Application of the Dice Similarity Coefficient (DSC) for Failure Detection of a Fully-Automated Atlas Based Knee MRI Segmentation Method

K. G. Baum[1], E. Schreyer[1], S. Totterman[1], J. Farber[1], J. Tamez-Peña[1], and P. González[1]
[1]4Qimaging, LLC, Rochester, NY, United States

Problem - Quantitative analysis of MRI images is providing new insight into and sensitivity to detect osteoarthritic progression[1], but is encumbered with the time, cost and variability associated with manual or semi-automated segmentation[2]. To address this, a fully-automated knee MRI segmentation and analysis method was developed and validated[3,4]. Although the method has proven to be robust, in a small percentage of cases (< 2%) underlying image quality or other anomalies may produce poor segmentation results. This study examines the feasibility of using the Dice Similarity Coefficient (DSC) as an objective, reproducible and automated method of accurately detecting segmentation failure.

Methods - The DSC is a measurement of spatial overlap used widely for comparing segmentation results[5]. The DSC, which can have a value ranging from zero to one, is defined as two times the volume of the intersection between two segmentations divided by the sum of the volumes of the two segmentations. The tested hypothesis was that successful segmentations would be significantly more similar to the anatomical atlas used for generating the segmentations than failed segmentations. This is intuitive as atlas based segmentation techniques rely on predefined shapes to properly segment similar ones. Thus, we expected that DSC values calculated between the atlas and successful segmentations would be significantly higher than that of the atlas and failed segmentations.

DSC measurements were calculated for a database of 2,500 knee segmentations of subjects in the Osteoarthritis Initiative public use dataset. The DSC calculation was limited to the bone regions that quantitative measurements are based on. Thirty cases which had DSC values uniformly distributed between the minimum and maximum value were randomly selected. The selected cases were presented for blinded reads by two experienced musculoskeletal radiologists who scored each segmentation result based on the quality of the segmentation in the weight bearing portions of the medial and lateral compartments. The radiologists' evaluation was used as the ground truth and a receiver operator curve (ROC) analysis was performed to evaluate the use of a DSC threshold value to identify failed segmentations.

Results - The distribution of DSC values for the 2,500 segmentations is shown in figure 1a. The large set of segmentations was necessary in order to provide enough samples in the leftmost tail so that images with a relatively uniform distribution of DCS values could be selected. Both radiologists' evaluations were in agreement for all images. Figure 1b shows a plot of the assigned "pass/fail" ratings versus the DSC value for the 30 images. Examples segmentations are shown in figure 2. Analysis resulted in an ideal ROC curve and a clear threshold below which all images had a failed segmentation and above which all images had an acceptable segmentation.

Conclusions - As is evident from the distribution of DSC values in figure 1a, segmentation failures are uncommon. From this experiment it is clear that the DSC can be used as an indicator to detect failed segmentations. When the segmentation is successful, the general shape of the segmented anatomical structures is preserved and a relatively high DSC value results. In situations where the segmentation fails, arbitrary boundaries with high curvature and shape dissimilar to any knee structures produce low DSC values. It is anticipated that in certain situations there may be an overlap of the DSC values corresponding to successful and failed segmentations, and that a single threshold will not be able to achieve a 100% accuracy. In such cases a tradeoff between specificity and sensitivity must be considered to select an appropriate threshold. As this study has shown, a fully-automated method for reliably detecting segmentation failures is feasible.

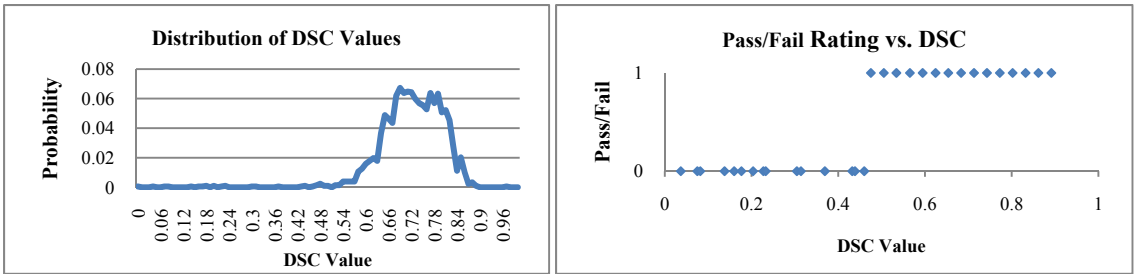

Fig. 1. (a) Distribution of DSC values from the 2,500 segmentations. (b) "Pass/fail" ratings assigned by the radiologists for the 30 randomly selected images used in the study. A value of one indicates a success and a value of zero indicates a failure.

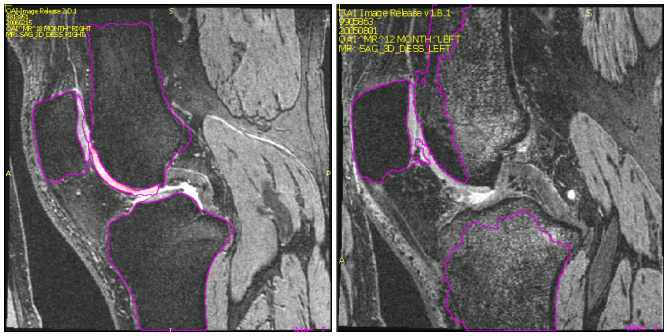

Fig. 2. (left) A typical segmentation labeled as passing by the expert radiologists. (right) A typical segmentation labeled as failing by the expert radiologists.

1. E. B. Dam, J. Folkesson, P. C. Pettersen, and C. Christiansen, "Automatic Morphometric Cartilage Quantification in the Medial Tibial Plateau from MRI for Osteoarthritis Grading," *Osteoarthritis Cartilage*, vol. 15, pp. 808-818, 2007.
2. J. Duryea et al., "Novel Fast Semi-Automated Software to Segment Cartilage for Knee MR Acquisitions," *Osteoarthritis Cartilage*, vol. 15, pp. 487-492, 2007.
3. S. Totterman, J. Tamez-Peña, E. Schreyer, P. González, and D. Hunter, "Cartilage-Bone Contrast Behavior in OAI Progression Sub-cohort; Correlation to WOMAC Scores," in *FDA-OARSI Biomarker Conference*, Bethesda, MD, 2009 Apr 23-24.
4. J. Tamez-Peña, V. Treviño, E. Schreyer, S. Totterman, and P. González, "Functional Characterization of a Fully Automated Knee Segmentation System on the OAI DESS Sequences," in *OARSI World Conference*, Montreal, Canada, 2009 Sep 10-13.
5. L. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, pp. 297-302, 1945.