Fully Automatic Segmentation of Head and Neck Organs using Active Appearance Models

Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, and Graham Vincent

Imorphics Ltd., Kilburn House, Manchester Science Park, Manchester, M15 6SE, UK.

Abstract. We present a fully automatic model based system for segmenting the mandible, parotid and submandibular glands, brainstem, optic nerves and the optic chiasm in CT images, which won the MICCAI 2015 Head and Neck Auto Segmentation Grand Challenge. The method is based on Active Appearance Models (AAM) built from manually segmented examples via a cancer imaging archive provided by the challenge organisers. High quality anatomical correspondences for the models are generated using a Minimum Description Length (MDL) Groupwise Image Registration method. A multi start optimisation scheme is used to robustly match the model to new images. The model has been cross validated on the training data to a good degree of accuracy, and successfully segmented all the test data.

1 Introducton

Intensity modulated radiotherapy requires detailed target and tissue sparing plans derived from accurate segmentations (or *contouring*) of key organs. Existing systems for the head and neck organ segmentation are dominated by atlas based segmentation as evaluated in [2], [5], [7], [12] and [13].

We have used a statistical modelling pipeline to build a model of the head and neck structures. Systems based on this pipeline have won previous challenges for knee bone and cartilage [11] and prostate [10]. The model fitting is based on Active Appearance Models (AAMs) [3] in which the statistics of shape and image information, and the correlations between them, are calculated from a training set of images. Various model hierarchies can be defined as appropriate. For example localisation using a model of the orbit can be followed by a specific optic nerve model. Variants of AAMs have been extensively developed (see [6] for a review).

The models were evaluated using Leave-One-Out Cross Validation (LOOCV) in which each patient case in turn is removed from the model training set, the statistical models re-built and then used to segment the test case. The resulting surface segmentation was turned into a partial volume image and thresholded at a partial volume of 0.5 to form a binary voxel segmentation. These were then compared with the reference segmentations using the standard DICE overlap measure and the 95th percentile Hausdorff distance, using the Plastimatch software [9].

Our interest in participating in the Head and Neck Auto Segmentation Grand Challenge was to evaluate how our standard AAM-based pipeline would perform on the head and neck organs against state of the art algorithms developed for that task. Our goal is to develop a system which is as generalisable as possible, and can be applied to a wide variety of clinical and research problems with minimal customised development.

2 Methods

2.1 Data

The models reported here are built using 33 CT scans from a cancer imaging archive provided through the ImagEngLab [1]. Voxel based reference segmentations for the mandible, parotid and submandibular glands, brainstem, optic nerves and the optic chiasm were provided by the organisers. These segmentations were turned into real valued 3D surfaces using the 0.5 valued iso-surface. The 3D surfaces and images form the input to the model building process described in the rest of this section.

The organisers also provided 10 cases without segmentations to test before the challenge (the "off-site" test data) and 5 test cases without segmentations on the day of the challenge (the "on-site" test data). The organisers, who have access to the reference segmentations, were able to analyse our results on these test datasets to generate DICE and Hausdorff distance measures for each case.

2.2 Generating surface correspondences

Statistical appearance models rely on a large set of anatomically equivalent landmarks (also known as *correspondences*) across the region of interest. Generating good quality correspondences is key to developing generalisable yet specific models.

To obtain the anatomical correspondences on the surfaces we used a variant of the Minimum Description Length approach to Groupwise Image Registration (MDL-GIR) [4]. The MDL-GIR method finds the set of deformations which register all the images together as efficiently as possible. This idea is made concrete by the use of Information Theory to define the amount of information required to encode a model using a particular set of deformations. The method is an optimisation to find the set of deformations requiring the least amount of information to encode. The output is a reference mean image and a set of deformations which map the mean image to each example image.

We applied the MDL-GIR method to the signed distance images derived from the segmented surfaces for each part independently. Initial registration is achieved by aligning the centre of gravity and scale. The MDL-GIR then proceeds with low parameter deformations (initially rigid rotations and translations) and then increasingly local deformations. The output reference mean image is, like the input images, a signed distance image and can be straight forwardly

segmented at the zero valued iso-surface. The mean surface is then propagated by the appropriate deformation field into the frame of each example. For each example the propagated surface lies close to the segmented surface and is projected onto it to generate correspondence points which are guaranteed to lie on the segmented surface.

In addition we generated some approximate correspondences for the left and right orbits by applying MDL-GIR to the CT images, segmenting the orbit in the mean image and propagating the mean orbit into the frame of each example. The orbit model allows for robust localisation for the optic nerve models.

The number of correspondence points output from this process varies from 64418 for the mandible to 1200 for each optic nerve. The mean model correspondences are shown in Figure 1.

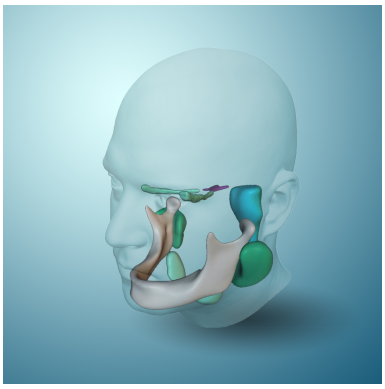

Fig. 1. Surface rendering of the mean model surfaces for the mandible, parotid and submandibular glands, brainstem, optic nerves and the optic chiasm.

2.3 Active appearance models

An appearance model is a statistical model of the shape of a structure and associated imaging information in and around the structure. It is useful to process the imaging information further to obtain feature response images such as gradients, corners and other points of interest [8]. We refer to all such imaging information and their derivatives as texture.

An appearance model has a set of parameters which control both the shape and the texture, and are *generative* i.e. a specific parameterisation can generate a realistic looking example of the shape and texture.

An AAM can match its appearance model to an image from a rough initial estimate, by optimising the model parameters to generate an example which matches the image as closely as possible (using the least squares sum of residuals). This can be made efficient by pre-computing the Jacobian describing the

average change in residuals with respect to changes in model parameters on a training set.

AAMs require an initial estimate of the model parameters including position, rotation and scale. We initialise multiple AAMs in a grid of starting points across the image. The grid of starting points are typically 20mm apart in all directions. This is done at a low image and model resolution with a small number of measured residuals to make it reasonably fast. The results of these searches are ranked according to the sum of squares of the residual, and a proportion (typically 75%) removed from consideration. The remaining search results are used to initialise models at a higher resolution, and so on. Finally, the single best result at the highest resolution gives the segmentation result.

Searching the entire image for everything independently is clearly inefficient. Therefore some structures are searched for in a geometric region which is defined relative to a previously searched structure. For example we initially search for the mandible and calculate its bounding box. We then predict a region to search for each of the other structures based on information learnt during the training phase. The predicted bounding boxes are very approximate and contain a large margin for error, but help make the process more efficient.

2.4 Segmentation pipeline

In summary, the segmentation proceeds according to the following pipeline:

- For each image:
  - Run N (typically *O(100) AAMs of the mandible from a grid of starting positions across the image at low resolution.*
    - Run the 25% best results at increased resolution
    - Repeat until at highest resolution
  - Choose best result
  - Initialise and run separate AAMs of left and right parotid glands, left and right submanidbular glands, using a search region relative to the mandible
  - Initialise and run separate AAMs of left and right orbit using a search region relative to the mandible, followed by AAMs of the optic nerves and optic chiasm

The pipeline takes approximately 30 minutes per image, most of which is taken up localising our reference structures, the mandible and the orbits. This stage could be substantially sped up by using standard imaging processing steps.

3 Results and Discussion

Table 1 shows the results of the cross validation on the training datasets, together with average results from recent papers in the literature, and the results for the off-site and on-site test data as reported by the organisers. The values for the literature are gathered from [2], [5], [7], [12] and [13]. Broadly speaking

our cross validation results compare well against the literature, being better or substantially better for every part, although we note that the results from the literature are indicative only - any more detailed comparison is not warranted here, as datasets can vary in size, consistency and difficulty, which is of course part of the motivation for the Grand Challenge.

The off-site test data presented similar characteristics to the training data. Since no segmentations were available to us, we only performed a visual check and considered that the results were comparable to the cross validation results. At this visual review two cases stood out as presenting particular problems, case 0522c0555 which presented with a variety of small structures, which we presume to be metastases. The automated segmentation performed well in this case, and was not misled by the multiple additional structures around the salivary glands. Case 0522c0746 which presented with a large inclusion inside or close to the parotid gland. In this case the automated segmentation identified the most likely shape for the parotid, and took no account of the inclusion. It is not clear what an automated segmentation should do in this situation, though we think that post-processing of the texture in the segmented structure should be able to robustly identify the inclusion/tumour.

Based on visual review, the automatic segmentation did not appear to be negatively affected by the amount of amalgam used to fill the teeth of some of the individual cases, which causes significant CT image artefacts in some cases.

The numerical results (DICE and 95th percentile Hausdorff distance) for the off-site and on-site test data as reported to us by the organisers are presented in Tables 1 and 2. The results are commensurate with each other and the cross validation results except for the optic nerve which scored a significantly lower value for the off-site test. This requires explanation. Estimating surfaces from thin slice based optic nerve segmentations gave a non-anatomic looking "stepped" structure. Prior to model building, we edited the input surfaces derived from the provided segmentation to make them look more locally anatomic. Unfortunately this effectively modelled a slightly different anatomical region which biased the model and degraded the numerical results against the original segmentations. For the on-site test data we reinstated a model built from the original surfaces.

The organisers also provided the rankings of the methods based on DICE and Hausdorff distance, and our method was the overall winner of the challenge, performing best on 5 out the 6 structures for both off-site and on-site test data[1], and second best on the sixth (the optic chiasm).

4 Conclusions

In this paper we have presented a fully automatic AAM based segmentation pipeline to segment the mandible, parotid and submandibular glands, brainstem, optic nerves and optic chiasm from CT images, built and cross validated on a public dataset. The system won the MICCAI 2015 Head and Neck Auto

[1] 2 teams submitted results after the on-site challenge workshop which affected some of the rankings

Table 1. Average DICE scores (standard deviation) for cross validation together with the off-site and on-site test data as reported by the organisers, and average values from the literature.

	BrainStem	Chiasm	Mandible	OpticNerve	Parotid	Submandibular
Cross validation	0.88 (0.03)	0.44 (0.21)	0.91 (0.02)	0.80 (0.05)	0.82 (0.10)	0.79 (0.05)
Off site	0.87 (0.04)	0.35 (0.16)	0.93 (0.01)	0.63 (0.05)	0.84 (0.07)	0.78 (0.08)
On site	0.89 (0.03)	0.45 (0.29)	0.92 (0.01)	0.78 (0.03)	0.84 (0.03)	0.78 (0.09)
Literature	0.81 (0.07)	0.37 (-)	0.90 (0.05)	0.71 (-)	0.76 (0.09)	0.71 (0.02)

Table 2. Average 95th percentile Hausdorff distance (standard deviation).

	BrainStem	Chiasm	Mandible	OpticNerve	Parotid	Submandibular
Off site	4.02 (2.02)	3.24 (0.42)	1.67 (0.62)	2.80 (0.59)	5.03 (2.43)	4.83 (1.84)
On site	3.36 (1.54)	3.95 (2.13)	2.59 (0.54)	1.83 (0.44)	6.84 (2.98)	5.08 (2.23)

Segmentation Grand Challenge. It is very encouraging that such models can be quickly built and validated on, what is for us, a completely new type of data and achieve results commensurate with the literature. A fully automated segmentation using this pipeline seems practical, though should probably be augmented with further post-segmentation processing to explicitly identify large tumours. The system appears robust to many of the metastases and inclusions found in patients in need of radiation therapy, and is not affected by the image artefacts caused by metal within the teeth.

References

1. A Public Domain Database for Computational Anatomy. `http://www.imagenglab.com/pddca_18.html`.
2. Susan Barley, Clare Antoine, Gareth Webster, Marie Tiffany, Navinah Nundlall, Rosemary Simmons, and Andrew Hartley. Head and Neck Cancer Radiation Therapy Balancing Accuracy with Efficiency in OnQ rts . *European Oncology & Haematology*, 10(2):98–101, 2014.
3. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
4. T. F. Cootes, C. V. Petrovi, R. Schestowitz, and C. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. *16th British Machine Vision Conference. Volume*, 2:879–888, 2005.
5. Jean-François Daisne and Andreas Blumhofer. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiation Oncology*, 8(1):154, 2013.
6. Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3d medical image segmentation: A review. *Medical Image Analysis*, 13(4):543–563, 2009.
7. Mariangela La Macchia, Francesco Fellin, Maurizio Amichetti, Marco Cianchetti, Stefano Gianolini, Vitali Paola, Antony J Lomax, and Lamberto Widesott. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiation Oncology*, 7(1):160, 2012.

8. I. M. Scott, T. F. Cootes, and C. J. Taylor. Improving appearance model matching using local image structure. In *In Information Processing in Medical Imaging, 18 th International Conference*, pages 258–269. Springer, 2003.
9. James Shackleford, Nadya Shusharina, Joost Verberg, Guy Warmerdam, Brian Winey, Markus Neuner, Philipp Steininger, Amelia Arbisser, Polina Golland, Yifei Lou, Chiara Paganelli, Marta Peroni, Marco Riboldi, Guido Baroni, Paolo Zaffino, MariaFrancesca Spadea, Aditya Apte, Ziad Saleh, JosephO Deasy, Shinichro Mori, Nagarajan Kandasamy, and GregoryC Sharp. Plastimatch 1.6 - current capabilities and future directions. *Int Conf Med Image Comput Comput Assist Interv*, 15(WS), 10 2012.
10. Graham Vincent, Gwenael Guillard, and Mike Bowes. Fully automatic segmentation of the prostate using active appearance models. *Proceedings of MICCAI 2012, Grand Challenge Workshop PROMISE12*, 2012.
11. Graham Vincent, Chris Wolstenholme, Ian Scott, and Mike Bowes. Fully automatic segmentation of the knee joint using active appearance models. *Proceedings of MICCAI 2010, Grand Challenge Workshop SKI10*, 2010.
12. Gary V Walker, Musaddiq Awan, Randa Tao, Eugene J Koay, Nicholas S Boehling, Jonathan D Grant, Dean F Sittig, Gary Brandon Gunn, Adam S Garden, Jack Phan, William H Morrison, David I Rosenthal, Abdallah Sherif Radwan Mohamed, and Clifton David Fuller. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiotherapy and Oncology*, 112(3):321–325, September 2014.
13. Mingyao Zhu, Karl Bzdusek, Carsten Brink, Jesper Grau Eriksen, Olfred Hansen, Helle Anita Jensen, Hiram a. Gay, Wade Thorstad, Joachim Widder, Charlotte L. Brouwer, Roel J H M Steenbakkers, Hubertus a M Vanhauten, Jeffrey Q. Cao, Gail McBrayne, Salil H. Patel, Donald M. Cannon, Nicholas Hardcastle, Wolfgang a. Tomé, Matthias Guckenberg, and Parag J. Parikh. Multi-institutional quantitative evaluation and clinical validation of Smart Probabilistic Image Contouring Engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography images in the head and neck, thorax, liver, and male. *International Journal of Radiation Oncology Biology Physics*, 87(4):809–816, 2013.