Subject-level Prediction of Segmentation Failure using Real-Time Convolutional Neural Nets

Robert Robinson[1], Ozan Oktay[1], Wenjia Bai[1], Vanya V. Valindria[1], Mihir M. Sanghvi[3,4], Nay Aung[3,4], José Miguel Paiva[3], Filip Zemrak[3,4], Kenneth Fung[3,4], Elena Lukaschuk[5], Aaron M. Lee[3,4], Valentina Carapella[5], Young Jin Kim[5,6], Bernhard Kainz[1], Stefan K. Piechnik[5], Stefan Neubauer[5], Steffen E. Petersen[3,4], Chris Page[2], Daniel Rueckert[1], and Ben Glocker[1]

[1]Biomedical Image Analysis Group, Imperial College London, London, UK
[2]Research & Development, GlaxoSmithKline, UK
[3]William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, UK
[4]Barts Heart Centre, Barts Health NHS Trust, London, UK
[5]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, UK
[6]Department of Radiology, Severance Hospital, Yonsei University College of Medicine, South Korea
r.robinson16@imperial.ac.uk

Abstract

Occasionally even the best automated method fails due to low image quality, artifacts or unexpected behaviour of black box algorithms. Being able to predict segmentation quality in the absence of ground truth is of paramount importance in clinical practice, but also in large-scale studies to avoid the inclusion of invalid data in subsequent analysis.

In this work, we propose two approaches of real-time automated quality control for cardiovascular MR segmentations using deep learning. First, we train a neural network on 12,880 samples to predict Dice Similarity Coefficients (DSC) on a per-case basis. We report a mean average error (MAE) of 0.03 on 1,610 test samples and 97% binary classification accuracy for separating low and high quality segmentations. Secondly, in the scenario where no manually annotated data is available, we train a network to predict DSC scores from estimated quality obtained via a reverse testing strategy. We report an $\mathrm{MAE} = 0.14$ and 91% binary classification accuracy for this case. Predictions are obtained in real-time which, when combined with real-time segmentation methods, enables instant feedback on whether an acquired scan is analysable while the patient is still in the scanner.

Introduction. Finding out that an acquired medical image is not usable for the intended purpose is not only costly but can be critical if image-derived quantitative measures should have supported clinical decisions in diagnosis and treatment. Real-time assessment of the downstream analysis task, such as image segmentation, is highly desired. Ideally, such an assessment could be performed while the patient is still in the scanner, so that in the case an image is not analysable, a new scan could be obtained immediately (even automatically). Such a real-time assessment requires two components, a real-time analysis method and a real-time prediction of the quality of the analysis result. This work proposes a solution to the latter with a particular focus on image segmentation as the analysis task. Specifically, we assess the quality of automatically generated segmentations of cardiovascular MR (CMR) from the UK Biobank (UKBB) Imaging Study [1].

Methods In this work we perform 2 experiments with data obtained from the UKBB. We take the Dice Similarity Coefficient (DSC) to be a measure of segmentation quality.

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

Data: Our initial dataset consists of 4,882 3D (2D-stacks) end-diastolic (ED) cardiovascular magnetic resonance (CMR) scans from the UK Biobank (UKBB) Imaging Study. All images have a manual segmentation which is unprecedented at this scale. We take these labelmaps as reference GT. Each labelmap contains 3 classes: left-ventricular cavity (LVC), left-ventricular myocardium (LVM) and right-ventricular cavity (RVC) which are separate from the background class (BG). In this work, we also consider the segmentation as a single binary entity comprising all classes: whole-heart (WH). After re-segmenting with a random forest, our dataset comprises 16,100 score-balanced segmentations with reference GT. From each segmentation we create 4 one-hot-encoded masks: masks 1 to 4 correspond to the classes BG, LVC, LVM and RVC respectively. At training time, our data-generator re-samples the UKBB images and our segmentations to have consistent shape of $[224, 224, 8, 5]$ making our network fully 3D with 5 data channels: the image and 4 segmentation masks. The images are normalized such that the entire dataset falls in the range $[0.0, 1.0]$. For comparison and consistency, we choose to use the same input data and network architecture for each of our experiments. We employ a 50-layer 3D residual network written in Python with the Keras library and trained on an 11GB Nvidia GeForce GTX 1080 Ti GPU. We use the Adam optimizer with learning rate of $1e^{-5}$ and decay of 0.005. Batch sizes are kept constant at 46 samples per batch. We run validation at the end of each epoch for model-selection purposes.

Experiment 1: Directly predicting DSC. Is it possible to directly predict the quality of a segmentation given only the image-segmentation pair? In this experiment we calculate, for each class, the DSC between our segmentations and the GT. These are used as training labels. We have 5 nodes in the final layer of the network where the output X is $\{X \in \mathbb{R}^5 \,|X \in [0.0, 1.0]\}$. This vector represents the DSC for each of the classes including background and whole-heart. We use mean-squared-error loss and also report mean-absolute-error between the output and GT DSC. We split our data 80:20:20 giving 12,880 training samples and 1,610 samples for each of the validation and test sets. Performing this experiment is costly as it requires the use of a large manually-labelled dataset which is not readily available in practice.

Experiment 2: Predicting RCA scores. Considering the promising results of the RCA framework [2, 3] in accurately predicting the quality of segmentations in the absence of large labelled datasets, can we use the predictions from RCA as training data to allow a network to give comparatively accurate predictions on a test-set? In this experiment, we perform RCA on all 16,100 segmentations. To ensure that we train on balanced scores, we perform histogram binning on the RCA scores and take equal numbers from each class. We finish with a total of 5363 samples split into training, validation and test sets of 4787, 228 and 228 respectively. The predictions per-class are used as labels during training. Similar to Experiment 1, we obtain a single predicted DSC output for each class using the same network and hyper-parameters, but without the need for the large, often-unobtainable manually-labelled training set.

Results The results from Experiment 1 in Table 1 show that our network is able to directly predict whole-heart DSC from the image-segmentation pair with MAE of 0.03 (SD = 0.04). We see similar performance on individual classes. For WH we report that 72% of the data have MAE less than 0.05 with outliers ($\mathrm{DSC} \geq 0.12$) comprising only 6% of the data. Results show excellent true-positive (TPR) and false-positive rates (FPR) on a whole-heart binary classification task with DSC threshold of 0.70. The reported accuracy of 97% is better than the 95% reported with RCA in [3]. Distributions of the MAEs for each class can be seen in Fig 1. Our results for Experiment 2 are also recorded in Table 1. It should be expected that direct predictions of DSC from the RCA labels are less accurate than in Experiment 1. The reasoning here is two-fold: first, the RCA labels are themselves predictions and retain inherent uncertainty and second, the training set here is much smaller than in Experiment 1. However, we report MAE of 0.14 (SD = 0.09) for the WH case and 91% accuracy on the binary classification task.

Conclusion and Discussion We recognize that our networks are prone to learning features specific to assessing the quality of random forest segmentations. We can build on this by training the network with segmentations generated from an ensemble of methods. However, we must reiterate that the purpose of the framework in this study is to give an indication of the *predicted quality* and not a direct one-to-one mapping to the reference DSC. Currently, these networks will correctly predict whether a segmentation is 'good' or 'poor' on some threshold, but will not confidently distinguish between two segmentations of similar quality.

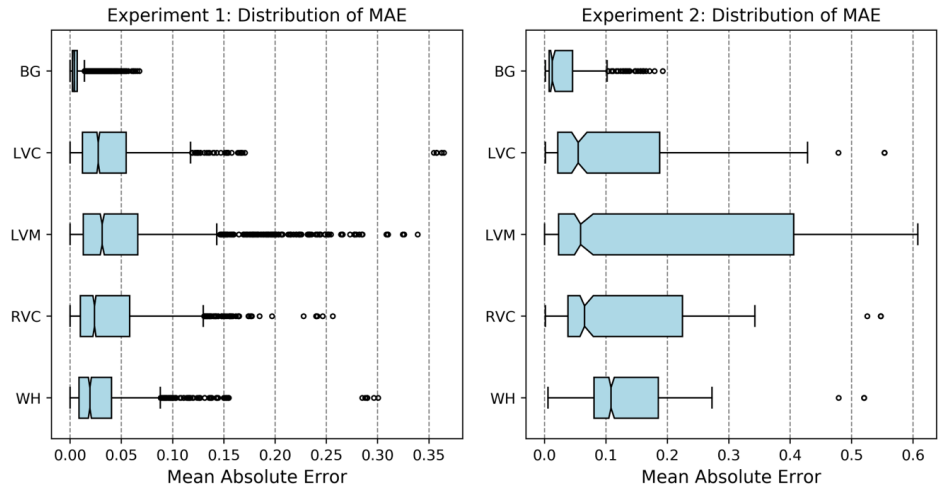

Figure 1: Distribution of the mean absolute errors (MAE) for Experiments 1 (left) and 2 (right). Results are shown for each class: background (BG), left-ventricular cavity (LV), left-ventricular myocardium (LVM), right-ventricular cavity (RVC) and for the whole-heart (WH).

Table 1: Mean absolute error (MAE) for on predicted segmentation quality over individual classes and whole-heart (WH). Standard deviations in brackets

Class	**Mean Absolute Error**	
	Experiment 1	Experiment 2
BG	0.008 (0.011)	0.034 (0.042)
LV	0.038 (0.040)	0.120 (0.128)
LVM	0.055 (0.064)	0.191 (0.218)
RVC	0.039 (0.041)	0.127 (0.126)
WH	**0.031 (0.035)**	**0.139 (0.091)**

The labels for training the network in Experiment 1 are not easily available in most cases. However, by performing RCA, one can automatically obtain training labels for the network in Experiment 2 and this could be applied to segmentations generated with other algorithms. The cost of using data obtained with RCA is an increase in MAE. This is reasonable compared to the effort required to obtain a large, manually-labeled dataset.

On average, the inference time for each network was of the order 600 ms on CPU and 40 ms on GPU. This is 14,000 times faster than with RCA (660 seconds) whilst maintaining good accuracy. In an automated image analysis pipeline, this method would deliver excellent performance at high-speed and at large-scale. When paired with a real-time segmentation method it would be possible to provide real-time feedback during image acquisition whether an acquired image is of sufficient quality for the downstream segmentation task.

References

[1] Steffen E. Petersen et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in caucasians from the UK biobank population cohort. *Journal of Cardiovascular Magnetic Resonance*, 19(1), 2017.

[2] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. *IEEE Transactions on Medical Imaging*, pages 1–1, 2017.

[3] Robert Robinson et al. Automatic quality control of cardiac mri segmentation in large-scale population imaging. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 720–727, Cham, 2017. Springer International Publishing.