Segmentation of Knee Images: A Grand Challenge

Tobias Heimann[1], Bryan J. Morrison[2], Martin A. Styner[3,4],
Marc Niethammer[4], and Simon K. Warfield[5]

[1] Div. Medical and Biological Informatics, German Cancer Research Center, Heidelberg, Germany
[2] Signature Imaging, Biomet Inc., Warsaw, IN
[3] Dept. of Psychiatry, University of North Carolina at Chapel Hill, NC
[4] Dept. of Computer Science, University of North Carolina at Chapel Hill, NC
[5] Dept. of Radiology, Children's Hospital, Boston, MA

Abstract. In this paper, we present an evaluation framework for the 3D segmentation of knee bones and cartilage from magnetic resonance images. The framework was established for one of the three challenges at the "Medical Image Analysis for the Clinic: A Grand Challenge" workshop held at the 2010 Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in Beijing, China. After this workshop, the framework will remain open to online submissions via `www.ski10.org`. We describe the motivation for this challenge, the preparation of training and test datasets, and the evaluation measures used to rate submitted results.

1 Introduction

Musculoskeletal diseases and articular disorders are one of the major health problems in developed countries and affect especially the aging population. The human knee joint is commonly affected by osteoarthritis (OA), a degenerative disease that is the primary cause of chronic disability in the United States [1]. OA leads to loss of articular cartilage, an effect that by now can be well-observed using magnetic resonance imaging (MRI) [2]. With this background, the segmentation of knee cartilage and the surrounding bones is a problem which has gained considerable importance in recent years. A major direction of research is to use cartilage segmentations for the development of biomarkers targeted at different stages of OA [3]. Moreover, segmentations of bones and cartilage are required for computer-based surgical planning of knee implants. Other applications include modeling of the knee by finite elements to predict joint kinematics [4] or the understanding of natural variation and physiological effects for healthy joints [5].

As usual in the field of medical image analysis, it is difficult to assess the suitability of published segmentation algorithms for the wide variety of images used in clinical practice. This is mainly due to the fact that every author evaluates his new algorithm on a different set of test images, often using different measures of accuracy. The “Grand Challenge” series of workshops, initiated at the 2007 Medical Image Computing and Computer Assisted Intervention (MICCAI) conference and since then repeated regularly, strives to remedy this problem. Each workshop consists of several challenges which provide open image databases and standardized evaluation procedures for a specific application. Applications that have been topic of challenges include liver segmentation [6], coronary artery centerline extraction [7], pulmonary nodule detection [8], and detection of microaneurysms in fundus photographs [9]. A constantly updated list of all challenges in medical image analysis is available at `http://www.grand-challenge.org`.

This year, one of the featured challenges is the segmentation of knee images, named SKI10 (`www.ski10.org`). Related to this challenge is the Osteoarthritis Initiative (OAI) [6], a large study of OA in the United States that aims to develop a public domain research resource for the evaluation of biomarkers. Like SKI10, OAI offers a large variety of knee images, unfortunately it does not (yet) include the reference segmentations of bones and cartilage that are required for an evaluation framework in the “Grand Challenge” tradition.

2 Data

2.1 Image data

Basis for the SKI10 challenge are 250 knee MRI images originating from the surgical planning program of Biomet, Inc. Cases of left and right knees are distributed approximately equally. The data was acquired at over 80 different centers in the USA, using machines from all major vendors, i.e. General Electric, Siemens, Philips, Toshiba, and Hitachi. All images were acquired in the sagittal plane with a pixel spacing of 0.4×0.4mm and a slice distance of 1mm. No contrast agents were used. Field strength was 1.5T in about 90% of the cases, the rest was acquired mostly at 3T, with some images acquired at 1T. The employed MRI sequences show a huge variety: the vast majority of images used T1-weighting, but some were also acquired with T2-weighting. Many images used gradient echo or spoiled gradient echo sequences, and fat suppression techniques were common as well.

After acquisition, all images were segmented interactively by experts at Biomet, Inc. Structures of interest were femur, femoral cartilage, tibia, and tibial cartilage. Images were processed slice by slice (sagittally), starting with a variable threshold to mark bones and cartilage. Subsequently, manual editing was used to clean up the resulting masks. As all images were used for surgery planning of partial or complete knee replacement, segmentations were created for a specific clinical goal, i.e. accuracy of contours is varying. On one side, in areas where implant position guides should be placed, accuracy of bone and cartilage segmentations is very high. On the other side, e.g. the exact boundary of the cartilage to the sides is not relevant for the planned surgery and may thus be quite far off the true location.

[6] `http://oai.epi-ucsf.org`

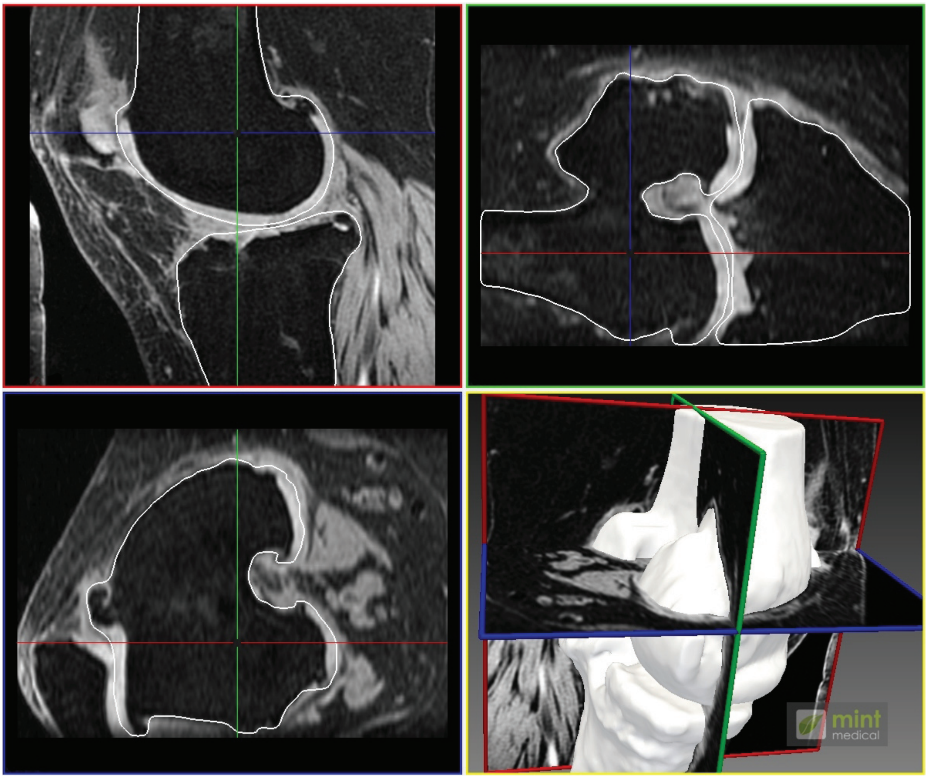

Fig. 1: One of the MRI knee images used in the SKI10 challenge, shown in different planes: sagittal (upper left), coronal (upper right), and transversal (lower left). Contours of the reference segmentations for combined bone and cartilage are displayed in white. The lower right window shows a 3D rendering of the scene.

All segmentations were exported as surfaces directly from the segmentation workstation. Technically, it was only possible to export combined bone and cartilage as one surface, and bone only as another surface. Before making available the 250 images, all patient information that could lead to identification was deleted from the DICOM headers. An example for a knee MRI is shown in Fig. 1.

2.2 Processing

The first task was to convert the combined surface segmentations to individual label masks of femur, femoral cartilage, tibia, and tibial cartilage. For the femur, the combined bone and cartilage segmentation was scan-converted to label mask f_{comb} and the bone segmentation was scan-converted to f_{bone}. After morphological closing of f_{bone}, the cartilage mask f_{cart} was generated by $f_{cart} = f_{comb} - f_{bone}$. To remove spurious cartilage voxels along the sides of the long bone that appeared due to small differences between the two surface meshes, a connected-component filter was run on f_{cart}, and components with less than 500 voxels (i.e. 80ml) were deleted. The same steps were taken for the tibia, leading to t_{bone} and t_{cart}. All individual masks were combined to a multi-label image with the following values: 0=background, 1=femur, 2=femoral cartilage, 3=tibia, 4=tibial cartilage.

The second task was to crop the images to a suitable region of interest around the joint area. Although many images extended a bit further proximally and distally from the joint, the available bone segmentations were limited to an area of less than 15cm around the joint. To save bandwidth when offering the data over internet, all images were cropped proximally and distally according to the extent of their respective bone segmentations. To the anterior and posterior sides, images were cropped to leave 50 ± 5 voxels background before reaching the closest bone. Finally, all images and multi-label masks were stored in a raw file format, stripping all DICOM information except image size and voxel spacing.

2.3 Organization

After processing, all datasets were checked for their suitability in a segmentation challenge. As mentioned above, the available segmentations were created for a specific clinical goal, and high accuracy was required only in certain areas. For an evaluation of segmentation algorithms as SKI10, however, reasonably accurate segmentations should be expected everywhere. With this goal, we selected the 100 most accurate segmentations from the set of 250 images. These 100 datasets were randomly divided into a training set of 60 images and a test set of 40 images.

After participants agreed to the rules for data usage, they could download the training set including corresponding multi-label masks as reference segmentations. This data can be used to train segmentation algorithms, but participants were also free to use their own training data instead, or in addition. No additional information as laterality or acquisition protocol was supplied. Registered participants were also given access to the test data, which consisted of the selected images only (without segmentations). Before a given deadline, each participating team had to submit automatically generated multi-label masks for these images.

3 Evaluation

All submitted results were evaluated by the workshop organizers with the same procedure, comparing the submitted segmentations to the hidden references.

3.1 Accuracy of bone segmentation

One aspect was the accuracy of bone segmentation, more exactly the combined bone and cartilage segmentation. Both structures together form the outer surface of the bone, which is of paramount importance for surgery planning. We

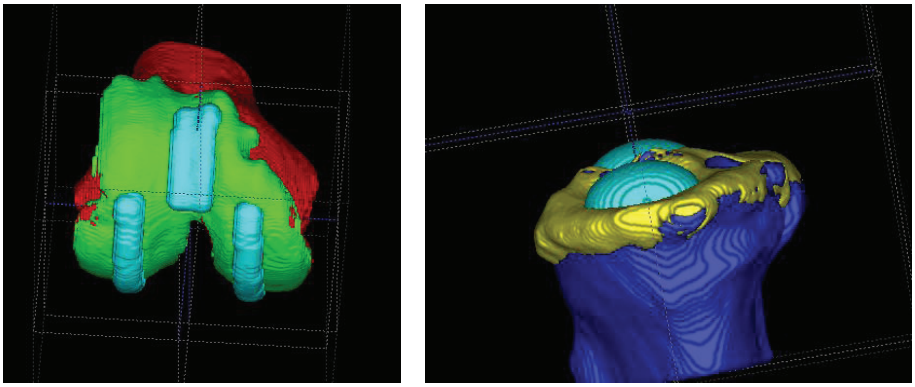

Fig. 2: Regions of interest for the evaluation of femoral (left) and tibial (right) cartilage.

evaluated the accuracy of this segmentation by two metrics: average symmetric surface distance (AvgD) and root-mean-square symmetric surface distance (RMSD). For these metrics, the boundary voxels of segmentation and reference have to be determined. To account for the anisotropic voxel spacing, segmentations were resampled first, doubling the number of slices (and leading to an approximately isotropic spacing of $0.4 \times 0.4 \times 0.5$mm). In the resampled segmentation, boundary voxels are defined as voxels which have at least one non-object voxel in their 18-neighbourhood. For each boundary voxel in the segmentation, the closest boundary voxel in the reference is searched and the corresponding Euclidean distance (in mm) is stored in a list. Subsequently, the same procedure is repeated for all boundary voxels in the reference, searching the segmentation. The average of all stored values is the average symmetric surface distance.

The root-mean-square symmetric surface distance is calculated similarly, except that all distances are squared before storing in the list, and the root is taken from the average value afterwards. In this latter metric, large deviations from the reference are penalized stronger, which corresponds to the paradigm that small segmentation errors are tolerable, while larger ones have to be avoided. Both distance measures are given in millimeters and reach 0 for a perfect segmentation.

3.2 Accuracy of cartilage segmentation

The second aspect is the accuracy of cartilage segmentation, which is of highest interest for applications in computer-aided diagnosis. Here, volume and thickness of cartilage are the most important parameters. As mentioned in Sec. 2.1, cartilage boundaries to the side were not always accurate in the reference segmentations. For this reason, we defined specific central regions of interest (ROIs), in which the segmentations are compared to their respective references.

The ROI for tibial cartilage consists of two elliptical areas around the contact surfaces to the femur. It is extracted automatically from the reference segmentation by the following procedure. First, the transversal slice with the largest area of tibia is determined, and the corresponding area is forwarded to a principal component analysis. The largest eigenvector defines the lateral axis of the knee joint and is used to determine the width of the tibia (taking into account the 3D geometry around the selected slice). On the line defined by this axis, condyles are expected at 25% and 75% of the total tibial width. Both condyles are projected upwards until the last voxels marked as tibia or tibial cartilage. At these two locations, two oriented ellipsoids of radius $15 \times 10 \times 10$mm are created, with the largest radius pointing along the anteroposterior axis. An example ROI for tibial cartilage is shown in Fig. 2 right.

The ROI for femoral cartilage consists of three areas: two at the condyles around the contact surfaces to the tibia, and one in the center around the contact surface to the patella. As contact surfaces move with varying flexion angle of the knee joint, all ROIs are elongated structures which are extracted automatically as follows. From the multi-label mask, all femoral cartilage voxels are selected which are within a certain distance from planes perpendicular to the lateral joint axis, located at the two estimated condyle positions and at the center. The required distance is 1.5mm for the condyle planes and 5mm for the central plane. Subsequently, selected voxels are cropped by a fixed-size bounding box positioned around the condyle centers. Different boxes are used for condyles and center plane. The remaining voxels are dilated by an ellipsoidal kernel of $15 \times 15 \times 5$ voxels, with the smaller side in the lateral direction (which also features a larger spacing). An example ROI for femoral cartilage is shown in Fig. 2 left.

Inside their respective ROIs, femoral and tibial cartilages are evaluated according to volumetric overlap error VOE and volumetric difference VD. The volumetric overlap error between the set of voxels of the segmentation S and the one of the reference R is given as $\text{VOE} = 100\left(1 - \frac{|S \cap R|}{|S \cup R|}\right)$. It is measured in percent and yields 0% for a perfect segmentation and 100% for no overlap at all. The volumetric difference between S and R is defined as $\text{VD} = 100\frac{|S|-|R|}{|R|}$. Since this measure is not symmetric, it is no metric. A value of 0% means that both volumes are identical. Please note that this does not imply that segmentation and reference are identical, they do not even need to overlap. In our case however, as cartilage segmentations are limited to the sides by their respective ROIs, cartilage volume is proportional to the average thickness. Thus, the VD indicates (at least approximately) the deviation from the average cartilage thickness.

3.3 Scoring and Presentation

To allow a ranking between different methods for automatic segmentation, the different measures for the different structures of interest have to be combined to one single value. We employ a scoring technique based on inter-observer variation for this purpose [6]. The principal idea is to take an independent second rater

Structure	Value 1	Value 2
Femur bone	AvgD = 0.45mm	RMSD = 0.77mm
Tibia bone	AvgD = 0.37mm	RMSD = 0.62mm
Femoral cartilage	VOE = 34.2%	VD = 7.1%
Tibial cartilage	VOE = 34.2%	VD = 7.1%

Table 1: Reference values for the accuracy of an independent second rater segmentation.

who manually outlines the respective structures and to compare his result to the hidden reference. On a range from 0 to 100 points, where 100 corresponds to a perfect segmentation, the second rater's outcome for each evaluation measure corresponds to 75 points. An algorithm that produces a result with an error twice as high is awarded 50 points, an error three times as high corresponds to 25 points. To prevent that a single unsuccessful segmentation ruins the entire score, there are no negative points, i.e. errors more than four times as high as the human observer's still correspond to 0 points. The scores for each metric are averaged per image, which results in a total score per image. Finally, the average score over all images is the single value on which all algorithms will be ranked. The second observer values used for score calculations are shown in Table 1.

To make the results from different teams easier to compare, the workshop organizers generated the same results table and figures for each submission. These tables and figures can be found in the results section of each SKI10 article.

4 Conclusions

In the weeks after the call for papers for the workshop was launched, 22 teams registered for the knee segmentation challenge and downloaded training and test datasets. Of these, 9 were from North America, 8 from Europe, 4 from Asia, and 1 from Australia. Finally, 6 teams submitted segmentation results for the supplied test images. Although the call was also open for interactive segmentation methods, all submitted results were generated by fully automatic approaches. This is probably due to the high number of test images and the complexity of segmenting four different structures of interest, which makes an interactive approach very time-consuming. After they received their respective evaluation results, all 6 teams submitted articles describing their methods and could be accepted to the workshop.

At the workshop in Beijing, we will organize an onsite competition with 10 new knee MRIs. Until then, teams can still work on improving their approaches. For further training, we have received an additional 250 knee MRIs, which will be processed as described in Sec. 2.2 and made available to participants several weeks before the workshop.

After the workshop, the website `www.ski10.org` will be extended to allow an online submission of results. This will allow interested researchers to test their methods against the collection of existing works described in these proceedings. From our experience with past challenges, we expect this website to grow over time and become a reference for cartilage and bone segmentation for knee MRIs. The large collection of images used in the SKI10 challenge shows the huge heterogeneity of clinical data for this application. Segmenting the different structures of interest from this data automatically is an extremely challenging task, but one that we have to face if we want to bring our methods into clinical practice.

Acknowledgements

We would like to thank Biomet, Inc. for supplying all dasets and segmentations used for the knee segmentation challenge. Without this data, this extensive comparison of methods would not have been possible. Also thanks to Ashley Rumple and Abeer Hadidi for providing additional segmentations in order to estimate inter-rater variability. Moreover, we would like to thank Siemens Healthcare, Philips Healthcare, and Chili GmbH for their financial support for the workshop. All images in this article were produced using MITK (`www.mitk.org`) and ITK Snap (`www.itksnap.org`).

References

1. CDC: Prevalence of disabilities and associated health conditions among adults – United States, 1999. MMWR Morb Mortal Wkly Rep **50** (2001)
2. Eckstein, F., Cicuttini, F., Raynauld, J.P., Waterton, J.C., Peterfy, C.: Magnetic resonance imaging (MRI) of articular cartilage in knee osteoarthritis (OA): morphological assessment. Osteoarthr Cartilage **14** (2006)
3. Folkesson, J., Dam, E.B., Olsen, O.F., Karsdal, M.A., Pettersen, P.C., Christiansen, C.: Automatic quantification of local and global articular cartilage surface curvatures: biomarkers for osteoarthritis? Magn Reson Med **59** (2008) 1340–1346
4. Baldwin, M.A., Langenderfer, J.E., Rullkoetter, P.J., Laz, P.J.: Development of subject-specific and statistical shape models of the knee using an efficient segmentation and mesh-morphing approach. Comp Meth Prog Bio **97** (2010) 232–240
5. Fripp, J., Crozier, S., Warfield, S.K., Ourselin, S.: Automatic segmentation and quantitative analysis of the articular cartilages from magnetic resonance images of the knee. IEEE Trans Med Imaging **29** (2010)
6. Heimann, T., van Ginneken, B., Styner, M.A., et al.: Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging **28** (2009) 1251–1265
7. Schaap, M., Metz, C.T., van Walsum, T., et al.: Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Med Image Anal **13** (2009) 701–714
8. van Ginneken, B., Armato, S.G., de Hoop, B., et al.: Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study. Med Image Anal **14** (2010) 707–722
9. Niemeijer, M., van Ginneken, B., Cree, M.J., et al.: Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans Med Imaging **29** (2010)