Validation of the KneeIQ segmentation framework on SKI10

Erik B. Dam[1,2]

[1]Biomediq A/S, Copenhagen, Denmark, and

[2]The D-BOARD European Consortium for Biomarker Discovery

Abstract. Several strong methods for knee MRI segmentation have been evaluated using the SKI10 collection. We evaluate a method combining multi-atlas pre-registration with a supervised classifier. The results on the training collection seem promising for cartilage segmentation. For bone segmentation, the method could likely be improved for scans with weak signal by incorporating a statistical shape model.

1. Introduction

Quantitative analysis of cartilage morphometry from MRI is a demanding task. Manual segmentation by a trained expert requires approximately an hour per scan per desired compartment. A typical phase III clinical trial can include 5,000 scans (1,000 participants, 3 visits, 1-2 knees per visit) and cover the main knee cartilages; ideally requiring around 25,000 expert hours of interaction. The large epidemiological osteoarthritis Initiative (OAI) study[1] includes 4,796 subjects with MRI and ultimo 2013 six main visits with in total 24,885 MRI visits and approximately 50,000 scans of left and right knees for a given sequence. Including five cartilage compartments, the around 250,000 expert interaction hours needed would correspond to around 125 working years. This reader bottleneck makes comprehensive analysis infeasible – if based on manual segmentation. Furthermore, more advanced 3D morphometric markers, such as cartilage surface smoothness[2] or joint congruity[3], will suffer from the inter-slice discontinuity artifacts caused by slice-wise delineation. Finally, inclusion of additional structures beyond cartilage such as bones and menisci increases the quantification challenge further.

Recently, several promising methods for automatic knee MRI segmentation have emerged and been evaluated using the SKI10 collection from the MICCAI 2010 segmentation challenge workshop[4]. The current top-scoring method at SKI10 (by February 2015) is fully automatic and based on an active appearance model (AAM), where the bone shape models have been built using the minimum description length (MDL) approach for optimizing correspondences between point distribution models (PDM), with average tibial and femoral cartilage volume overlap errors at 23.4% and 23.9%[5]. The second-scoring at SKI10 is also based on bone shape models, but the cartilages are segmented using a multi-object graph optimization method that results in cartilage overlap errors of 25.8% and 26.4%, respectively[6].

In other medical image segmentation tasks, a lot of attention has been given to registration-based approaches. Non-linear registration methods have been matured allowing elastic, diffeomorphic, and other geometric transformation classes[7], and the implementations are now feasible in terms of computation time. However, for joint segmentation, it is not obvious to follow this trend. Firstly, joints are a mixture of rigid and soft tissue, meaning that standard elastic registration approaches may be too flexible. Secondly, the important cartilage compartments are so small that a standard registration objective function may implicitly prioritize registration accuracy in the bones rather than the cartilages. Finally, cartilage denudation results in topological changes that are problematic in diffeomorphic registration models. It would appear that dedicated methods are needed for robust and accurate joint registration – such as articulated-rigid bone registration followed by locally elastic registration in cartilage areas. The third-ranking SKI10 method actually applies a multi-atlas non-rigid registration approach to produce an initial segmentation that is used as input to a graph-cut method, resulting in cartilage overlap errors of 28.3% and 27.6% for tibial and femoral cartilages (from UPMC, previous version presented[8]). The fifth-ranked method[9] also applies multi-atlas registration using patch-based label fusion (previously demonstrated to be very applicable for brain MRI segmentation) followed by a specialized three-class classification. This demonstrates the potential for registration-based methods, but also the need for combination with other approaches.

The objective of this SKI10 submission was to evaluate the knee image quantification (KneeIQ) framework for fully automatic segmentation of knee joints from MRI. The KneeIQ framework combines some of the appealing aspects from the previous approaches by using a voxel classification framework heavily inspired by Folkesson[10], but combining with a multi-atlas pre-registration step. The rigid pre-registration allowed the selected classification features to be more specific to the individual structures since the majority of the background is effectively identified and the structures well aligned. The framework allows varying structure sets and can be used for different collections with different configurations of available bone/cartilage/meniscus compartments training data.

2. Methods

The KneeIQ segmentation framework combines rigid multi-atlas registration with voxel classification in a multi-structure setting[11]. The voxel classification step includes a region-of-interest analysis and feature selection for each structure. This is a generalization of the method proposed by Folkesson[10] that was implemented for the two medial cartilage compartments. Central is in the update from Folkesson we the generalized to arbitrary compartments and the added registration step to improve the performance.

Pre-processing

Prior to using the KneeIQ framework, three pre-processing steps were performed.

Firstly, the scanners and sequences used for the scans in SKI10 are very different. This makes the scans less homogeneous than most collections of knee MRI. Due to the resulting variation in scan appearance, the intensity was naïvely normalized by division with the intensity mean.

Secondly, we sub-divided the manual cartilage compartments for the training scans. The medial and lateral tibial compartments were marked with four points located at the internal/external/anterior/posterior of the tibial plateau in the compartments. These four points defined a hyper-ellipse that was used to define medial/lateral tibial sub-compartments in the manual training segmentations. The femoral training segmentations were divided by marking a plane separating medial and lateral sub-compartments approximately at the trochlea (the intercondylar groove).

Finally, we labeled the scans by the laterality (left/right).

Segmentation

The segmentation framework using multi-atlas registration and supervised classification was trained without modification using the 60 training cases without evaluation ROIs. The masks for anatomical structures were defined by the provided SKI10 masks for Tibia and Femur bones directly, and the pre-processed mask for medial/lateral tibial/femoral cartilages.

The training segmentation method was then tested on the 40 training cases with evaluation ROIs.

Finally, the 50 evaluation cases were segmented.

Following segmentation, the resulting cartilage sub-compartments were trivially merged into one tibial and one femoral compartment, for each case.

Evaluation Metrics

The optimization of the voxel classification method uses Dice volume overlap as the objective function, where Dice overlap for two sets A and B is defined as $Dice(A,B) = 2*|A \cap B|/(|A|+|B|)$. The SKI10 evaluation uses the volume overlap error defined as $VOE(A,B) = 100*(1 - |A \cap B|/|A \cup B|)$. Dice volume overlap and volume overlap error are directly related: $VOE = 100*(1-Dice/(2-Dice))$.

Thereby, we may expect that the method will perform slightly better for the SKI volume overlap error performance metric (used for the cartilage segmentation evaluation) than the metrics related to surface distances (used for the bone segmentation evaluation).

3. Results

We firstly recorded the training performance using the Dice volume overlap resulting from the segmentation framework using the cartilage sub-compartments. For the two bone compartments, the Dice scores were 0.97. For the cartilage compartments, the Dice scores were 0.64-0.73.

Using the same set of segmentations, the training set with evaluation ROIs was then scored using the SKI10 evaluation metrics. These results we are given in Table 1.

For the official challenge results on the 50 evaluation cases, see www.ski10.org.

4. Discussion

The volume overlaps on the training set (Dice scores at 0.97 for bone and 0.64-0.73 for cartilages) are lower than what we have seen on other collections, particularly for the cartilages. Most often, we see Dice scores around 0.8 for the cartilages. There are three likely reasons for this:

- The scan quality is challenging for some of the scans.
- The intensity ranges vary a lot between the scans. We applied a simple normalization where we divided by the mean intensity. In other collections with homogeneous scans, we have seen that this normalization degrades segmentation performance. Due to the inhomogeneity of the scans here, the normalization improves performance, but it is still quite naïve and a more robust normalization would likely improve performance further.
- Our annotations of the tibial and femoral cartilage sub-compartments are not ideal and likely include a lot of variation between the scans. This challenges the optimization and has an impact on training performance. It is unclear what the impact on the evaluation performance will be.

However, the results on the training set with evaluation ROIs generally compares quite well with most methods that are currently evaluated on the SKI10 website. Visual inspection shows that the typical areas for large segmentation errors are the tibial and femoral shafts toward the boundary of the scans – particularly for cases with poor image quality in these regions. Most likely, a shape model would allow more robust segmentation for these cases.

5. Acknowledgements

We gratefully acknowledge financial support from the Danish Research Foundation ("Den Danske Forskningsfond").

The research leading to these results has received funding from the D-BOARD consortium, a European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 305815.

6. References

1. “The Osteoarthritis Initiative (OAI),” Publicly available at www.oai.ucsf.edu, 2012, <www.oai.ucsf.edu>.
2. J. Folkesson et al., “Automatic quantification of local and global articular cartilage surface curvature: biomarkers for osteoarthritis?,” Magn Reson. **59**(6), 1340–1346 (2008) [doi:10.1002/mrm.21560].
3. S. Tummala et al., “Automatic Quantification of Tibio-Femoral Contact Area and Congruity,” IEEE Trans. Med. Imaging **31**(7), 1404–1412 (2012) [doi:10.1109/TMI.2012.2191813].
4. T. Heimann et al., “Segmentation of Knee Images: A Grand Challenge,” in Proc. MICCAI Workshop on Medical Image Analysis for the Clinic, pp. 207–214 (2010).
5. G. Vincent et al., *Fully Automatic Segmentation of the Knee Joint using Active Appearance Models*, http://www.ski10.org/data/2011-01-27-1131.pdf; published January 27, 2011; downloaded March 14, 2011. (2011).
6. H. Seim et al., *Model-based Auto-Segmentation of Knee Bones and Cartilage in MRI Data*, http://www.ski10.org/data/2010-11-09-1131.pdf; published November 9, 2010; downloaded March 14, 2011 (2010).
7. M. Holden, “A Review of Geometric Transformations for Nonrigid Body Registration,” IEEE Trans. Med. Imaging **27**(1), 111–128 (2008).
8. J.-G. Lee et al., “Fully-automated Segmentation of Cartilage from the MR Images of Knee Using a Multi-atlas and Local Structural Analysis Method,” in RSNA Abstract Book **SSG13-07** (2013).
9. L. Shan et al., “Automatic atlas-based three-label cartilage segmentation from MR knee images,” Med. Image Anal. **18**(7), 1233–1246 (2014) [doi:j.media.2014.05.008].
10. J. Folkesson et al., “Segmenting Articular Cartilage Automatically Using a Voxel Classification Approach,” IEEE Transon Med. Imaging **26**(1), 106–115 (2007).
11. E. B. Dam et al., “Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative,” J. Med. Imaging **2**(2), 024001–024001 (2015) [doi:10.1117/1.JMI.2.2.024001].

Table 1: Segmentation accuracy for the training set with 40 cases including evaluation ROIs. Accuracy is given using SKI10 evaluation metrics: average surface distance (AvgD), root mean squared distance (RMSD), volume overlap error (VOE), volume difference (VD), scores as defined at www.ski10.org.

	Femur Bone		Tibia Bone		Femur cartilage		Tibia cartilage		
Test case	AvgD [mm]	RMSD [mm]	AvgD [mm]	RMSD [mm]	VOE [%]	VD [%]	VOE [%]	VD [%]	Total Scr
1	0.65	1.11	1.33	2.38	34.4	-1.7	28.63	5.40	58.7
2	0.62	0.98	0.44	0.80	26.8	-7.5	25.06	-8.88	70.8
3	0.63	1.19	0.37	0.65	22.2	-18.8	16.73	-6.94	67.3
4	1.02	1.90	0.41	0.58	33.4	38.9	30.07	18.17	52.8
5	0.60	1.06	0.52	0.98	21.2	9.4	20.31	0.51	73.3
6	0.56	0.98	0.36	0.65	21.6	-8.3	13.69	4.42	76.1
7	0.56	1.12	0.41	0.74	30.9	-7.7	34.61	6.31	70.9
8	0.62	1.02	0.56	0.97	21.4	5.1	25.51	5.11	72.7
9	0.40	0.71	0.42	0.88	21.8	15.9	15.51	-11.79	68.0
10	0.47	0.88	0.37	0.63	19.0	4.8	19.74	-13.20	74.4
11	0.41	0.67	0.36	0.66	21.9	7.9	31.09	16.73	70.2
12	0.78	1.49	0.60	1.15	34.2	-11.4	36.13	-14.08	58.2
13	1.32	2.23	0.55	1.03	48.0	45.9	28.99	-18.92	44.5
14	0.62	1.13	0.47	0.79	30.4	-12.9	29.23	-6.18	67.4
15	0.56	0.94	0.48	0.83	25.4	4.9	23.75	-20.33	67.0
16	0.49	1.06	0.40	0.75	28.0	-11.3	17.55	-5.07	72.3
17	0.84	1.57	0.87	1.80	26.1	-4.5	29.35	4.63	62.4
18	0.91	1.88	0.42	0.76	24.8	-7.5	21.96	-7.68	67.7
19	0.41	0.76	0.36	0.68	18.7	1.4	18.21	-3.56	81.9
20	1.03	1.70	0.54	0.95	27.6	15.9	23.58	21.63	53.4
21	0.55	1.08	0.28	0.56	24.6	-1.0	30.41	23.04	69.8
22	0.88	1.64	0.52	0.92	25.0	2.5	20.70	1.91	72.9
23	0.74	1.39	0.35	0.62	21.7	0.5	20.22	-5.17	77.4
24	0.60	1.18	0.50	1.08	26.9	-16.1	26.21	22.13	56.6
25	0.78	1.47	0.42	0.70	25.5	3.8	20.91	19.38	66.5
26	0.66	1.09	0.49	0.85	34.3	9.3	32.85	-5.79	68.3
27	0.48	0.83	0.60	1.15	22.7	-14.9	23.36	-10.73	64.5
28	0.42	0.84	0.28	0.47	28.5	6.7	15.86	-7.36	77.4
29	1.18	1.96	0.51	0.89	29.4	6.1	21.77	-0.67	68.3
30	0.45	0.86	0.50	0.81	26.0	-3.6	42.29	39.28	64.1
31	1.04	1.96	0.63	1.08	38.2	13.0	30.41	5.16	58.8
32	0.60	1.10	0.52	1.02	29.0	6.1	28.48	-5.65	70.6
33	0.78	1.32	0.77	1.29	29.5	-3.8	27.50	-14.38	62.4
34	0.55	0.99	0.41	0.71	21.6	12.0	19.25	6.38	71.8
35	0.90	1.68	0.69	1.46	30.8	-11.8	36.64	-23.40	50.0
36	0.57	1.03	0.60	1.11	22.3	4.3	16.58	-10.51	70.6
37	0.54	0.89	0.42	0.69	23.2	-4.9	29.22	2.78	76.8
38	0.61	1.42	0.47	0.93	16.9	-4.8	16.10	5.85	73.7
39	0.98	1.92	0.37	0.64	31.3	-14.8	23.83	-0.72	66.4
40	0.48	0.82	0.36	0.69	31.9	-14.7	31.51	8.48	68.1
Avg	0.68	1.25	0.50	0.91	26.9	0.8	25.10	0.41	67.1