Assessing the Accuracy and Inter-rater Agreement of Segmenting Articular Cartilage

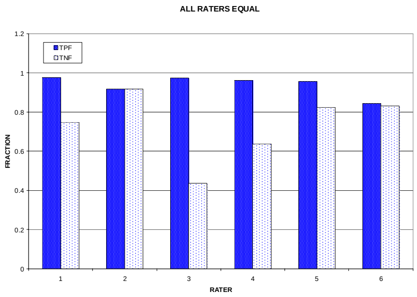
P. Thacker[1], M. Downs[1], J. W. Jaromczyk[1], P. A. Hardy[2]

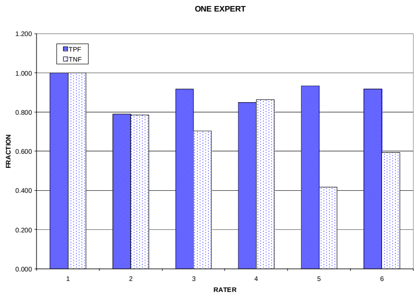
[1]Computer Science, University of Kentucky, Lexington, Kentucky, United States, [2]Biomedical Engineering, University of Kentucky, Lexington, Kentucky, United States

Introduction The segmentation of articular cartilage from 3D MR images is useful for following the progression of subjects with osteoarthritis. The thinness of the cartilage and the presence of tissues adjacent to it which have similar image intensity causes difficulty for humans and computer algorithms to reproducibly and accurately segment cartilage. Evaluating the accuracy of different segmentation algorithms is difficult because of the complexity of the decisions to be made and especially because the algorithms require some user intervention and thus the final result depends on the operator. We have developed a method of determining the agreement among different raters in segmenting articular cartilage. The method will find use in evaluating different manual, semi-automated and fully automated segmentation algorithms.

Methods We modified a statistical method outlined by S.K. Warfield to calculate the true positive fraction (TPF) and the true negative fraction (TNF) of the performance of different raters in segmenting 3D images.[1] Six raters segmented one image set consisting of 64 slices of a 3D image of the knee of a subject with mild osteoarthritis. Each subject used a custom-developed segmentation algorithm to semi-automatically segment the femoral articular cartilage and to produce a ROI file containing the segmented regions. The analysis algorithm examined the six ROI files to determine for each pixel the probability that the rater correctly selected it or excluded it from the ROI. A cumulative TPF and TNF was calculated for each rater.

Results The algorithm calculated the probability a given pixel should be included assuming first, that all the raters were equal and second, that rater #1 was an expert and their choices should be taken as truth. The performance of the raters on one data set is shown in figure 1. The performance of the raters assuming rater #1 was the expert is shown in figure 2.





Discussions & Conclusions The results of figures 1 and 2 show the raters were highly consistent in selecting the regions to include as articular cartilage, i.e. high values of TPF. However, there was wide variation in the TNF among the different raters. The TNF, which relates the ability of the rater to discriminate pixels to exclude, is perhaps indicative of the difficulty of correctly excluding adjacent tissues. The method provides a rigorous approach for evaluating the performance or skill of human raters and machine algorithms for segmenting images. As such it will be very useful in testing and improving automated algorithms for segmenting articular cartilage from MR images.

References

[1] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Validation of Image Segmentation and Expert Quality with an Expectation-Maximization Algorithm. In *MICCAI 2002: Fifth International Conference on Medical Image Computing and Computer-Assisted Intervention; 2002 Sep 25-28; Tokyo, Japan*, pages 298-306, Heidelberg, Germany, 2002. Springer-Verlag.

Acknowledgements

This study was supported by a Biomedical Science grant from the Arthritis Foundation and a grant from EPSCoR KY NSF Summer Research for Undergraduates.