASURE, THE VALUES OF THE THRESHOLD u, THE BEST SCORE ACHIEVED $max\{x_i\}$ AND THE ESTIMATED RETURN LEVEL x_M ARE REPORTED. M = 5,000

	DSC			HD (mm)			AD (mm)			RMSD (mm)		
OAR	u	Best	x_M	u	Best	x_M	u	Best	x_M	u	Best	x_M
Brain	0.987	0.991	0.999	6.00	3.97	3.69	0.42	0.37	0.37	0.80	0.59	0.58
Brainstem	0.900	0.935	0.979	4.00	2.96	2.90	0.90	0.62	0.53	1.20	0.81	0.77
Cochlea L	0.900	1.000	1.000	0.80	0.42	0.42	0.20	0.12	0.12	0.25	0.16	0.16
Cochlea R	0.900	1.000	1.000	0.80	0.33	0.33	0.20	0.11	0.10	0.25	0.15	0.11
Oral Cavity	0.900	0.928	0.977	9.00	5.47	5.02	1.50	1.10	1.04	2.20	1.47	1.36
Parotid L	0.870	0.913	0.966	7.00	4.26	3.70	1.30	0.81	0.55	1.80	1.10	0.86
Parotid R	0.870	0.903	0.905	7.00	4.67	4.63	1.30	0.96	0.94	1.80	1.26	1.23
Spinal Cord	0.890	0.921	0.921	3.00	2.08	2.07	0.55	0.38	0.33	0.70	0.48	0.38
Submand. Gland L	0.870	0.912	0.916	4.00	2.89	2.89	0.90	0.62	0.61	1.20	0.84	0.84
Submand. Gland R	0.870	0.914	0.913	4.00	3.09	3.07	0.90	0.64	0.62	1.20	0.87	0.87
Oesophagus	0.760	0.856	0.869	10.00	4.97	4.94	1.40	0.86	0.83	2.00	1.16	1.11
Heart	0.935	0.960	0.962	12.00	7.88	7.80	1.80	1.29	1.27	2.80	1.86	1.86
Lung L	0.986	0.992	1.000	20.00	11.32	10.31	0.55	0.35	0.35	1.30	0.67	0.66
Lung R	0.985	0.991	1.000	20.00	10.41	8.38	0.50	0.35	0.34	1.30	0.58	0.50
Mediastinum Env.	0.940	0.963	0.964	20.00	11.76	11.65	1.80	1.27	1.26	3.00	1.84	1.84
Spinal Cord	0.910	0.947	0.946	8.00	3.01	3.02	0.55	0.42	0.42	0.90	0.58	0.58

Threshold approach states the value that is expected to be exceeded at least once every return period, but it does not state by how much it might be exceeded.

B. Performance Expectation for Multi-Atlas Segmentation

The expected performance of multi-atlas segmentation given a database of 5,000 atlases was consistently higher than for single-atlas segmentation for all organs.

To provide an overview of the results, without directly comparing between different OARs, the expected segmentation performance in the head and neck region was in excess of a DSC of 0.90 for all organs. While for the thoracic region, all organs had a DSC in excess of 0.85.

Full results are provided in the Table IV. Similarly to single-atlas segmentation, the threshold values, the maximum observed values and the estimated performance return levels are reported for each segmentation performance measure and each OAR.

IV. DISCUSSION

A. Validity of the Use of Extreme Value Theory

The assumption of i.i.d. samples is made when estimating return values using POT. This must be examined.

The assumption of identical distribution may be questioned in that the distribution of segmentation results of all atlases for one patient might be considered a different distribution to the results with another patient. This objection would be valid if the EVT block maxima method were applied patient by patient. However, using the POT method, all the patient-atlas results are considered as a single distribution representative of the patient population, and the points over the threshold are drawn from this distribution. Furthermore, the same process, i.e. registration and fusion, was employed to process all cases in the databases.

Regarding the assumption of sample independence, it might be argued that applying case A as an atlas to case B will generate a similar result to applying case B to atlas A and therefore not all results are independent. However, if the sampling of atlas and patient cases is regarded as a random sampling experiment, sampling with replacement means that the first experiment does not influence the outcome of the second and as such the two experiments are independent samples which have similar results. To support this statement; the EVT analysis has been repeated on the upper and lower triangular part of the leave-one-out result matrix separately. In this way, only the segmentation results in one direction

were included for analysis, but this also cuts the number of observations used to fit the model by half. It has been observed that, in general, the estimated segmentation performance for a database of 5000 atlases is very similar. The same can be argued of the multi-atlas segmentation experiment, while multiple results shared the same input cases, these cases are sampled with replacement thus the samples are, although similar, statistically independent.

B. Registration and Fusion Methods

The impact of choice of registration method and contour fusion were not evaluated in this research, but their impact is worth consideration.

Taking the concept of atlas selection to an (unrealistic) extreme, it could be expected that the selected atlas would be identical to the patient in the presence of perfect atlas selection. In that scenario, there is no need for DIR. However, in practice such an infinite database is not available, thus DIR is required to correct for differences between the atlas and the patient. While currently state-of-the-art registration cannot fully correct for differences in inter-subject registration tasks, it is known that the smaller the non-rigid deformation required, the lower the error [31]. As the size of the database increases and the difference between patient and atlas is reduced, the impact of DIR is reduced. Thus, in the extreme value scenario, the choice of DIR becomes less important. It is reasonable to assume that replacing the DIR method with another choice may have a small impact on the absolute performance of auto-contouring, but that the general trend regarding extreme performance will be similar.

The same argument applies to contour fusion methods. While the absolute performance may vary slightly, the relative impact on performance of various methods should be unaffected, and the general finding in the extreme scenario should remain the same. Nevertheless, the impact of various contour fusion methods on the segmentation performance would be an interesting experiment to perform in the future. In particular, locally weighted patch-based approaches may yield valuable contributions towards segmentation performance [32], [33].

C. Clinical Interpretation of Expected Contouring Quality

The extreme value theory experiments suggest that, while improvement can be expected with an increasing database size, "perfect" segmentation may not be achieved routinely i.e. DSC scores of less than 1 and distances of greater than 0 found as the expected performance.

A limitation of atlas-based segmentation techniques in general that needs to be discussed is the use of the reference contours. As there is no objective measure to test whether the contours are an accurate representation of the in-vivo morphology, it is necessary to note that the used reference contours may vary from the ground truth. With regards to intra- & inter-observer variability it has been shown that variability in contouring decreases in the presence of well-defined guidelines [34]. The clinical contours used for this study have been made according to the institutional guidelines, and all contours were of sufficient quality for clinical use.

It should also be noted that such quantitative measures are blunt tools when considering clinical impact of segmentation errors. Small segmentation errors in some locations may have high clinical significance, while gross errors in other locations may not affect a treatment plan at all.

The expected performance appears similar to reported inter-observer variability of manual contouring [3], meaning that, assuming perfect atlas selection, atlas-based segmentation can provide accuracy performance at level similar to clinical contouring.

However, the failure to achieve perfect segmentation cannot be attributed to inter-observer variability, since this experiment considers the best segmentation performance measures. Thus, lower performance measures resulting from differences in contouring are unlikely to contribute to exceedance values. It is noted that lower performance is found for more anatomically variable and poorly defined (on CT imaging) structures, such as the parotids or the oesophagus. This suggests that the "less than perfect" performance may be explained from the variability of shapes observed between patients rather than contouring variation. If this is the case, the results imply that even a very large database would not be sufficient to capture the anatomical variation within the population to deliver "perfect" auto-contouring. This is suggested by the almost linear to log-log scale dependence of the return level function of the database size as can be seen on Fig. 4.

It is noted, that in the absence of perfect single-atlas auto-contouring, multi-atlas fusion remains a useful algorithmic improvement.

V. Conclusion

A method to estimate a clinically achievable expected performance in atlas-based auto-contouring has been proposed and applied to OARs in the head and neck and thoracic body regions. This approach suggested that, in the presence of perfect atlas selection, atlas-based auto-contouring could reach clinical performance levels given a large database of atlases. However, the range of variability between subjects means that perfect segmentation is not achieved using an ambitious but clinically achievable database size. As a consequence multi-atlas fusion remains beneficial.

Acknowledgments

The authors would like to thank Mareli Grady, Department of Statistic, University of Oxford, for directing us to the Extreme Value Theory as an avenue for this investigation.

References

[1] 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.

[2] 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, 2011.

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

[4] B. W. K. Schipaanboord *et al.*, "OC-0068: Can atlas-based auto-contouring ever be perfect?" *Radiotherapy Oncol.*, vol. 119, pp. S30–S31, Apr. 2016.

[5] 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.

[6] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, "*Quo Vadis*, atlas-based segmentation?" in *Handbook of Biomedical Image Analysis*, 2005, pp. 435–486.

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

[8] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, Jr., "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.

[9] 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.

[10] O. Commowick and G. Malandain, "Efficient selection of the most similar image in a database for critical structures segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 10, no. 2, 2007, pp. 203–210.

[11] 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.

[12] 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.

[13] J. M. P. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.

[14] L. Ramus, O. Commowick, and G. Malandain, "Construction of patient specific atlases from locally most similar anatomical pieces," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 13, no. 3, 2010, pp. 62–155.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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, p. 091701, 2013.

[22] 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.

[23] 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.

[24] 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.

[25] S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer Series in Statistics), 1st ed. London, U.K.: Springer, 2001.

[26] M. Ribatet and C. Dutang. (2016). POT: Generalized Pareto distribution and peaks over threshold. R Package Version 1.1-6. [Online]. Available: https://CRAN.R-project.org/package=POT

[27] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[28] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[29] 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, 2009.

[30] A. Larrue, D. M. Gujral, C. M. Nutting, T. Kadir, and M. Gooding, "PD-0132: Impact of the number of available atlases on the performance of adaptive multi-atlas contouring," *Radiotherapy Oncol.*, vol. 111, Suppl. 1, pp. S51–S52, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167814015302371, doi: 10.1016/S0167-8140(15)30237-1.

[31] N. Kadoya *et al.*, "Evaluation of various deformable image registration algorithms for thoracic images," *J. Radiat. Res.*, vol. 55, no. 1, pp. 175–182, 2014.

[32] F. Rousseau, P. A. Habas, and C. Studholme, "A supervised patch-based approach for human brain labeling," *IEEE Trans. Med. Imag.*, vol. 30, no. 10, pp. 1852–1862, Oct. 2011.

[33] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen, "A generative probability model of joint label fusion for multi-atlas based brain segmentation," *Med. Image Anal.*, vol. 18, no. 6, pp. 881–890, 2014.

[34] S. K. Vinod, M. Min, M. G. Jameson, and L. C. Holloway, "A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology," *J. Med. Imag. Radiat. Oncol.*, vol. 60, no. 3, pp. 393–406, 2016.