Similarity clustering based atlas selection for pelvic CT image segmentation

Angel Kennedy[1], Jason Dowling[2,4,7,8], Peter B Greer[3,4], Lois Holloway[5,6,7,8], Michael G Jameson[5,6,7], Dale Roach[6,7], Soumya Ghose[9], David Rivest-Hénault[2], Marco Marcello[10], Martin A Ebert[1,8,10,11]

1. Radiation Oncology, Sir Charles Gairdner Hospital, Nedlands, WA 6009, Australia
2. Australian e-Health Research Centre, CSIRO, Royal Brisbane and Women's Hospital, QLD 4029, Australia
3. Calvary Mater Newcastle Hospital, Newcastle, NSW 2298, Australia
4. School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, NSW 2308, Australia
5. Ingham Institute for Applied Medical Research, Sydney, NSW 2170, Australia
6. Liverpool Cancer Therapy Centre, Liverpool Hospital, Sydney, NSW 2170, Australia
7. South Western Sydney Clinical School, University of New South Wales, Sydney, NSW 2052, Australia
8. Centre for Medical Radiation Physics, University of Wollongong, Wollongong, NSW 2522, Australia
9. Department of Biomedical Engineering, Case Western University, Cleveland, Ohio, OH 44106, United States
10. School of Physics and Astrophysics, University of Western Australia, Crawley, WA 6009, Australia
11. 5D Clinics, Claremont, WA 6010, Australia

Correspondence address:

Ms Angel Kennedy

Level G, B Block, Hospital Ave

Nedlands, Western Australia 6009

Australia

Tel: +61 8 6457 4931

Email: Angel.Kennedy@health.wa.gov.au

Abstract

Purpose: To demonstrate selection of a small representative subset of images from a pool of images comprising a potential atlas pelvic CT set to be used for autosegmentation of a separate target image set. The aim is to balance the need for the atlas set to represent anatomical diversity with the need to minimise resources required to create a high quality atlas set (such as multi-observer delineation), whilst retaining access to additional information available for the potential atlas image set.

Methods: Pre-processing was performed for image standardisation, followed by image registration. Clustering was used to select the subset that provided the best coverage of a target dataset as measured by post-registration image intensity similarities. Tests for clustering robustness were performed including repeated clustering runs using different starting seeds and clustering repeatedly using 90% of the target dataset chosen randomly. Comparisons of coverage of a target set (comprising 711 pelvic CT images) were made for atlas sets of 5 images (chosen from a potential atlas set of 39 pelvic CT and MR images) (a) at random (averaged over 50 random atlas selections), (b) based solely on image similarities within the potential atlas set (representing prospective atlas development), (c) based on similarities within the potential atlas set and between the potential atlas and target dataset (representing retrospective atlas development). Comparisons were also made to coverage provided by the entire potential atlas set of 39 images.

Results: Exemplar selection was highly robust with exemplar selection results being unaffected by choice of starting seed with very occasional change to one of the exemplar choices when the target set was reduced. Coverage of the target set, as measured by best normalised cross correlation similarity of target images to any exemplar image, provided by five well-selected atlas images (mean=0.6497) was more similar to coverage provided by the entire potential atlas set (mean=0.6658) than randomly chosen atlas subsets (mean=0.5977). This was true both of the mean values and the shape of the distributions. Retrospective selection of atlases (mean=0.6497) provided a very small improvement over prospective atlas selection (mean=0.6431). All differences were significant ($p<1.0E-10$).

Conclusions: Selection of a small representative image set from one dataset can be utilised to develop an atlas set for either retrospective or prospective autosegmentation of a different target dataset. The coverage provided by such a judiciously selected subset has the potential to facilitate propagation of numerous retrospectively defined structures, utilising additional information available with multi-modal imaging in the atlas set, without the need to create large atlas image sets.

Key words: Autosegmentation, clustering, image-registration, image-atlas

A. Introduction

Over the last couple of decades considerable effort has been invested into compiling radiotherapy treatment planning datasets, collected through multicentre clinical trials, for retrospectively identifying predictors of treatment outcome[1,2]. More recently, infrastructure has been developed to undertake such analyses on clinical databases in situ in geographically-distributed clinics[3]. One specific process being utilised in such analyses is autosegmentation of regions of interest ('structures') enabling the rapid retrospective segmentation of possibly complex structures, while ensuring segmentation consistency.

Atlas-based autosegmentation is the process of propagating structures, manually delineated in the space of one or more atlas images, to the space of a target image via registration. There are many variations on atlas-based segmentation methods[4,5]. The success of atlas-based segmentation is dependent on, amongst other factors, features of the application domain[6], and both the quality of the initial expert segmentation and the quality of the registration of the atlas to the target images[5,7].

For application domains such as pelvic CT low contrast borders can increase variability in manual segmentations between experts. This makes it desirable to have atlases segmented by multiple experts so that "consensus" structures can be produced for each atlas image and to provide a measure of inter-rater reliability[4,5]. The lack of soft tissue contrast can make it desirable to form an atlas from image sets where co-registered images from a different modality with high soft tissue contrast such as MR are available. The resources required to generate a multi-structure, multi-image, multi-observer atlas can make it necessary to compromise on the number of atlas images used.

One approach to maximising the utility of a small number of atlas images is to select images that best represent the variation in the target population[4,7,8]. An advantage of developing an atlas set for retrospective segmentation is that the entire target population is available before atlas selection. Identifying a representative subset from a new target population to use as an atlas for that population can lead to improved segmentation over using an atlas chosen from a different population without reference to the target set[9]. However, in cases where additional information, such as multi-modal imaging is available only for a dataset from a separate population it can be desirable to select atlas images from outside the target set.

Successfully identifying representative images requires a good metric for how suitable one image is to act as an atlas for another. Several features of the pelvic CT domain contribute to making this challenging. High quality registrations between these images are difficult to find. Aside from the low contrast borders between organs such as the bladder, prostate and rectum, high levels of variation often exist in bladder and rectal filling leading to large differences and deformations of structures between patient images[4]. Features such as air pockets, calcifications and seeds implanted to aid localisation of the prostate during treatment will often not have direct correspondences in the images of different patients. Consequently local measures of image similarity are generally better for the selection of appropriate atlas images for the segmentation of each individual structure than more global ones[4,5,10], as would be required when selecting representative images for segmentation of multiple structures covering a large area. In addition the intensity similarity metric that best predicts post segmentation structure overlap is not always the same between structures[4]. Finally when selecting atlas images from a different population to the target dataset differences in image quality and acquisition protocols

must be considered and may reduce the ability of the atlas set to represent the variation in the target set.

The aim of this paper is to present a method developed for judicious atlas image selection in the context of pelvic CT segmentation to improve the ability of a small number of atlas images, selected from one population, to be used to autosegment multiple structures in a different large heterogeneous target population. More specifically, procedures are presented to (a) standardise the image sets and minimise heterogeneity before calculating image similarity and (b) perform clustering based selection of exemplars from a pool of potential atlas images that are representative of variation in a target population. To the best of our knowledge this is the first paper exploring the use of clustering methods to select representative exemplars for atlas based segmentation of CT images or to explore clustering based selection of exemplars from one image population to generate an atlas for segmenting another.

B. Materials and Methods

A. Datasets

The potential atlas (PA) set consisted of pre-registered pelvic MR and CT datasets for 39 prostate cancer patients attending the Calvary Mater Newcastle Hospital, New South Wales, Australia[11]. Prostate, bladder and rectum segmentations, manually delineated by 3 observers and combined into gold standard segmentations were also available for this dataset. The target (T) set consisted of CT for 711 prostate cancer patients treated across 23 centres during the TROG 03.04 RADAR trial[12].

B. Constraints

The need to undertake analysis on the RADAR trial dataset meant specific constraints applied to the selection of exemplar images:

- Atlas images could only be selected from the PA set due to the availability of corresponding MR images, required for the segmentation of some soft tissue structures.
- Image registrations and atlas selection could only be performed on CT images as MR images were not available for the T set.
- In order to avoid any dependence on inconsistently defined[12] structures already present in the T set, structure-guided registration and similarity-assessment were excluded from consideration.
- The need to propagate multiple extensive structures meant accurate registration and intensity similarity calculation was required over the entire pelvic region.

C. Pre-processing

There are many potential sources of image heterogeneity in multi-centre trial CT data. These can include: variations in CT intensities due to variable image-acquisition, post-processing and scaling information[13]; variations in fields of view; variations in slice thickness, consistency of slice thicknesses and the superior-inferior extent of image slices; the presence of image artefacts and inconsistencies, such as inclusion of support systems and immobilisation devices. Image heterogeneity can negatively impact image

registration and consequently atlas based segmentation by effectively adding noise to the intensity similarity metric.

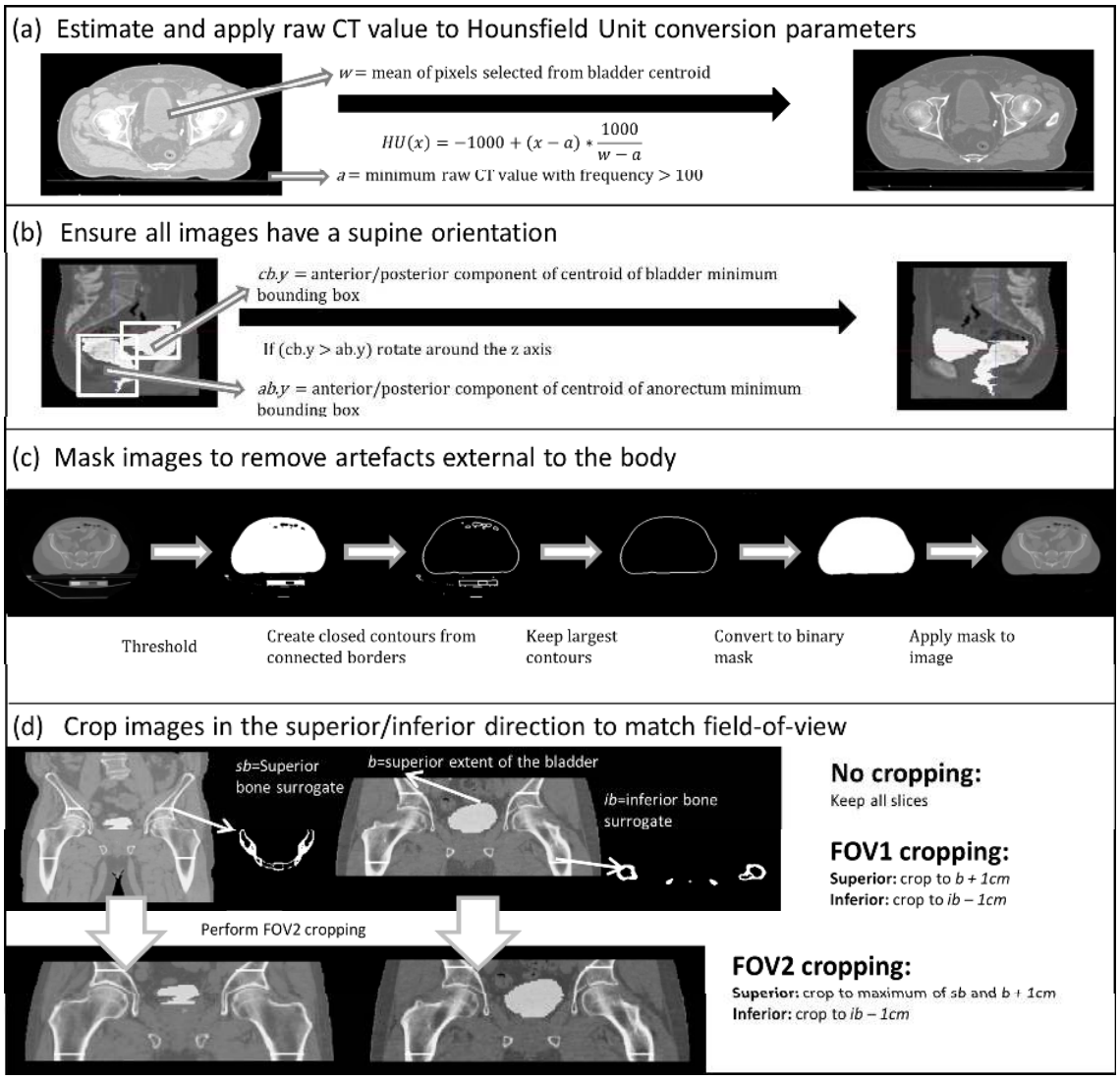

Figure 1. Image pre-processing steps aimed at reducing irrelevant image similarities. (a) Estimate raw pixel value to Hounsfield Unit conversion parameters from known image regions of air and water when these parameters are missing from exported image headers. (b) Ensure matching orientation by checking comparative location of organ centroids and applying a rotation if necessary. (c) Create and apply body region masks to remove external artefacts. (d) Improve field-of-view match by cropping to organs and/or identified bone regions of interest. Field-of-view 1 (FOV1) defines the cropping region by the bladder only in the superior direction, whereas field-of-view 2 (FOV2) uses a bone surrogate as well. Further details of methods and evaluation are available in the supplementary material sections a-d.

Figure 1 provides an illustration of the pre-processing steps undertaken to reduce image heterogeneity. Some of these steps were required due to loss or inaccuracy of image metadata from archived data[13]. The pre-processing steps shown in Figure 1 were:

a. Estimation and application of raw CT to Hounsfield unit (HU) conversion parameters (for further detail see [14]). An evaluation of the impact of estimation inaccuracies is provided in supplementary material section A.
b. Patient imaging orientation was automatically assessed and images rotated if not supine. Supplementary material section B provides further detail.

c. Image artefacts external to the patient body were removed by creating and applying a body region mask. Details of the algorithm used are presented in supplementary material section C.
d. Image sets were cropped to a common field of view to reduce the size of the region to be registered and to improve consistency between image sets. Two definitions for the cropping region were evaluated, with FOV2 used in pre-processing for the clustering presented here. Further information on the methods used to automate the process is presented in supplementary material section D. An evaluation of the impact of different methods of field-of-view cropping on post-registration image similarity is presented in supplementary material section E.

Patients with large artefacts caused by hip prostheses in the central pelvic region were excluded due to extreme degradation of the soft tissue regions of interest.

D. Registration and Similarity Calculation

Pairwise rigid and deformable registrations of the 39 potential atlas CT images were performed along with rigid and deformable registrations of each potential atlas CT image to each of the 711 target CT images. The Normalised Cross Correlation (NCC) similarity was then calculated for each registered pair.

A robust, inverse consistent, block matching rigid registration algorithm called Mirorr[15] was used to provide the initial mapping of the potential atlas images to target images. This was followed by an Insight Segmentation and Registration Toolkit[16] (ITK) based implementation of the non-parametric diffeomorphic demons deformable registration algorithm[17]. Diffeomorphic demons was selected due to the need to use a non-parametric registration algorithm and the previously demonstrated superiority of the algorithm for deformable registration of the pelvic regions for atlas propagation[4]. Previous atlas pre-selection methods have employed rigid registration without deformable registration to reduce computation time[5,9] but given that in this case the primary limiting factor was the time required to manually generate high quality atlases it was important to optimise the quality of the image similarity calculation. A brief description is provided below of a comparison of rigid only versus rigid and deformable registration on the relationship between intensity and structure overlap similarity metrics.

We chose the normalised version of cross-correlation (NCC) to calculate post-registration image similarity because: normalisation provides a linearly-independent metric that is insensitive to errors in exported image pixel value to HU conversion parameter estimation; cross-correlation has previously compared favourably against normalised mutual information (NMI) and the sum-of-squared differences (SSD) for the selection of pelvic CT atlas images when calculated locally to each structure[4]. It was not, however, found to be the best intensity similarity metric for all structures examined.

A comparison of intensity similarity metrics (NCC, NMI, SSD, and the L2 norm), calculated globally, in terms of their Pearson product-moment correlation coefficient (correlation coefficient) to the Dice Similarity Coefficient (DSC) measure of structure similarity was performed. This was done for the three soft tissue segmentations that were available for the PA set after either rigid only or rigid and deformable registration. The highest correlation coefficient between any global intensity similarity metric and average DSC for the three structures after rigid registration only was 0.0534 indicating a need for the better registration afforded by deformable registration for this application. The highest

correlation coefficient overall was between NCC after deformable registration and DSCs after deformable registration though this was still low at 0.27. Further details of both methods and results are presented in supplementary material section F.

The overall suitability of NCC as a similarity metric for determining how well one image can act as an atlas for another is likely to depend on many factors including: the structures to be segmented[4]; the quality of the images; the region of the image that similarity is calculated over; the quality of the image pre-processing and registration; the atlas propagation algorithm used. The comparison of intensity metrics in terms of their correlation coefficient to structure overlap presented in supplementary material section F provides only an approximation of the suitability of the metric for this application as most of the intended structures were not available for evaluation and the atlas propagation algorithm has not been selected. An additional qualitative evaluation of the suitability of NCC for our application, based on visual inspection and relating image features to post registration NCC, was therefore performed and is presented in supplementary material section F.

E. Clustering

The aim of clustering was to approximate the diversity in the T set using only a small number of representative members ("exemplars") of the PA set. The selection of the exemplars can be achieved using an algorithm that can cluster the combined T and PA sets, based on the pre-determined similarities for all combinations of image pairs, whilst ensuring that PA images will be selected as exemplars. The affinity propagation algorithm was used as implemented in the apcluster Cran R package[18,19]. Affinity propagation is able to accommodate sparse similarity matrices so only relevant similarities are considered and has been demonstrated to be a low-error efficient clustering method[19]. As a result of its message passing algorithm used to determine, at each iteration, the suitability of any one dataset to be an exemplar for any other dataset, affinity propagation returns the dataset selected to be the best exemplar for each cluster along with the clusters themselves. Consequently no further steps were required to select cluster exemplars to comprise the atlas set.

Affinity propagation is initialised with similarity weights connecting pairs of data-points and with an initial preference value for each data-point to be chosen as an exemplar. The weights and initial preference values ultimately determine both the number of clusters and the choice of exemplars. The similarity matrix was constructed using asymmetric similarities s of all potential atlas images pa to each other (s_{pai_paj}) as well as all potential atlas images to all target images t (s_{pai_tj}). Initial preference values pr for the target images were set to a very low value (pr_t=-100) compared to potential atlas (pr_{pa}=min(s)) images to ensure that they would not be chosen as exemplars.

Selection of five clusters was enforced by iteratively adjusting starting preferences for the potential exemplars by ±α, depending on whether too few or too many clusters were chosen, until the algorithm achieved the desired number of clusters. Initially α was set to 2*(median(s)-min(s)) and was then halved each time the direction of change reversed (e.g. if the previous run resulted in too many clusters and the current run had too few). Five clusters were selected as an achievable number (given available resources for the creation of high quality, multi-observer, atlases) for use in a subsequent multi-structure inter-observer delineation exercise. The impact of changing the number of clusters selected was explored with results presented in the Supplementary material section H.

Random selection of five exemplars was performed 50 times. Aggregated similarities of randomly selected exemplars to target images from the 50 runs were used as a base-line for comparison with exemplars selected by clustering.

C. Results

A. Clustering Robustness

Repeated runs of affinity propagation on the same similarity set using different random starting seeds resulted in the same cluster assignment and exemplar selection indicating robustness. To test robustness with small changes to the target dataset, clustering was performed after removing similarities to 10% of the target images, chosen randomly. Four of the five exemplars were chosen every time. The fifth was chosen 97 out of 100 times.

B. Coverage of the Target Dataset

Figure 2 presents a comparison of frequency distributions of NCC calculated between registered target images (T) and different atlases. Distributions are shown for similarities (S) of the single most similar atlas image (max; representing the case where only the most similar atlas image is used to segment a target image) and each target image. These atlases comprise: the entire potential atlas (PA) set of 39 images (representing the best possible coverage achievable if the entire PA set was used); 5 randomly selected exemplars, with results aggregated over 50 runs (RE); 5 exemplar images chosen when clustering only within the PA set ("PAE", replicating *prospective* use of the PA set, blinded to similarities with the T set) and; 5 exemplar images clustered on the basis of similarities within the PA set as well as between the PA and T sets ("PATE", representing *retrospective* use of similarities from the PA set to the T set).

When the 5 exemplars are chosen based on clustering (series 3; mean=0.6431, sd=0.0876, and series 4, mean=0.6497, sd=0.0857) coverage of the target data set more closely approximates that of the best possible coverage when the entire dataset is used (series 1; mean=0.6658, sd=0.0785) than the coverage provided by random selection of exemplars (series 2; mean=0.5977, sd=0.0963). This is true of the shape of the distributions (as shown in figure 2) as well as the overall mean. The decrease in average best similarity of T images to any exemplar when only 5 exemplars selected based on the T set in comparison to using all 39 potential atlas images is small but significant ($p=1.17E-35$)

Taking into consideration similarities to target images as opposed to only potential atlas images during clustering resulted in a small significant improvement for the best similarity ($p=1.8E-8$ based on a paired, one-tailed, Student's t-test) of target images to any atlas image.

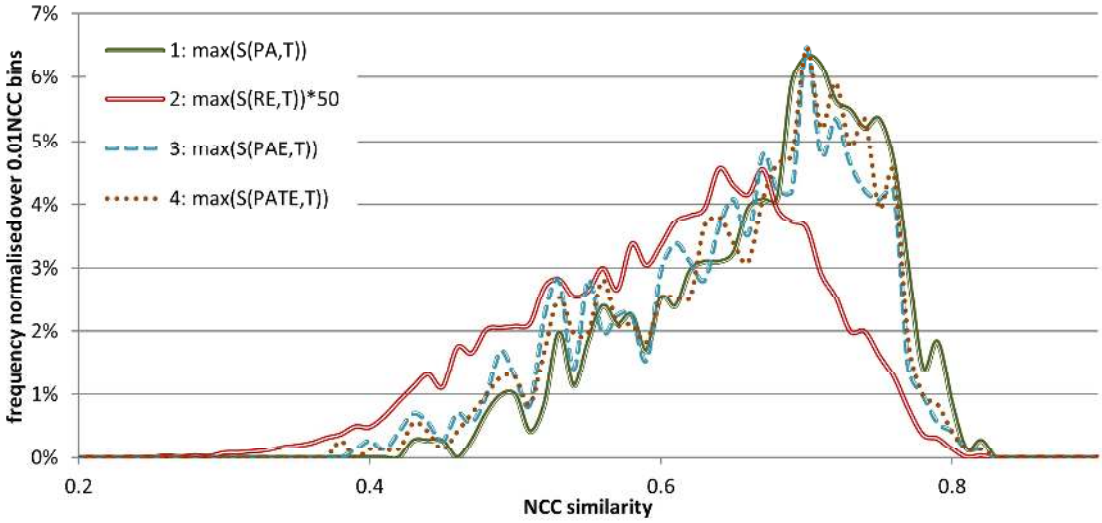

Figure 2: Normalised frequency distributions of NCC similarity ("S") of all target images to (**1**) the most similar of the entire PA image set, (**2**) the most similar of a randomly selected image set R aggregated over 50 runs, (**3**) the most similar of the PAE image set, (**4**) the most similar of the PATE image set.

A similar pattern of results was found when similarities from all exemplars to all target images were considered (representing the case where multiple atlas images are used to segment a single target image). The mean of similarities for all exemplars to all target images when exemplars were selected randomly was 0.5143 (sd=0.1096) compared to 0.5416 (sd=0.1188) for exemplars selected based on the PA set alone vs 0.5589 (sd=0.1132) for exemplars based on the PA and T sets. All differences were significant ($p < 1.0E-10$).

D. Discussion

Retrospective segmentation of large older multi-trial image sets has the potential to result in improved quality datasets that can be useful in multiple applications such as retrospective identification of relationships between planning and treatment outcome or preparation of a training set for a machine learning model. Retrospective segmentation is a challenging task when faced with limited resources, made more difficult due to issues of image quality and image diversity. These difficulties are compounded in the case of pelvic CT images by high levels of deformability and low contrast boundaries of several of the structures of interest[4]. The higher soft tissue contrast and detail visible on MR can enable segmentation of regions that are not clearly visible on CT making it desirable to select atlas CT images for which corresponding MR images are available. By performing clustering we were able to select a small subset of potential atlas images (those with corresponding MR images) that provides coverage, as measured by global NCC, of the target set (without available MR images) that is almost as good as the coverage available from using the entire potential atlas set and that is far superior to a random selection of the atlas subset. When considering NCC frequency distributions in Figure 2 it is seen that the NCC distribution based on using the entire PA (series 1) set is very similar to that based on five judiciously selected exemplars (series 3 and series 4). As such, selecting the five most representative PA images minimises the impact of reducing the number of available atlas images as measured by NCC. Figure 2 shows that using clustering based exemplar selection results in a noticeable improvement in the NCC similarity distribution when compared to a random selection of five atlas images.

A small but significant improvement in coverage of the Target set was seen when exemplar selection was based on similarities to the T set as well as the PA set (retrospective clustering) as opposed to the PA set only (prospective clustering). This was true when either similarities to all exemplars were considered or only the best similarity for each target image to any exemplar. This improvement demonstrates some advantage of retrospective clustering with an available target set (as available when using clinical trial or collated clinical data) as opposed to prospective target sets (as might be obtained during a treatment planning process for new patients). The fact that prospective clustering came close to providing the coverage achieved using retrospective clustering suggests that an atlas selected using this method has potential to be used for organ segmentation in a clinical setting or for segmenting other trial datasets. The time and processing required for segmentation in these settings would depend on the exact methods selected for atlas propagation and the hardware and software utilised.

One factor likely to influence the amount of improvement that can be achieved by considering the T set during clustering is the relative distributions of variations in features important to registration in the PA vs the T set. Differences in imaging characteristics between datasets, for example, motivated Doshi et al's strategy of creating a new atlas set selected from members of the target dataset when segmenting new datasets[9]. If, for a particular feature, the distributions are different between the datasets but there is enough variation of the feature in the PA set then it should be possible to select exemplars that better represent the distribution of variations of that feature in the T set. If on the other hand the distributions are similar or there is little to no variation in the PA set for that feature then the inclusion of similarities between the PA and T sets during selection of exemplars has limited potential to improve the extent to which the selected exemplars represent the variation of that feature within the T set.

Visual inspection of registrations with different levels of NCC similarity (supplementary material F) led to the identification of several features including image slice-widths, patient BMI and bladder volume that appeared to have an impact on registration success. Examination of means and standard deviations showed that distributions of bladder volume and BMI between the datasets were very similar but that there was a much higher variation in slice gaps in the T set (mean=3.21mm, sd=1.07mm) than in the PA set (mean=2.41mm, sd=0.19mm). Moderate correlation coefficients to post registration similarities existed for differences between registered images in BMI and bladder volume but not slice gaps within the PA set. For post registration similarities between the PA and T sets the strongest correlation coefficient is for differences in slice gaps. The correlation coefficient with BMI is reduced and the correlation coefficient with bladder volume disappears (see supplementary material sections F and G for further detail). Differences in variation in slice gaps may have reduced the effectiveness of taking similarities to the T images into consideration during exemplar selection both because the PA set cannot match the variation in the T set and because the impact on registration success can obscure relationships with other features.

Another factor that can influence the impact of retrospective as opposed to prospective clustering is the number of clusters being selected. An investigation into the way coverage changed as the number of clusters increased from one to ten (see supplementary material H) showed that for cluster numbers less than four the coverage offered by prospective clustering compared to the retrospective clustering was unstable. For cluster numbers four or higher the more clusters used the more similar the coverage created by prospective and retrospective clustering (retrospective is guaranteed to be at least as good as prospective). This suggests that if only a small number of exemplars are used in

comparison to the size of the PA set retrospective selection of exemplars can be more advantageous. Interestingly, with one exception, increasing the number of clusters resulted in an exemplar being added but none being replaced.

A limitation of this work is the low correlations found within the PA set between global measures of intensity similarity calculated over the entire pelvic region and local structure overlap, despite the use of deformable registration and the pre-processing performed to reduce irrelevant image differences (see supplementary material section F). The difficulties involved in performing registration of large image regions containing multiple highly deformable organs can mean that an atlas image that is suitable for segmenting one structure within an image is not necessarily suitable for segmenting another. Future work could examine the possible benefit of splitting images up into smaller regions and calculating intensity similarity separately for each region. Clustering would then need to account for how well exemplars covered the variation in target images for each separate region.

The impact of the size of an atlas set on the accuracy of subsequent contour propagation has previously been investigated. Isgum et al presented a method for automatically identifying a subset of atlases expected to yield maximal performance on new data[20]. This method required that the potential atlas set and the target set used for atlas selection already have ground truth segmentations, which is not the case for the application presented here. Awate and Whitaker presented a method for quantifying the number of atlas images required to achieve a particular level of accuracy given the target dataset and atlas propagation methodology[6]. Their method also relied on having a number of pre-existing segmented images. An inability to determine or provide the optimal number of atlases may mean that a suboptimal number is created and used. Awate and Whitaker used ten atlases as a minimum in all their tests of their model and Isgum et al found that eight was the optimal number for their application. Given this and the difficulty of pelvic CT registration and segmentation it is unlikely that a set of five atlas images is optimal for this application. However, this may be compensated for to an extent by judicious selection of atlas propagation strategies such as those described below.

Analysis of the relationship of image similarity and image characteristics may provide us with information that is useful in the selection of atlas propagation strategies or pipelines. For example, if slice-gap differences are impacting registration quality it may be possible to compensate for this to an extent by matching image resolutions using down-sampling of higher resolution images[21] or improved slice gap interpolation methods when up-sampling lower resolution images[22] prior to registration of atlas images to target images. In cases where there are significant differences between populations for particular features it may be helpful to employ algorithms that make use of learned information about structure shape and intensity profiles such as active shape models[23] or algorithms such as LEAP[7], which are designed to overcome population differences by employing image similarity information and intensity refinement to iteratively expand the set of atlas images, using images already segmented by the atlas. The combination of such tools with the judicious selection of an initial set of images with high quality segmentations should enable sufficiently accurate retrospective segmentation with limited resources.

E. Conclusions

Clustering of image similarities can be used to select an image subset that is representative of population variation within a target dataset. The extent to which this representativeness will translate into successful atlas contour propagation depends on several factors including: how well the similarity measure reflects the registration success

of the images within the region of the contours to be propagated; how closely the variation in the potential atlas set matches that of the target set; and the methods used to propagate atlas contours. The image pre-processing and registration pipe-line presented here removes some of the irrelevant image differences in an attempt to improve the quality of image intensity similarity as a measure of how well one image can act as an atlas for another. In the context examined here, where a small atlas is being selected for segmentation of large, highly variable, and often low contrast regions, improvements in atlas selection are most likely to come from refinements in the methods of image registration and similarity calculation.

Acknowledgements

This work was supported by the National Health and Medical Research Council of Australia (NHMRC Project Grant 1077788). We are grateful to Prof James Denham and other investigators of the TROG 03.04 RADAR trial and to the developers of the CSIROs in house image processing software MILXView.

F. Disclosure of conflicts of interest

The authors have no relevant conflicts of interest to disclose.

References

1. Ebert MA, Haworth A, Kearvell R, et al. Detailed review and analysis of complex radiotherapy clinical trial planning data: Evaluation and initial experience with the SWAN software system. *Radiother Oncol.* 2008;86(2):200-210.
2. Deasy JO, Blanco AI, Clark VH. CERR: A computational environment for radiotherapy research. *Medical Physics.* 2003;30(5):979-985.
3. Jochems A, Deist TM, van Soest J, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. *Radiother Oncol.* 2016;121(3):459-467.
4. Acosta O, Dowling J, Drean G, Simon A, de Crevoisier R, Haigron P. Multi-Atlas-Based Segmentation of Pelvic Structures from CT Scans for Planning in Prostate Cancer Radiotherapy. In: El-Baz AS, Saba L, Suri J, eds. *Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective.* doi: 10.1007/978-1-4614-8498-1_24 Boston, MA: Springer US; 2014:623-656.
5. J.E. Iglesias MRS. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis.* 2015;24(1):205-219.
6. Awate SP, Whitaker RT. Multiatlas Segmentation as Nonparametric Regression. *IEEE Transactions on Medical Imaging.* 2014;33(9):1803-1817.
7. Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D. LEAP: Learning embeddings for atlas propagation. *NeuroImage.* 2010;49(2):1316-1325.
8. Langerak TR, Berendsen FF, Van der Heide UA, Kotte ANTJ, Pluim JPW. Multiatlas-based segmentation with preregistration atlas selection. *Medical Physics.* 2013;40(9):091701-n/a.
9. Doshi J, Erus G, Ou Y, Gaonkar B, Davatzikos C. Multi-Atlas Skull-Stripping. *Academic Radiology.* 2013;20(12):1566-1576.
10. Zaffino P, Ciardo D, Raudaschl P, et al. Multi atlas based segmentation: should we prefer the best atlas group over the group of best atlases? *Physics in Medicine & Biology.* 2018;63(12):12NT01.

11. Dowling JA, Sun J, Pichler P, et al. Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences. *Int J Radiat Oncol Biol Phys.* 2015;93(5):1144-1153.
12. Kearvell R, Haworth, A., Ebert, M. A., Murray, J., Hooton, B., Richardson, S., Joseph, D. J., Lamb, D., Spry, N. A., Duchesne, G. and Denham, J. W. Quality improvements in prostate radiotherapy: Outcomes and impact of comprehensive quality assurance during the TROG 03.04 'RADAR' trial. *J Med Imag Radiat Oncol.* 2013;57:247-257.
13. Chang D, Joseph DJ, Ebert MA, et al. Effect of androgen deprivation therapy on muscle attenuation in men with prostate cancer. *J Med Imag Radiat Oncol.* 2014;58(2):223-228.
14. Kennedy A, Dowling JA, Greer P, Ebert MA. Estimation of Hounsfield Unit conversion parameters for pelvic CT images. *Australasian Physical & Engineering Sciences in Medicine.*In press.
15. Rivest-Hénault D, Dowson N, Greer PB, Fripp J, Dowling JA. Robust inverse-consistent affine CT–MR registration in MRI-assisted and MRI-alone prostate radiation therapy. *Medical Image Analysis.* 2015;23(1):56-69.
16. Yoo TS, Ackerman,Michael J., Lorensen,William E., Schroeder,Will, Chalana,Vikram, Aylward,Stephen, Metaxas,Dimitris, Whitaker,Ross Engineering and algorithm design for an image processing API: A technical report on ITK - The Insight Toolkit. In: James D. Westwood HMH, Richard A. Robb, Don Stredney, ed. *Medicine Meets Virtual Reality 02/10.* Vol 85. Amsterdam, The Netherlands: IOS press; 2002:586-592.
17. Vercauteren T, Pennec X, Perchant A, Ayache N. Non-parametric Diffeomorphic Image Registration with the Demons Algorithm. In: Ayache N, Ourselin S, Maeder A, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007: 10th International Conference, Brisbane, Australia, October 29 - November 2, 2007, Proceedings, Part II.* doi: 10.1007/978-3-540-75759-7_39 Berlin, Heidelberg: Springer Berlin Heidelberg; 2007:319-326.
18. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. *Bioinformatics.* 2011;27(17):2463-2464.
19. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. *Science.* 2007;315(5814):972.
20. Isgum I, Staring M, Rutten A, Prokop M, Viergever MA, van Ginneken B. Multi-Atlas-Based Segmentation With Local Decision Fusion—Application to Cardiac and Aortic Segmentation in CT Scans. *IEEE Transactions on Medical Imaging.* 2009;28(7):1000-1010.
21. Zhao C, Carass A, Jog A, Prince JL. Effects of Spatial Resolution on Image Registration. *Proceedings of SPIE--the International Society for Optical Engineering.* 2016;9784:97840Y.
22. Frakes DH, Dasi,Lakshmi P., Pekkan,Kerem, Kitajima,Hiroumi D. Sundareswaran,Kartik, Yoganathan,Ajit P. Smith,Mark J T. A new method for registration-based medical image interpolation. *IEEE Transactions on Medical Imaging.* 2008;27(3):370-377.
23. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding.* 1995;61(1):38-59.