Automatic hippocampal subfield segmentation in ultra-high resolution MRI using Convolutional Neural Network

- Article

Groupnr: 17gr10407
Group members: Anne Krogh Nøhr,
Bolette Dybkjær Hansen, Nina Jacobsen

Master's Thesis
Biomedical Engineering and Informatics

AALBORG UNIVERSITY
DENMARK

frontiers

Automatic hippocampal subfield segmentation in ultra-high resolution MRI using Convolutional Neural Network

Nina Jacobsen [1*] **Anne Krogh Nøhr** [1] **Bolette Dybkjær Hansen** [1]

[1] *Department of Health Science and Technology, Aalborg University, Aalborg, Denmark*

All authors contributed equally to this work

Correspondence*:
Nina Jacobsen
njacob12@student.aau.dk

Keywords: Segmentation, deep learning, convolutional neural network, hippocampal subfield, ultra-high resolution, MRI.

ABSTRACT

Objective: The hippocampal subfields are of great interest within research due to a diagnostic connectivity to neurodegenerative diseases, and increased access to ultra-high field MR imaging has made segmentation of these subfields feasible. Automatic methods for segmentation of the hippocampal subfields have been proposed. However, these methods show limitations with respect to segmentation of smaller subfields and they are very time consuming. In current research, Convolutional Neural Networks (CNN) have been used to segment brain structures and lesions, and the approach have shown fast and accurate segmentations. The aim of this study was to develop a CNN based automatic method for hippocampal subfield segmentation (DeepHSS) and explore the impact of a small dataset.

Method: DeepHSS consists of a preprocessing pipeline followed by a 12 layer, two pathways 3D CNN trained using dense training. The training set includes 7T TSE MR images acquired from 15 subjects and associated manual labels delineating six hippocampal subfields and Entorhinal Cortex. DeepHSS was tested with 10 subjects. The segmentations by DeepHSS were compared to the corresponding manual labels using the Dice Similarity Coefficient (DSC).

Results: DeepHSS showed fast segmentation of hippocampal subfields with an average foreground DSC at 0,83$\pm$0.03. Highest average segmentation DSC was achieved for Subiculum (0.80$\pm$0.05) and Dendate Gyrus (0.77$\pm$0.05) whilst the lowest average segmentation accuracy was found for Entorhinal Cortex (0.49$\pm$0.14). The accuracy of the segmentation increased with the number of training subjects.

Conclusion: Using our automatic hippocampal subfield segmentation method, DeepHSS, we demonstrated CNNs as an efficient method for automatic hippocampal subfield segmentation despite utilisation of a small dataset. The results were comparable with results obtained using ASHS.

1 INTRODUCTION

The human hippocampal formation is an anatomically complex brain structure, composed of two convoluted sheets of gray matter and located centrally in the medial temporal lobe (Boutet et al., 2014; Giuliano et al., 2017). This structure is of great interest within research areas concerning neurodegenerative diseases (Wisse et al., 2017). Hippocampal volumetry and morphology are important biomarkers for both Alzheimer's Disease (AD) (Maruszak and Thuret, 2014; Small et al., 2011; Boutet et al., 2014), and Temporal Lobe Epilepsy (TLE) (Coras et al., 2014), since this brain structure is target for structural changes in each condition (Maruszak and Thuret, 2014; Small et al., 2011; Boutet et al., 2014; Coras et al., 2014). In AD, the hippocampus is affected approximately 5,5 years before clinical diagnosis, which leaves potential for early diagnosis and initialization of preventive treatment (Maruszak and Thuret, 2014; Panegyres et al., 2016). The hippocampal formation consists of several subfields, all composed of seven structural layers but with different cellular composition (Boutet et al., 2014). During AD disease progression, the subfields are differently affected (Boutet et al., 2014; Small et al., 2011), which makes the monitoring of their structural properties highly desirable (Small et al., 2011; Giuliano et al., 2017). The hippocampal subfields are small structures and some of the subfield boundaries can not be consistently visualized in low-resolution MRI (Wisse et al., 2017). In recent years, ultra-high field (7T) MRI has emerged, enabling detailed in vivo visualization of the hippocampal subfields (van der Kolk et al., 2013; Boutet et al., 2014; Wisse et al., 2017). This has induced an increased interest for segmentation of this brain region and several manual delineation protocols have been published for hippocampal subfield segmentation (Giuliano et al., 2017). The hippocampus is often divided into the following subfields: Cornu Ammonis (CA1-CA4), Dentate Gyrus (DG), presubiculum, and Subiculum (Sub) (Boutet et al., 2014). However, due to the complexity of the problem, manual delineation protocols are not consistent; the subfields' boundaries are differently defined and hippocampus might also include further or fewer subdivisions (Wisse et al., 2017; Giuliano et al., 2017). In general, T2w MR images are the most widely used contrast for segmentation of the hippocampal subfields, since it is easier to separate DG and CA in T2w images compared to T1w (Wisse et al., 2017). On the other hand, T1w images allow easier perception of outer boundaries since the contrast difference between gray and white matter is more distinct (Wisse et al., 2017; Winterburn et al., 2013).

The majority of hippocampal subfield volume measurements are based on manual segmentations, however the disagreement between segmentation protocols limits comparative studies (Giuliano et al., 2017). Furthermore, manual delineation is both laborious and time consuming (Wisse et al., 2017; Giuliano et al., 2017). This substantiates a need of an accurate automatic hippocampal subfield segmentation method (Giuliano et al., 2017). To date, two different approaches for automatic hippocampal subfield segmentation have been developed using ultra-high field MRI (Wisse et al., 2016; Iglesias et al., 2015; Giuliano et al., 2017). The first method, FreeSurfer 6.0, is a learning-based method (Iglesias et al., 2015), which utilizes a generative MRI model and probabilistic atlas of the hippocampal anatomy in combination with Bayesian inference to learn the hippocampal subfields from manually delineated labels (Iglesias et al., 2015). The atlas was built with both low-resolution in vivo and ultra-high resolution ex vivo MRI (Iglesias et al., 2015), which potentially reduces application possibilities as it includes prior information specific for an elderly population (Giuliano et al., 2017). Furthermore, the method has reported difficulties with some subfield boundaries (Iglesias et al., 2015). The second method, ASHS, is based on multi-atlas joint label fusion, utilizing topological atlases with ultra-high resolution in vivo MRI and manually delineated labels, and an adaptive boosting algorithm for output correction (Yushkevich et al., 2015; Wisse et al., 2016). ASHS is a re-trainable tool-box, which entails this method to be generalizable to different populations (Giuliano et al., 2017). ASHS is the only automatic hippocampal subfield segmentation method validated using ultra-high

resolution MRI, and has obtained high accuracy for the larger subfields but low accuracy for smaller subfields (Wisse et al., 2016). Both of the aforementioned methods are time consuming (FreeSurfer$\geq$10 hours on single CPU, ASHS $\geq$ 24 hours on single CPU) (Iglesias et al., 2015; Wisse et al., 2016).

Convolutional Neural Networks (CNN) have shown fast and accurate segmentations of brain structures and lesions in neuroimaging compared to corresponding manual and automatic segmentation methods (Choi and Jin, 2016; Kamnitsas et al., 2017; Nie et al., 2016; Rajchl et al., 2017; Kleesiek et al., 2016; Havaei et al., 2017; Moeskops et al., 2016; Pereira et al., 2016). Additionally, a CNN architecture can be trained to fit different segmentation problems, e.g. Rajchl et al. (2017) segmented structures in images of premature lungs and brains using the same network (Rajchl et al., 2017). In contrary to the aforementioned automatic hippocampal subfield segmentation methods (González-Villá et al., 2016; Giuliano et al., 2017), an approach using a CNN is not dependent on manual feature modelling nor complicated image processing such as non-linear registration, as it learns the relevant image features automatically through convolutions (LeCun et al., 2015; Prasoon et al., 2013). The performance of a CNN is influenced by several factors including, data preprocessing, architecture design, and training strategy (LeCun et al., 2015; Pereira et al., 2016). One discussed factor is the amount of data needed for a CNN to achieve reasonable results (Choi and Jin, 2016; Moeskops et al., 2016). This issue was previously approached by Cho et al. (2015) who aimed to classify various 2D CT images and found the training progress' learning curve to converge towards a steady state after training the CNN with more than 200 images (Cho et al., 2015). To our knowledge, the amount of data required to train a CNN to segment brain structures in 3D MR images has not been investigated.

The aim of this study is to develop an automatic method for hippocampal subfield segmentation (DeepHSS) in in vivo ultra-high resolution MR images, based on a convolutional neural network. Moreover, the impact of a small training dataset is explored.

2 METHOD

The proposed hippocampal segmentation method, DeepHSS, is composed of two main components; a data preprocessing pipeline and a CNN. The CNN is based on the framework DeepMedic, which was developed by Kamnitsas et al. (2017).

2.1 Data

The dataset used in this study was previously utilized by Wisse et al. (2016) for validation of ASHS. The dataset comprises brain scans acquired from 26 healthy subjects (46 % men, mean age: 59 $\pm$9 years, median Mini Mental Examination score (Folstein et al., 1975) 29, 25-30). For each subject a 7T T2w TSE image (0.7 x 0.7 x 0.7 mm^3 voxel size) and a manual delineated segmentation were available. The following hippocampal subfields were contained in the manually delineated segmentations; CA1, CA2, CA3, DG, Sub, Entorhinal Cortex (ERC), tail, and cyst (Wisse et al., 2016). 15 subjects were used for training, 10 were used for test.

2.2 Preprocessing

All 7T TSE MR images were run through a preprocessing pipeline. The pipeline consisted of three steps. To reduce the amount of excess image information, all images were initially masked with a whole brain mask generated using BET2 (Jenkinson et al., 2015). Subsequently, to avoid covariate shifts in the CNN, the masked images were normalized to have zero mean and unit variance. Finally, all images were bisected by removing the left hemisphere. This was done to focus on the right hippocampus.

2.3 CNN architecture and training

The CNN embedded in DeepHSS was developed using the DeepMedic framework described by(Kamnitsas et al., 2017). An illustration of the CNN is present in Figure 1. The CNN was implemented as a 3D architecture with two pathways. The two pathways were similarly constructed with eight convolutional layers and kernels of size [3 x 3 x 3]. The number of kernels in each layer were [30 30 40 40 40 40 50 50] and layer 4, 6, and 8 were residual layers. The pathways were fed the same segments, but one of the pathways were fed a version where all dimensions were down sampled by three. The purpose of the down-sampled pathway was to find coarse features and use them to e.g. find the position of the hippocampal subfields, letting the other pathway focus on finer details (Kamnitsas et al., 2017) such as the subfield boarders and textual information. The convolutional layers were followed by two fully connected layers, connecting both pathways, with 150 kernels of size 3 x 3 x 3 in each layer. The activation function applied in both the convolutional and fully connected layers were PReLU. A softmax layer was applied to classify the input segments' voxels into nine classes, eight hippocampal subfields and background. For post-processing, a CRF layer was implemented. This layer aimed to improve the softmax layers classifications by refining weak predictions and utilizing voxel neighbourhood relations.

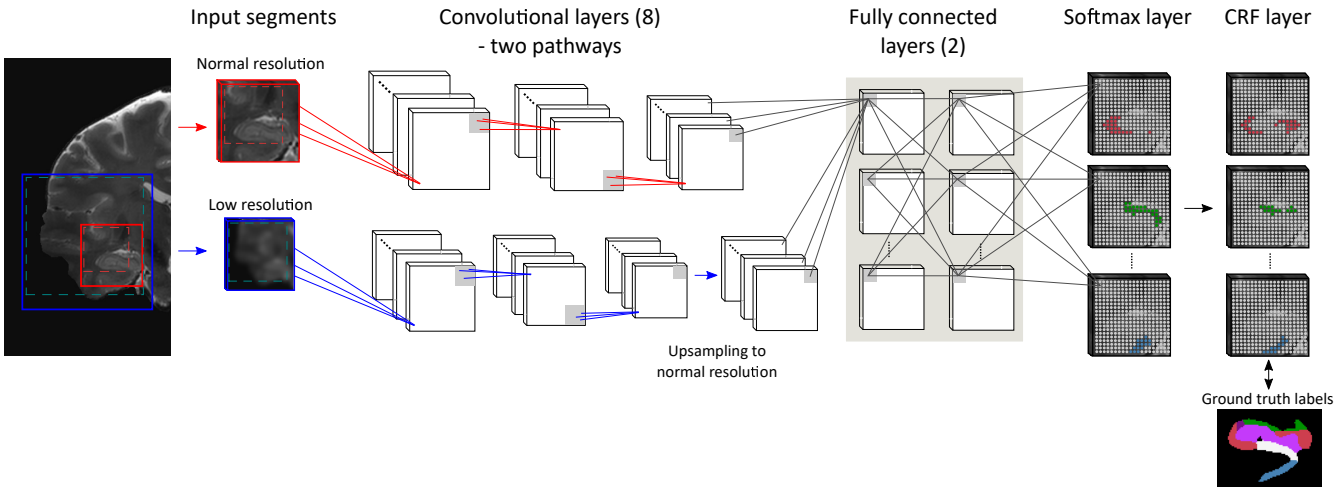


Figure 1. Schematic illustration of the CNN embedded in DeepHSS. The CNN consists of 12 layers in total, with eight convolutional layers in each of two pathways, two fully connected layers, and two classification layers.

The training strategy implemented was dense training, which was first suggested by Kamnitsas et al. (2017). Dense training is a method based on dense inference, which is particularly relevant for training a 3D CNN, seeking an accurate and fast training. Dense training entail training with batches, which consist of segments extracted with 50 % probability of having their centre voxel in foreground or background, which alleviate class imbalance (Kamnitsas et al., 2017). The training consisted of 22 epochs and each epoch included 20 subepochs. The optimizer algorithm implemented was RMSProp. Weight initialization, batch normalization, and regularization methods were applied as described by Kamnitsas et al. (2017).

2.4 Testing of DeepHSS

The performance of DeepHSS was qualitatively evaluated by means of visual inspection, and quantified by comparing automatically generated segmentations to corresponding manually delineated segmentations with the similarity metric Dice Similarity Coefficient (DSC). DSC is a measure of how much two segmentations overlap and was computed using the following equation (Dice, 1945):

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{1}$$

As overall performance measures, a foreground DSC and an average subfield DSC were calculated. Results for subfield CA1, CA2, CA3, DG, Sub and ERC were included in both measures, and combined to one class in the calculations of foreground DSC.

3 RESULTS

The performance of DeepHSS in relation to the number of training subjects is presented in Figure 2 and Table 1. The average foreground DSC was 0.73 $\pm$ 0.07 when DeepHSS was trained with 3 subjects and the foreground DSC improved until it reached 0.83 $\pm$ 0.03 for 15 subjects. The foreground DSC increased exponentially as the number of training subjects increased until 9 subjects, were the curve, seen in Figure 2, bends and converge towards a steady state.

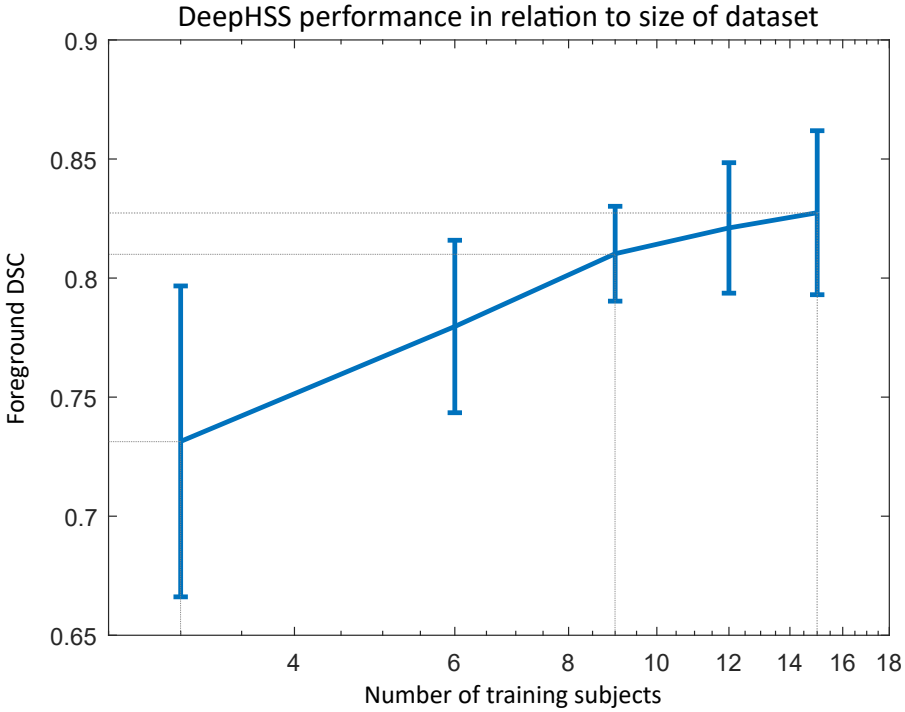


Figure 2. Foreground DSC between manual segmentations and segmentations obtained using DeepHSS as a function of the number of subjects used for training of the CNN. The network was trained using 3, 6, 9, 12, and 15 subjects. The graph illustrates an exponential correlation until 9 subjects where the learning curve bends and converge towards a steady state.

The DSC for all automatic hippocampal subfield segmentations using DeepHSS including overall average measures are presented in Table 1. Training with 15 subjects, the highest subfield DSC was obtained for DG (0.77 ± 0.05) and Sub (0.80 ± 0.05), and the average subfield DSC was 0.66 ± 0.08.

Table 1. Hippocampal subfield DSC, average subfield DSC and foreground DSC calculated between segmentations obtained using DeepHSS and their corresponding manual delineated segmentations. The network was trained using 3, 6, 9, 12, and 15 subjects.

	3 subjects	**6 subjects**	**9 subjects**	**12 subjects**	**15 subjects**
CA1	0.48 ± 0.08	0.54 ± 0.07	0.56 ± 0.04	0.58 ± 0.07	0.61 ± 0.06
CA2	0.60 ± 0.06	0.64 ± 0.07	0.69 ± 0.05	0.69 ± 0.06	0.69 ± 0.05
DG	0.66 ± 0.08	0.72 ± 0.06	0.76 ± 0.04	0.77 ± 0.04	0.77 ± 0.05
CA3	0.54 ± 0.14	0.58 ± 0.12	0.60 ± 0.13	0.59 ± 0.18	0.59 ± 0.19
Sub	0.72 ± 0.11	0.77 ± 0.06	0.80 ± 0.05	0.80 ± 0.06	0.80 ± 0.05
ERC	0.42 ± 0.15	0.44 ± 0.11	0.46 ± 0.14	0.49 ± 0.17	0.49 ± 0.14
Average subfield DSC	0.57 ± 0.08	0.61 ± 0.06	0.64 ± 0.05	0.65 ± 0.07	0.66 ± 0.08
Foreground DSC	0.73 ± 0.07	0.78 ± 0.04	0.81 ± 0.02	0.82 ± 0.03	0.83 ± 0.03

Figure 3 presents a visualization of the best and worst predicted hippocampal subfield segmentation obtained by DeepHSS trained with 15 subjects and the associated manual segmentations, all superimposed onto the corresponding preprocessed 7T TSE MR image. The segmentations are presented in the coronal and sagittal view.

The best hippocampal subfield segmentations obtained using DeepHSS and the corresponding manual labels are visually very similar. However, small deviations are present due to disagreement of subfield boundary. Two examples are marked by arrows in Figure 3. The worst hippocampal subfield segmentations obtained using DeepHSS are affected by under-segmentations (see white arrows in first and third column, Figure 3). Moreover, a few anatomical misdetections are observed outside of the hippocampus. As an example, ERC has been incorrectly detected in the parahippocampal gyrus (see arrow in fourth column, Figure 3).

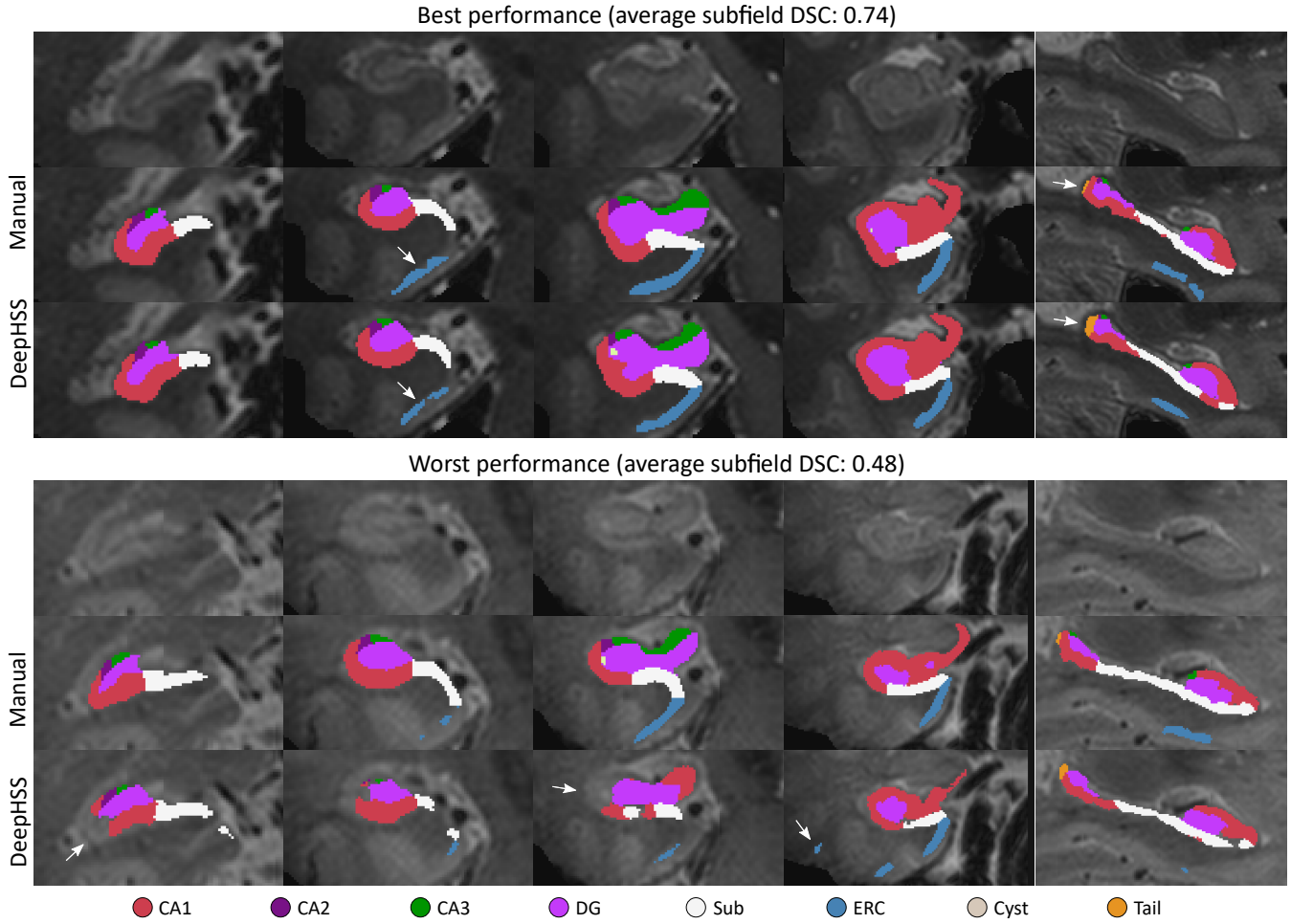


Figure 3. Coronal and sagittal views of the best and worst predicted hippocampal subfield segmentations obtained using DeepHSS compared to the corresponding manual labels. The best performance has an average subfield DSC at 0.74 whilst the worst performance scores an average subfield DSC at 0.48. Arrows point out differences between the manual delineated segmentation and the automatic segmentation obtained using DeepHSS.

4 DISCUSSION

This study demonstrates CNNs as an efficient method for automatic hippocampal subfield segmentation despite utilazation of a small dataset. The proposed method, DeepHSS, consists of a preprocessing pipeline and a two-pathway deep CNN, which was developed using the Deepmedic framework by Kamnitsas et al. (2017). DeepHSS was trained and validated using in vivo T2w ultra-high field MR images and associated manual segmentations, which have previously been used to validate ASHS. DeepHSS demonstrated a fast segmentation process (1.25 h) and achieved segmentation accuracy comparable with manual segmentations and segmentations performed by ASHS. The performance of the method demonstrates the flexibility of CNNs, and supports this as an easy adaptive method for various segmentation problems, as previously stated by Rajchl et al. (2017).

The performances of DeepHSS and ASHS are presented in Tabel 2. It is notable that DeepHSS performs better for the small subfields (CA2 and CA3) than ASHS, whereas ASHS performs better for the largest subfield (CA1) than DeepHSS.

Table 2. Dice similarity measure for right hippocampus subfields. A comparison between results obtained using DeepHSS and ASHS (Wisse et al., 2016).

Hippocampal subfield	DeepHSS (DSC)	ASHS (DSC)
CA1	0.61 ± 0.06	0.83 ± 0.02
CA2	0.69 ± 0.05	0.65 ± 0.09
DG	0.77 ± 0.05	0.84 ± 0.03
CA3	0.59 ± 0.19	0.54 ± 0.13
Sub	0.80 ± 0.05	0.78 ± 0.04
ERC	0.49 ± 0.14	0.75 ± 0.06

Wisse et al. (2016) obtained the lowest accuracy for CA2, CA3 and ERC, and explained this by a correspondingly low manual interrater repeatability for these subfields ($0.27 < \text{ICC} < 0.88$) (Wisse et al., 2016). Additionally, the anterior and posterior boundaries of these subfields are based on geometric rules rather than visibility of textual changes (Wisse et al., 2016). DeepHSS is affected by the same limitations of the training data, which is exemplified by misclassification of ERC in the hippocampal gyrus (see Figure 3). Misclassification outside of the hippocampal formation was found for CA1 as well, which in combination with an observed CA1 under-segmentation explain its low DSC. Since CA1 is the largest subfield it is hypothesized that the segmentation of this subfield is more dependant on textual features than the smaller subfields.

The number of necessary training subjects to obtain accurate segmentations was investigated by training the network in parallel using 3, 6, 9, 12, and 15 subjects. We found a relation between an increased number of subjects and an increased segmentation accuracy converging towards a steady state. This finding is consistent with existing research by Cho et al. (2015), who found the training progress' learning curve to converge towards a steady state after training the CNN with more than 200 images, classifying 2D CT images of various body parts (Cho et al., 2015). Hence, the optimum size of training sets vary for different classification problems. Our segmentation problem can be considered homogeneous with a standardised acquisition protocol between subjects, which is in contrary to Cho et al. (2015), where images were acquired using different settings. Moreover, the anatomy of hippocampus is homogeneous between subjects compared to e.g Kamnitsas et al. (2017) who segmented brain tumours and lesions. However, due to unclear boundaries between each hippocampal subfield, and high manual interrater variability it can be difficult for a CNN to distinguish between each class.

We presented a deep learning approach for resolvement of the complex hippocampal subfield segmentation problem, which compared to multi-atlas based methods has potential to be more robust, since it is independent on complex non-linear registrations and feature engineering (González-Villá et al., 2016). This approach share properties with the multi-atlas based method, ASHS, which also benefits from a type of supervised machine learning. As previously discussed in a review by González-Villá et al. (2016), multi-atlas techniques show good results within brain segmentation, since the atlases provide spatial information and adds robustness, whereas supervised classifiers are able to model structure appearance and improve initial segmentations (González-Villá et al., 2016). A supervised CNN also utilizes both of these properties, however, as the feature extraction is in-depended on human domain expertise and found by automatic optimization, a combination of the most prone features is found in the training dataset and

used for segmentation (LeCun et al., 2015). This makes it highly desirable in problems where complex underlying correlations are present.

Neither DeepHSS or ASHS has yet been challenged by the problem of distinguishing patients from controls. Feeding the CNN more information could give it a base for further development of connectivities between features, and possibly achieve higher accuracy. Moreover, as investigated by Çitak-Er et al. (2017), a CNN could be utilized as a diagnostic tool finding wider relations between biomarkers.

REFERENCES

Boutet, C., Chupin, M., Lehéricy, S., Marrakchi-Kacem, L., Epelbaum, S., Poupon, C., et al. (2014). Detection of volume loss in hippocampal layers in alzheimers disease using 7t mri - a feasibility study. *NeuroImage: Clinical* 5, 341–348. doi:10.1016/j.nicl.2014.07.011

Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?

Choi, H. and Jin, K. H. (2016). Fast and robust segmentation of the striatum using deep convolutional neural networks. *Journal of Neuroscience Methods* 274, 146–153. doi:10.1016/j.jneumeth.2016.10.007

Çitak-Er, F., Goularas, D., and Ormeci, B. (2017). A novel convolutional neural network model based on voxel-based morphometry of imaging data in predicting the prognosis of patients with mild cognitive impairment. *Journal of Neurological Sciences* 32, 52–69

Coras, R., Milesi, G., Zucca, I., Mastropietro, A., Scotti, A., Figini, M., et al. (2014). 7t mri features in control human hippocampus and hippocampal sclerosis: An ex vivo study with histologic correlations. *Epilepsia* 55, 2003–2016. doi:10.1111/epi.12828

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi:10.2307/1932409

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "mini-mental state". *Journal of Psychiatric Research* 12, 189–198. doi:http://dx.doi.org/10.1016/0022-3956(75)90026-6

Giuliano, A., Donatelli, G., Cosottini, M., Tosetti, M., Retico, A., and Fantacci, M. E. (2017). Hippocampal subfields at ultra high field MRI: An overview of segmentation and measurement methods. *Hippocampus* 27, 481–494. doi:10.1002/hipo.22717

González-Villá, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., and Lladó, X. (2016). A review on brain structures segmentation in magnetic resonance imaging. *Artificial Intelligence in Medicine* 73, 45–69. doi:http://dx.doi.org/10.1016/j.artmed.2016.09.001

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis* 35, 18–31. doi:10.1016/j.media.2016.05.004

Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., et al. (2015). A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *NeuroImage* 115, 117–137. doi:http://dx.doi.org/10.1016/j.neuroimage.2015.04.042

Jenkinson, M., Pechaud, M., and Smith, S. (2015). Bet2: Mr-based estimation of brain, skull and scalp surfaces. *In Eleventh Annual Meeting of the Organization for Human Brain Mapping*

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis* 36, 61–78. doi:10.1016/j.media.2016.10.004

Kleesiek, J., Urbana, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., et al. (2016). Deep mri brain extraction: A 3d convolutional neural network for skull stripping. *NeuroImage* 129, 460–469. doi:10.1016/j.neuroimage.2016.01.024

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444

Maruszak, A. and Thuret, S. (2014). Why looking at the whole hippocampus is not enough - a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for alzheimer's disease diagnosis. *Frontiers in Cellular Neuroscience* 8. doi:10.3389/fncel.2014.00095

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J. N. L., and Isgum, I. (2016). Automatic segmentation of mr brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging* 35, 1252–1261. doi:10.1109/TMI.2016.2548501

Nie, D., Wang, L., Gao, Y., and Sken, D. (2016). Fully convolutional networks for multi-modality isointense infant brain image segmentation. *Proc IEEE Int Symp Biomed Imaging* doi:10.1109/ISBI.2016.7493515

Panegyres, P., Berry, R., and Burchell, J. (2016). Early dementia screening. *Diagnostics* 6, 6. doi:10.3390/diagnostics6010006

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging* 35, 1240–1251. doi:10.1109/TMI.2016.2538465

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., and Nielsen, M. (2013). Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *International Conference on Medical Image Computing and Computer-Assisted Intervention* 16, 246–253. doi:10.1007/978-3-642-40763-5_31

Rajchl, M., Lee, M. C. H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., et al. (2017). Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE TRANSACTIONS ON MEDICAL IMAGING* 36, 674–683

Small, S. A., Schobel, S. A., Buxton, R. B., Witter, M. P., and Barnes, C. A. (2011). A pathophysiological framework of hippocampal dysfunction in ageing and disease. *Nature Reviews Neuroscience* 12, 585–601. doi:10.1038/nrn3085

van der Kolk, A. G., Hendrikse, J., Zwanenburg, J. J., Visser, F., and Luijten, P. R. (2013). Clinical applications of 7t MRI in the brain. *European Journal of Radiology* 82, 708–718. doi:10.1016/j.ejrad.2011.07.007

Winterburn, J. L., Pruessner, J. C., Chavez, S., Schira, M. M., Lobaugh, N. J., Voineskos, A. N., et al. (2013). A novel in vivo atlas of human hippocampal subfields using high-resolution 3t magnetic resonance imaging. *NeuroImage* 74, 254–265. doi:http://dx.doi.org/10.1016/j.neuroimage.2013.02.003

Wisse, L. E., Daugherty, A. M., Olsen, R. K., Berron, D., Carr, V. A., Stark, C. E., et al. (2017). A harmonized segmentation protocol for hippocampal and parahippocampal subregions why do we need one and what are the key goals? *HIPPOCAMPUS* 27, 3–11. doi:http://dx.doi.org/10.3174/ajnr.A4659

Wisse, L. E. M., Kuijf, H. J., Honingh, A. M., Wang, H., Pluta, J. B., Das, S. R., et al. (2016). Automated hippocampal subfield segmentation at 7T MRI. *American Journal of Neuroradiology* 37, 1050–1057. doi:10.3174/ajnr.A4659

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., et al. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping* 36, 258–287. doi:10.1002/hbm.22627

Automatic hippocampal subfield segmentation in ultra-high resolution MRI using Convolutional Neural Network

- Worksheets

Groupnr: 17gr10407
Group members: Anne Krogh Nøhr,
Bolette Dybkjær Hansen, Nina Jacobsen

Master's Thesis
Biomedical Engineering and Informatics

AALBORG UNIVERSITY
DENMARK

AALBORG UNIVERSITY
STUDENT REPORT

Department of Health Science and Technology
Aalborg University, Denmark
http://www.smh.aau.dk

Title:
Automatic hippocampal subfield segmentation in ultra-high resolution MRI using Convolutional Neural Network

Theme:
Master's Thesis

Project Period:
Spring semester 2017

Project Group:
17gr10407

Participants:
Anne Krogh Nøhr
Bolette Dybkjær Hansen
Nina Jacobsen

Supervisors:
Lasse Riis Østergaard
Steffen B Petersen

External supervisor:
Steffen Bollmann

Page Numbers: 67

Date of Completion:
June 6, 2017

Danish summary

Demens er et globalt problem, som ca. 46,8 millioner mennesker levede med i 2015. Den hyppigste årsag til demens er alzheimers, og risikoen for at udvikle alzheimers i løbet af livet ligger på ca. 10.5 %. Alzheimers udvikles over årtier og starter med en lang asymptomatisk periode, efterfulgt af kognitiv funktionsnedsættelse, og, i de sene stadier af sygdommen, demens. Tidlig diagnose af Alzheimers er vigtig, da der er evidens for, at behandling har bedre effekt, hvis den påbegyndes på et tidligt stadie. Alzheimers diagnosticeres i dag via en række biomarkører, heriblandt hippocampus atrofi. Allerede ca. 5,5 år før alzheimers diagnosen stilles, begynder hippocampus at degenerere forskelligt fra normal aldersbetinget degenerering. Degenereringen hos alzhimers patienter sker asynkront i de forskellige underregioner af hippocampus, og i forlængelse af, at ultrahøjtopløselig MRI har gjort det muligt at observere disse underregioner mere detaljeret end tidligere, er det blevet foreslået at anvende volumetriske målinger af disse underregioner som biomarkører for alzhimers.

På nuværende tidspunkt opfattes manuel segmentering af regionerne som gold standard. Dog er de manuelle segmenteringsprotokoller som udgangspunkt subjektive, tidskrævende og stiller store krav til optegnerens erfaring. For at imødekomme disse problemer, er der blevet udviklet automatiske segmenterings metoder. Der eksisterer to metoder til segmentering af hippocampus regioner i ultrahøjtopløselig MRI; ASHS og FreeSurfer. Disse metoder har vist lovende resultater, dog har ASHS problemer med at finde de små regioner, imens udviklerene af FreeSurfer har vurderet at de volumetriske mål for regionerne fundet med FreeSurfer skal fortolkes forsigtigt.

Convolutional Neural Networks (CNNs) er blevet foreslået som en alternativ tilgang til segmenterings problemer og har vist lovende resultater for segmentering af forskellige hjernestrukturer og læsioner i MRI billeder. En af fordelene ved disse netværker er, at de automatisk lærer de features, som karakteriserer problemet. Yderligere undgås avanceret preprocessering af billederne som ikke lineær registrering, hvilket de førnævnte automatiske segmenterings metoder benytter.

Formålet med dette projekt er, at udvikle en CNN basseret automatisk metode til segmentering af underregionerne i hippocampus i ultrahøjtopløselige MR billeder.

En automatisk metode til segmentering af hippocampus regioner (DeepHSS) blev udviklet på baggrund af et litteraturstudie og initielle test. DeepHSS omfatter preprocessering og et 2 pathway, 12 lag dybt, 3D CNN. CNNet i DeepHSS er baseret på det allerede udviklede skelet, DeepMedic. Den første del af netværket er to ens parallelle pathways, som består af 8 convolutional layers. Begge pathways får de samme input-billedsegmenter, men den ene pathway får en downsamplet version. De to pathways er efterfulgt af to fully-connected layers. I både de convolutional layers og de fully-connected layers benyttes aktiveringsfunktionen PReLU. Efter de fully-connected layers klassificeres hvert voxel til en af de 8 klasser ved brug af et softmax lag. Det sidste lag er et CRF lag, der benyttes til postprocessering. Netværket var trænet ved at bruge en metode, der hedder Dense træning.

Da DeepHSS var færdigudviklet blev det trænet på et datasæt bestående af 7T T2w TSE billeder af 15 raske forsøgspersoner samt tilhørende manuelle segmenteringer, hvori de følgende underregioner af hippocampus er segmenteret: Corneas Ammonis (CA1, CA2, CA3), Dentate Gyrus (DG), Subiculum (Sub), Entorhinal Cortex (ERC), tail, and cyst. Efterfølgende blev DeepHSS testet på et tilsvarende datasæt indeholdende data for 10 raske forsøgspersoner.

DeepHSS blev evalueret ved sammenligning med manuelle segmenteringer via dice score og visuel inspektion. DeepHSS opnåede en forgrund dice score på 0,83 $\pm$ 0.03. De højeste dice scores blev fundet for subiculum (0,80 $\pm$ 0.05) og Dendate Gyrus (0.77 $\pm$ 0.05), mens den laveste dice score blev fundet for Entorhinal Cortex (0.49 $\pm$ 0.04). Sammenhængen mellem træningsdatasættets størrelse og præstationen af DeepHSS blev også undersøgt, og der blev observeret en tydelig sammenhæng mellem størrelsen af datasættet og DeepHSS præstation.

DeepHSS demonstrerede en hurtig segmenteringsproces og opnåede en segmenteringsnøjagtighed sammenlignelig med manuel segmentering og segmenteringer opnået ved brug af eksisterende automatiske metoder. Den gode præstation viser, at CNNer er fleksible, og at metoden er let adaptiv til andre segmenteringsproblemer. Sammenlignet med ASHS havde DeepHSS bedre dice score for de små regioner, mens ASHS havde en markant bedre dice for den største region. En af årsagerne til, at DeepHSS har en forholdsvis lav dice score for den største region er, at DeepHSS klassificerer områder andre steder i hjernen som dette subfield. Ved visuel inspektion blev det vurderet, at DeepHSS gennerelt fandt de korrekte grænser for den største region.

Contents

Chapter 1
Introduction

Dementia is a rising issue worldwide and in 2015 was the estimated number of individuals living with dementia 46.8 million. Due to an increasing elderly population, this number will double every 20 years onwards, reaching 131.5 million affected people in 2050. [Prince et al., 2015] Alzheimer's Disease (AD) is the most common cause of dementia, and during a lifetime, the average risk of evolving AD is 10.5 percent. AD evolves over decades, and starts with a long asymptomatic period, the preclinical period. [Sperling et al., 2011] This is followed by Mild Cognitive Impairment (MCI), either amnestic (aMCI) or non-amnestic, and for many of the affected patients, the final stage of the disease is AD related dementia. Patients with aMCI are more likely to evolve AD dementia. Symptoms seen in MCI patients are characterized by cognitive and functional decline, such as memory and languages impairment. Dementia emerges when the progression of decline becomes severe enough to interfere with daily living activities. [Panegyres et al., 2016]

The first vague diagnosis criteria for AD were based on observed cognitive skills, but definite diagnosis was solely stated after post-mortem examinations. Early diagnosis of AD is important, since evidence for higher potential for a better outcome when treatment is initiated in the preclinical stage is present. New diagnostic criteria were formed in 2011, in which various biomarkers were included. Thereby, the importance of finding the most reliable and precise biomarkers for diagnosis of AD during the preclinical stage was recognized. [Panegyres et al., 2016]

Pathologically, AD is characterized by the development of amyloid plaques and neurofibrillary tangles, which cause neuronal loss and eventually neuronal death. One of the earliest affected brain structures is the hippocampus, for which reason hippocampal atrophy has been incorporated as a diagnostic criteria. In most cases, hippocampus is considerably damaged at the time of diagnosis. However, the rate of atrophy in hippocampus diverge from normal approximately 5.5 years before AD diagnosis is given, thus there is a potential for utilization of this biomaker. [Maruszak and Thuret, 2014]

The hippocampal formation is a complex structure which can be divided into several subfields [Wisse et al., 2017]. It have been suggested that volumetric measurements of these subfields are more promising biomarkers than whole hippocampus volume [Small et al., 2011; Maruszak and Thuret, 2014; Boutet et al., 2014]. Additionally, segmentation of the subfields are important for other neurodegenerative diseases [Small et al., 2011] e.g. drug-resistant temporal lobe epilepsy [Coras et al., 2014; Boutet et al., 2014; Small et al., 2011]. However, segmentation of especially the smaller subfields can be difficult [Wisse et al., 2017]. 3T MRI is widely used in clinical practise, however, an increased number of researchers have access to 7T MRI, also considered ultra-high resolution MRI [van der Kolk et al., 2013]. With the increasing accessibility to ultra-high resolution MR images, the possibility for investigation of smaller anatomical structures in the brain in vivo are evolving [van der Kolk et al., 2013].

To date, manual segmentation of hippocampal subfields is the "golden standard" and several manual protocols have been proposed [Wisse et al., 2017]. However, manual segmentation is labour intensive and time consuming, especially in high resolution images where one segmentation takes approximately 50 hours. For this reason, a lot of resources are required to carry out a study involving manual segmentation, which can be a problem for laboratories with limited resources. [Iglesias et al., 2015] Another issue is that manual segmentation can be inconsistent when performed by different experts [Choi and Jin, 2016]. To overcome

these issues automatic segmentation methods have been proposed. [Iglesias et al., 2015] However, when the automatic methods for segmentation of hippocampal subfields are compared to the manual methods limited validity is shown. Consequently, a need for an automatic segmentation method for hippocampal subfields is still present. [Wisse et al., 2017]

Initial problem statement

What properties describe the problem of hippocampal subfield segmentation, how is this problem approached so far and what methods could serve as alternatives?

Chapter 2
Background

In order to segment hippocampal subfields, knowledge about the anatomical structure of the subfields, how these are affected by AD, and their appearance in MR images are essential. Consequently, these topics are addressed in the first section of the problem analysis. The next section concerns the present automatic segmentation methods for hippocampal subfields segmentation. Finally, an alternative segmentation approach is discussed, and studies using this approach for other segmentation problems are reviewed.

2.1 Hippocampus and Alzheimer's Disease

The human brain contains two hippocampi, which are located in the medial temporal lobes of the brain, and includes the subiculum, dendate gyrus and hippocampus proper. They play a key role in consolidation of information from both short term and long term memory, associative learning, and spatial processing. [Burgess et al., 2002; Hartley et al., 2013] The hippocampus can be divided into several subfields, which usually contains: Corneas Ammonis 1-4 (CA1-CA4), Presubiculum (Pre), Subiculum (Sub) and Dendate Gyrus (DG), however, more subfields can be distinguished [Giuliano et al., 2017]. Each subfield consists of 7 layers of different cellular composition. [Boutet et al., 2014] Figure 2.1 illustrates the hippocampali in a 7T single subject MR image.

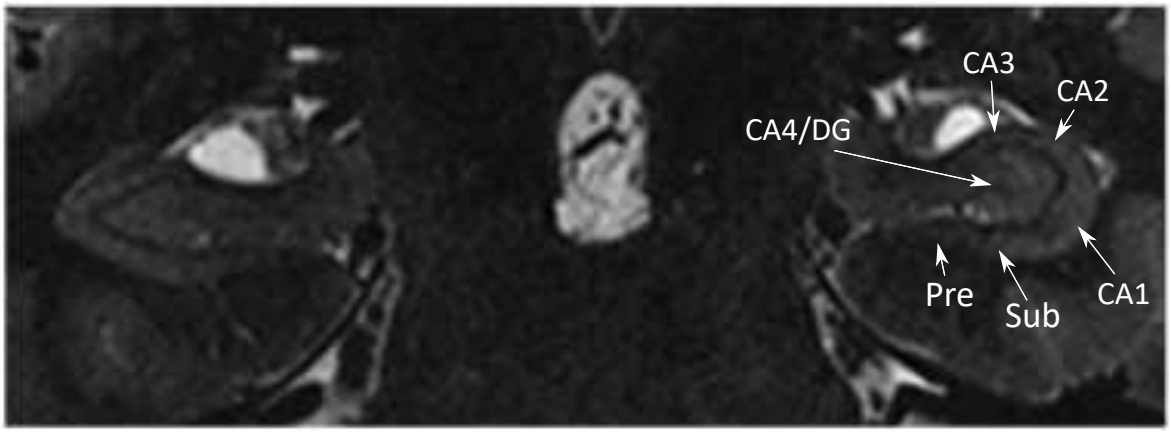


Figure 2.1: In vivo T2w 7T MRI slice capturing the hippocampi in a coronal view. Position of subfield CA1-4, DG, Pre, and Sub are marked by arrows. Inspired by [Wisse et al., 2016].

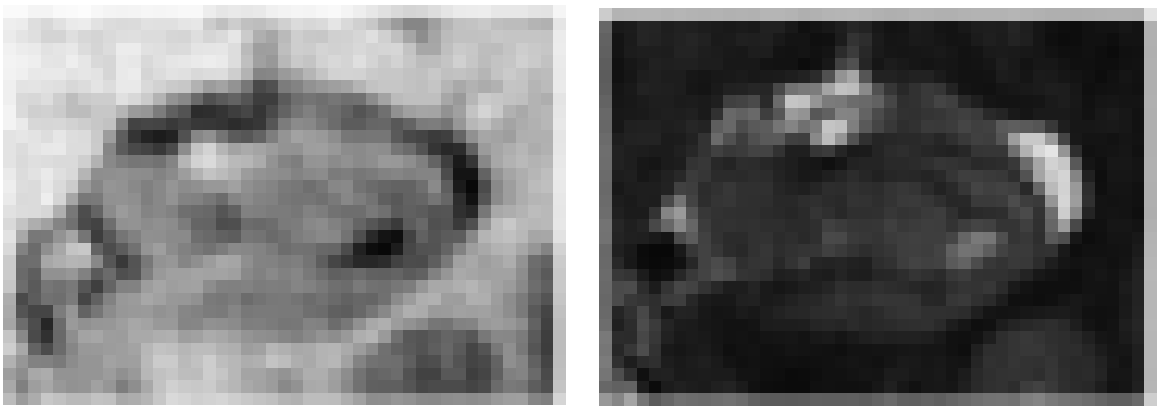
To date, it is difficult to study the hippocampal subfields in neuroimaging due to a disagreement between segmentation protocols in current literature. A variety of segmentation protocols are proposed based on different definitions and terminology of the subfields' boundaries. Consequently, comparison between study results are unprecise. It is even difficult to compare results between studies investigating similar phenomena and populations. Consequently, a need of a standardised method is present. [Wisse et al., 2017]

2.1.1 Degeneration of the hippocampal subfields

Neurofibrillary tangles, amyloid plaques and consequently hippocampal atrophy are, as described in section 1, features characterizing AD. In early stages of AD, years before cognitive deficits are visible, the hippocampal subfields are affected. When the neurofibrillary tangles target the hippocampus, the first affected subfield is CA1, then subiculum, and next CA2 and CA3. Parallel to neurofibrillary tangles, synapse and neuronal loss are observed, and the most prominent neuronal loss is seen for CA1. The size of CA1 is hardly reduced in normal ageing individuals but for an individual with AD the atrophy diverge from the normal rate years before diagnosis. [Maruszak and Thuret, 2014] Thereby, CA1 has been suggested as a biomarker for AD [Joie et al., 2013; Maruszak and Thuret, 2014; Boutet et al., 2014]. Thinkness meassurements of layers within CA1 have been found to distinguish AD and non-AD subjects better than whole hippocampus volume [Joie et al., 2013]. Apart from showing a decreased size of CA1 for AD patients, the literature also shows a smaller volume of the CA1-2 transition zone and subiculum. These findings correlate with the observed level of neural loss for these structures. [Maruszak and Thuret, 2014]

2.1.2 Appearance of hippocampal subfilds at MRI

Both T1-weighted (T1w) and T2w MRI are used for hippocampal subfield segmentation. In figure 2.2 a 7T T1w and a 7T T2w image of hippucampus are illustrated. The most commonly used contrast for segmentation is T2w, since the boundary between the CA and the DG is more distinct. [Wisse et al., 2017] The contrast difference between subfields are in general more distinct in T2w images compared to T1w. However, the difference in contrast between gray and white matter is unclear in T2w images, for which reason segmentation of the whole hippocampus is easier in T1w, where the low intensity gray matter in hippocampus stands out against the surrounding high intensity white matter and the lower intensity cerebral spinal fluid. [Winterburn et al., 2013] Segmentations of hippocampus based on multimodal MRI, where information from both modalities are used, can thereby be an advantage leading to a more accurate segmentation. [Iglesias et al., 2015]

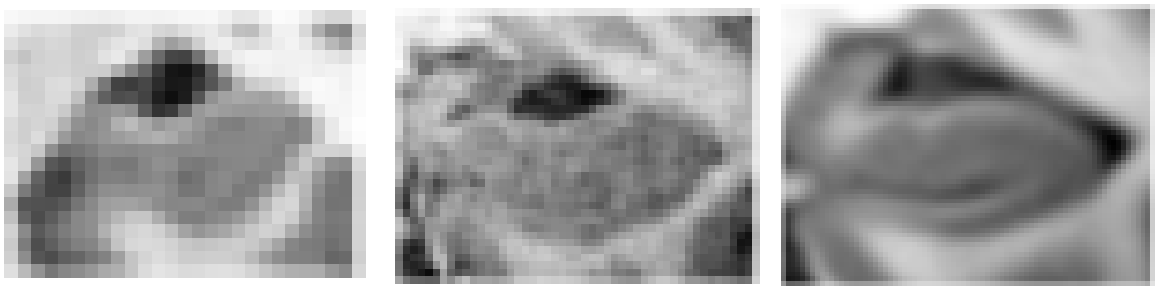


(a) **(b)**

Figure 2.2: Coronal views of hippocampus proper in a slice from a T1w 7T MR image (a) and a T2w 7T MR image (b). The T2w image was resampled to the T1w image.

Another feature highly influencing the accuracy of the segmentation is the image resolution. The higher field strength of the MR scanner, the higher spacial resolution can be obtained. Compared to both 1,5T and 3T, 7T MRI can provide a more anatomically detailed

visualization of the hippocampus, since the contrast between white and grey matter also increases with the resolution. [van der Kolk et al., 2013] Even though 7T MRI can provide a resolution of the hippocampus allowing identification of layers within the different subfields [Boutet et al., 2014], more detailed maps would allow even better visualization. In order to achieve a higher resolution than possible by 7T single subject MRI, a minimum deformation average (MDA) MRI model may be generated. This model improves structural delineation and image contrast [Ullmann et al., 2013], and the signal-to-noise ratio increases for every new subject included in the model generation [Evans et al., 2012]. Furthermore, the model captures the mean and variability of the population, and can therefore be seen as a population prior [Janke and Ullmann, 2015]. Further methodical information about MDA is found in Appendix A. In figure 2.3 a hippocampus is illustrated in a 3T MRI, 7T MRI and 7T MDA MRI model.



(a) **(b)** **(c)**

Figure 2.3: Coronal views of hippocampus proper in a slice from a 3T T1w single subject (a), 7T T1w single subject (b), and a T1w 7T MDA MRI model (c). The image resolution is 1mm isotropic, 0.5mm x 0.533mm x 0.533mm, and 0.3mm isotropic, respectively. The 3T MRI and the 7T MRI are acquired from the same subject.

2.2 Methods for segmentation of hippocampal subfields

To date, three open source automated methods for hippocampal subfield segmentation exist [Giuliano et al., 2017; Pipitone et al., 2014; Wisse et al., 2016; Iglesias et al., 2015]. Two of these methods were developed using 7 T MRI [Wisse et al., 2016; Iglesias et al., 2015], whilst one of these were validated on 7 T data [Wisse et al., 2016]. All three methods are based on atlases. Table 2.1 presents an overview of the three segmentation methods.

Table 2.1: Automatic methods for hippocampal subfield segmentation. All methods are altas-based, but rely on different datasets and methodical approaches.

Source	Data	Segmentation approach
Pipitone et al. [2014]	Developed with in vivo 3T T1w images, validated on 1.5T and 3T T1w images	Multi-atlas segmentation (MAGet-Brain)
Iglesias et al. [2015]	Developed with in vivo 1.5T and ex vixo 7T. Validated on T1w and T2w MRI 1,5 and 3T.	Combines a probabilistic atlas and Bayesian inference. (FreeSurfer version 6.0) Build on previous version descriped by Leemput et al. [2009].
Yushkevich et al. [2015]; Wisse et al. [2016]	Developed and validated using in vivo 7T T1w and T2w images	Multi-atlas based method combined with machine learning (ASHS). ASHS was described by [Yushkevich et al., 2015] and later validated by [Wisse et al., 2016].

2.2.1 Methods developed with 7T MRI

The method developed by Iglesias et al. is called FreeSurfer 6.0, and is a parametric atlas based method developed on the basis of FreeSurfer 5.3 by Leemput et al.. The atlases utilized for this method were constructed with both in vivo and ex vivo MRI data. The in vivo data was T1w and obtained from 39 subjects using a 1,5T MR scanner, and had thirty-six brain structures delineated by labels, including the left and right hippocampus. The structures around hippocampus were delineated with the purpose of adding contextual information to the atlas. From 15 subjects the ex vivo MRI data was acquired using a 7T scanner less than 24h post-mortem. Thirteen labels distinguishing the hippocampal subfields and layers were generated by seven manual labellers using all ex vivo scans and based on a protocol made for the purpose. The constructed atlas described the prior probabilities of label occurrences and were used to segment MRI scans using a generative model. The generative model uses a likelihood distribution to predict how a labelled image (each voxel is attributed to a unique label) is translated into an MRI image (each voxel has an intensity). The segmentation is an optimization problem and bayesian inference is used to solve it, by searching for the most likely label given the atlas and the observed intensities of the image. The method was validated using in vivo MRI scans with different resolutions (1.5T or 3T) and contrasts (T1 and/or T2) from three databases. Iglesias et al. reported visual replicable automatic hippocampal subfield segmentations when compared qualitatively to manual segmentations from Winterburn et al.. Furthermore, Iglesias et al. stated that utilization of both T1w and T2w MRI for the validation improved the segmentations. Iglesias et al. [2015] had issues with blurring of some of the boundaries e.g. the boundaries between CA subfields. Figure 2.4 is an illustration of three coronal slices with the associated automatic (using both T1w and T2w MRI) and manual segmentation results used for the qualitative evaluation. Results for automatic hippocampal subfield segmentations at standard resolution (1mm) MRI were described as volumetric results which should be interpreted with caution due to low contrast and resolution of the input data. [Iglesias et al., 2015] Using FreeSurfer the computation time for one subject was 20-40 minutes for the segmentation, and 10 hours for preprocessing, when calculated on a single core cpu. [Iglesias et al., 2015]

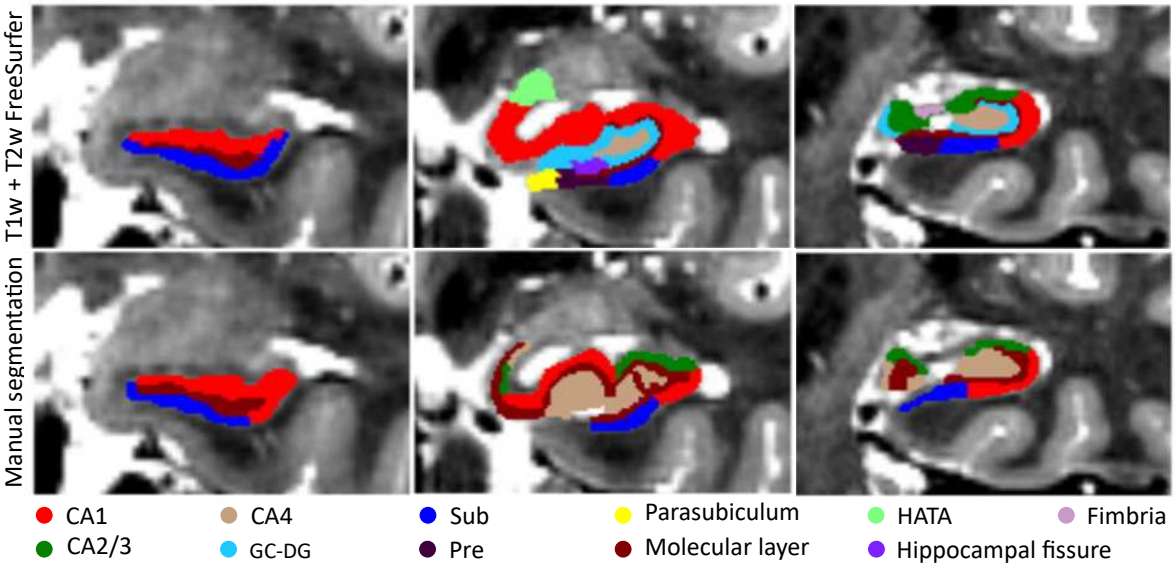


Figure 2.4: Three coronal slices with the associated automatic and manual hippocampal subfield segmentation results. Labeling protocols used for FreeSurfer and manual segmentations, respectively, did not match. 11 out of 13 subfields were detected in present views. The Modified from Iglesias et al. [2015].

Yushkevich et al. suggested a non-parametric multi-atlas approach, refereed to as ASHS [Yushkevich et al., 2015]. It was originally developed using 3T in-vivo MR images, but later modified to run with 7T in-vivo MR images both for training and validation [Wisse et al., 2016]. ASHS was build with two pipelines; a training pipeline and a segmentation pipeline. For the training, ASHS includes 26 atlases each consisting of one T1w, one T2w MR image, and one set of manual labels drawn from the T2w image, delineating 6 hippocampal subfields and entorhinal cortex (ERC). In the training leave-one-out joint label fusion was utilized, running through all atlases, and for each obtained automatic segmentation, this was compared to the corresponding manually generated segmentation. A classifier was then trained to find systematic patterns of errors between the two segmentations. The atlases and the parameters of the trained classifier were utilized as input for the segmentation pipeline, which was used for segmentation of a new, unlabelled subject. Joint label fusion was employed in order to obtain a multi-atlas segmentation for the unlabelled subject. Subsequently, the trained classifiers were applied for correction of the segmentation. [Yushkevich et al., 2015; Wisse et al., 2016] ASHS enables new users to apply their own atlas and retrain the software. Wisse et al. validated ASHS using 7T data. Wisse et al. compared manual and automatic segmentations, and achieved the best segmentation with a generalized dice score at 0,85, whilst the worst segmentation had a generalized dice score at 0,75. Figure 2.5 illustrates a segmentation result with a median generalized dice score (0.80). The validation showed that ASHS generated high accuracy results for larger subfields (CA1, DG and Sub), whilst it had difficulties with the smaller subfields (CA2, CA3) and ERC. However, the authors concludes these errors to be comparable to disagreements between manual raters. ASHS required >24 hours for segmentation of one new subject on a single core cpu. [Wisse et al., 2016]

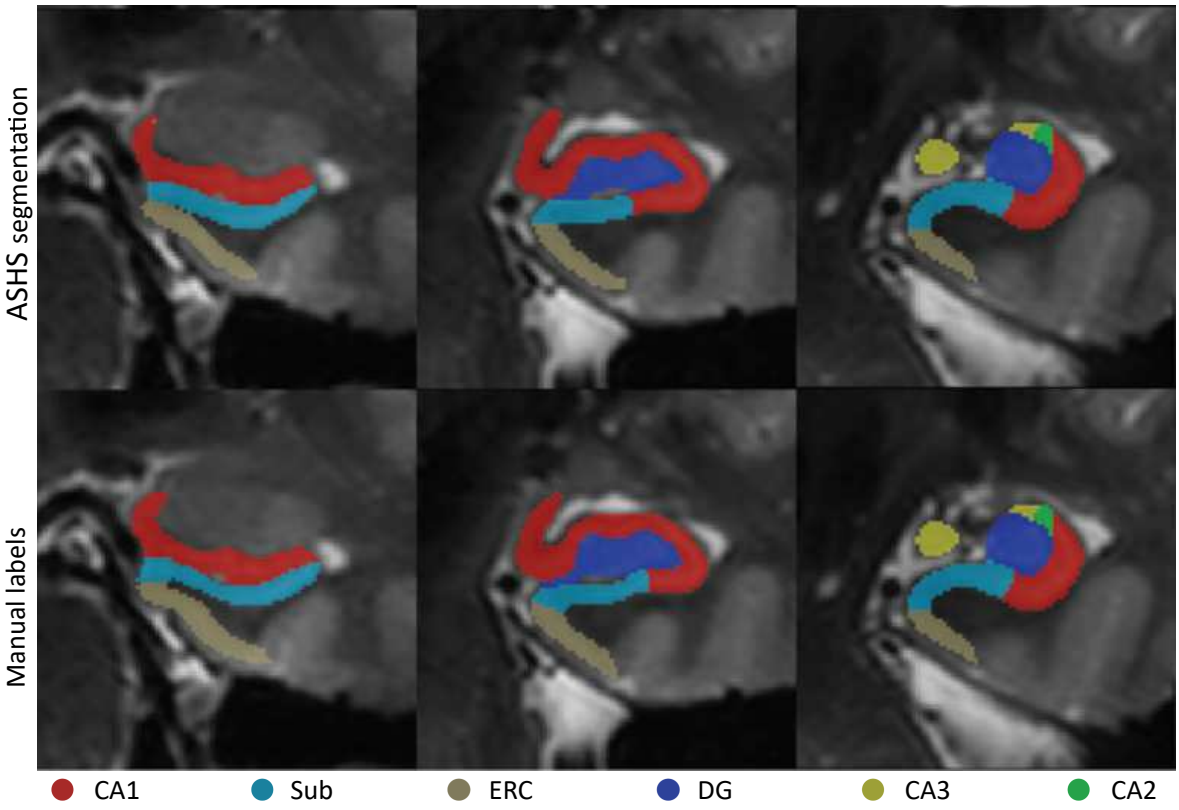


Figure 2.5: Three coronal slices with the associated automatic and manual hippocampal subfield segmentations from ASHS. Results illustrating the median performance of ASHS. Dice score was at 0.80. 6 out of 7 subfields was detected in present views. Modified from Wisse et al. [2016].

The aforementioned methods are atlas based and when compared to manual segmentation methods, the atlas based methods show promising results for the larger subfields, but for the smaller structures the methods show limitations [Wisse et al., 2017; Iglesias et al., 2015; Giuliano et al., 2017]. FreeSurfer has been criticised for being ungeneralisable to younger populations, since its atlas was constructed from an elderly mortal population. Moreover, even though ASHS is able to be re-trained with new datasets, and thereby generalisable to various populations, it does not contain the ability to include other contrasts than T2w an T1w MR, like susceptibility weighted imaging (SWI) or quantitative susceptibility mapping (QSM), for which reason, it does not exploit the full potential of ultra-high field imaging. [Giuliano et al., 2017]

2.3 Convolutional Neural Network

In research, neural networks are getting more popular and have been proposed as an alternative approach for segmentation problems, being able to include a wide number of contrasts and modalities. Especially Convolutional Neural Networks (CNN) have shown promising results in segmentation of brain structures and lesions in biomedical imaging. The CNNs presented in the literature differ in architecture and training strategy, but they all follow some basic characteristics. [Choi and Jin, 2016; Kamnitsas et al., 2017; Nie et al., 2016; Rajchl et al., 2017; Kleesiek et al., 2016; Havaei et al., 2017; Moeskops et al., 2016; Pereira et al., 2016] In this Section are these characteristics elaborated.

2.3.1 Convolutional Neural Network architecture

A CNN architecture contains at least one sequence of layers, and some of the basic layers are; convolutional layers, fully-connected layers, and a classification layer, as illustrated in figure 2.6. [Gu et al., 2017; LeCun et al., 2015]

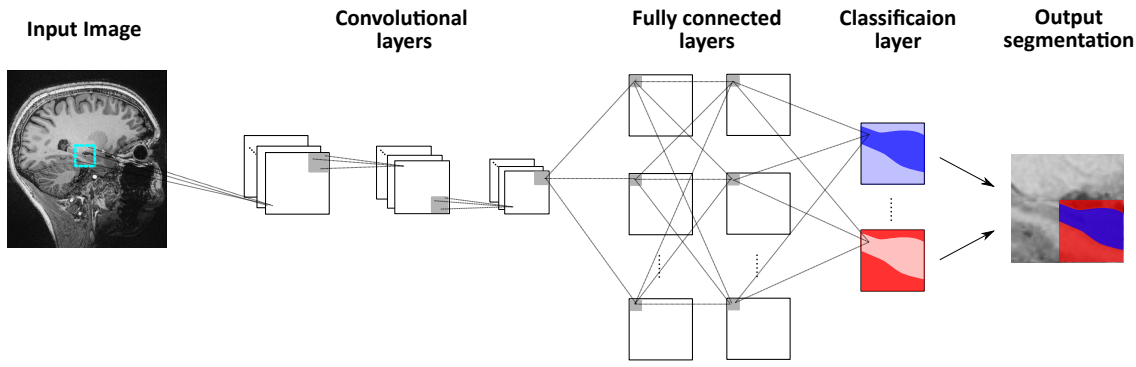


Figure 2.6: Illustration of a convolutional neural network with three types of layers; convolutiolan layer, fully-connected layers, and a classification layer. Modified from Mamoshina et al. [2016].

Typically, the first layers in a CNN are convolutional layers, which are used to extract features. These layers are typically locally connected.[LeCun et al., 2015] When a layer is locally connected each node in the layer is connected to a single node in the previous layer [Vieira et al., 2017]. In each node of a convolutional layer a feature map is created using convolution and a filter. [LeCun et al., 2015] Fully-connected layers often follow the convolutional layers. These layers merge the information from the extracted features to find global information. [LeCun et al., 2015] The final layer in a CNN is a classification layer classifying each of the pixels/voxels. [Gu et al., 2017] Each layer contain several nodes. The

nodes of a layer is connected to the nodes of a previous layer through weights, referred to as filter banks. In each node a weighted sum is calculated from the inputs from the previous layer and the result is passed through a non-linear function, called an activation function. [LeCun et al., 2015]

2.3.2 Training Convolutional Neural Network

CNNs need to be trained to identify the weights in the network. The training is a global optimization problem. Thereby an error function needs to be minimized to find the optimal set of weights. [Gu et al., 2017] A common method used to train the networks is backpropagation. The first step in this approach is to initialize all unknown weights to a random value. Subsequently, a training image is passed through the network and an output is calculated. This step is called forward propagation. Since the output is calculated based on random parameters the output will also be random. To quantify the total error of the output in relation to the real output an error function is used. Next, backprobagation starts, and this procedure works backwards calculating the gradients of the error in relation to the weights in all the layers of the network starting from the output layer to the input layer. Based on their contribution to the total error a new set of weights are calculated to reduce the output error. The forward propagation and backpropagation is repeated for all the training images to find the set of weights most suitable for this training set.

When a network is trained it can be too closely fitted to the training set resulting in overfitting, and this might cause poor classification of the test data [Arbabshirani et al., 2016]. Overfitting is especially a problem in neuro-imaging, where the number of voxels in each image often is higher than the number of training sets. [Arbabshirani et al., 2016] Overfitting can be prevented using regularization methods [Gu et al., 2017].

2.4 Segmenting brain structures and lesions using Convolutional Neural Networks

Segmentation problems involving structures and lesions in the brain have recently been approached using CNNs [LeCun et al., 2015; Choi and Jin, 2016]. The CNNs presented in the literature aim to segment; striatum [Choi and Jin, 2016], lesions[Kamnitsas et al., 2017], haemorrhage [Kamnitsas et al., 2017], tumors [Kamnitsas et al., 2017; Havaei et al., 2017; Pereira et al., 2016], infant brain structures [Nie et al., 2016], and different tissues [Moeskops et al., 2016]. Table 2.4 presents an overview of the CNNs and contains; study reference, objective of the study, data used, and architecture of the CNN.

Table 2.2: State-of-the-art methods for segmentation of brain structures based on CNNs. The table content provides a coarse overview of the CNNs objective, applied data, and architecture.

Study	Objective	Data	Architecture of neural network
Choi and Jin [2016]	Segmentation of striatum	The data set contained 20 T1w MRI scans of healthy subjects. 15 scans were used for training and validation. 5 scans were used for test.	A global network was used to find the rough location of the Striatum and chopping the original image. The chopped output image of the global network was given as input to a local network, which found the detailed structures of striatum. Both networks were build of six 3D convolutional layers and the final layer for classification was a softmax layer (striatum or background).

Kamnitsas et al. [2017]	Segmentation of brain haemorrhage, lesion and tumour	Trained and tested on 3 different datasets one for each disease, with 274>N>28. In all three cases multiple MRI contrasts were used.	Dual pathway, 11-layers 3D CNN. The two pathways were used to gain both local and larger contextual information, and they are composed of convolutional layers. The pathways were merged and followed by, fully connected layers, a softmax layer, and lastly a CRF layer. Code available online.
Nie et al. [2016]	Segmentation of gray matter, white matter, and cerebrospinal fluid in images of infants brain	A small dataset of T1, T2, and fractional anisotropy images were used for training and validation.	For each modality a CNN was made. The CNNs were composed of three convolutional layer groups followed by two de-convolutional layers. The final layer was a softmax unit, classifying the input to one of the four categories. The CNN was trained for each of the modalities. Finally, the high-layer features from three CNNs were fused together to make the output segmentation.
Kleesiek et al. [2016]	Discriminate between non-brain and brain tissue in images from clinical routine and with several modalities.	T1 images of healthy subjects from three databases (N=135) were used for testing. Additionally, one multimodal (non-enhanced and contrast enhanced T1, T2 and T2-FLAIR) dataset (N=53) with recurrent brain tumours was used for testing and training.	3D CNN consisting of seven convolutional layers. Max-pooling was applied after the first convolutional layer. The final layer was a softmax layer, which classified if a voxel was brain or non-brain tissue.
Rajchl et al. [2017]	Segmentation of objects from bounding box annotations	For testing T2 MR scans (1.5T) acquired from 55 fetal subjects (30 healthy and 25 with intrauterine growth restriction) were used. For training an equal amount of patches (10^5) from a training database was used for each class.	The CNN was composed of two convolutional layers both followed by a max-pooling layer. Next was a fully-connected layer followed by the output layer making the classification to foreground or background. DeepCut was used as the optimisation approach.
Havaei et al. [2017]	Brain tumour segmentation	T1, T1C, T2 and Flair scans from 65 patients with different grade of tumour.	Two pathway CNN; a local pathway with a small receptive field and a global with a large receptive field. Both the local and global pathways were composed of convolutional layers, pooling layers, the activation function max-out, and a softmax output layer.
Moeskops et al. [2016]	Tissue classification and segmentation of five healthy groups ranging from infants to adults (6-8 tissue classes)	5-20 subjects in each test class (neonatal, young adults and adults), T2 or T1w, 3T or 1,5T MR scans.	CNN with three pathways all composed of three convolutional layers followed by a fully connected layer. After the fully connected layers the three pathways were connected to a softmax layer for the classification. For two of the three pathways max-pooling was used after the convolutional layers.
Pereira et al. [2016]	Segmentation of two types of brain tumors: Low Grade Gliomas and High Grade Gliomas	T1, T1 with contrast, T2 and FLAIR scans of 65 different subjects.	Two CNNs were build, one for each of the tumor types. The CNNs got 4 images as input. One of the CNN's was composed of 6 convolutional layers, where layer three and five were followed by max pooling and the last three layers were fully connected layers. The other CNN was composed of 9 layers. The structure was the same as for the previous described CNN except for the removal of two convolutional layers.

2.4.1 Architecture of the state-of-the-art CNNs

The architecture of the state-of-the-art CNNs presented in table 2.4 differ. Most of the CNNs employed a 2D architecture [Nie et al., 2016; Rajchl et al., 2017; Havaei et al., 2017; Moeskops et al., 2016; Pereira et al., 2016] whereas Choi and Jin, Kamnitsas et al., and Kleesiek et al. proposed 3D architectures [Choi and Jin, 2016; Kamnitsas et al., 2017; Kleesiek et al., 2016]. The advantage of the 3D CNN architecture is the inclusion of 3D contextual information. However, compared to the 2D CNNs the 3D CNNs have an increased number of parameters and require significant memory. Additionally, the computational requirements of the 3D CNNs are larger due to the computationally expensive 3D convolutions. There are different approaches to overcome these issues, e.g. Kamnitsas et al. applied a technique called dense-inference, which reduce inference time by avoiding repeated convolutions on the same voxels. [Kamnitsas et al., 2017]
Some of the CNNs are structured conventionally with a single sequence of layers [Pereira et al., 2016; Kleesiek et al., 2016; Rajchl et al., 2017], as described in Section 2.3.1, while other CNNs are made of multiple pathways [Kamnitsas et al., 2017; Nie et al., 2016; Havaei et al., 2017; Moeskops et al., 2016] or a serial of networks [Choi and Jin, 2016]. The reasoning behind using multiple pathways was often to archive different feature levels [Kamnitsas et al., 2017; Havaei et al., 2017; Moeskops et al., 2016; Choi and Jin, 2016]. Kamnitsas et al. and Havaei et al. both proposed a dual pathway CNN, with the purpose of getting detailed features from one of the pathways, and obtain higher level features (e.g. location) from the other pathway [Kamnitsas et al., 2017; Havaei et al., 2017]. Kamnitsas et al. obtained detailed features by feeding one of the pathways a MRI segment with the original resolution, and the higher level features were obtained using a down-sampled version of the same MRI segment for the other pathway [Kamnitsas et al., 2017]. Havaei et al. achieved the different levels of features by applying different receptive fields. To obtain detailed features a 7 x 7 receptive field was used in one of the pathways, whereas a 17 x 17 receptive filed was used in the pathway obtaining the higher level features [Havaei et al., 2017]. Moeskops et al. also proposed a multiple pathway CNN to obtain different levels of features [Moeskops et al., 2016]. Just as Havaei et al., receptive fields of different size were applied. However, in this case a three pathway CNN was suggested with receptive fields of 25 x 25, 51 x 51, and 75 x 75 voxels [Havaei et al., 2017]. Similarly, Choi and Jin exploited information from features of different levels. In this approach a two serial CNN was build consisting of a global and a local CNN. First the global CNN found the approximate location of the striatum and based on the result the input image was cropped. Subsequently, the cropped image was fed to the local CNN where a more detailed structure was segmented. The purpose of the two serial CNN was to reduce the computational burden and avoid segmentation of brain structures similar to striatum in the whole images. [Choi and Jin, 2016] Nie et al. made a three pathway network, one pathway for each of the modalities they found interesting. The pathways were trained separately. Thereby, biases and weights were specially optimized for each modality. [Nie et al., 2016]

The state-of-the-art CNNs were composed of a different number of layers and features, but convolutional layers and fully connected layers were always applied. Pooling layers were frequently applied [Nie et al., 2016; Kleesiek et al., 2016; Rajchl et al., 2017; Havaei et al., 2017; Moeskops et al., 2016; Pereira et al., 2016] and the output layer translating the networks' output to classification predictions was in most cases a softmax output layer [Choi and Jin, 2016; Kamnitsas et al., 2017; Nie et al., 2016; Kleesiek et al., 2016; Havaei et al., 2017; Moeskops et al., 2016; Pereira et al., 2016]. The softmax layer was used for two class problems [Choi and Jin, 2016; Kleesiek et al., 2016; Rajchl et al., 2017; Kamnitsas et al., 2017] and multi-class ($4 \geq classes \leq 9$) problems [Nie et al., 2016; Havaei et al., 2017; Pereira et al., 2016;

Moeskops et al., 2016; Kamnitsas et al., 2017]. In order to optimize the predicted classification both Kamnitsas et al. and Rajchl et al. applied a Conditional Random Field (CRF) algorithm as post processing [Kamnitsas et al., 2017; Rajchl et al., 2017]. This algorithm incorporates pixel/voxel neighbourhood pattens for correction of the classification predictions from the softmax layer [Sutton and McCallum, 2012]. In practise, this improves the fineness of the segmentation [Zheng et al., 2016]. The most frequently used activation function was Rectified Linear unit (ReLU) [Choi and Jin, 2016; Nie et al., 2016; Kleesiek et al., 2016; Rajchl et al., 2017; Moeskops et al., 2016; Pereira et al., 2016] but also Parametric Rectified Linear Unit (PReLU) [Kamnitsas et al., 2017] and max-out was used [Havaei et al., 2017].

2.4.2 Training of the state-of-the-art CNNs

Training of a CNN is an optimization problem, where a loss function is used to fit the CNN to the training data (see Section 2.3.2) [Gu et al., 2017]. The loss functions used for training of the state-of-the-art CNNs were in most cases based on cross entropy [Kamnitsas et al., 2017; Moeskops et al., 2016], or a variant of cross entropy such as the categorical cross entropy [Rajchl et al., 2017; Pereira et al., 2016] or Kullback-Leibler [Kleesiek et al., 2016]. Choi and Jin suggested a more simple loss function based on mean square error [Choi and Jin, 2016]. For the training Nie et al. applied an already developed method called Caffe [Nie et al., 2016]. Caffe is a deep learning framework, which can be applied for training, testing, fine-tuning, and deploying of CNNs [Jia et al., 2014].

To avoid overfitting different regularization methods can be applied. One regularization method is application of more training data than parameters in the model. If only a small dataset is available, other regularization methods can be used. These include among other early stopping of the training before the parameters are fitted 100 percent to the training data, reducing the number of parameters by shared weights, use of a validation set to find the optimal time to stop the training [Nowlan and Hinton, 1992], L1 and L2 regularization [Kamnitsas et al., 2017], and dropouts [Srivastava et al., 2014]. In the dropout method random nodes are removed in each training session, and consequently the network is not dependent on the specific features describing each of the training set. [Srivastava et al., 2014]

The regularization method applied by the state-of-the-art CNNs was in most cases dropout [Choi and Jin, 2016; Kamnitsas et al., 2017; Rajchl et al., 2017; Havaei et al., 2017; Moeskops et al., 2016; Pereira et al., 2016]. In adition to dropouts, Kamnitsas et al. [2017] used L1 and L2 regularization. In some cases the regularisation method was not described [Nie et al., 2016; Kleesiek et al., 2016].

2.4.3 Factors influencing segmentation performance

Several factors influence the segmentation performance of a CNN, and one factor is network architecture. Havaei et al. investigated if a 3D architecture would improve the segmentation performance in comparison with a 2D architecture. The performance did not improve when using a 3D architecture, but the computation time increased significantly [Havaei et al., 2017]. In contrary, when Kamnitsas et al. applied a 3D network the average dice score of the segmentation performance increased from 61,5 % to 66,6 % [Kamnitsas et al., 2017].

CNN architectures including both a local and a global pathway result in better segmentations than single pathway architectures, and joint training of the two pathways lead to better results compared to separate training of the two pathways [Havaei et al., 2017]. Kamnitsas et al. argue, that the advantage of a two pathway architecture is the down-sampled pathway which finds the global structures allowing the other pathway to focus on detailed patterns associated with ambiguous areas and fine structures [Kamnitsas et al., 2017].

Another architectural factor influencing the performance is the size of the kernels. Pereira et al. suggest a deeper network with smaller kernels, since this architecture has shown to perform better in more cases compared to a less deep network with larger kernels. A network applying smaller kernels, typically with a size of 3x3x3, has fewer weights and allows utilization of a deeper network architecture. A more non linear segmentation is thereby possible. [Kamnitsas et al., 2017; Pereira et al., 2016] Kamnitsas et al. and Rajchl et al. experienced increased performance when adding a CRF layer for post processing [Kamnitsas et al., 2017; Rajchl et al., 2017]. Small architecture details might not be important for the performance of a CNN, since the CNN optimize the weights and biases by itself [Moeskops et al., 2016].

A second factor influencing the performance of CNNs are the choice of preprocessing. Pereira et al. investigated the influence of several factors such as preprocessing, data augmentation, activation functions and the effect of smaller kernels in deeper networks, and the factor increasing the accuracy of the CNN the most was preprocessing [Pereira et al., 2016]. Kamnitsas et al. also suggested preprocessing to be an important factor, which could increase the performance of their developed CNN [Kamnitsas et al., 2017]. Pereira et al. suggested to use an intensity normalization method for preprocessing to overcome variation in the intensities of the same tissues across acquisitions and subjects [Pereira et al., 2016].

A third factor influencing the performance of a CNN is the data set applied for training [Kleesiek et al., 2016]. Havaei et al. had issues using manually segmented training data due to large variations in the manual segmentations [Havaei et al., 2017]. Havaei et al. and Kleesiek et al. suggested to train with segmentations from several segmentation methods [Havaei et al., 2017; Kleesiek et al., 2016]. Kamnitsas et al. observed a drop in the performance of the their system, when the data being segmented was clinical data from databases different from the training data set, and they suggested to allow for this in a data augmentation step [Kamnitsas et al., 2017]. Moeskops et al. suggested to use more training data with higher diversity to obtain a more generalized CNN, but they also considered this might result in lower performance of the segmentations [Moeskops et al., 2016]. Pereira et al. also suggested the size of the training data set to be an important factor in overfitting, and suggested that a larger dataset can reduce overfitting. Data augmentation can be obtained by rotating an original patch to generate new patches [Pereira et al., 2016].

2.4.4 Evaluation of the state-of-the-art CNNs

The CNNs presented in Table 2.4 obtained a segmentation performance with a better or similar accuracy compared to conventional manual methods [Choi and Jin, 2016; Kamnitsas et al., 2017; Rajchl et al., 2017] and conventional automatic methods such as FreeSurfer [Choi and Jin, 2016; Kleesiek et al., 2016; Rajchl et al., 2017; Havaei et al., 2017]. In comparison with the conventional methods CNNs have the advantage, that features are automatically learned and consequently manual feature engineering is avoided [Rajchl et al., 2017; Kleesiek et al., 2016]. Additionally, the same CNN can have the ability to fit different types of problems, i.e Rajchl et al. utilized the same network for both segmentation of premature lung and brain structures [Rajchl et al., 2017; Choi and Jin, 2016; Kamnitsas et al., 2017; Rajchl et al., 2017]. CNNs require less computational time for one segmentation than conventional methods e.g. the computation time using the CNN developed by Choi and Jin was approximately 1.5 min applying a CPU and only a few seconds when a GPU was applied, whereas the computation time was approximately 10 h using FreeSurfer. [Choi and Jin, 2016; Kamnitsas et al., 2017; Rajchl et al., 2017]. One of the disadvantages of CNNs are that the training process is computationally heavy and time consuming. However, several parameters in the architecture and preprocessing can reduce the required training time. These include e.g. smaller kernels

[Kamnitsas et al., 2017; Pereira et al., 2016], reduced search area [Rajchl et al., 2017; Moeskops et al., 2016], implementation on GPU [Havaei et al., 2017], and usage of 2D architectures instead of 3D architectures [Moeskops et al., 2016], normalization of input data, and data augmentation [Pereira et al., 2016].
In conclusion; throughout the presented papers a range of different proposals for how the CNNs should be developed, trained, and tested have been presented. However, the state-of the-art CNNs also have some similar features and they all show promising results and form a basis for further development.

Chapter 3
Project aim

Dementia is a world wide problem, with almost 50 million people suffering from the disease. No cure for dementia exists, but preventive treatment can postpone impairment. To exploit the treatment best, the patient should be treated as early as possible, which requires early diagnosis. For AD patients, changes in the hippocampi occur approximately 5.5 years before a diagnosis is given, and these changes are i.e. present as a volume decrease of different subfields within hippocampus, among these CA1, CA2, and subiculum. During the last decade 7T MRI has made it possible to observe these subfields in vivo, which opens the opportunity of earlier AD diagnosis.

Several protocols for manual segmentation of hippocampal subfields in 7T MR images have been developed, however, the diverseness of the protocols induce high variation between the segmentation results obtained through studies, and the method is highly resource demanding. The time and labour requiring perspectives in manual segmentation have resulted in several attempts to develop automatic segmentation methods for hippocampal subfields. These automatic methods show promising results, but have problems identifying smaller subfields accurately and are very time demanding.

In current research, Convolutional Neural Networks (CNNs) have shown accurate and fast segmentations of brain structures and lesions in MR images compared to manual and automatic atlas based segmentation methods. A segmentation method based on a CNN is not dependent on manual feature modelling nor complicated image processing such as non-linear registration, as it learns the relevant image features automatically through convolutions. Moreover, it has been demonstrated, that one CNN can be used for several segmentation problems.

Aim

The aim of this study is to develop an automatic method for hippocampal subfield segmentation (DeepHSS) in in vivo ultra-high resolution MR images, based on a convolutional neural network. Furthermore, the potential of this approach is explored in relation to existing methods for hippocampal subfield segmentation.

- Determine optimal configurations for the CNN applied in DeepHSS.
- Investigate if automatic segmentation methods can be applied to generate ground truth labels and used to train the CNN in DeepHSS.
- Validate hippocampal subfield segmentations obtained using DeepHSS and compare to segmentations obtained using excising methods.

3.1 Scope

The proposed automatic method for hippocampal subfield segmentation is named DeepHSS. The method was developed mainly to segment the subfields within hippocampus prober and subiculum, since these subfields according to literature are first affected in AD (see Section 2.1.1), and results were compared to results obtained using the exciting methods ASHS and FreeSurfer (see Section 2.2). The main component of DeepHSS is a supervised CNN,

for which reason a dataset containing labelled single subject 7T MR images was needed for training. Since, manual ground truth (GT) labels were not available in the initial stages of DeepHSS development, these had to be generated automatically. It was decided to use one of the aforementioned open source segmentation tools (see Section 2.2) for generation of GT labels. To provide best possible conditions, the segmentation was performed on 7T MDA models with both T1w and T2w contrast, and subsequently warped into subject space. Later in the development of DeepHSS manual GT labels became available and were applied in the final stages of DeepHSS development.

DeepHSS was developed using the public available CNN framework DeepMedic, described by Kamnitsas et al. [2017]. DeepMedic is a well structured network, which allows creation of CNNs for segmentation of structures through a set of configuration files. The development of the CNN within DeepHSS required considerable memory and had to be run on a GPU. For this reason, the supercomputer Abacus was utilized as computational tool. Hardware and software specifications are described at: *https://abacus.deic.dk/setup/hardware*.

The following chapters contain a description of DeepHSS. Chapter 4 is documentation of the method behind DeepHSS. Chapter 5 contains initial tests of DeepHSS conducted through the development process, and Chapter 6 contains the final test of DeepHSS. Two different datasets were used for the initial tests and the final test due to known availability at the time. Both datasets are described within the chapter of the corresponding test.

Chapter 4
Proposed solution: DeepHSS

This Chapter contains a description of DeepHSS, which was developed for hippocampal subfield segmentation in ultra-high resolution MRI. The description covers preprocessing of the data, network architecture, and network training schedule. An overview of the chapter can be seen in figure 4.1.

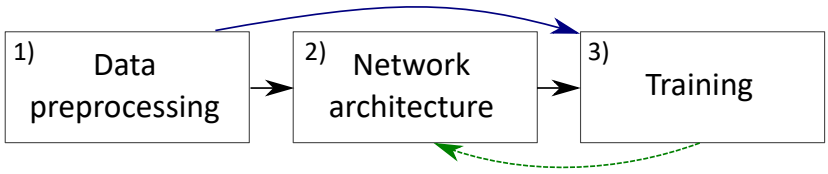


Figure 4.1: Flowchart illustrating the development of DeepHSS. The input data is firstly preprocessed and fed to step three, training. Secondly, the network architecture is build. Thirdly the network's weights are determined through an iterative training process, which is denoted by the green arrow.

In the first step of DeepHSS, the data is preprocessed in order to standardise the data and reduce runtime. In the second step, the network architecture is set up. As described in Section 3.1, the proposed network is based on the open source CNN framework, DeepMedic. In step three of the network development, the build network's weights are determined through an iterative training process, denoted by the green arrow in Figure 4.1.

4.1 Data preprocessing

The data was preprocessed through three steps: Firstly, whole brain masks were generated using the Brain Extraction Tool (BET2) [Jenkinson et al., 2015] and applied to the 7T MR images using the tool mincmath in order to reduce the amount of excess image information, and thereby make the network computational lighter.

Subsequently, to deal with covariate shifts the input MR images were normalized to have zero mean and unit variance using MATLAB. Covariate shifts, which are changes in the distribution (e.g mean, standard division) of an image, can occur to images processed in neural networks if the network is fed images with different distributions. When covariate shifts occur this can lead to a reduced inference of the network [Shimodaira, 2000].

Finally, the 7T MR images and their respective GT label images were bisected by removal of the image part corresponding to left brain hemisphere. Bisection was performed to avoid unwanted segmentation due to symmetric properties of the brain. This issue was discovered and is described through the initial tests of DeepHSS, described in Chapter 5. The bisection was performed using the tool mincreshape.

4.2 Network architecture

The CNN in DeepHSS is build as a 3D network with two pathways both containing eight convolutional layers. The convolutional layers are followed by two fully connected layers and a classification layer developed to classify the voxels into nine classes. The classifications are finally post-processed by a CRF layer, using the GT labels and neighbourhood relations. The network architecture is illustrated in figure 4.2.

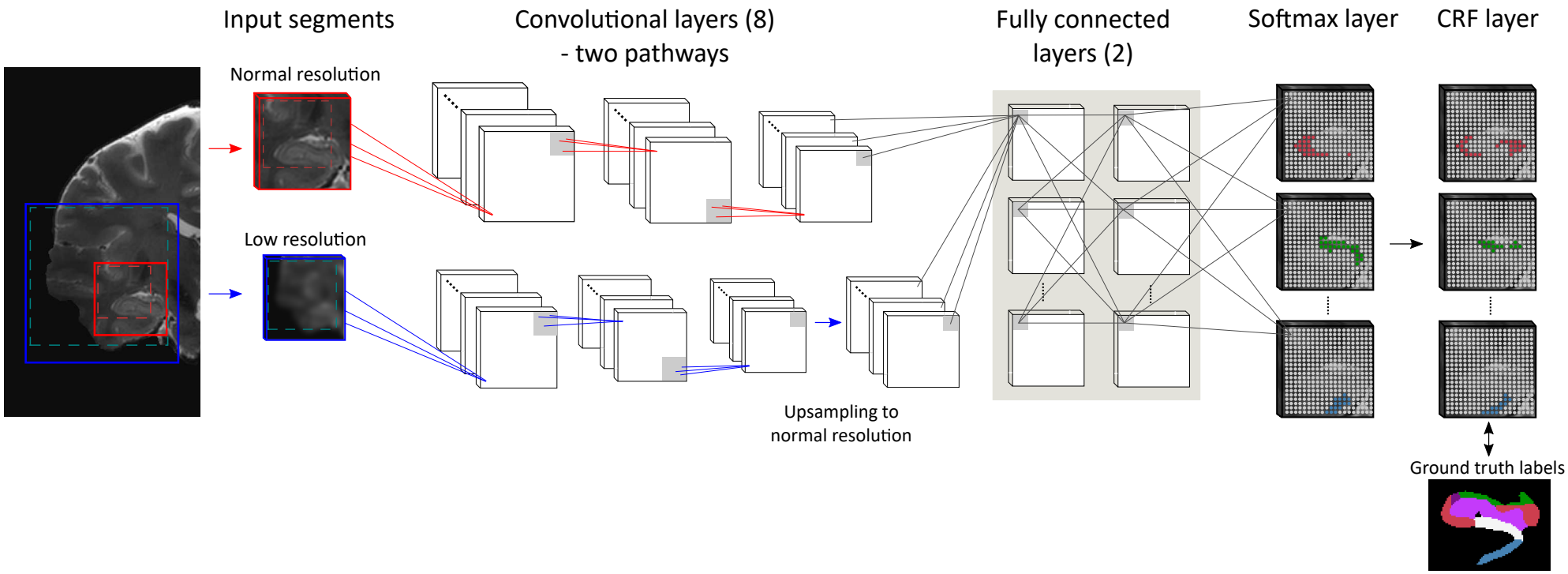


Figure 4.2: Schematic illustration of the 3D CNN imbedded in DeepHSS. The input to the network is 7T TSE MRI volumes, from where input segments are extracted. Segments of full resolution are fed into the upper pathway of the network, whereas downsampled segments are fed to the lower pathway of the system. After eight convolutoonal layers, the data is processed trough two fully connected layers and classified by a softmax layer. The classification is finally corrected with a CRF layer at the end of the network, using the GT labels and neighbourhood relations. Inspired by Kamnitsas et al. [2017].

The CNN architecture is divided into five parts; Overall network configurations, convolutional layers, fully connected layers, classification layers, and activation functions. All of these are elaborated in the following sections.

4.2.1 Overall network configurations

With the objective of building a network, which has the potential of gaining the highest performance possible, the CNN in DeepHSS was build with a 3D architecture. As described in Section 2.4.4, Havaei et al. and Kamnitsas et al. dissagree whether the benefits of a 3D network make up for the increased computation time compared to a 2D network. However, since deepMedic form the base for the CNN embedded in DeepHSS, and Kamnitsas et al. [2017] found DeepMedic to gain higher performance with a 3D architecture, a 3D architecture was chosen. The CNN in DeepHSS was build using a two-pathway approach. This approach was picked with the purpose of incorporating both coarse and fine details of the input images. A two-pathway approach has, as stated in Section 2.4.4, been found to be beneficial for localization of the region of interest. Since a CNN, which is able to handle half brain images, is desired in order to avoid image dependent, unnecessary preprocessing, the two-pathway approach was seen as useful for hippocampal subfield segmentation. The two pathways were designed to be similar, however, with inspiration from Kamnitsas et al., the input to the sub-sampled pathway was down sampled by three.

4.2.2 Convolutional layers

The pathways in DeepHSS were both composed of convolutional layers. The first convolutional layer in each pathway was build to handle images preprocessed as described in Section 4.1. As described in section 2.4.4, it is beneficial to use small kernel sizes and deeper networks when employing a 3D architecture. Consequently, the size of the kernels were chosen to be 3x3x3, and the number of convolutional layers was chosen to be eight. The number of kernels per layer was chosen to be [30 30 40 40 40 40 50 50]. The network was build with tree residual layers; 4, 6, and 8. These specific values were selected with inspiration from Kamnitsas et al..

4.2.3 Fully connected layers

The convolutional layers were followed by two fully connected layers, which had the purpose of combining all prior information. The fully connected layers were implemented as a linear classifier, which performs global reasoning and generates global semantic information [Gu et al., 2017]. With inspiration from Kamnitsas et al. the CNN implemented in DeepHSS was implemented with 150 kernels of size 3x3x3 in each fully-connected layer.

4.2.4 Classification layers

In order to translate the output from the fully connected layers to a statistical classification of each voxel in the input image, a softmax layer was implemented. As presented in Section 2.4, several of the state-of-the-art CNNs contain a softmax layer as the classification layer, for which reason this exact type of classifier was chosen.

For post-processing, a CRF layer was implemented as the final layer in the CNN. As stated in Section 2.4, a CRF layer can be beneficial since it is able to refine coarse and weak voxel-level label predictions, and therefore corrects the initial classification and effectively improves the fineness of the segmentation.

4.2.5 Activation function

The activation function applied in all nodes of the layers were a rectifier of the type PReLU. This function is a refinement of the ReLU function, where all negative values are multiplied by a learnable parameter. PReLU was chosen, since it has shown reduced errors compared with ReLU. Additionally, networks applying rectifiers are in general easier to train compared to networks using sigmoid-like activation functions. [He et al.]

As described in Section 4.1, covariate shifts can occur when the input images to a network have different distributions. However, the input from one layer to another can vary too, which induce internal covariate shifts. [Ioffe and Szegedy, 2015] To handle internal covariate shifts a method called batch normalization, proposed by Ioffe and Szegedy [2015], was applied to all hidden layers of the network. When applying batch normalization the input to the activation functions in the network is normalized by whitening it to have zero mean and unit variance. This is done in mini-batches, as it would be impractical to do it for the whole layer. [Ioffe and Szegedy, 2015]

4.3 Network training

The CNN in DeepHSS was trained using the training method dense training, which was proposed by Kamnitsas et al. [2017]. The first step in the training is initialisation of the network's weights. Next, the weights are determined by using an optimization algorithm to minimize a cost function, as described in Section 2.3. Due to the overall training method, dense training, each epoch of the optimization algorithm contains a number of subepochs, in which image segments are randomly extracted from the training images. The segments are fed to the network, which classify each voxel. Subsequently, the cost function calculates the cost between the CNN's classifications and GT labels. When the training has iterated through the subepochs in one epoch an overall cost for the epoch is calculated. Next, the learning rate is lowered and the CNNs weights are updated utilizing the optimization algorithm. The training also includes regularization to avoid overfitting. An illustration of the training process is presented in Figure 4.3.

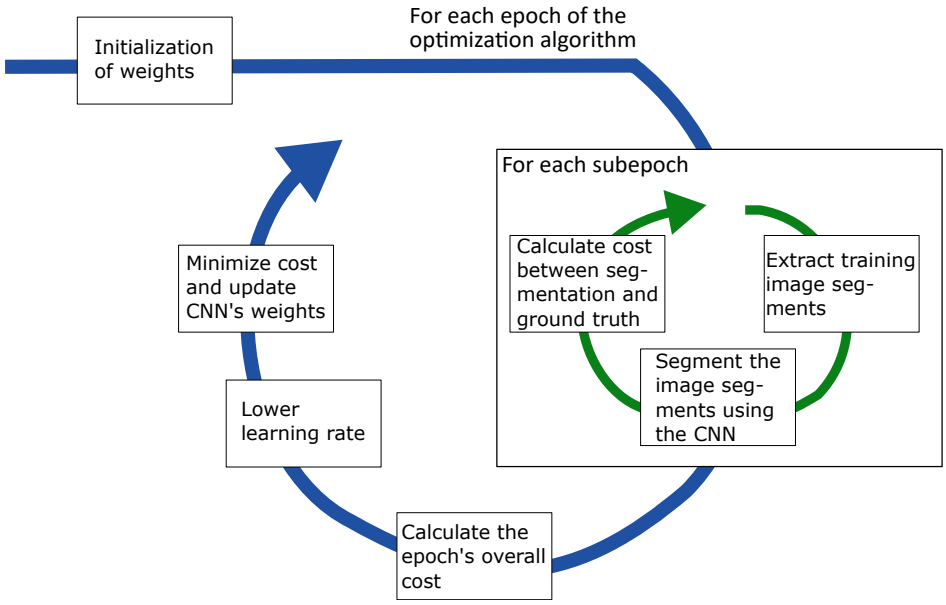


Figure 4.3: Broad overview of the training cycle. Each training epoch is composed of several subepochs, in which several segments are subtracted from the training images. The CNN makes a classification of the image segments, and a cost function is used to calculate the cost between the CNN's classification and the GT labels. Next, the overall cost is calculated for the whole epoch, the learning rate is lowered, the cost function is minimized, and the network's weights are updated.

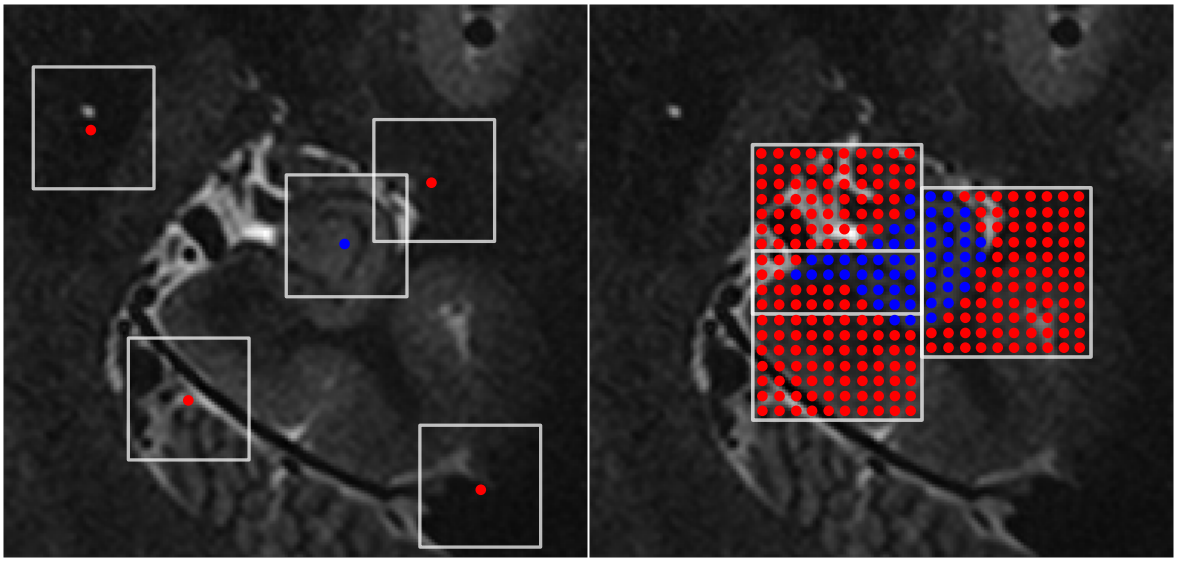
In this Section the components composing the training is elaborated.

4.3.1 Training approach: Dense training

The training method dense training is an alternative approach to the common training method, called patch-wise training. The main difference between the two methods is how the image is fed to the network. In the patch-wise training approach, the images are divided into patches, which have the size of the CNN's receptive field. All patches are extracted randomly from the training images to form a batch. In one training epoch the whole batch is processed, and the output is a prediction of the center voxel of each patch. The principal behind patch-wise training is illustrated in Figure 4.4a. However, when handling large 3D networks, the patch-by-patch method becomes computationally heavy, especially since large batches are desired to increase the accuracy of the predictions. This issue is met by applying dense inference, which is the base of dense training. [Kamnitsas et al., 2017]

Dense inference is a technique, which can be applied to fully-convolutional networks when the size of the input to the network is larger than the size of the CNN's receptive field. When this is the case, the dimensions of the feature maps increase and likewise does the dimensions of the classification feature maps, and for each map an output will be generated for each stride of the receptive field of the input. Thereby, a prediction of several voxels positioned next to each other in the image can be obtained simultaneously. Using this technique, repeated convolutional calculations of the same voxel in overlapping patches are avoided, as would be the case using the patch-wise training method, and this reduces memory loads and calculation costs. The most optimal performance of dense inference is when the whole image is processed in one forward pass. However, this is computational expensive especially for large 3D networks and large datasets. To address this issue multiple image segments are extracted and used as input. To ensure the representation of the region of interest, in this case each hippocampal subfield, the segments are extracted with the content being with 50% probability of belonging to either background or foreground, foreground being one of the hippocampal subfields, as

illustrated in figure 4.4b. In multi-class problems, like this, the frequency of which the foreground's classes are extracted comply with real distribution of the classes relative to the background. [Kamnitsas et al., 2017]



(a) (b)

Figure 4.4: Figure (a) and (b) are illustrations of how a 7T TSE MR image is fed differently to a network dependent on the training type. (a) is an example of five randomly extracted patches, which are applied as input for patch-wise training. (b) shows three image segments used as input to a network when dense training is applied.

4.3.2 Weight initialisation

Proper initialisation of the weights in a network is an important factor for convergence of the weights. Deep networks are often initialized by attributing the network's weights a random value from a gaussian distribution which has a fixed standard deviation. However, when this initialization method is applied to deep networks they have difficulties converging. To overcome this issue some authors have developed initialization methods taking variations in standard distribution into account. [He et al.] developed an initialization method, which base the initialization of weights on the variance of the response for each layer. [He et al.] This method shows good results for networks using non linear activation functions [He et al.], which are the type of activation function applied in DeepHSS's CNN (see 4.2). Consequently, this method was used for initialization of the CNN's weights.

4.3.3 Optimisation algorithm

It was chosen to use an optimization algorithm with adaptive learning rate, since these methods perform better compared to non-adaptive methods such as stochastic gradient descent. The reason for this is, that adaptive methods find the learning rate automatically, which induce avoidance of getting trapped in a local suboptimal minima, and consequently, the algorithms find the minima faster. A few methods with adaptive learning rate exist. One of them is ADAM and this method was initially implemented, as it takes the advantages from two other adaptive learning rate methods named AdaGrad and RMSProp. Additionally ADAM has shown slightly better results compared to these. [Ruder, 2016] However, as described in Chapter 5, when testing the CNN using ADAM, the training progress contained

some unacceptable fluctuations, for which reason RMSProp was used instead. For more information about advantages and disadvantages for non-adaptive and adaptive learning rate methods see Appendix C.

4.3.4 Cost function

The CNN's weights, θ, were determined by minimizing the cost function in each epoch of the optimization algorithm. When dense training is applied the cost function can be described by: [Kamnitsas et al., 2017]

$$J(\Theta; I_s, c_s) = -\frac{1}{B \cdot V} \sum_{s=1}^{B} \sum_{V=1}^{V} log(p_{c_s^v}(x^v)), \tag{4.1}$$

where B is the number of image segments composing a batch, I_s and c_s are the true labels of the V voxels in the s^{th} segment in the batch, c_s^v is the true label of voxel v in segment s, $p_{c_s^v}$ is the predicted posterior probability for class c_s^v, and x^v is the position of the class in the classification feature map.

4.3.5 Regularisation

As described in Section 2.4.4, overfitting often occurs in neuroimaging and can be avoided using regularisation methods. The regularisation methods; dropout, L1 regularisation, and L2 regularisation were applied in DeepHSS's CNN.

As described in Section 4.2, the last four layers of the CNN are two fully connected layers, a softmax layer, and a CRF layer. 50 % dropout was applied to the second fully connected layer and the softmax layer. The regularization method dropout is elaborated in Section 2.4.2.

The usage of L1 and L2 regularisation implies, that two extra terms are added in the cost function, one for L1 and one for L2. In L1 regularisation the term is the sum of the weights, and in L2 regularization the term is the sum of the square of the weights. When the T1 and T2 terms are added to an cost function, the weights decrease, causing less influence by some features, and for the T1 term several weights will be equal to zero, resulting in a sparse weight vector. Thereby, several features are ignored which is believed to be beneficial. [Ng, 2004] The two therms are scaled by an adjustable parameter, which was set to 0.000001 for L1 and 0.0001 for L2.

Chapter 5
Initial data and test

This Chapter contains a description of the initial tests of DeepHSS. In the first Section is the dataset used for the tests described including the acquisition details and the method for generation of the ground truth labels. The next Section concern the initial tests, in which the purpose, method, results and discussion of each test is presented.

5.1 Data

A T1w and T2w 7T MRI acquired from three healthy subjects were available for this project. Additionally, a T1w and a T2w MDA MRI model where available for hippocampal subfield label generation.

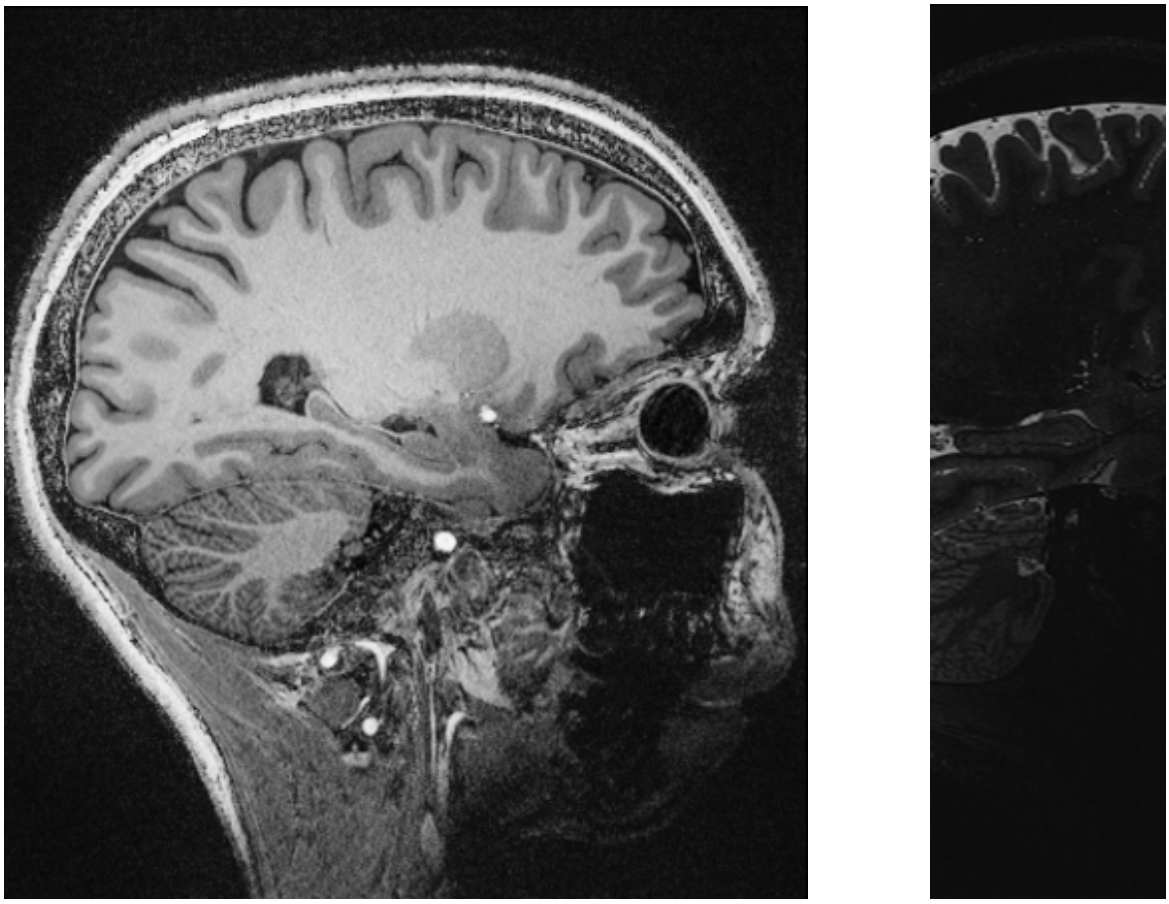
5.1.1 Data acquisition

The data was acquired using a 7T whole body MRI scanner (Siemens Healthcare, Erlangen, Germany). The gradient system of the scanner was a SC72 with a slew rate of 200 mT/m/s and a maximum gradient strength of 70 mT/m. For the radio frequency transmission and reception a 7T Tx/32 channel Rx head array (Nova Medical, Wilmington, MA, USA) was used. [Janke et al., 2016]

The T1w images were whole brain scans obtained by a prototype MP2RAGE sequence (WIP 900) with the resolution 0,5mm x 0,533mm x 0,5mm. During both acquisitions the following parameters remained constant: TR = 4330ms, TI1/TI2 = 750/2370ms, TE = 2.8ms, flip angles = 5 and 6 degrees, and GRAPPA = 3. [Janke et al., 2016]

The T2w images were orthogonal to the main axis of the hippocampus and were obtained by a 2D Turbo Spin Echo (TSE) sequence. Both subjects were imaged three times and to reduce the extend of data to be processed and increase SNR the three acquisitions per subject were averaged. The following parameters remained constant: resolution = 0.2x0.2x0.8mm, flip angle = 134 degree, and TR = 10.3s. [Janke et al., 2016]

In Figure 5.1 (a) a sagittal slice from a 7T single subject scan with MP2RAGE contrast is presented, and the corresponding sagittal slice from the same subject's 7T single subject scan with TSE contrast is presented by Figure 5.1 (b).

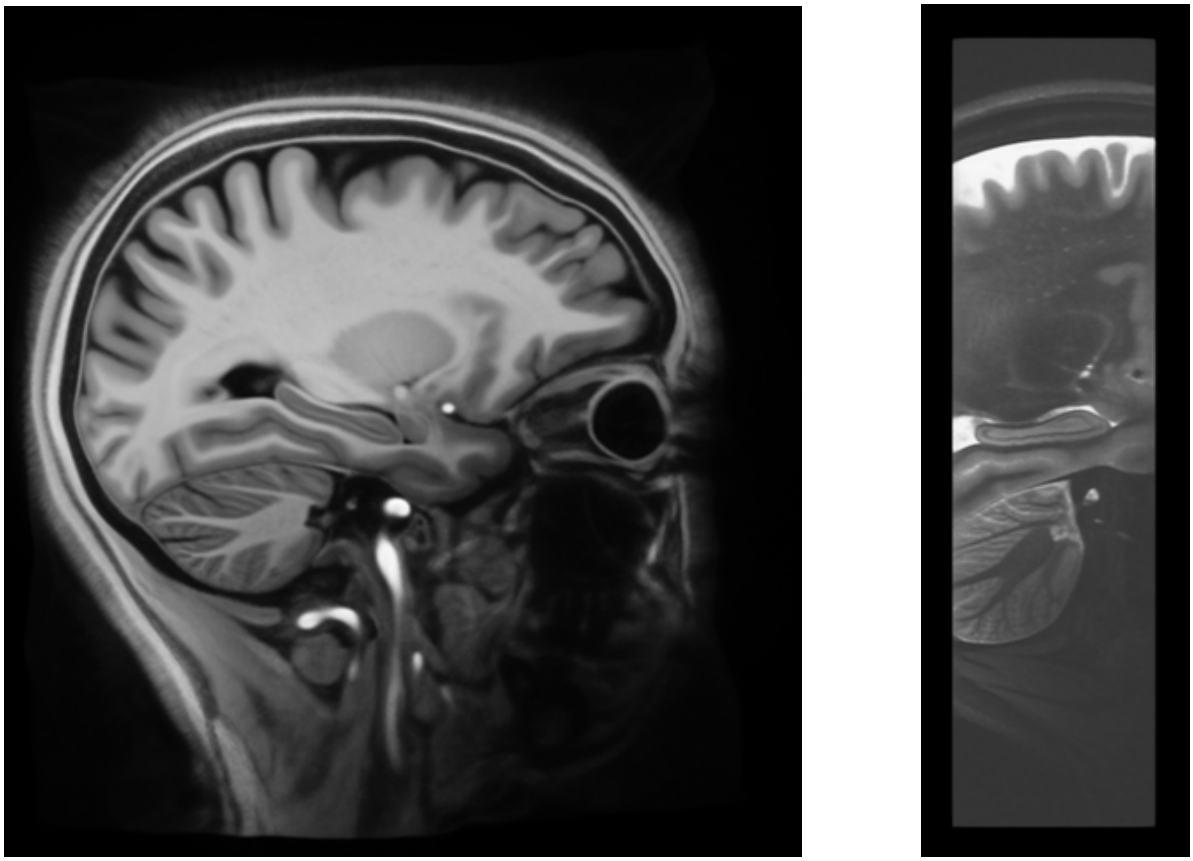


(a) **(b)**

Figure 5.1: Figure (a) and (b) are sagittal slices from a single subject 7T MRI. Figure (a) is with MP2RAGE contrast whilst Figure (b) is a TSE contrast. The 7T MP2RAGE single subject scan covers the whole brain whereas the 7T TSE single subject scan covers a slice of the brain including both hippocampi.

5.1.2 7T MDA MRI models

The two MDA MRI models are included with the purpose of constructing ground truth labels for the single subject scans. Both models are probabilistic and generated using the method described by Janke and Ullmann and Grabner et al.. One of the models are with MP2RAGE contrast and the other model is with TSE contrast. The 7T MP2RAGE MDA model are based on 48 single subject 7T MP2RAGE scans, whilst the TSE 7T MDA model are based on 26 single subject 7T TSE scans. [Janke et al., 2016] Acquisition of the utilized single subject 7T scans are described in Section 5.1.1. A general description of how the MDA models were generated can be found in Appendix A. In Figure 5.2 (a) a sagittal slice of the 7T MP2RAGE MDA model is shown, whilst a sagittal slice of the 7T TSE MDA model is presented by Figure 5.2 (b).



(a) (b)

Figure 5.2: Figure (a) and (b) are sagittal slices from the 7T MRI MDA models. The model illustrated in Figure (a) is with MP2RAGE contrast and the model illustrated in Figure (b) is with TSE contrast. The 7T MP2RAGE MDA model covers the whole brain whereas the 7T TSE MDA model covers a part of the brain including the hippocampus.

The models are freely available at *http://imaging.org.au/Human7T/TSE* and *http://imaging.org.au/Human7T/MP2RAGE*

5.1.3 Ground truth hippocampal subfield labels

In order to gain a complete input data set for the DeepHSS, ground truth (GT) labels for each single subject 7T MR image are needed. In the initial tests the GT labels were generated using three methods:

1. Segmenting the single subject 7T MR images using ASHS.
2. Segmenting the MDA MRI models using FreeSurfer version 6.0 and warping the segmentations from model space to subject space
3. Segmenting the MDA MRI models using ASHS and warping the segmentations from model space to subject space

For method 1. the GT labels were simply generated by applying ASHS to the single subject 7T MR images. For both method 2. and 3. the MP2RAGE and the TSE 7T MDA model were utilized by the automatic segmentation methods to achieve as reliable hippocampal subfield labels as possible. Subsequently, the labels were warped from model space into subject space using the inverse transformations obtained when the single subject 7T MR images were wrapped to the model space during the MDA model generation. The

labelling process is illustrated in Figure 5.3, and the shell script used for the process is present in Appendix B.

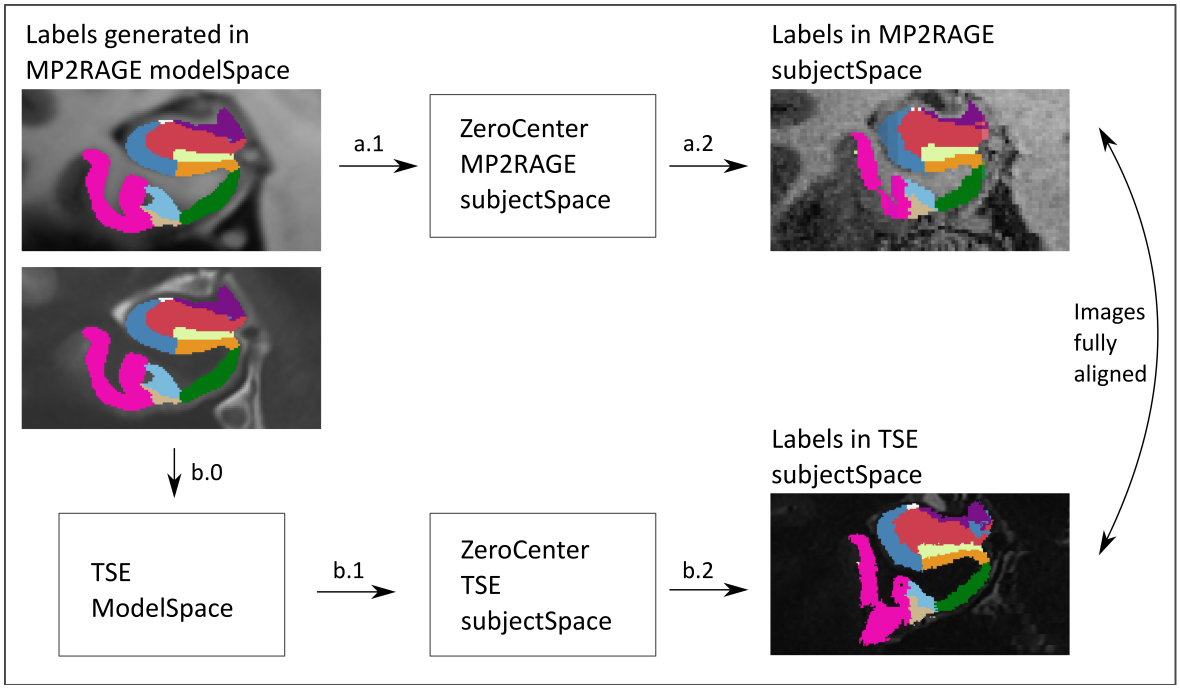


Figure 5.3: Illustration of how the ground truth labels are generated. The 7T MRI MDA models, which are coregistered in the MP2RAGE model space, are labelled using FreeSurfer. Subsequently, both the labels for the TSE contrast and the labels for the MP2RAGE contrast image are warped into their respective subject spaces. Due to the nature of MDA model generation, this takes a few steps. a.1: inverse non-linear transformation are applied to the labels in order to warp the MP2RAGE labels from model to subject space. b.0: the TSE labels are warped from MP2RAGE model space to TSE model space using both inverse non-linear and linear transformation. b.1: inverse non-linear transformation are applied to the TSE labels in order to warp those from model to zero center subject space. a.2 and b.2: the coordinate systems of the label images are alternated to fit the related information of the TSE image and MP2RAGE image in subject space.

Figure 5.4 presents a comparison between labels constructed using ASHS on MDA models and then warped into subject space and labels constructed directly in subject space with ASHS. Figure 5.5 is an illustration of the labels obtained using FreeSurfer and ASHS, respectively, superimposed onto the 7T TSE MDA model.

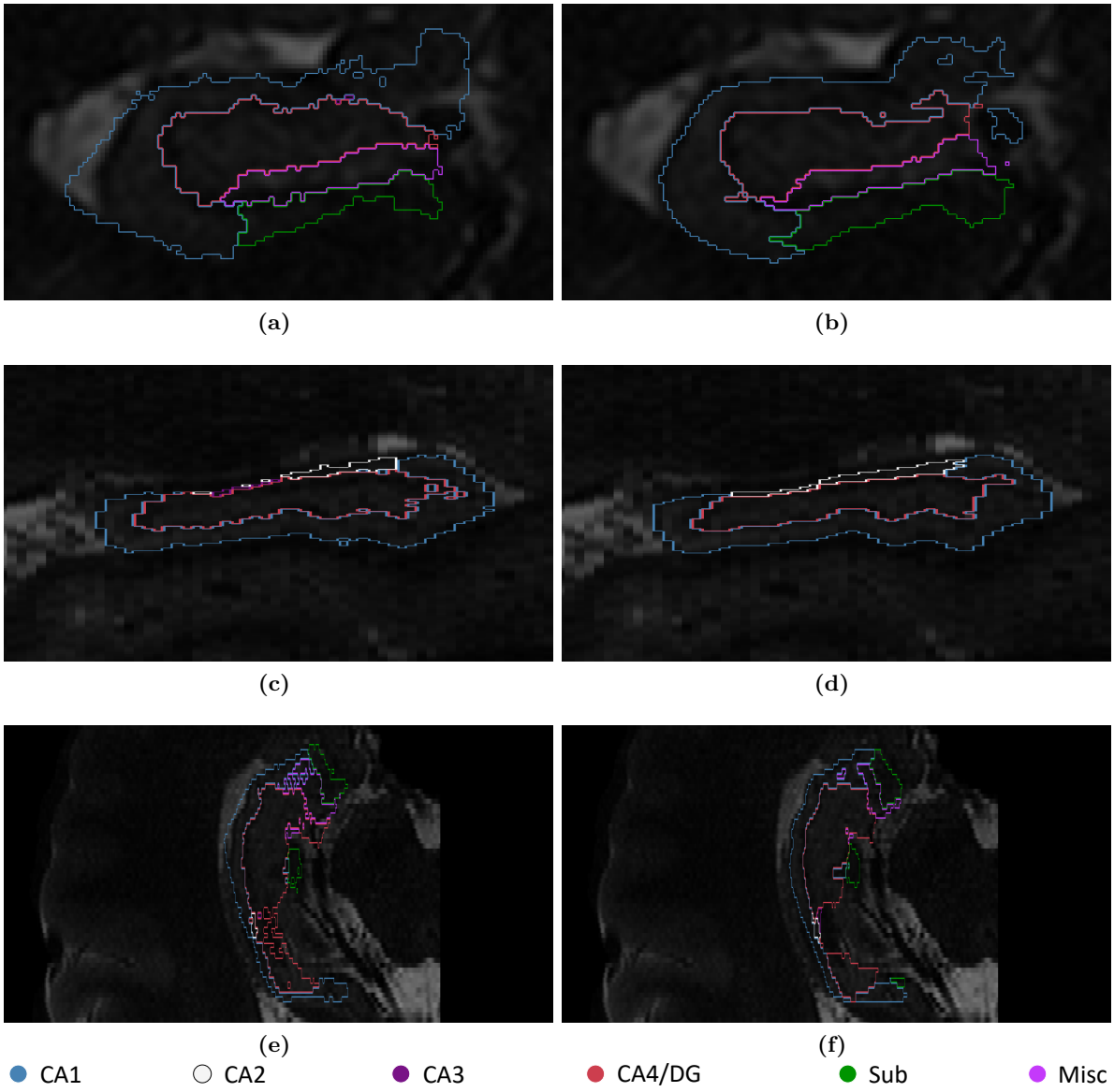


Figure 5.4: Coronal (a-b), sagittal (c-d) and transverse (e-f) view of the hippocampal subfield labels obtained using ASHS at the MDA models and warped into subject space (a,c,e) and directly in subject space (b,d,f). All labels are superimposed onto the corresponding 7T TSE MRI. Outer boundaries of labels generated directly in subject space follow the hippocampus smoother, this is visually clear in the coronal view.

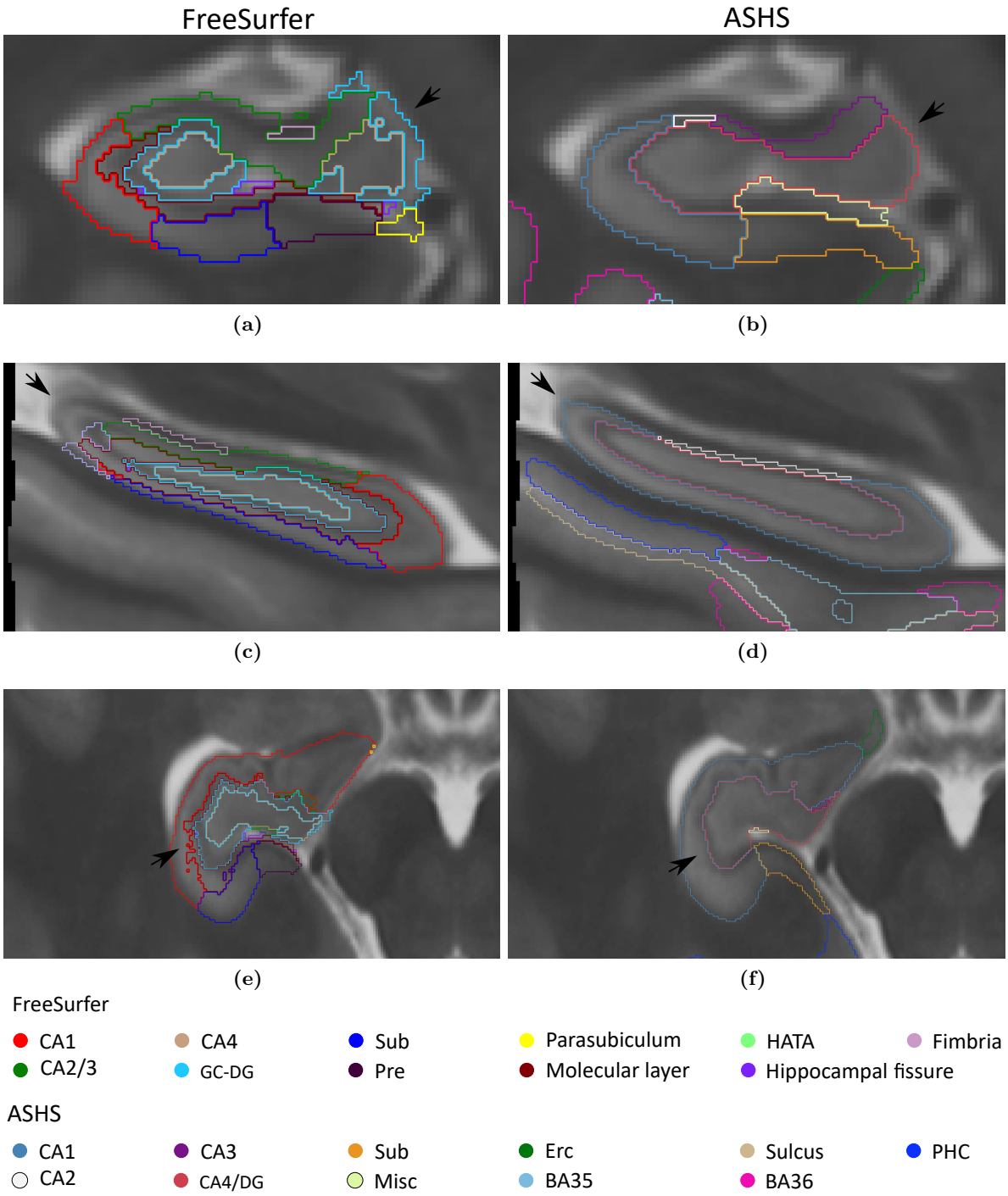
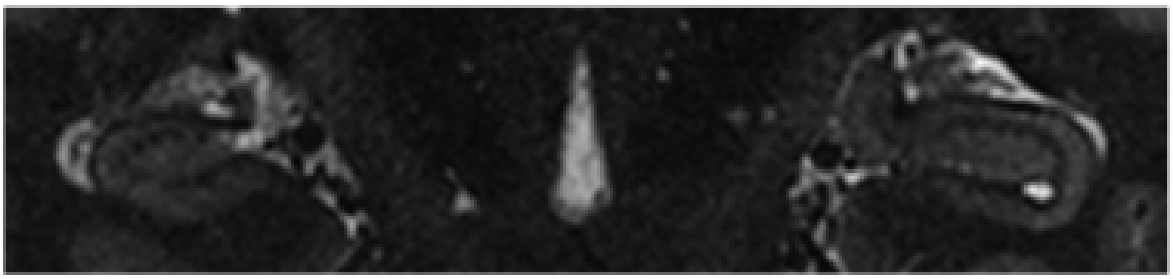


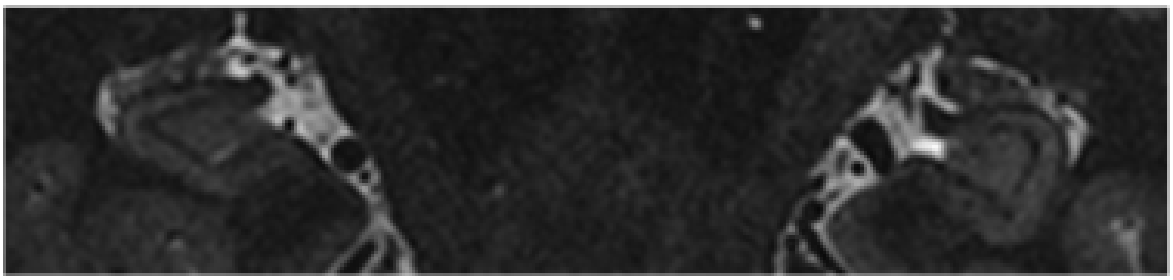
Figure 5.5: Hippocampal subfield labels obtained by FreeSurfer (a, c, e) and ASHS (b, d, f), respectively, in the coronal (a, b), sagittal (c, d), and the transverse view (e, f) superimposed onto the 7T TSE MDA model. FreeSurfer divides the hippocampus into 11 subfields whilst ASHS devides the hippocampus into 6 subfields and adds further 5 subfields outside of hippocampus proper. Most of these are beyond the cropped view. The labels obtained using ASHS does in general have more smooth boundaries than the labels obtained using FreeSurfer, and some subfields are more accurately delineated by edges following changes in the TSE contrast. Examples of this are pinpointed with black arrows in all three views.

5.1.4 Data augmentation

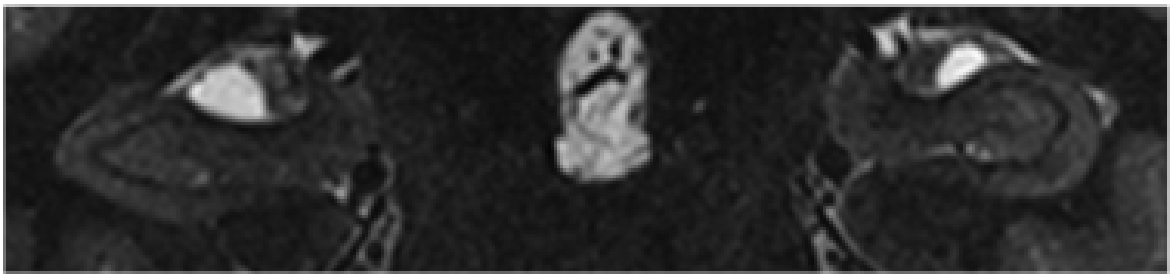
Data augmentation was utilized, due to the limited size of the dataset. The augmentation exploits that each subject has two hippocampi which are diverse from each other by mirroring each subjects left hippocampus ground truth labels and each subject's 7T MR images. Figure 5.6 illustrates both the right and the mirrored left hippocampus in the same cropped coronal 7T MP2RAGE MR slice.



(a)



(b)



(c)

Figure 5.6: Coronal view of all three subjects (a, b, and c) at their respective 7T TSE MR image slice. The slices captures both hippocampi, which are of different shapes.

5.2 Initial tests

The initial tests aim to investigate; if DeepHSS is capable of segmenting hippocampal subfields, the influence of input labels, and if it is beneficial to train the CNN in DeepHSS with GT labels generated using existing automatic segmentation methods on MDA models.

For the initial tests a small dataset (N=6) was available, as explained in Section 5.1. Consequently, 6-fold cross-validation was applied for evaluation of the performance in all initial tests. In each fold, the CNN was trained with five subjects and tested with the last subject, resulting in six subtests. To quantify DeepHSS performance in the tests a foreground dice score (DSC), an overall score for all subfields, was calculated between DeepHSS segmentations and the manual segmentations. To better understand how the training progressed the

foreground DSC was in all tests calculated for every second epoch throughout the training.

5.2.1 Test 1: Segmentation of hippocampal subfields and experience with optimizer algorithm

This test aims to investigate if DeepHSS is capable of segmenting hippocampal subfields.

Method

The CNN in DeepHSS was set up to take one modality, MP2RAGE, as input. The GT labels were produced using FreeSurfer on the 7T MDA models whereupon the labels were warped from model space to subject space, as described in Section 5.1.3. The MP2RAGE MR images were masked with a whole brain mask.

Results

Figure 5.7 illustrates the segmentations using DeepHSS in relation to the GT labels. All labels are superimposed onto the corresponding MR image with MP2RAGE contrast.

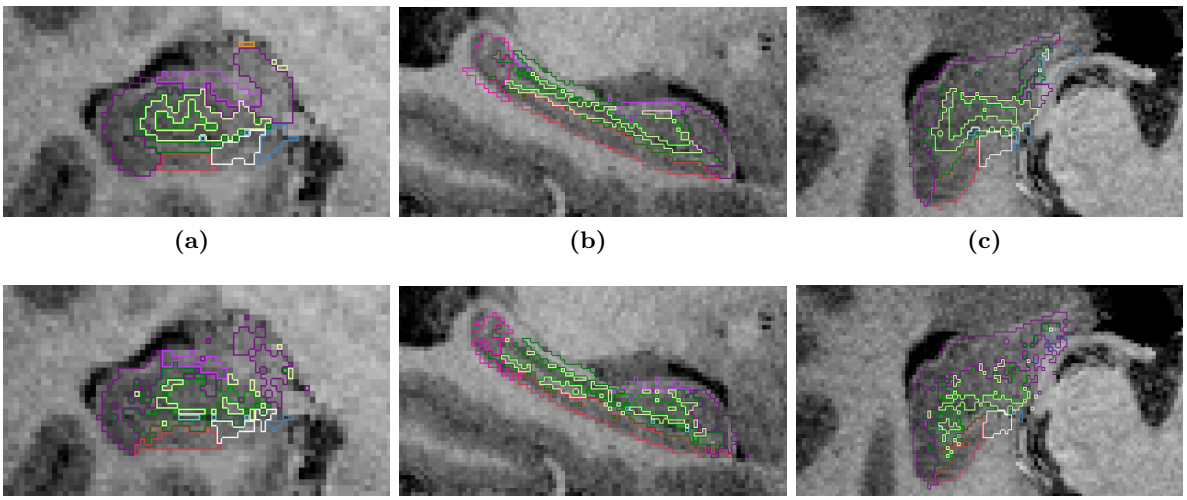


Figure 5.7: Coronal, sagittal, and transverse view of the hippocampal subfield segmentation obtained using multispectral segmentation with FreeSurfer via model space (a-c) and the predicted segmentations (d-f), all superimposed onto the corresponding 7T MP2RAGE MR image.

As seen in Figure 5.7 DeepHSS manages to segment the hippocampal subfields. The foreground DSC between the segmentations obtained using DeepHSS and the manual segmentations for every second epoch of the training process is presented in Figure 5.8.

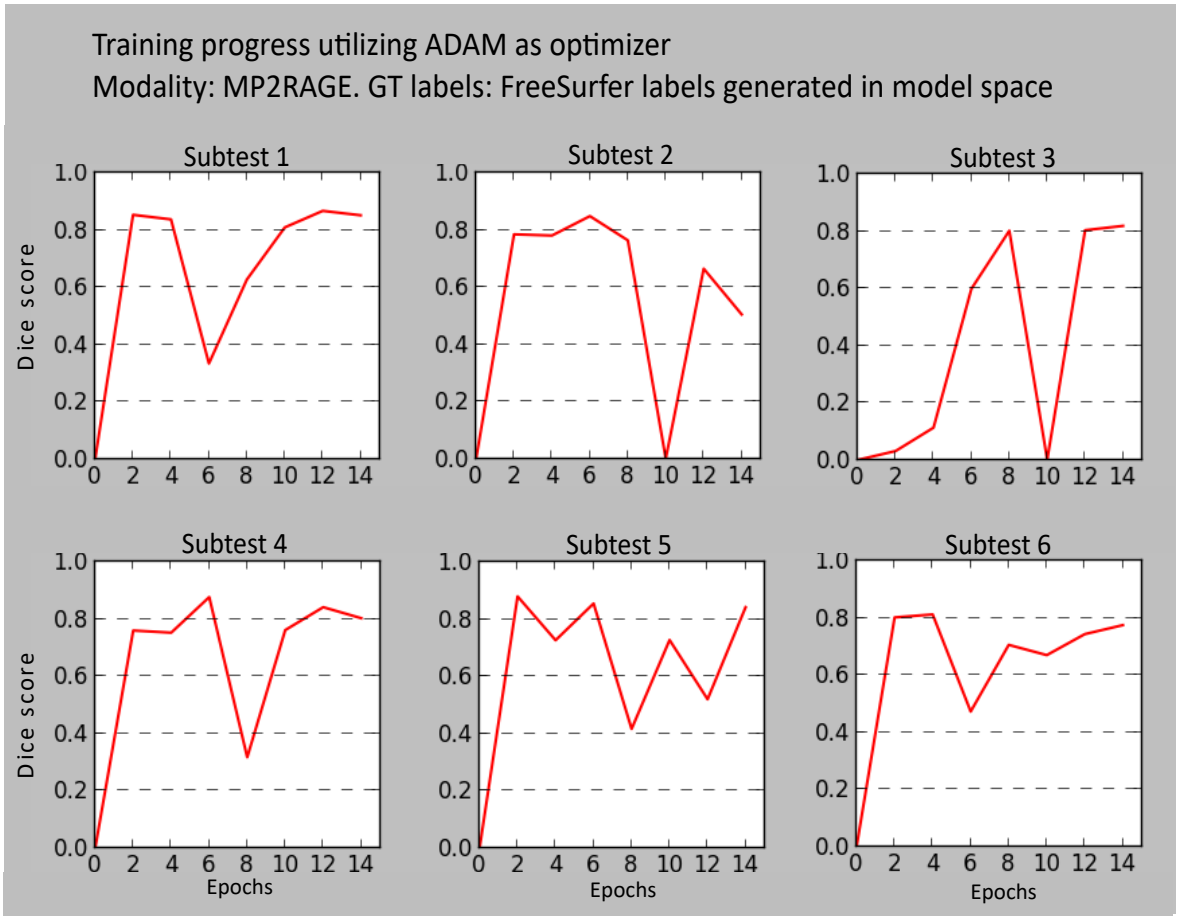


Figure 5.8: Six graphs illustrating foreground DSC as a function of number of training epochs. The graphs are fluctuating and does not represent a continuous increasing DSC. The optimizer algorithm used was ADAM.

As seen in Figure 5.8, the foreground DSC between the CNN's segmentations of the test data and the GT segmentation fluctuates as the network is trained with more epochs.

Discussion

It was possible to find subfields within hippocampus, however, the graphs illustrating the training process were oscillating unexpectedly. The foreground DSC was expected to converge towards a steady state as the number of training epochs increased. These fluctuations in the DSC might be due to the optimization algorithm ADAM having difficulties finding the global minimum. To overcome this issue, the optimizer was changed to RMSProp, since RMSProp has shown a performance comparable to ADAM (see section C).

To investigate whether the CNN became more stable by changing the optimizer, this test was redone after changing the optimization algorithm. The foreground DSC between the DeepHSS segmentations and the GT segmentation for every second epoch are illustrated in Figure 5.9.

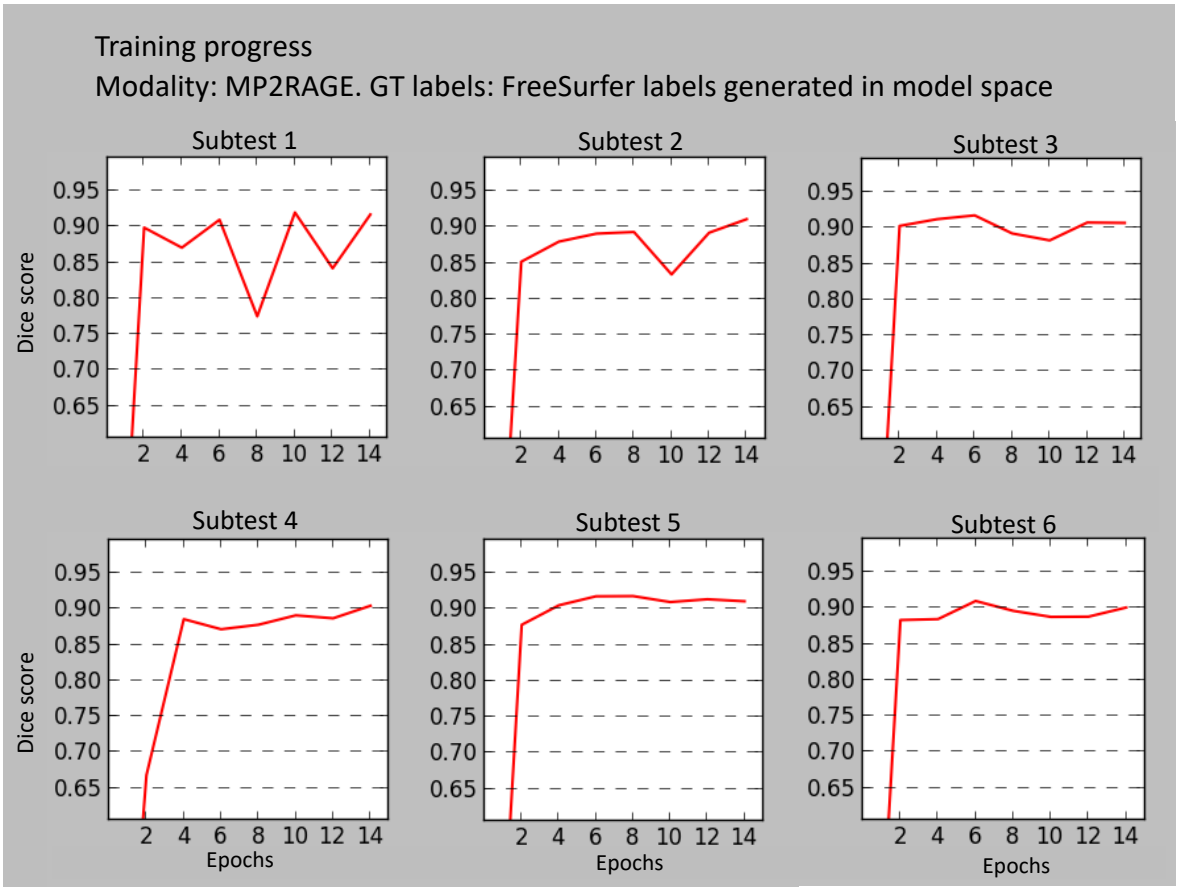


Figure 5.9: Six graphs illustrating foreground DSC as a function of number of training epochs. The optimizer algorithm used was RSMprops.

As seen in Figure 5.9, the graphs reaches a plateau after 2-4 training epochs. The training progressions in Figure 5.9 fluctuates less compared to 5.8, thereby a more stable training progress is achieved when the optimization algorithm is RMSprops compared to ADAM. Consequently, RMSprop was applied as the CNN's optimizer.

5.2.2 Test 2: Optimal inputlabels and image modalities

The aim of this test is to investigate which automatic method can generate the most optimal GT labels and if TSE or MP2RAGE images should be applied for training of the CNN in DeepHSS: FreeSurfer on MDA models warped to subjectspace or ASHS on MDA models warped to subject space.

Method

For this test the CNN in DeepHSS was trained using four different versions of GT labels and associated single subject 7T MR images with different modalities:

1. FreeSurfer labels generated in model space and warped to subject space and single subject 7T MP2RAGE MR images
2. FreeSurfer labels generated in model space and warped to subject space and single subject 7T TSE MR images
3. ASHS labels generated in model space and warped to subject space and single subject 7T MP2RAGE MR images

4. ASHS labels generated in model space and warped to subject space and single subject 7T TSE MR images

Results

The training progress of the CNN trained with FreeSurfer labels and MP2RAGE modality, FreeSurfer labels and TSE modality, ASHS labels and MP2RAGE modality, and ASHS labels and TSE modality can be seen in Figure 5.9, 5.10, 5.11, and 5.12, respectively.

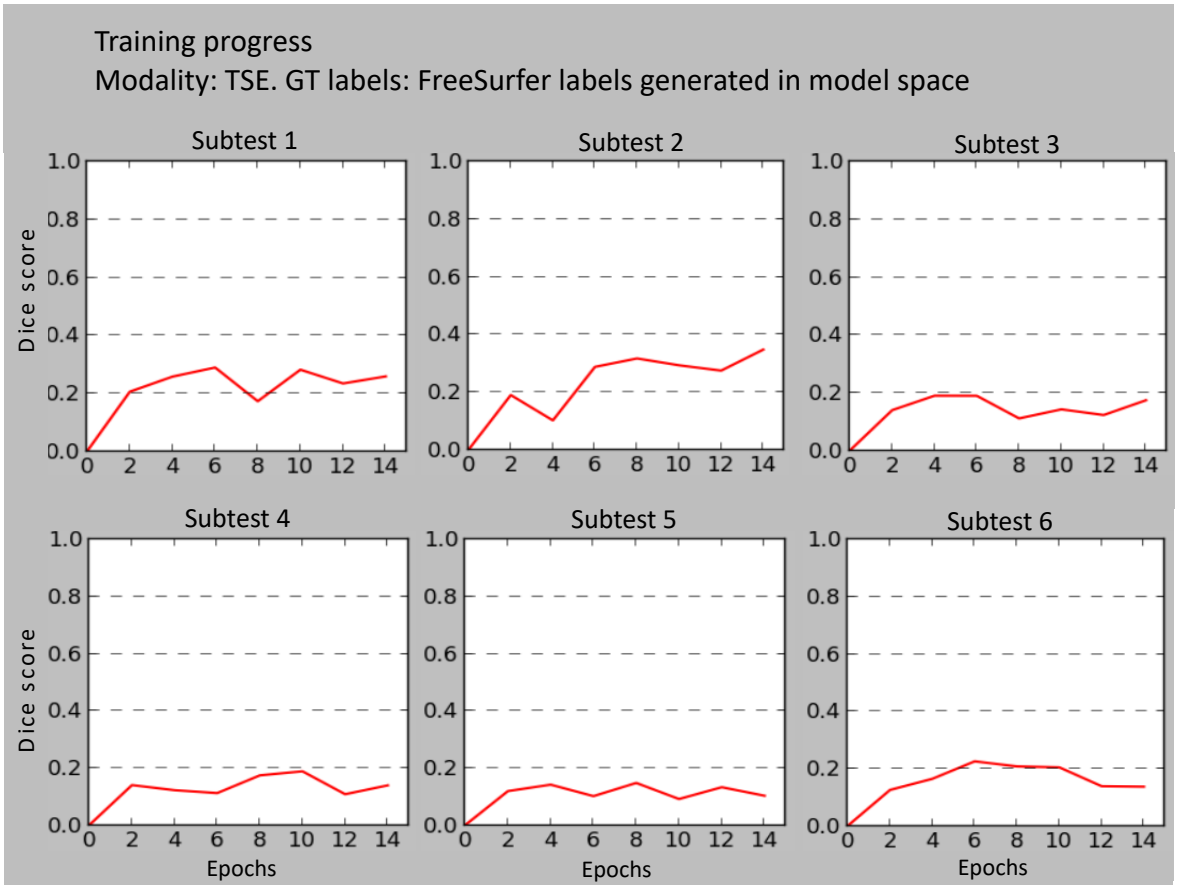


Figure 5.10: Six graphs illustrating the foreground DSC as a function of number of training epochs. The CNN in DeepHSS was trained using FreeSurfer labels generated in model space and warped to subject space and single subject 7T TSE MR images.

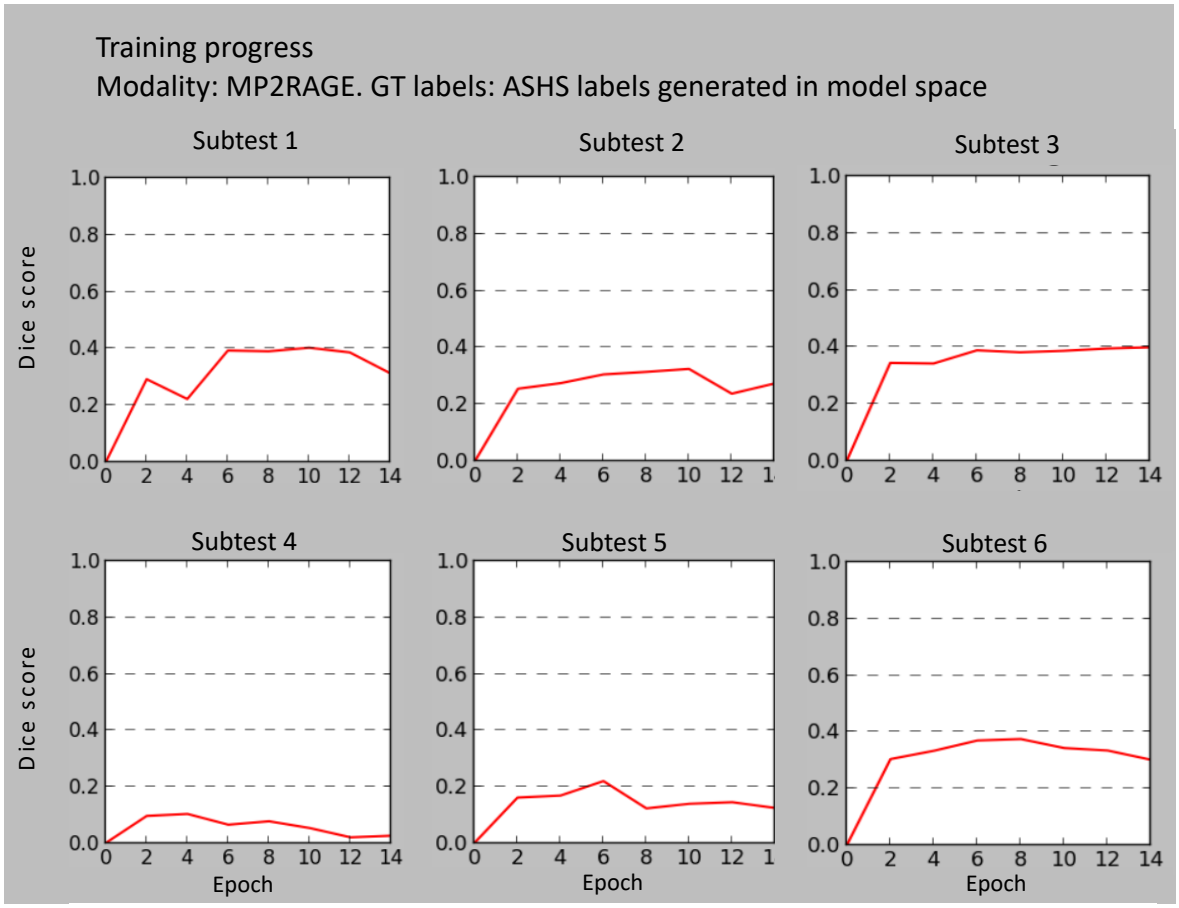


Figure 5.11: Six graphs illustrating the foreground DSC as a function of number of training epochs. The CNN in DeepHSS was trained using ASHS labels generated in model space and warped to subject space and single subject 7T MP2RAGE MR images.

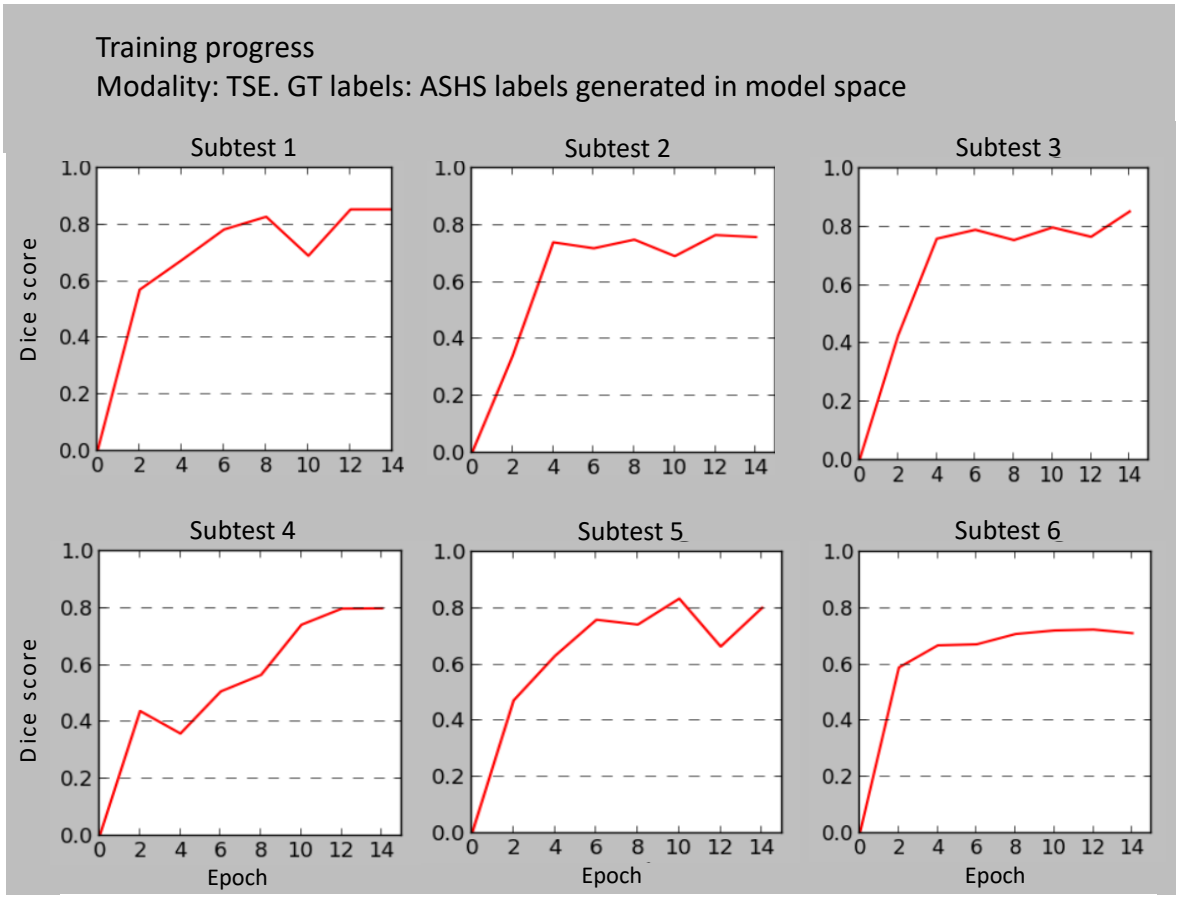


Figure 5.12: Six graphs illustrating the foreground DSC as a function of number of training epochs. The CNN in DeepHSS was trained using ASHS labels generated in model space and warped to subject space and single subject 7T TSE MR images.

The two best performances were observed when the CNN in DeepHSS was trained using FreeSurfer generated labels and MP2RAGE images and ASHS generated labels and TSE images. The average DSC for segmentations obtained utilizing FreeSurfer and MP2RAGE images was 0.64 ± 0.02 whereas it was 0.63 ± 0.08 when ASHS and TSE images were applied. It is observed that the CNN trained with FreeSurfer labels and MP2RAGE images overall obtains the highest DSC. When visualising the segmentations it was noticed that the network segmented parts of the opposite hippocampus as seen in Figure 5.13.

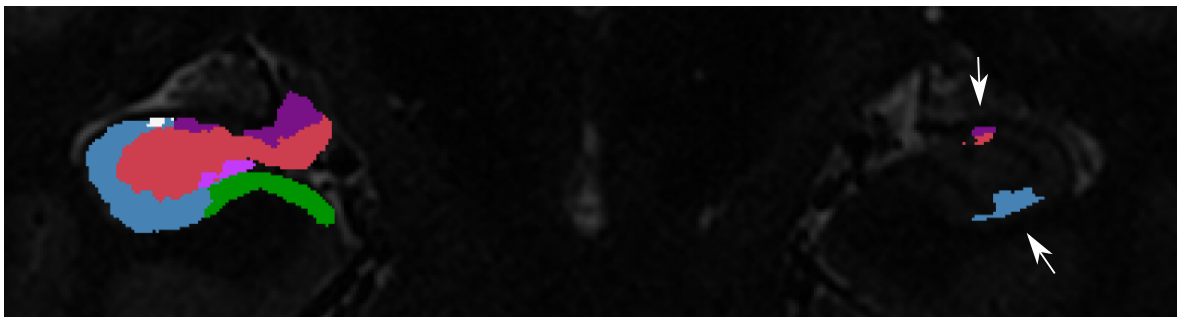


Figure 5.13: Coronal view of the hippocampal subfield segmentation obtained using DeepHSS trained with 7T TSE images acquired from 5 subjects and GT labels obtained using FreeSurfer and MDA models.

Discussion

As it appears from the results, the network trained with GT labels generated using FreeSurfer performs best on MP2RAGE MR images whilst the network trained with ASHS generated labels performs best on TSE scans. Furthermore, when the FreeSurfer generated GT labels and MP2RAGE images were applied a slightly better performance was obtained compared to when the ASHS generated GT labels and TSE images were applied.

As described in Section 2.2.1 the segmentation performed by ASHS was compaired to manual segmentation with DSC whereas FreeSurfer was validated by visual inspection. Additionally, Giuliano et al. [2017] states that the volumetric results for each subfield should be carefully interpreted when FreeSurfer is used for segmentation. For this reason it was chosen to use ASHS for labelling despite DeepHSS performed slightly better when using labels generated by FreeSurfer.

It was considered suspicious that the performance of ASHS and FreeSurfer were almost similar. For this reason the input labels were investigated, and an error in the warping from model space to subject space was found as illustrated in Figure 5.4. For this reason it was decided to use ASHS direct on the subject for future tests.

During the tests segmentation of the opposite hippocampus was observed, and it was unclear if these incorrect segmentations were caused by the data mirroring when the data was augmented, or if it was caused by similarity between the left and right hippocampus' local features (e.g. textual features). To clarify this, a network was trained with two of the non mirrored subjects and tested on the last subject. The results of this test can be seen in Figure 5.14 and as the arrows indicates the network still segments parts of the opposite hippocampus. Thereby, the segmentations in the opposite hippocampus must be due to the network's use of local features. For this reason, the scans were bisected to only contain the right hippocampus in future tests.

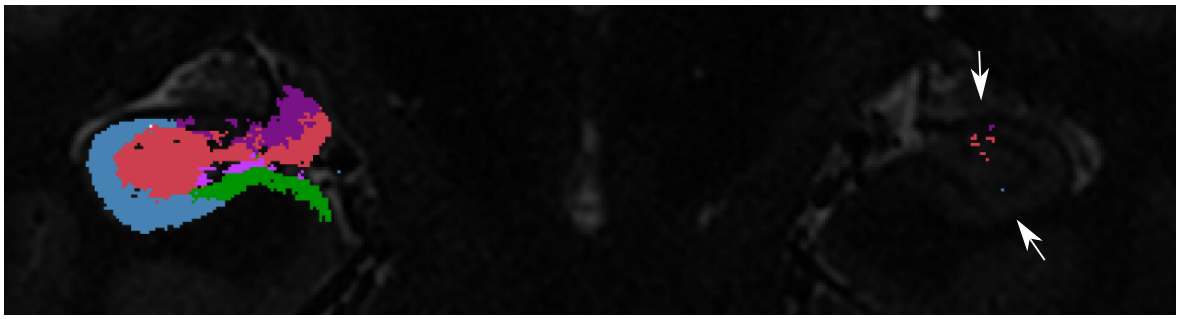


Figure 5.14: Coronal view of the hippocampal subfield segmentations obtained using DeepHSS trained with 2 training subjects.

5.2.3 Test 3: Adjusted DeepHSS

The aim of this test is to evaluate the performance of DeepHSS when implementing the parameters discussed in Section 5.2.2.

Method

ASHS was applied to 7T TSE MR images in subjectspace to obtain GT labels. The GT labels are illustrated in Section 5.1.3. In contrary to previous tests the CNN in DeepHSS was fed scans containing exclusively the right part of the brain.

Results

The training progress of the CNN contained in DeepHSS trained with TSE images and GT labels generated using ASHS is illustrated in Figure 5.15.

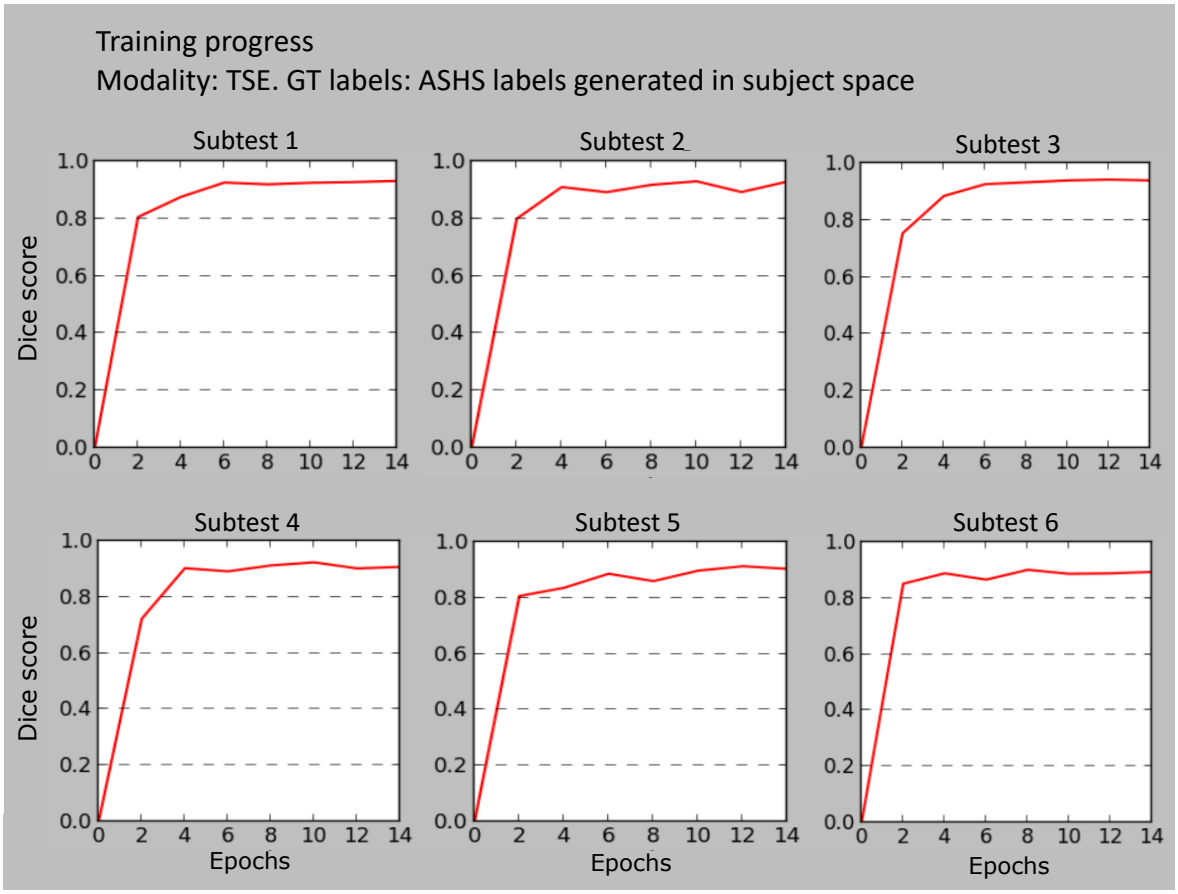


Figure 5.15: Six graphs illustrating the training process for all subtests. The CNN was trained with 7T TSE MR images and GT labels obtained using ASHS in subject space.

As seen in Figure 5.15 the network obtained a foreground DSC of approximately 0.9 in all six subtests after 8 training epochs. An example of the segmentations obtained using DeepHSS and the associated GT label is illustrated in Figure 5.16.

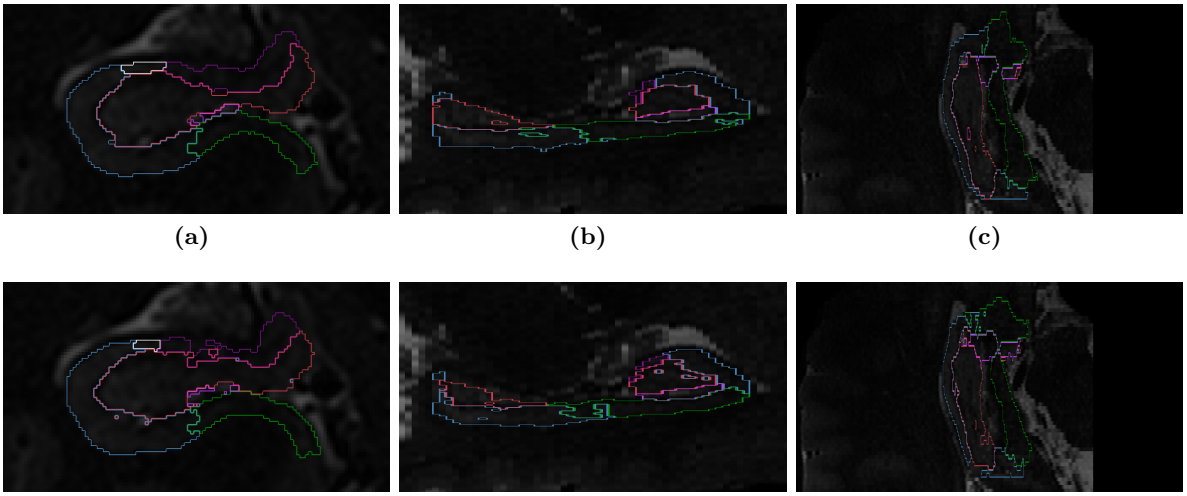


Figure 5.16: Coronal, sagittal, and transverse view of the hippocampal subfield segmentation obtained using multispectral segmentation with ASHS directly in subject space (a-c), and predicted segmentations using DeepHSS, all superimposed onto a 7T TSE MR image.

As seen in Figure 5.16 the segmentation obtained by the network are comparable to the GT labels generated by ASHS. The average DSC for the subfields in this test was 0.75 $\pm$ 0.05.

Discussion

The results obtained in this test, where GT labels are generated using ASHS directly in subjects space, were better compared to the results in test 2 (see Section 5.2.2), when ASHS was used to generate GT labels in model space. The average DSC for the individual subfields increased from 0.63 $\pm$ 0.08 in test 2 to 0.75 $\pm$ 0.05 in this test. This is the opposite result of what was expected, since utilization of the 7T MDA models, should have improved the quality of the GT labels. The reason might be the potential error in the warping process from model space to subject space, described in Section 5.2.2, or bisection of the images in this test.

Chapter 6
Final data and test

This Chapter presents the final test of DeepHSS. First, the dataset and the GT labels used for the test are described. Subsequently, the test are elaborated and the results are discussed.

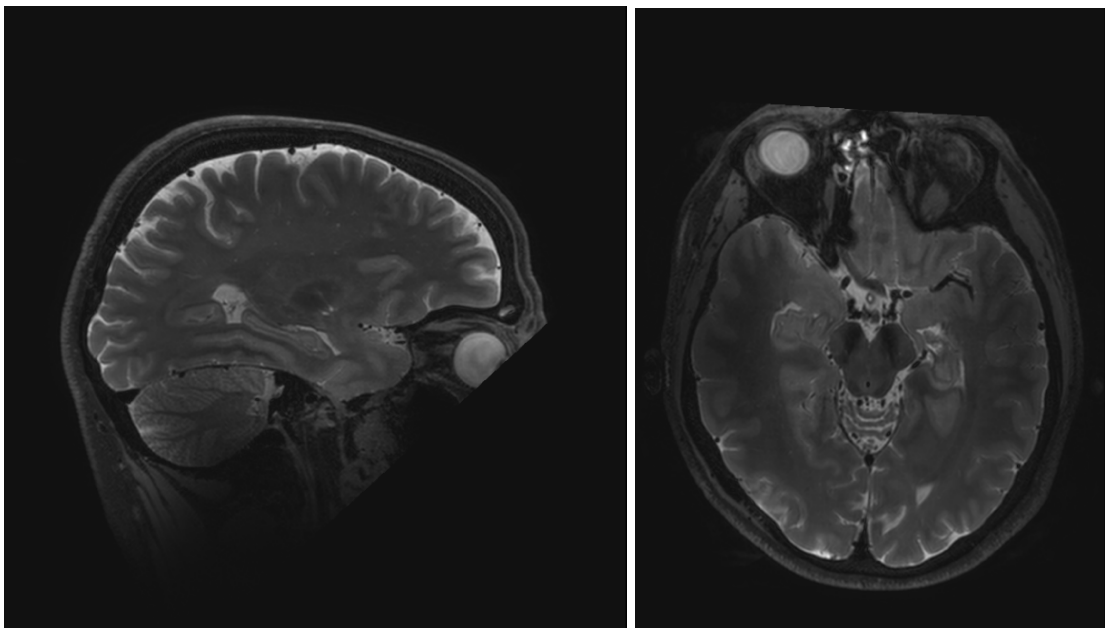
6.1 Data

A dataset from [Wisse et al., 2016] became available late in the process of DeepHSS's development. The dataset contains 7T T2w MRI images acquired from 26 subjects (46 % men, mean age: 59 ±9 years, median Mini Mental Examination score [Folstein et al., 1975] 29, 25-30), and manual delineated GT labels for each MRI.

6.1.1 Data acquisition

The images were acquired using a 7T MRI scanner (Philips Healthcare, Best, the Netherlands) with a 16-channel receive coil and a volume transmit coil (Nova Medical, Wilmington, Massachusetts) [Wisse et al., 2016].

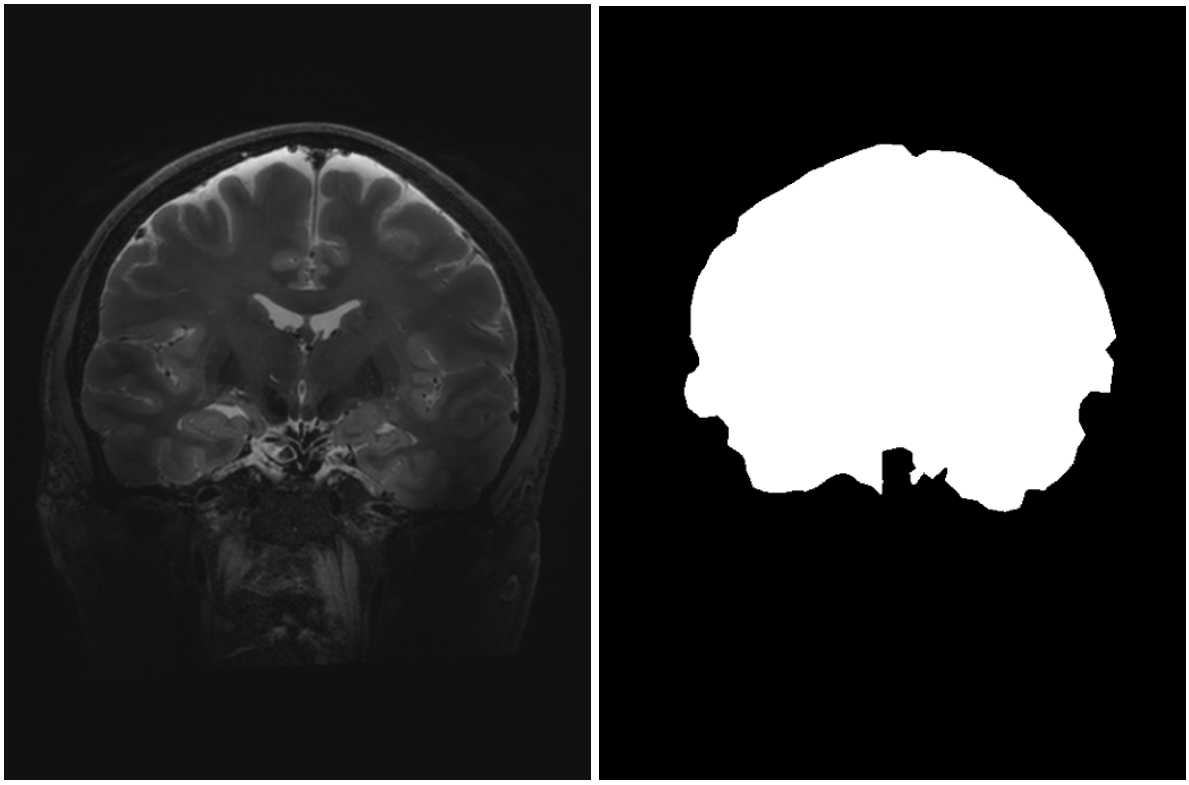
The T2w MR images were obtained by a 3D TSE (TSE factor 182) sequence. The following parameters remained constant: resolution = 0.7x0.7x0.7mm, flip angle = 120 degree, TR = 3.158s, matrix size of 356x357x272, nominal TE of 0.301s, and a 2D sensitivity encoded with the acceleration factors 2.0 (anterior-posterior) x 2.8 (right-left). [Wisse et al., 2016] A sagittal and transverse slice from a 7T MR image with TSE contrast is presented in Figure 6.1.



(a) **(b)**

Figure 6.1: Sagittal (a) and transverse (b) slice from a 7T MRI with TSE contrast.

As described in Chapter 4, the MR images were masked and bisected. A coronal view of one full 7T TSE MR image with corresponding brain mask is illustrated in Figure 6.2.



(a) (b)

Figure 6.2: Coronal slice from a 7T MRI with TSE contrast, and the corresponding mask. Due to low contrast in the medial temporal lobes, these are barely included by the mask.

6.1.2 Ground truth hippocampal subfield labels

For GT labels, manual segmentations were available for each of the subjects. The following hippocampal subfields were segmented in the GT labels; Corneas Ammonis (CA)1,(CA2), CA3, Dentate Gyrus (DG), Subiculum (Sub), Entorhinal Cortex (ERC), tail, and cyst. For the manual labelling Wisse et al. [2016] used an in-house developed software based on MeVisLab (MeVis Medical Solutions, Bremen, Germany23). [Wisse et al., 2016]

Figure 6.3 illustrates one manual label in all three anatomical views superimposed onto the corresponding bisected, masked 7T TSE MR image.

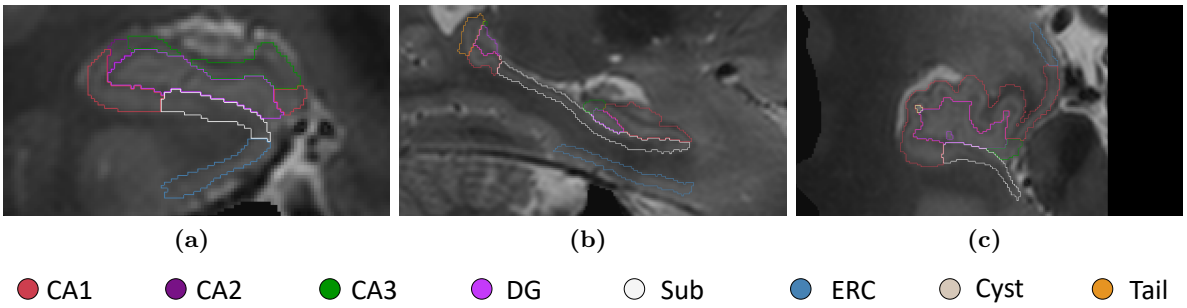


Figure 6.3: Manual hippocampal subfield labels from Wisse et al. [2016] illustrated in the coronal (a), sagittal (b), and the transverse view (d). Eight hippocampal subfields were delineated.

6.2 Test

A new and larger dataset (N=26) with manual segmentations became available late in the development process of DeepHSS, as described in Section 6.1. Besides allowing for training with more subjects, the size of the new dataset allowed for investigation of the required amount of training subjects, a factor influencing the performance of a CNN and the risk of overfitting (see Section 2.4.3).

6.2.1 Method

The CNN in DeepHSS was trained five times using a different number of training subjects; 3, 6, 9, 12, and 15, resulting in five subtests. The number of training epochs was set to 22 for all training sessions. DeepHSS was tested with 10 subjects in every subtest, and the performance was quantified by comparing predicted segmentations to the corresponding GT labels using DSC.

6.2.2 Results

To investigate the training progress of the CNN in DeepHSS a DSC for the whole foreground (all hippocampal subfields) was calculated between the segmentations obtained using DeepHSS and the GT labels for every second epoch for one test subject. In Figure 6.4 the foreground DSC as a function of training epochs for the subject is illustrated when the network was trained with 3, 6, 9, 12, and 15 subjects. As illustrated in 6.4, the foreground DSC increased as the training proceeded, but the learning rate reached a steady state around 14^{th} epoch.

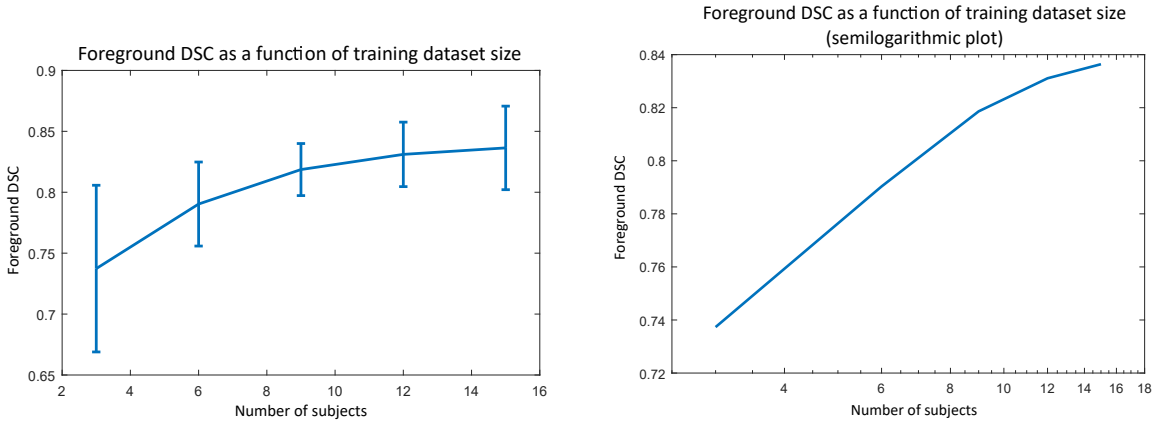
Figure 6.4: Foreground DSC for one test subject between segmentations obtained using DeepHSS and the GT labels at every second training epoch. Each graph represents the training progress for the network trained using 3, 6, 9, 12, and 15 subjects, respectively

DSC for all hippocampal subfields are presented in Table 6.1. Results from all subtests are represented.

Foreground DSC as a function of training subjects is illustrated in Figure 6.5.

Table 6.1: DSC between segmentations obtained using DeepHSS and manual segmentations, when the network in DeepHSS is trained using 3, 6, 9, 12, and 15 subjects.

	3 subjects	**6 subjects**	**9 subjects**	**12 subjects**	**15 subjects**
CA1	0.48 ± 0.08	0.54 ± 0.07	0.56 ± 0.04	0.58 ± 0.07	0.61 ± 0.06
CA2	0.60 ± 0.06	0.64 ± 0.07	0.69 ± 0.05	0.69 ± 0.06	0.69 ± 0.05
DG	0.66 ± 0.08	0.72 ± 0.06	0.76 ± 0.04	0.77 ± 0.04	0.77 ± 0.05
CA3	0.54 ± 0.14	0.58 ± 0.12	0.60 ± 0.13	0.59 ± 0.18	0.59 ± 0.19
Sub	0.72 ± 0.11	0.77 ± 0.06	0.80 ± 0.05	0.80 ± 0.06	0.80 ± 0.05
ERC	0.42 ± 0.15	0.44 ± 0.11	0.46 ± 0.14	0.49 ± 0.17	0.49 ± 0.14
Average subfield DSC	0.57 ± 0.08	0.61 ± 0.06	0.64 ± 0.05	0.65 ± 0.07	0.66 ± 0.08
Foreground DSC inclusive tail and cyst	0.74 ± 0.07	0.79 ± 0.03	0.82 ± 0.02	0.83 ± 0.03	0.84 ± 0.03
Foreground DSC excluding tail and cyst	0.73 ± 0.07	0.78 ± 0.04	0.81 ± 0.02	0.82 ± 0.03	0.83 ± 0.03



(a) (b)

Figure 6.5: Foreground DSC between segmentations obtained using DeepHSS and the associated GT labels as a function of training dataset size. (b) is made using the same data as (a), but is plotted in a semilogarithmic coordinate system.

The lowest performance of DeepHSS was obtained when training with 3 subjects (foreground DSC = 0.74 ± 0.07, average DSC = 0.57 ± 0.08) and the highest performance was obtained when the network trained using 15 subjects (foreground DSC = 0.84 ± 0.03, average DSC = 0.66 ± 0.08). The average DSC of DeepHSS increased as the number of training subjects increased and the DSC for the individual subfields mainly increased, but were stagnant in some cases.

Training with 15 subjects, the hippocampal subfields with the highest DSC were Sub and DG (> 0.77), whilst the lowest DSC were achieved for subfield CA3 and ERC (> 0.49). Average subfield DSC was 0.66±0.08. Tail and cyst were omitted from this calculation for comparative reasons. Foreground DSC including tail and cyst were 0.84±0.03, excluding tail and cyst the foreground DSC were 0.83±0.03.

Figure 6.6 presents a visualization of the best and worst automatic segmentation of hippocampal subfields obtained by DeepHSS trained with 15 subjects and the associated manual segmentations. By comparison of the best hippocampal subfield segmentations obtained using DeepHSS and the corresponding manual GT labels illustrated, those are found very similar. However, small deviations are present due to disagreement of subfield boundary, which are marked by arrows in Figure 6.6. The worst hippocampal subfield segmentations obtained using DeepHSS are affected by under-segmentations. Moreover, a few anatomical misdetections

are observed. As an example, ERC has been incorrectly detected in the parahippocampal gyrus (see Figure 6.6, arrow in fourth column, Figure 3).

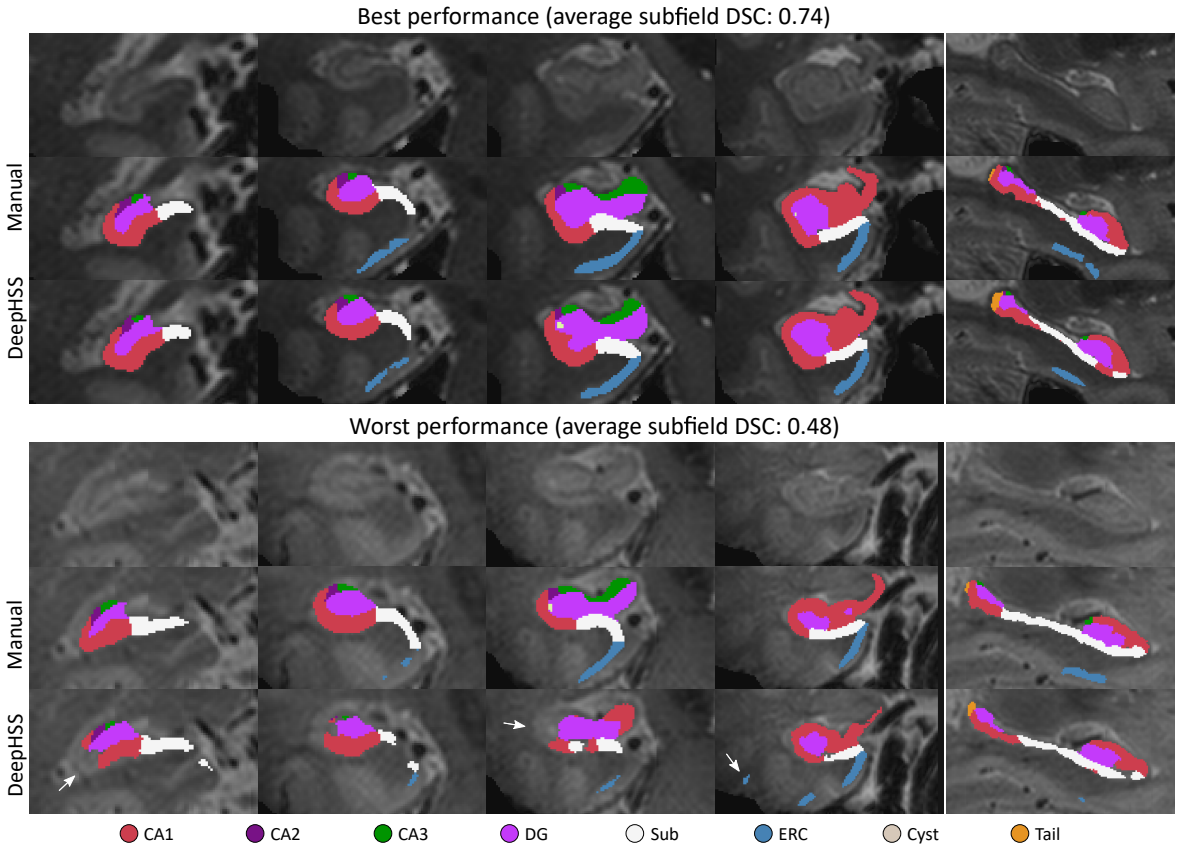


Figure 6.6: Coronal and sagittal views of the best and worst predicted hippocampal subfield segmentations obtained using DeepHSS compared to the corresponding GT labels. The best performance has an average subfield DSC at 0.74 (exclusive tail and cyst) whilst the worst performance scores an average subfield DSC at 0.48 (exclusive tail and cyst). Arrows point out differences between the segmentations.

6.2.3 Discussion

The proposed automatic hippocampal subfield segmentation method, DeepHSS, consists of a preprocessing pipeline and a two-pathway deep CNN, which was developed using the Deepmedic framework [Kamnitsas et al., 2017]. DeepHSS was trained and validated using ultra-high field in vivo T2w MR images and associated manual segmentations, which have previously been used to validate ASHS. DeepHSS demonstrated a fast segmentation process (an hour and a quarter) and achieved high segmentation accuracy comparable to manual segmentations and segmentations obtained using existing automatic methods. The high performance of the method demonstrates the flexibility of CNNs, and supports this as an easy method to apply for various segmentation problems.

Number of training subjects

The foreground DSC for all hippocampal subfields improved as the number of training subjects increased. However, it was observed that the foreground DSC increased less as more training subjects were applied, indicating the learning rate was converging towards a steady state. This finding is consistent with existing research by Cho et al. [2016], who reached a steady state after applying 200 images in the training of a CNN, classifying CT images of

various body part [Cho et al., 2016].
Approaching a steady state using only 15 training subjects could be explained by the homogenety of the segmentation problem at hand, as low variability of the hippocampus anatomy between subjects are observed. For example Wisse et al. [2016] measured the volume of the manually segmented hippocampal subfields for the same dataset applied in this test, and the subfield with the largest standard deviation was CA1, which was only 1.53±0.23 mL. If the dataset included a wider population, e.g. AD patients, more training subjects might be necessary to obtain accurate segmentations, since the hippocampus properties would vary more. If AD patients were included the volume of the subfileds would differ more, due to the atrophy of hippocampal subfields [Maruszak and Thuret, 2014].

Another reason why a steady state was approched applying a small dataset is, that the images in the dataset were acquired similarly using the same scanner and acquisition settings. Cho et al. [2016] reached a steady state after 200 training images using a dataset composed of CT images, which were acquired using different acquisition settings, such as radiation dosage, and different image reconstruction filters [Cho et al., 2016]. When images are acquired differently more training images might be necessary for the network to learn and compensate for these differences.

The quality of the dataset and associated labels might also influence the required number of training subjects. In the initial tests (see Section 5.2.3) a performance of approximately 0.9 was obtained when training with only 6 subjects. However, the quality of the dataset applied in the initial test were higher compared to the dataset used in this Section, as illustrated in Figure 6.7.

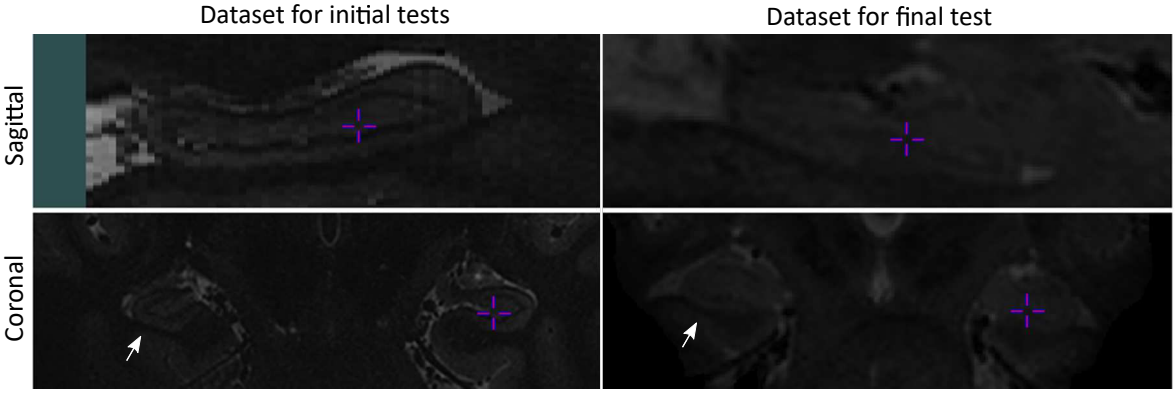


Figure 6.7: Coronal and sagittal views of two subjects from the dataset used in initial tests and dataset used in this final test, respectively.

Higher image quality results in e.g. more clear borders, by which the subfields appear more clear. Additionally, the labels used in the initial tests were based on automatic segmentation of the subfields. Automatic segmentation methods find the statistical most likely position of the subfields and is not influenced by subjective observations. Thereby it might be easier for an other automatic method to find similar labels.

DeepHSS's segmentation performance

When the CNN in DeepHSS was trained with 15 subjects it was possible to segment the hippocampal subfields ($0.49 < \text{DSC} < 0.80$).

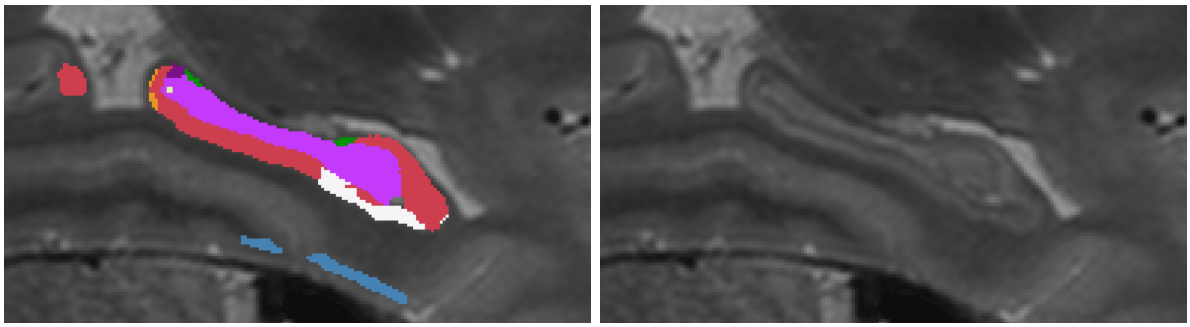
The performances of DeepHSS and ASHS were compaired and are presented in Tabel 6.2. It is notable that DeepHSS performs better for the small subfields (CA2 and CA3) than ASHS, whereas ASHS performs better for the largest subfield (CA1). Wisse et al. [2016] obtained the lowest accuracy for CA2, CA3 and ERC, and explained it by a correspondingly

low interrater repeatability for these subfields ($0.27 < \text{ICC} < 0.88$). Additionally, the anterior and posterior boundaries of these subfields are based on geometric rules rather than visibility of textual changes. [Wisse et al., 2016]

Table 6.2: Dice similarity measure for right hippocampal subfields. A comparison between results obtained using DeepHSS, and ASHS by Wisse et al. [2016]. The difference between scores were estimated by the formula. * one subject was removed from the test dataset

Hippocampal subfield	DeepHSS (DSC)	DeepHSS (DSC)*	ASHS (DSC)
CA1	0.61 ± 0.06	0.61 ± 0.06	0.83 ± 0.02
CA2	0.69 ± 0.05	0.70 ± 0.04	0.65 ± 0.09
DG	0.77 ± 0.05	0.78 ± 0.03	0.84 ± 0.03
CA3	0.59 ± 0.19	0.63 ± 0.02	0.54 ± 0.13
Sub	0.80 ± 0.05	0.81 ± 0.04	0.78 ± 0.04
ERC	0.49 ± 0.14	0.53 ± 0.09	0.75 ± 0.06

DeepHSS is affected by the same limitations of the training data as Wisse et al. [2016]. The accuracy of CA1 when segmented using DeepHSS was low. By visual inspection of the CA1 segmentation it was noticed that DeepHSS overall performs accurate segmentation of CA1 compared to the manual segmentation. The low accuracy of CA1 can be explained by undersegmentation, as seen in Figure 6.6, and other parts of the brain outside hippocampus being misclassified as CA1, as illustrated in Figure 6.8.



(a) **(b)**

Figure 6.8: The Figure illustrates a sagittal slice from a single subject 7T MRI with TSE contrast (a) without labels and (b) with labels obtained using DeepHSS. The red area left to the hippocampus are voxels misclassified as CA1.

Another factor reducing the overall average performance of DeepHSS is a notably whose performance for one subject. As seen in Table 6.1 the average subfield DSC when training the CNN in DeepHSS with 15 subjects is 0.66 ± 0.08, and the subject having a notably whose performance has a average DSC of 0.48, which is far lower than the standard deviation. As presented in Table 6.2, removal of this subject increase the average DSC and reduce the standard deviation for all but one subfield. The poor segmentation for this subject can be explained by variation in the dataset, i.e. the scan of this subject seems to have a different contrast compared to the scan of the subject for which DeepHSS performs the best, see Figure 6.6.

Bibliography

M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage*, 2016.

C. Boutet, M. Chupin, S. Lehéricy, L. Marrakchi-Kacem, S. Epelbaum, C. Poupon, C. Wiggins, A. Vignaud, D. Hasboun, B. Defontaines, O. Hanon, B. Dubois, M. Sarazin, L. Hertz-Pannier, and O. Colliot, "Detection of volume loss in hippocampal layers in alzheimers disease using 7t mri - a feasibility study," *NeuroImage: Clinical*, vol. 5, pp. 341–348, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2014.07.011

N. Burgess, E. A. Maguire, and J. O'Keefe, "The human hippocampus and spatial and episodic memory," *Neuron*, vol. 35, 2002.

J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" 2016.

H. Choi and K. H. Jin, "Fast and robust segmentation of the striatum using deep convolutional neural networks," *Journal of Neuroscience Methods*, vol. 274, pp. 146–153, dec 2016. [Online]. Available: https://doi.org/10.1016%2Fj.jneumeth.2016.10.007

R. Coras, G. Milesi, I. Zucca, A. Mastropietro, A. Scotti, M. Figini, A. Ḿuhlebner, A. Hess, W. Graf, G. Tringali, I. Bĺumcke, F. Villani, G. Didato, C. Frassoni, R. Spreafico, and R. Garbelli, "7t mri features in control human hippocampus and hippocampal sclerosis: An ex vivo study with histologic correlations," *Epilepsia*, vol. 55, no. 12, pp. 2003–2016, nov 2014. [Online]. Available: http://dx.doi.org/10.1111/epi.12828

A. C. Evans, A. L. Janke, D. L. Collins, and S. Baillet, "Brain templates and atlases," *NeuroImage*, vol. 62, no. 2, pp. 911–922, aug 2012. [Online]. Available: https://doi.org/10.1016%2Fj.neuroimage.2012.01.024

M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state"," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, nov 1975.

A. Giuliano, G. Donatelli, M. Cosottini, M. Tosetti, A. Retico, and M. E. Fantacci, "Hippocampal subfields at ultra high field MRI: An overview of segmentation and measurement methods," *Hippocampus*, vol. 27, no. 5, pp. 481–494, feb 2017. [Online]. Available: https://doi.org/10.1002%2Fhipo.22717

G. Grabner, A. L. Janke, M. M. Budge, D. Smith, J. Pruessner, and D. L. Collins, "Symmetric Atlasing and Model Based Segmentation: An Application to the Hippocampus in Older Adults," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*. Springer Science + Business Media, 2006, pp. 58–66. [Online]. Available: http://dx.doi.org/10.1007/11866763_8

J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liub, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," *Computer Vision and Pattern Recognition*, 2017, arXiv:1512.07108v5.

T. Hartley, C. Lever, N. Burgess, and J. O'Keefe, "Space in the brain: how the hippocampal formation supports spatial cognition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1635, pp. 20 120 510–20 120 510, dec 2013.

M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, jan 2017. [Online]. Available: https://doi.org/10.1016%2Fj.media.2016.05.004

K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification."

J. E. Iglesias, J. C. Augustinack, K. Nguyen, C. M. Player, A. Player, M. Wright, N. Roy, M. P. Frosch, A. C. McKee, L. L. Wald, B. Fischl, and K. V. Leemput, "A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI," *NeuroImage*, vol. 115, pp. 117–137, jul 2015. [Online]. Available: https://doi.org/10.1016%2Fj.neuroimage.2015.04.042

S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

A. L. Janke and J. F. Ullmann, "Robust methods to create ex vivo minimum deformation atlases for brain mapping," *Methods*, vol. 73, pp. 18–26, feb 2015.

A. L. Janke, K. O'Brien, S. Bollmann, T. Kober, L. Marstaller, and M. Barth, "A 7t human brain microstructure atlas by minimum deformation averaging at 300um," in *24th Annual ISMRM Scientific Meeting and Exhibition, Singapore*, 2016.

M. Jenkinson, M. Pechaud, and S. Smith, "Bet2: Mr-based estimation of brain, skull and scalp surfaces." *In Eleventh Annual Meeting of the Organization for Human Brain Mapping*, 2015.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," 2014.

R. L. Joie, A. Perrotin, V. de La Sayette, S. Egret, L. Doeuvre, S. Belliard, F. Eustache, B. Desgranges, and G. Chételat, "Hippocampal subfield volumetry in mild cognitive impairment alzheimer's disease and semantic dementia," *NeuroImage: Clinical*, vol. 3, pp. 155–162, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2013.08.007

K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, feb 2017. [Online]. Available: https://doi.org/10.1016%2Fj.media.2016.10.004

D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.

J. Kleesiek, G. Urbana, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep mri brain extraction: A 3d convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, 2016.

Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

K. V. Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl, "Automated segmentation of hippocampal subfields from ultra-high resolution in vivo mri," *Hippocampus*, vol. 19, no. 6, pp. 549–557, jun 2009. [Online]. Available: https://doi.org/10.1002%2Fhipo.20615

P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Molecular pharmaceutics*, vol. 13, pp. 1445–1454, 2016.

A. Maruszak and S. Thuret, "Why looking at the whole hippocampus is not enough - a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for alzheimer's disease diagnosis," *Frontiers in Cellular Neuroscience*, vol. 8, mar 2014.

P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Isgum, "Automatic segmentation of mr brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, may 2016. [Online]. Available: https://doi.org/10.1109%2Ftmi.2016.2548501

A. Y. Ng, "Feature selection, l1 vs. l2 regularization and rotational invariance," 2004.

D. Nie, L. Wang, Y. Gao, and D. Sken, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," *Proc IEEE Int Symp Biomed Imaging*, apr 2016. [Online]. Available: https://doi.org/10.1109%2Fisbi.2016.7493515

S. J. Nowlan and G. E. Hinton, "Simplifying neural networks by soft weight-sharing," *Neural Computation*, 1992.

P. Panegyres, R. Berry, and J. Burchell, "Early dementia screening," *Diagnostics*, vol. 6, no. 1, p. 6, jan 2016. [Online]. Available: http://dx.doi.org/10.3390/diagnostics6010006

S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, may 2016. [Online]. Available: https://doi.org/10.1109%2Ftmi.2016.2538465

J. Pipitone, M. T. M. Park, J. Winterburn, T. A. Lett, J. P. Lerch, J. C. Pruessner, M. Lepage, A. N. Voineskos, and M. M. Chakravarty, "Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates," *NeuroImage*, vol. 101, pp. 494–512, nov 2014. [Online]. Available: https://doi.org/10.1016%2Fj.neuroimage.2014.04.054

M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, "World alzheimer report 2015. the global impact of dementia. an analysis of prevalence, incidence, cost and trends," Alzheimer's Disease International, Tech. Rep., 2015.

M. Rajchl, M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE TRANSACTIONS ON MEDICAL IMAGING*, vol. 36, no. 2, pp. 674–683, 2017.

S. Ruder, "An overview of gradient descent optimization algorithms," 2016.

H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, oct 2000.

S. A. Small, S. A. Schobel, R. B. Buxton, M. P. Witter, and C. A. Barnes, "A pathophysiological framework of hippocampal dysfunction in ageing and disease," *Nature Reviews Neuroscience*, vol. 12, no. 10, pp. 585–601, sep 2011. [Online]. Available: https://doi.org/10.1038%2Fnrn3085

R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, D. C. Park, E. M. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V.

Wagster, and C. H. Phelps, "Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 280–292, may 2011. [Online]. Available: https://doi.org/10.1016%2Fj.jalz.2011.03.003

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Reasearch*, 2014.

C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields*. Now Publishers, 2012, vol. 4, no. 4. [Online]. Available: https://doi.org/10.1561%2F2200000013

J. F. P. Ullmann, C. Watson, A. L. Janke, N. D. Kurniawan, G. Paxinos, and D. C. Reutens, "An MRI atlas of the mouse basal ganglia," *Brain Structure and Function*, vol. 219, no. 4, pp. 1343–1353, may 2013. [Online]. Available: https://doi.org/10.1007%2Fs00429-013-0572-0

A. G. van der Kolk, J. Hendrikse, J. J. Zwanenburg, F. Visser, and P. R. Luijten, "Clinical applications of 7t MRI in the brain," *European Journal of Radiology*, vol. 82, no. 5, pp. 708–718, may 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ejrad.2011.07.007

S. Vieira, W. H. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders methods and applications," *Neuroscience and Biobehavioral Reviews*, 2017.

J. L. Winterburn, J. C. Pruessner, S. Chavez, M. M. Schira, N. J. Lobaugh, A. N. Voineskos, and M. M. Chakravarty, "A novel in vivo atlas of human hippocampal subfields using high-resolution 3t magnetic resonance imaging," *NeuroImage*, vol. 74, pp. 254–265, jul 2013.

L. E. M. Wisse, H. J. Kuijf, A. M. Honingh, H. Wang, J. B. Pluta, S. R. Das, D. A. Wolk, J. J. M. Zwanenburg, P. A. Yushkevich, and M. I. Geerlings, "Automated hippocampal subfield segmentation at 7T MRI," *American Journal of Neuroradiology*, vol. 37, no. 6, pp. 1050–1057, feb 2016. [Online]. Available: https://doi.org/10.3174%2Fajnr.a4659

L. E. Wisse, A. M. Daugherty, R. K. Olsen, D. Berron, V. A. Carr, C. E. Stark, R. S. Amaral, K. Amunts, J. C. Augustinack, A. R. Bender, J. D. Bernstein, M. Boccardi, M. Bocchetta, A. Burggren, M. M. Chakravarty, M. Chupin, A. Ekstrom, R. de Flores, R. Insausti, P. Kanel, O. Kedo, K. M. Kennedy, G. A. Kerchner, K. F. LaRocque, X. Liu, A. Maass, N. Malykhin, S. G. Mueller, N. Ofen, D. J. Palombo, M. B. Parekh, J. B. Pluta, J. C. Pruessner, N. Raz, K. M. Rodrigue, D. Schoemaker, A. T. Shafer, T. A. Steve, N. Suthana, L. Wang, J. L. Winterburn, M. A. Yassa, P. A. Yushkevich, and R. la Joie, "A harmonized segmentation protocol for hippocampal and parahippocampal subregions why do we need one and what are the key goals?" *HIPPOCAMPUS*, vol. 27, pp. 3–11, 2017.

P. A. Yushkevich, J. B. Pluta, H. Wang, L. Xie, S.-L. Ding, E. C. Gertje, L. Mancuso, D. Kliot, S. R. Das, and D. A. Wolk, "Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment," *Human Brain Mapping*, vol. 36, no. 1, pp. 258–287, sep 2015. [Online]. Available: https://doi.org/10.1002%2Fhbm.22627

S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," *ArXiv*, 2016.

Appendix A
Minimum deformation averaging

Minimum deformation averaging (MDA) is a method used to create models visualizing the average morphology of a population. From the chosen population the model only includes consistent structures which can be aligned using non-linear registration. Before the MDA model can be generated the dataset representing the population is pre-processed by manual inspection to identify and remove obvious artefacts. Additionally, the intensities of the dataset is normalized to remove very low or high intensities. From the pre-processed dataset a MDA model is build by iterating through a sequence of four steps, as illustrated in figure A.1. [Janke and Ullmann, 2015]

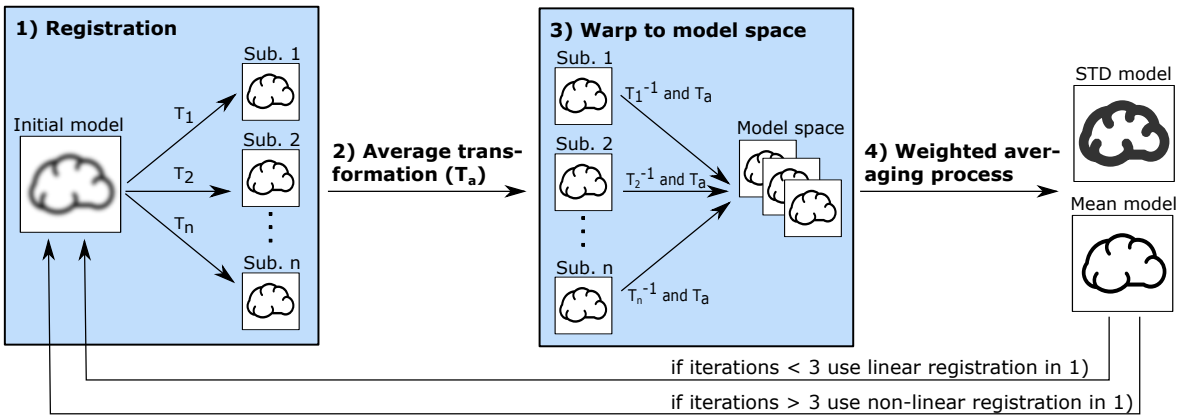


Figure A.1: Illustration of how a MDA model is generated through a iterative process of four steps. In step 1) is the initial model aligned to each subject using registration and in step 2) is the transformations used to achieve the alignments averaged. In step 3) is the subject images transformed into model space. Lastly, the mean and the STD model is obtained using a weighted averaging process. Modified from [Janke and Ullmann, 2015]

To enable the first step in the first iteration an initial model must be chosen, e.g. the subject with the highest signal-to-noise ratio. The initial model is then symmetrically aligned and blurred to remove the features of the subject. In the first step, the initial model is aligned to each subject using affine registration and normalized cross correlation as the similarity measure. On the basis of the alinements model-to-subject transformations are obtained and an average transformation can be determined in the second step. In the third step, each image is transformed into the model space by applying the inverted model-to-subject transformations and later the average transformation. In the final step, the mean and the standard deviation (STD) model is created by applying a weighted averaging process to the STD image. This process compare each voxel of the subject image to the corresponding voxel in the mean STD image and quantify the match using a z-score. Finally, masks are made for each of the subject images by using a z-score threshold and these masks are used to make the first version of the mean and the STD model. In the next iteration the first version is applied as the initial model, and the four steps are repeated minimum three times until the model is stable. The process is repeated again using non-linear registration to align larger features such as large brain structures and the overall brain shape. To track the fitting process and determine when the final model is achieved the STD model can be used. [Janke and Ullmann, 2015]

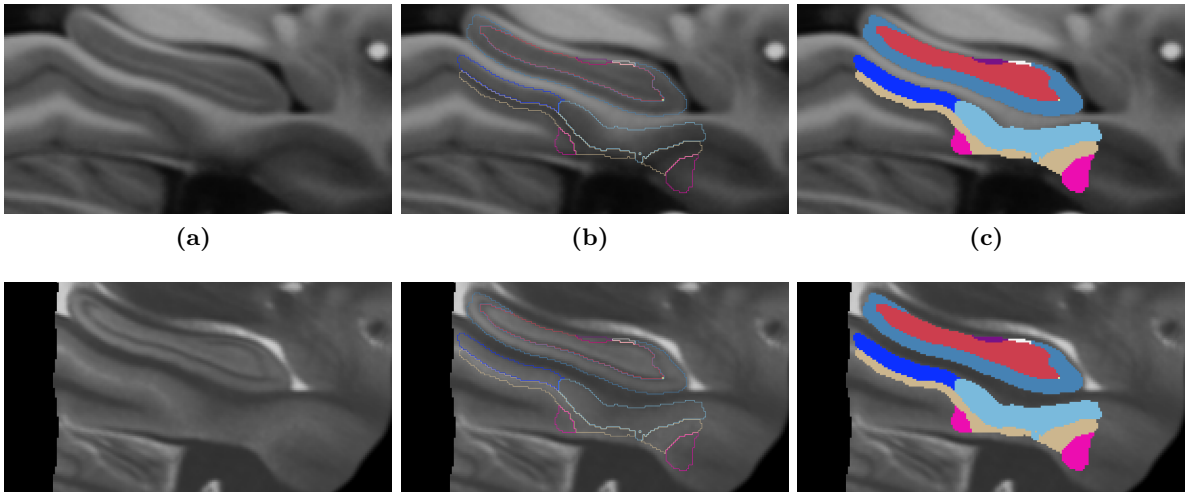
Appendix B
MDA model based generation of ground truth labels

```
#!/bin/bash
#
#SBATCH --account aauhpc_fat       # account
#SBATCH --nodes 1-6                # number of nodes
#SBATCH --time 15:00:00            # max time (HH:MM:SS)
#SBATCH --job-name=DataLabelling
#System variables

export FREESURFER_HOME=/path/to/freesurfer
source $FREESURFER_HOME/SetUpFreeSurfer.sh
export SUBJECT_DIR1=/path/to/subject/

echo "Setup done"
echo "Running segmentation - recon-all -all"

recon-all -i MP2RAGEmodel.nii -s <name>
recon-all -s <name> -all
recon-all -s <name> -hippocampal-subfields-T2 TSEmodel.nii t2only
recon-all -s <name> -hippocampal-subfields-T1T2 TSEmodel.nii t1t2
echo "Segmentation done"

echo "Use volcentre to warp Subject into Zero subject space \
for resampling proterties"

volcentre -zero_dircos $SUBJECT_DIR1/s16_sb_subjectSpace_MP2RAGE.mnc \
$SUBJECT_DIR1/s16_sb_subjectSpace_MP2RAGE_ZeroCenter.mnc \

volcentre -zero_dircos $SUBJECT_DIR1/s16_sb_subjectSpace_TSE.mnc \
$SUBJECT_DIR1/s16_sb_subjectSpace_TSE_ZeroCenter.mnc \


echo "Warp segmentations into subject space"
echo "Running mincresample MP2RAGE model space to \
MP2RAGE zero centre space, lh and rh"

mincresample -nearest_neighbour -transformation \
$SUBJECT_DIR1/0005-s16_sb_20150518_150218_6_mri.xfm \
$SUBJECT_DIR1/lh.hippoSfLabels-T1-t1t2.v10.mnc \
$SUBJECT_DIR1/lh.hippoSfLabels_s16_MP2RAGE_zeroCentreSpace.mnc \
-like $SUBJECT_DIR1/s16_sb_subjectSpace_MP2RAGE_ZeroCenter.mnc
```

```
mincresample -nearest_neighbour -transformation \
$SUBJECT_DIR1/0005-s16_sb_20150518_150218_6_mri.xfm \
$SUBJECT_DIR1/rh.hippoSfLabels-T1-t1t2.v10.mnc \
$SUBJECT_DIR1/rh.hippoSfLabels_s16_MP2RAGE_zeroCentreSpace.mnc \
-like $SUBJECT_DIR1/s16_sb_subjectSpace_MP2RAGE_ZeroCenter.mnc


echo "Running antsApplyTransforms and mincresample, MP2RAGE model space to
TSE model space"
echo "TSE lh"

antsApplyTransforms -n NearestNeighbor -t \
$SUBJECT_DIR1/CC_zchopped_original_tse0_inverse_NL.xfm -i \
$SUBJECT_DIR1/lh.hippoSfLabels-T1-t1t2.v10.mnc -r \
$SUBJECT_DIR1/tseModel_L15_hippocampus-TSE-7T-sym-mincanon_v0.8.mnc -o \
$SUBJECT_DIR1/lh.hippoSfLabels_s16_TSE_TSEmodelSpaceOnlyCC.mnc

mincresample -nearest_neighbour -transformation \
$SUBJECT_DIR1/tseModelSpace2mp2rage_Modelspacelsq6_manualinit.xfm \
$SUBJECT_DIR1/lh.hippoSfLabels_s16_TSE_TSEmodelSpaceOnlyCC.mnc \
$SUBJECT_DIR1/lh.hippoSfLabels_s16_TSE_TSEmodelSpace.mnc \
-like $SUBJECT_DIR1/tseModel_L15_hippocampus-TSE-7T-sym-mincanon_v0.8.mnc \
-invert_transformation

echo "TSE rh"

antsApplyTransforms -n NearestNeighbor -t \
$SUBJECT_DIR1/CC_zchopped_original_tse0_inverse_NL.xfm -i \
$SUBJECT_DIR1/rh.hippoSfLabels-T1-t1t2.v10.mnc -r \
$SUBJECT_DIR1/tseModel_L15_hippocampus-TSE-7T-sym-mincanon_v0.8.mnc -o \
$SUBJECT_DIR1/rh.hippoSfLabels_s16_TSE_TSEmodelSpaceOnlyCC.mnc

mincresample -nearest_neighbour -transformation \
$SUBJECT_DIR1/tseModelSpace2mp2rage_Modelspacelsq6_manualinit.xfm \
$SUBJECT_DIR1/rh.hippoSfLabels_s16_TSE_TSEmodelSpaceOnlyCC.mnc \
$SUBJECT_DIR1/rh.hippoSfLabels_s16_TSE_TSEmodelSpace.mnc \
-like $SUBJECT_DIR1/tseModel_L15_hippocampus-TSE-7T-sym-mincanon_v0.8.mnc \
-invert_transformation

echo "Mincresample TSE model space to zero centre model space"
echo "TSE lh"

mincresample -nearest_neighbour -transformation \
$SUBJECT_DIR1/s16_sb/0012-average-normStepSize-norm-clamped-norm-iso.xfm \
$SUBJECT_DIR1/lh.hippoSfLabels_s16_TSE_TSEmodelSpace.mnc \
$SUBJECT_DIR1//lh.hippoSfLabels_s16_TSE_zeroCentreSpace.mnc \
-like $SUBJECT_DIR1/s16_sb_subjectSpace_TSE_ZeroCenter.mnc

echo "TSE rh"
```

```
mincresample -nearest_neighbour -transformation \
$SUBJECT_DIR1/s16_sb/0012-average-normStepSize-norm-clamped-norm-iso.xfm \
$SUBJECT_DIR1/rh.hippoSfLabels_s16_TSE_TSEmodelSpace.mnc \
$SUBJECT_DIR1/rh.hippoSfLabels_s16_TSE_zeroCentreSpace.mnc \
-like $SUBJECT_DIR1/s16_sb_subjectSpace_TSE_ZeroCenter.mnc
```

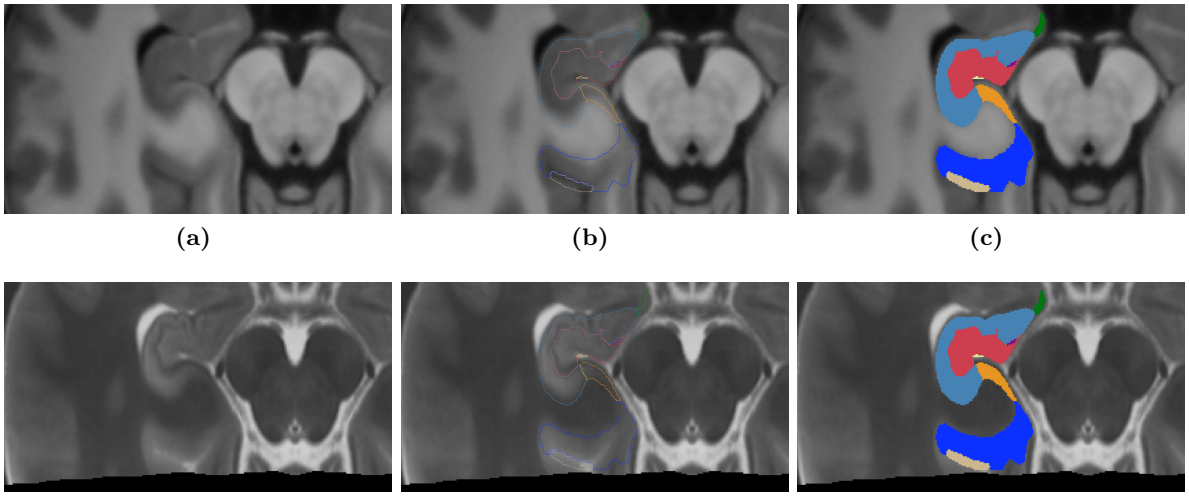
Illustration of ground truth: Sagittal and transverse view



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure B.1: Sagittal view of the hippocampal subfield segmentation obtained using multispectral segmentation with ASHS. The segmentation was generated using both the T1w (a-c) and T2w (d-f) 7T MDA MRI model. a and e has the outline of the labels superimposed unto them. c and f has the filled labels superimposed unto them. The white arrows in b and c points out the posterior end of hippocampus, where the labels does not fully cover.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure B.2: Transverse view of the hippocampal subfield segmentation obtained using multispectral segmentation with ASHS. The segmentation was generated using both the T1w (a-c) and T2w (d-f) 7T MDA MRI model. a and e has the outline of the labels superimposed unto them. c and f has the filled labels superimposed unto them.

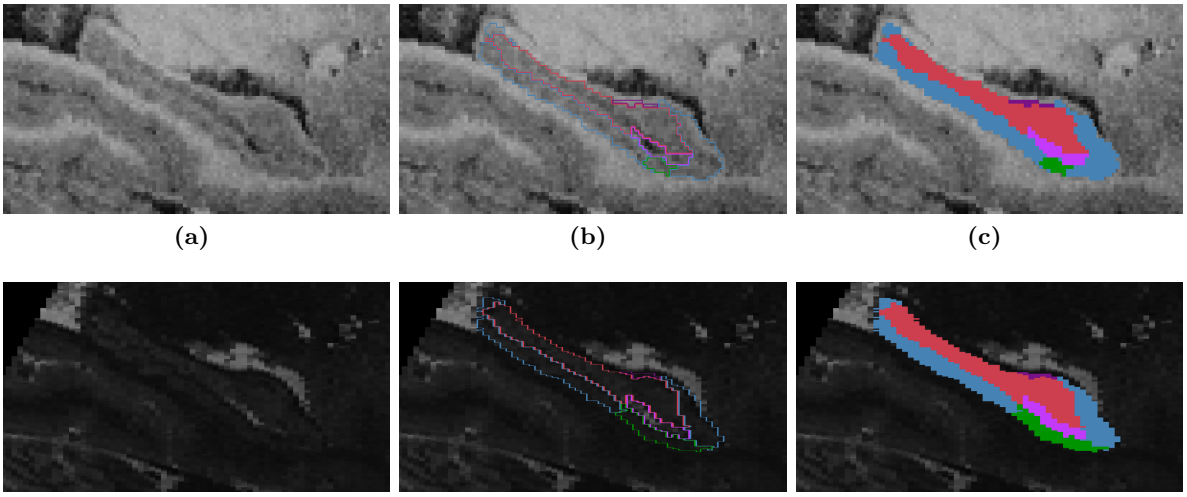


Figure B.3: Sagittal view of the hippocampal subfield segmentation obtained using multispectral segmentation with ASHS. The segmentation was generated using both the T1w and T2w 7T MDA MRI model and warped into subject space. a-c is a slice capturing the hippocampus at a T1w 7T scan, whilst d-f is a slice capturing the hippocampus at a T2w 7T scan. a and e has the outline of the labels superimposed unto them. c and f has the filled labels superimposed unto them. All labels has been post processed to have 6 labels.

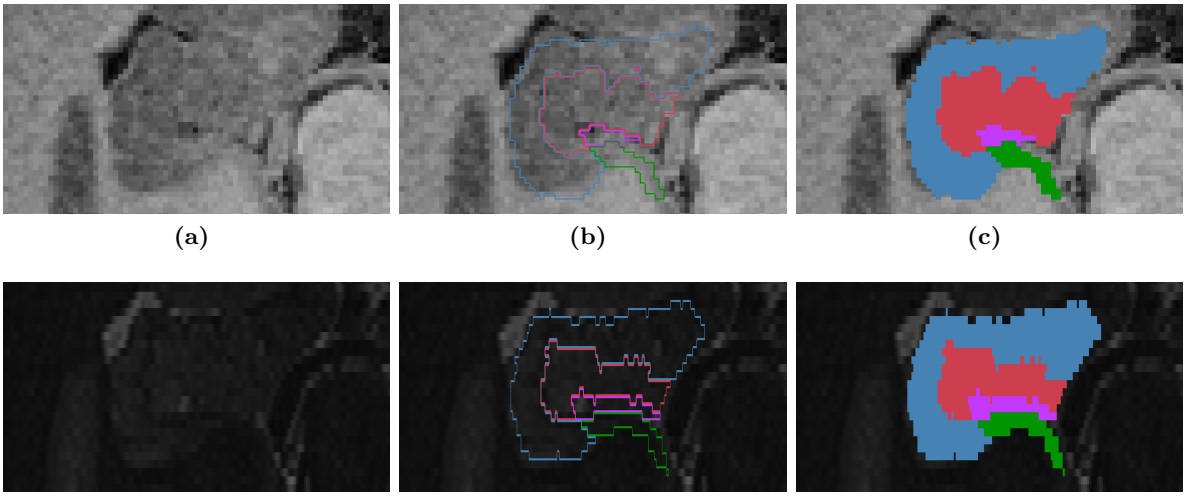


Figure B.4: Transverse view of the hippocampal subfield segmentation obtained using multispectral segmentation with ASHS. The segmentation was generated using both the T1w and T2w 7T MDA MRI model and warped into subject space. a-c is a slice capturing the hippocampus at a T1w 7T scan, whilst d-f is a slice capturing the hippocampus at a T2w 7T scan. a and e has the outline of the labels superimposed unto them. c and f has the filled labels superimposed unto them. All labels has been post processed to have 6 labels.

Appendix C
Optimization algorithms and performance measures

Optimization algorithms

The most commonly used optimization algorithm to minimize an error function and find a NN's parameters is Gradient Decent (GD). Overall, there exist three versions of GD; Batch Gradient Decent(BGD), Stochastic Gradient Decent(SDG), Mini-batch gradient descent. These optimizers differ in the amount of data they use to make the computations and the amount of data determine a trade-off between computation time and the accuracy of the network's parameters. BGD computes the error function's gradient for the parameters for the whole training set, whereas SGD updates the parameters for each training data in the training set. Consequently, BGD preform redundant calculations and SDG is usually faster than BGD. On the other hand, SGD can cause fluctuations of the error function, which can complicate convergence to the minimum. These issues are not experienced for BGD. [Ruder, 2016]

Mini-batch gradient descent takes the best from both methods by calculating the gradient of the error function for N mini-batches. This approach compute the gradient for the mini-batches very efficient and lead to a more stable convergence. [Ruder, 2016]

The GD methods are frequently applied in research as optimizer. However, challenges are experienced using these methods, among these are:

1. The learning rate must be manually chosen
2. Application on small datasets
3. Can get trapped in a suboptimal local minima
4. Slow at finding the minima.

To accommodate for these challenges adaptive learning rate methods were developed. Among these are AdaGrad, which is gradient based and adapt the learning rate to the frequency of the parameters. Consequently, the method is good when data is sparse. Additionally, the method increase the robustness of SGD and have been used to train large NN with success [Ruder, 2016]. In the GD methods the parameters of a network was all updated at once and the learning rate was the same during training. In the contrary, at time step t AdaGrad apply different learning rates of each of the parameters and the learning rates are modified at each time step. Thereby, manual adjustment is avoided. The main weakness of AdaGrad is, that the learning rate always decrease and at some point become infinitesimally small. [Ruder, 2016] A method developed to solve this problem is RMSProp, which is also an adaptive learning rate method. RMSProp solve AdeGrad's diminishing learning rate by dividing the learning rate by an exponentially decreasing average of squared gradients. [Ruder, 2016]

A third adaptive learning rate method is Adam, which take the advantages from AdaGrad and RMSProp. [Kingma and Ba, 2015] Adam's pseudo-code is illustrated in figure C.1. Like RMSprop, Adam applies an exponentially decreasing average of past squared gradients (v_t), but in addition Adam also uses an exponentially decreasing average of past gradients (m_t). m_t and v_t correspond to the estimate of the gradients' first moment (mean) and second

moment (variance), respectively. Another difference between RMSprop and Adam is, that Adam added a bias-correction. [Ruder, 2016]

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
  $m_0 \leftarrow 0$ (Initialize 1st moment vector)
  $v_0 \leftarrow 0$ (Initialize 2nd moment vector)
  $t \leftarrow 0$ (Initialize timestep)
  while θ_t not converged **do**
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
    $\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
    $\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
  end while
  return θ_t (Resulting parameters)

Figure C.1: Pseudo-code of how Adam function. In algorithm the operations are all elementwise. Kingma and Ba [2015] suggest $\epsilon = 10^{-8}$, $\alpha = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. From Kingma and Ba [2015].

Overall, the adaptive learning rate methods are better compared to the GD, due to the four points listed above. Among the adaptive learning rate methods the best choice of optimizer might be Adam. RMSprop and Adam are very similar and they both compensate for AdaGrad's weakness. However, compared to RMSprop Adam have shown a slightly better performance in the end of the optimization when the gradients are sparser. [Ruder, 2016]

Performance measures

To evaluate the performance the network the accuracy, sensitivity, specificity and dice score was calculated on test data. This was done regularly throughout the training procedure, allowing to follow the performance of the CNN as a function of the number of training epochs.

Accuracy was used to evaluate the performance of the proposed CNN and is the number of correct classifications of the validation data voxels and the formula can be seen in equation C.1.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{C.1}$$

where TP is the number of voxels for a subfield classified true, TN is the number of background voxels classified true, FP is the number of voxels for a subfield classified false, and FN is the number of background voxels classified false.

Sensitivity was calculated as well to give a measure of the proposed CNN's ability to correctly classify the voxels of a hippocampal subfield. Sensitivity is calculated using equation C.2

$$Sensitivity = \frac{TP}{TP + FN} \tag{C.2}$$

To quantify the CNN's ability to correctly classify the background's voxels specificity was calculated too as precented in equation C.3

$$Specificity = \frac{TN}{FP + TN} \tag{C.3}$$

Lastly, dice score was calculated to evaluated the performance, since this method is often used to compare brain structure segmentations [Choi and Jin, 2016; Kamnitsas et al., 2017; Nie et al., 2016; Kleesiek et al., 2016; Havaei et al., 2017; Moeskops et al., 2016; Rajchl et al., 2017; Pereira et al., 2016]. Dice score measures the overlap between two segmentations and the formula can be seen in equation C.4

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{C.4}$$

where GT is the ground true segmentation and Seg is the segmentation performed by the proposed CNN. [Havaei et al., 2017]

Appendix D

Abstract to ESMRMB 2017 annual scientific meeting

This abstract was submitted to ESMRMB the 23rd of May 2017.

#655 3D convolutional neural network for hippocampal subfiels segmentation in ultra-high resolution MRI

N. Jacobsen[1], B.D. Hansen[1], A.K. Noehr[1], L.R. Oestergaard[1], S.B. Petersen[1], S. Bollmann[2]

[1] Aalborg, DK, Aalborg University, Department of Health Science and Technology
[2] St Lucia, AU, The University of Queensland, Centre for Advanced Imaging

Machine Learning
Scientific Session

Purpose / Introduction

Recent studies suggest that volumetric measurements of hippocampal subfields could deliver biomarkers for an early detection of Alzheimer's Disease [1,2,3,4,5,6], and increased access to 7T MRI have made segmentation of these subfields feasible [7,1]. Atlas-based methods for automatic segmentation of the hippocampal subfields have been proposed. However, these methods show limitations with respect to segmentation of smaller subfields and they are very time consuming [8,9,10]. In current research, Convolutional Neural Networks (CNN) have been used to segment brain structures and lesions, and the approach shows fast and accurate segmentations [11,12,13,14,15,16,17,18]. This study aims to explore CNNs as a fast and reliable method for hippocampal subfield segmentation.

Subjects and method

We used MP2RAGE and TSE contrast minimum deformation average models [18,19,20] to generate high-quality training labels. To achieve this we segmented the subfields in the high-resolution model space using FreeSurfer and transferred the labels to 3 healthy subjects utilizing the transformation matrices obtained during the average model construction. To further augment the dataset we flipped the 3 subjects and the corresponding labels.

The proposed segmentation method is a supervised 3D CNN developed using the framework DeepMedic [11]. The network has two parallel pathways consisting of 8 convolutional layers, 2 fully-connected layers, a softmax layer, and a CRF layer. The network was trained using an approach based on dense inference [11]. The network was trained on 5 subjects and tested on the left-out subject (Leave-one-out cross validation) to achieve a segmentation performance measurement. The predicted labels were compared to the training labels using dice scores.

Results

Figure F.1 represents average dice score through a training period of 15 epochs for each of the 6 subtests. The total average dice score was 0.9031±0.0182. Training time was between 19 and 20 hours. Figure F.2 illustrates the predicted segmentations for subtest 5 after 15 training epochs, compared to ground truth labels.

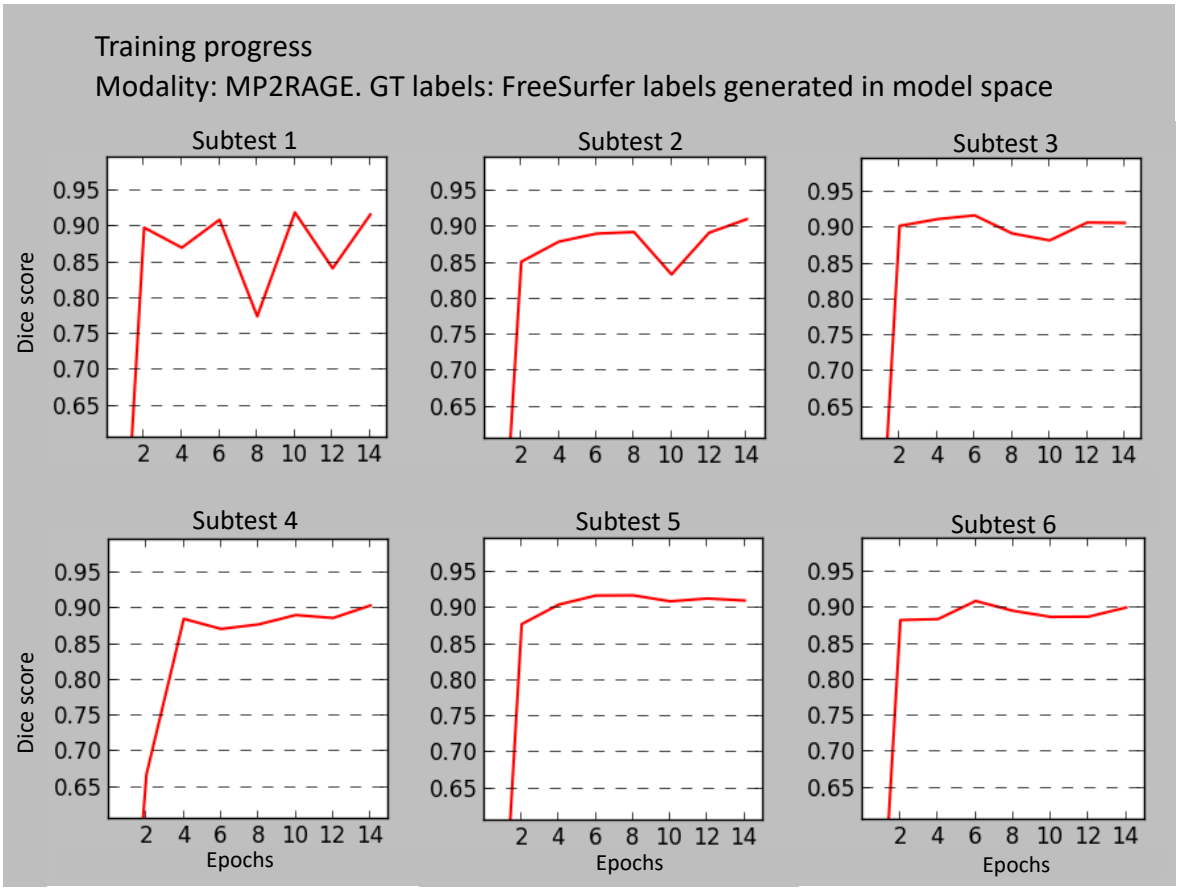


Figure D.1

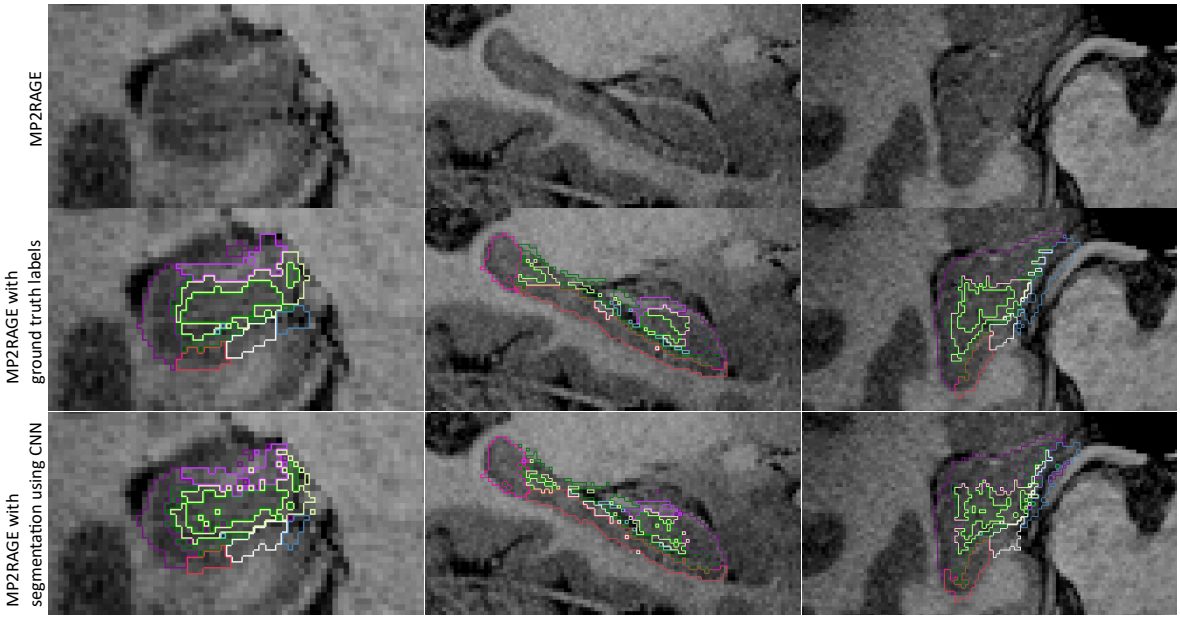


Figure D.2

Discussion / Conclusion

It was possible to segment hippocampal subfields with an average dice score of 0.9031 ± 0.0182. As illustrated in figure 2 the labels used as ground truth show some inconsistency in the boundaries for the subfields, reducing the requisite of the CNNs training and thereby the quality of the predicted segmentations.

The high precision regardless of the ground truth labels shows that CNNs with the right configurations and regularization methods are highly adaptable to new tasks, and perform well on segmentation tasks despite small datasets. Increasing the training dataset could reduce errors further and increase accuracy.

References

[1] C. Boutet, M. Chupin, S. Lehéricy, L. Marrakchi-Kacem, S. Epelbaum, C. Poupon, C. Wiggins, A. Vignaud, D. Hasboun, B. Defontaines, O. Hanon, B. Dubois, M. Sarazin, L. Hertz-Pannier, and O. Colliot, "Detection of volume loss in hippocampal layers in alzheimers disease using 7t mri - a feasibility study," NeuroImage: Clinical, vol. 5, pp. 341–348, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2014.07.011

[2] G. A. Kerchner, C. P. Hess, K. E. Hammond-Rosenbluth, D. Xu, G. D. Rabinovici, D. A. C. Kelley, D. B. Vigneron, S. J. Nelson, and B. L. Miller, "Hippocampal CA1 apical neuropil atrophy in mild alzheimer disease visualized with 7T MRI," Neurology, vol. 75, no. 15, pp. 1381–1387, oct 2010. [Online]. Available: http: //dx.doi.org/10.1212/WNL.0b013e3181f736a1

[3] R. L. Joie, A. Perrotin, V. de La Sayette, S. Egret, L. Doeuvre, S. Belliard, F. Eustache, B. Desgranges, and G. Chételat, "Hippocampal subfield volumetry in mild cognitive impairment alzheimer's disease and semantic dementia," NeuroImage: Clinical, vol. 3, pp. 155–162, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2013.08.007

[4] A. Maruszak and S. Thuret, "Why looking at the whole hippocampus is not enough - a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for alzheimer's disease diagnosis," Frontiers in Cellular Neuroscience, vol. 8, mar 2014

[5] J. Pluta, P. Yushkevich, S. Das, and D. Wolk, "In vivo analysis of hippocampal subfield

atrophy in mild cognitive impairment via semi-automatic segmentation of t2-weighted mri," Journal of Alzheimer's Disease, vol. 31, no. 1, pp. 85–99, 2012.

[6] A. G. van der Kolk, J. Hendrikse, J. J. Zwanenburg, F. Visser, and P. R. Luijten, "Clinical applications of 7t MRI in the brain," European Journal of Radiology, vol. 82, no. 5, pp. 708–718, may 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ejrad.2011.07.007

[7] L. E. M. Wisse, H. J. Kuijf, A. M. Honingh, H. Wang, J. B. Pluta, S. R. Das, D. A. Wolk, J. J. M. Zwanenburg, P. A. Yushkevich, and M. I. Geerlings, "Automated hippocampal subfield segmentation at 7T MRI," American Journal of Neuroradiology, vol. 37, no. 6, pp. 1050–1057, feb 2016. [Online]. Available: https://doi.org/10.3174

[8] J. E. Iglesias, J. C. Augustinack, K. Nguyen, C. M. Player, A. Player, M. Wright, N. Roy, M. P. Frosch, A. C. McKee, L. L. Wald, B. Fischl, and K. V. Leemput, "A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI," NeuroImage, vol. 115, pp. 117–137, jul 2015. [Online]. Available: https://doi.org/10.1016

[9] A. Giuliano, G. Donatelli, M. Cosottini, M. Tosetti, A. Retico, and M. E. Fantacci, "Hippocampal subfields at ultra high field MRI: An overview of segmentation and measurement methods," Hippocampus, vol. 27, no. 5, pp. 481–494, feb 2017. [Online]. Available: https://doi.org/10.1002

[10] H. Choi and K. H. Jin, "Fast and robust segmentation of the striatum using deep convolutional neural networks," Journal of Neuroscience Methods, vol. 274, pp. 146–153, dec 2016. [Online]. Available:

[11] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," Medical Image Analysis, vol. 36, pp. 61–78, feb 2017. [Online]. Available: https://doi.org/10.1016

[12] D. Nie, L. Wang, Y. Gao, and D. Sken, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," Proc IEEE Int Symp Biomed Imaging, apr 2016. [Online]. Available: https://doi.org/10.1109

[13] M. Rajchl, M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," IEEE TRANSACTIONS ON MEDICAL IMAGING, vol. 36, no. 2, pp. 674–683, 2017.

[14] J. Kleesiek, G. Urbana, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep mri brain extraction: A 3d convolutional neural network for skull stripping," NeuroImage, vol. 129, pp. 460–469, 2016

[15] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," Medical Image Analysis, vol. 35, pp. 18–31, jan 2017. [Online]. Available: https://doi.org/10.1016

[16] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Isgum, "Automatic segmentation of mr brain images with a convolutional neural network," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1252–1261, may 2016. [Online]. Available: https://doi.org/10.1109

[17] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1240–1251, may 2016. [Online]. Available: https://doi.org/10.1109

[18] Bollmann, Steffen, Andrew Janke, Lars Marstaller, David Reutens, Kieran O'Brien, and Markus Barth. "MP2RAGE T1-Weighted Average 7T Model," January 1, 2017. doi:10.14264/uql.2017.266.

[19] Munk, J., N. Jacobsen, M. Plocharski, L. R. Østergaard, M. Barth, A. Janke, and S. Bollmann. "Contrast Matching of Ultra-High Resolution Minimum Deformation Averaged

MRI Models to Facilitate Computation of a Multi-Modal Model of the Human Brain." In Proc. Intl. Soc. Mag. Reson. Med. 25, 1352. Honolulu, 2017.
http://indexsmart.mirasmart.com/ISMRM2017/PDFfiles/1352.html.

[20] Janke, Andrew L., and Jeremy F. P. Ullmann. "Robust Methods to Create Ex Vivo Minimum Deformation Atlases for Brain Mapping." Methods, Spatial mapping of multi-modal data in neuroscience, 73 (February 2015): 18–26. doi:10.1016/j.ymeth.2015.01.005.