Hough-CNN: Deep Learning for Segmentation of Deep Brain Regions in MRI and Ultrasound

Fausto Milletari*, Seyed-Ahmad Ahmadi*, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötzel, Nassir Navab

Abstract—In this work we analyse the performance of Convolutional Neural Networks (CNN) on medical data by benchmarking the capabilities of different network architectures to solve tasks such as segmentation and anatomy localisation, under clinically realistic constraints. We propose several CNN architectures with varying data abstraction capabilities and complexity, and we train them with different amounts of annotated image patches in order to analyse their performance when training sets of different size are used. Our implementation is capable of performing convolutions on volumetric data using 3D kernels, which enables further performance analysis when 2D, 2.5D and 3D data are employed during training and testing. Our novel segmentation approach does not only make use of voxel-wise classification outcomes produced by the network, but also considers high level features obtained from the second-last fully connected layer of the CNNs in a voting-based strategy. Through Hough voting and patch-wise back-projection from a multi-atlas, anatomy localisation and segmentation are performed and the region of interest is delineated in 3D by a smooth and regular contour. Our method has been tested on volumetric medical images acquired in two modalities for segmentation of deep-brain structures, such as the midbrain in ultrasound and a set of 26 regions in MRI.

Index Terms—Convolutional neural networks, deep learning, segmentation, Hough voting, Hough CNN, ultrasound, MRI.

I. INTRODUCTION

Recent research has shown the ability of convolutional neural networks (CNN) to deal with complex machine vision problems: unprecedented results were achieved in tasks such as classification [1], [2], segmentation, and object detection [3], [4], often outperforming human accuracy [5]. CNNs have the ability of learning a hierarchical representation of the input data without requiring any effort to design handcrafted features [6]. Different layers of the network are capable of different levels of abstraction and capture different amount of structure from the patterns present in the image [7]. Due to the complexity of the tasks and the very large number of network parameters that need to be learned during training, CNNs require a massive amount of annotated training images in order to deliver competitive results. As a consequence, significant performance increase can be achieved as soon as faster hardware and higher amount of training data become available [1].

In this work we investigate the applicability of convolutional neural networks to medical image analysis in order to perform segmentation of volumetric ultrasound (US) and MRI scans. For this reason we perform an unprecedented study on parameter variations and network architectures while we propose a novel segmentation framework based on Hough voting to delineate regions of interest for clinically relevant purposes.

Clinical applicability of these methods is a special focus of this paper. The paradigm-shifting results delivered by CNNs in computer vision were in parts accomplished with the help of web-scale training datasets and significant computational resources, both of which are largely unrealistic in clinical environments, at least so far. The amount of available data might significantly increase in future, in light of recent studies on crowd-based labeling of medical images [8] and successes with “gamification” [9]. Clinical research, however, still requires training on only few labeled examples, until larger datasets can be acquired through multi-centre studies or by leveraging...
crowd knowledge. Computational resources are another issue due to budget and space constraints in clinics. Outsourcing to third-party CPU or GPU clusters conflicts with confidentiality issues and closed IT networks in hospitals. Therefore, in this study, we perform all training and testing of CNN networks on clinically realistic dataset sizes, using a high-performance, but stand-alone PC workstation.

Segmentation of brain structures in US and MRI has widespread clinical relevance, but it is challenging in both modalities. In MRI, the segmentation of basal ganglia is a relevant task for medical diagnosis, treatment and clinical research. A concrete application is pre-operative planning of Deep Brain Stimulation (DBS) neurosurgery in which basal ganglia, like the sub-thalamic nucleus (STN) and globus pallidus internal (GPI), are targeted for treatment of symptoms of Parkinson’s disease (PD) and dystonia, respectively [10]. Accurate localisation and outlining of these nuclei can be challenging, even when performed manually, due to their weak contrast in MRI data. Moreover, fully manual labelling of individual MRIs into multiple regions in 3D is extremely time-consuming and often prohibitive. Clinicians rely on segmentations performed through deformable registration of the patient’s brain with a fully annotated atlas and subsequent transfer of anatomic labels using inverse deformation fields. In DBS planning for example, atlas-based segmentation is state-of-the-art in both research [10], [11] and commercial solutions [12]. Atlas-based methods have several disadvantages. First, the time necessary to perform a single registration ranges from 30 to 90 minutes on a regular PC using a state-of-the-art toolkit [13]. Second, atlas-building is inflexible, requiring a new atlas for every new modality or carefully engineered similarity measures for multi-modal registration. Third, when multiple transformations resulting from registration are concatenated, errors accumulate leading to relevant mistakes. In this work, we demonstrate that we can train a CNN-based classifier to detect and localise anatomic regions at much higher speed and with much more flexibility than using deformable registration (cf. sections IV-D and IV-C). We train our method on 45 volumes which were previously atlas-segmented in order to create a bronze-standard ground-truth [14], that was carefully inspected and validated by clinicians. Our method learns the shape and anatomical context of 26 regions including basal ganglia, in a similar way as they are represented by the atlas. We demonstrate multi-region segmentation on 10 previously unseen volumes despite the small size and low contrasts of some structures.

Apart from multi-region segmentation of basal ganglia in MRI, we demonstrate segmentation of the midbrain using Hough-CNN in transcranial ultrasound (TCUS). TCUS can be used to scan deep brain regions non-invasively through the temporal bone window. Using TCUS, hyper-echogenicities of the Substantia Nigra (SN) can be analysed, gaining valuable information to perform differential [15] and early [16] diagnosis of Parkinson’s Disease (PD). A crucial step towards computer assisted diagnosis of PD is midbrain segmentation [17], [18]. This task is reportedly challenging even for human observers [19]. In order to penetrate the skull, low frequencies need to be applied resulting in an overall reduction of the resolution and in the presence of large incoherent speckle patterns. Scanning through the bone, moreover, attenuates a large part of the ultrasound energy, leading to overall reduction of signal-to-noise ratio, as well as low contrast and largely missing contours at anatomic boundaries. Additionally, the higher speed of sound in the bone leads to phase aberration [20] and de-focussing of the ultrasound beam which causes further lowering of the image quality. A variety of image quality, anatomical visibility and 3D ultrasound fan geometry can be seen in Figure 3. Registration methods, in particular non-linear registration, are very difficult under these conditions. Therefore, atlas-building and atlas-based segmentation methods tend to fail in ultrasound. In this work, we demonstrate that our proposed method, which is registration-free, is directly applicable to a dataset of TCUS volumes with manual segmentations of the midbrain. We train our method on 40 TCUS volumes, and demonstrate highly robust localisation and accurate segmentation of the midbrain on more than 100 previously unseen TCUS volumes. To the best of our knowledge, this is the first work employing CNNs to perform ultrasound segmentation.

Our work features several contributions:

- We propose Hough-CNN, a novel segmentation approach based on a voting strategy similar to [18]. We show that the method is multi-modal, multi-region, robust and implicitly encoding priors on anatomical shape and appearance. Hough-CNN delivers results comparable or superior to other state-of-the-art approaches while being entirely registration-free. In particular, it is better than methods based on voxel-wise classification.
- We propose and evaluate several different CNN architectures, with varying numbers of layers and convolutional kernels per layer. In this way we acquire insights on how different network architectures cope with the amount of variability present in medical volumes and image modalities.
- Each network is trained with different amounts of data in order to evaluate the impact of the number of annotated training examples on the final segmentation result. In particular, we show how complex networks with higher parameter number cope with relatively small training datasets.
- We adapted the Caffe framework [21] to perform convolutions of volumetric data, preserving its third dimension across the whole network. We compare CNN performance using 3D convolution to the more common 2D convolution, as well as to the 2.5D approach [22].

In Section V we discuss the potential of extending our method by increasing its flexibility via transfer learning towards multi-modal segmentation and computer-aided diagnosis.

1By “Bronze-standard” [14] we indicate the ground-truth segmentation that can be obtained by using (1) a state-of-the-art atlas-based approach and (2) validation of the results by expert observers. It is infeasible to obtain manually annotated “gold-standard” boundaries for whole volumes in the MRI dataset. Our comparison in section IV-C reveals that some regions cannot be confidently outlined by humans due to lack of contrast.
II. RELATED WORKS

In this section we give an overview of existing approaches that employ CNNs to solve problems from both computer vision and medical imaging domain.

In the last few years CNNs became very popular tools among the computer vision community. Classification problems such as image categorisation [1], [2], object detection [23] and face recognition [24] as well as regression problems such as human pose estimation [25], and depth prediction from RGB data [26] have been addressed using CNNs and unprecedented results have been presented to the community. In order to cope with the challenges present in natural images, such as scale changes, occlusions, deformations different illumination settings and viewpoint changes, these methods needed to be trained on very large annotated datasets and required several weeks to be built even when powerful GPUs were employed. In medical imaging, however, it is difficult to obtain even a fraction of this amount of resources, both in terms of computational means and amount of annotated training data.

Almost all works we encountered employing deep learning could be trained only on few dozen of training images (e.g. [27], [28], [29], [30], [31]). Most networks were applied to tasks that could be solved by interpreting the images patchwise in a sliding window fashion. In this case, several thousands of annotated training examples could be obtained from just a few images. Dataset augmentation techniques, such as random patch rotation and mirroring, were also applied when the analysed objects were invariant to these transformations [22], [27], [28], [29]. This is the case for cell nuclei, lymph nodes and tumour regions, but not for anatomical structures with regular size and local context, such as regions of the brain or abdomen. Another way to deal with little training data is to embed CNNs as core components into other methods which are well known to the community and have been previously applied to the same class of problems. A deep variational model is proposed in [33]. Their CNN is embedded into a global inference model, i.e. the CNN outputs are treated as unary potentials on a graph and the segmentation is solved via minimum s-t cuts on the predicted graph. In [34] the CNN performs 3D regression to predict an affinity graph, which can be solved via graph partitioning techniques or connected components in order to segment neuron boundaries. Active shape models are realised with CNNs in [35] via regression of multi-template contributions and object location. Variational Deep Learning was realised in [30] by combining shape-regularised levelset methods with Deep Belief Networks (DBN) for left ventricle segmentation in cardiac MRI.

In this work, we propose a novel Hough-CNN detection and segmentation approach which utilises CNNs at its core to efficiently process medical volumes patch-wise and obtains voxel-wise classifications along with high level features – used to retrieve votes – that are descriptive of the object of interest. Generalised Hough voting has been proposed in the past to address problems related with object detection and tracking. Recent works such as [36], [37] performed Hough voting using a CNN. Their respective aim is to obtain head poses and cell locations in 2D by using the network to perform simultaneous classification and vote regression. In this work we propose a more flexible voting mechanism based on neighbourhood relations in feature space. On one hand, this allows us to cast, for each patch, a variable amount of votes that can be associated with additional information such as segmentation patches, therapeutic indications or diagnostic information which may be added or modified at any time without requiring re-training. On the other hand, by using votes collected from annotated training images and stored as displacement vectors, our method does not experience any unpredictable behaviour when inputs that strongly differ from those observed during training are presented to the network.

Compared to computer vision which performs Deep Learning mostly on 2D images, medical images often deal with volumes acquired through scanners such as MRI or CT. In our literature review, most works have continued working in 2D by approaching 3D scans in a slice-by-slice fashion (e.g. [27], [28], [29], [30], [31], [32], [33], [34]). The advantage is high speed, low memory consumption and the ability to utilise pre-trained nets such as AlexNet [1], either directly or via transfer learning. The obvious disadvantage is that anatomic context in the directions orthogonal to the image plane are entirely discarded. Some groups who employed 3D convolutions found that computational tractability was an issue, and classification was either impossible [22] or suffered in accuracy since compromises on patch-size had to be made [31]. Other groups have applied 3D convolution successfully for Alzheimer’s disease detection from whole-MRI [41] or regression of affinity graphs from 3D convolution [24]. A different approach that was applied to full-brain segmentation from MRI in [22] combined small 3D patches with larger 2.5D ones that include more context. The 2.5D patches, in particular, consisted of a stack of three 2D patches extracted respectively from the sagittal, coronal and transversal planes. All patches were assembled into eight parallel CNN pathways in order to achieve high-quality segmentation of 134 brain regions from whole brain MRI.

In this work, we evaluate the performance of our network when 2D, 2.5D and 3D patches are employed. In particular, we supply rather long-range 3D patches which retain a large amount of anatomical context.

Another important issue in CNN-related research is the search for optimal CNN network architecture: we have found very little literature that addresses this issue systematically. Although several networks were analysed in [27], [28], we have found only one study on “very deep CNN” [42], in which the number of convolutional layers was varied systematically (8-16) while keeping kernel sizes fixed. The study concluded that small kernel sizes in combination with deep architectures can outperform CNNs with few layers and large kernel sizes.

In this work we propose and benchmark six network architectures, including one very deep network having 8 convolutional layers as shown in Table [1].
We propose six different convolutional neural network architectures trained with patches extracted from annotated medical volumes. We optimise our models to correctly categorise data-points into different classes. The volumes were acquired in two different modalities, US and MRI, and depict deep structures of the human brain. Accurate segmentation of the desired regions has been achieved though a Hough voting strategy, inspired by [18], which was employed to simultaneously localise and segment the structures of interest.

A. Convolutional neural networks

A CNN consists of a succession of layers which perform operations on the input data. Convolutional layers (symbol $C_k$) convolve the images $I_{size}$ presented to their inputs with a predefined number ($k$) of kernels, having a certain size $s$, and are usually followed by activation units which rescale the results of the convolution in a non linear manner. Pooling layers (symbol $P_{size}^{\text{stride}}$) reduce the dimensionality of the responses produced by the convolutional layers through downsampling, using different strategies such as average-pooling or max-pooling. Finally, fully connected layers (symbol $F_{\#\text{neurons}}$) extract compact, high level features from the data. The kernels belonging to convolutional layers as well as the weights of the neural connections of the fully connected layers are optimised during training through back-propagation. The network architecture is specified by the user, by defining the number of layers, their kind, and the type of activation unit. Other relevant parameters are: the number and size of the kernels employed during convolution, the amount of neurons in the fully connected part and the downsampling ratio applied by the pooling layers. We propose six network architectures that are described in Table I.

<table>
<thead>
<tr>
<th>Name</th>
<th>Network Architecture</th>
<th>Activation function</th>
<th>Initialisation</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-3-3-3-3</td>
<td>$F_{31} \cdot C_{64}^{7} \cdot P_{2} \cdot C_{64}^{5} \cdot C_{64}^{3} \cdot C_{64}^{3} \cdot C_{2}^{3} \cdot C_{1}^{3} \cdot F_{128} \cdot F_{128} \cdot F_{#\text{regions}}$</td>
<td>PReLU</td>
<td>MSRA</td>
<td>$F$ use drop-out (ratio 0.5)</td>
</tr>
<tr>
<td>3-3-3-3-3-3-3-3-3</td>
<td>$F_{31} \cdot C_{64}^{7} \cdot C_{64}^{5} \cdot C_{64}^{3} \cdot C_{64}^{3} \cdot C_{2}^{3} \cdot C_{1}^{3} \cdot F_{128} \cdot F_{128} \cdot F_{#\text{regions}}$</td>
<td>PReLU</td>
<td>MSRA</td>
<td></td>
</tr>
<tr>
<td>5-5-5-5-5</td>
<td>$F_{31} \cdot C_{64}^{7} \cdot C_{64}^{5} \cdot C_{64}^{3} \cdot C_{64}^{3} \cdot C_{2}^{3} \cdot C_{1}^{3} \cdot F_{128} \cdot F_{128} \cdot F_{#\text{regions}}$</td>
<td>PReLU</td>
<td>MSRA</td>
<td></td>
</tr>
<tr>
<td>7-5-5-5-5</td>
<td>$F_{31} \cdot C_{64}^{7} \cdot C_{64}^{5} \cdot C_{64}^{3} \cdot C_{64}^{3} \cdot C_{2}^{3} \cdot C_{1}^{3} \cdot F_{128} \cdot F_{128} \cdot F_{#\text{regions}}$</td>
<td>PReLU</td>
<td>MSRA</td>
<td></td>
</tr>
<tr>
<td>9-7-5-3-3</td>
<td>$F_{31} \cdot C_{64}^{7} \cdot C_{64}^{5} \cdot C_{64}^{3} \cdot C_{64}^{3} \cdot C_{2}^{3} \cdot C_{1}^{3} \cdot F_{128} \cdot F_{128} \cdot F_{#\text{regions}}$</td>
<td>PReLU</td>
<td>MSRA</td>
<td></td>
</tr>
<tr>
<td>Small Alex</td>
<td>$F_{31} \cdot C_{64}^{31} \cdot P_{2} \cdot C_{5}^{31} \cdot C_{2}^{31} \cdot C_{1}^{31} \cdot F_{128} \cdot F_{128} \cdot F_{#\text{regions}}$</td>
<td>PReLU</td>
<td>MSRA</td>
<td></td>
</tr>
</tbody>
</table>

In our approach we made use of parametric rectified linear units $\sigma$ (PReLU) as our activation functions.

$$PReLU(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases}$$ (1)

The parameter $\alpha$ in the PReLU activation function is learnt during training, along with other network weights. In this context we initialise the network parameters using MSRA $\sigma$ as it is an appropriate choice when employing PReLU activation units.

Many authors [11, 43] reported that the tendency of the network to overfit can be decreased by using a technique called “drop-out” during training which inhibits the outputs of a random fraction of the neurons of the fully connected layers in each iteration. In this way it is possible to limit their excessive specialisation to specific tasks, which is believed to be at the origin of overfitting in CNNs.

Finally, we employ max-pooling layers to reduce the dimensionality of the data as it traverses the network. The input of the pooling layer is exhaustively subdivided into sub-patches having fixed size and overlapping by an amount controlled by the “stride” parameter. Only the maximal value in each sub-patch is forwarded to the next layer. This procedure is known to incorporate a spatial invariance to the network which contradicts the desired localisation accuracy required for segmentation. For this reason we limit the usage of pooling layers to the minimum amount required to meet the existing hardware constraints.

B. Voxel-wise classification

A set $\mathbf{T} = \{p_1, ..., p_N\}$ of square (or cubic) patches having size $p$ pixels is extracted from $J$ annotated volumes $V_j$ with $j \in \{1,...,J\}$ along with the corresponding ground truth labels $\mathbf{Y} = \{y_1, ..., y_N\} \in \mathbb{R}$. Based on this training set CNNs are optimised to categorise the patches correctly. The resulting trained networks are capable of performing voxel-wise classification, also called semantic segmentation, of volumes by interpreting them in a patch-wise fashion. However, due to the lack of regularisation and enforcement of statistical priors this approach delivers sub-optimal results (Figure 7). For this reason we introduce a novel segmentation method that is based on simultaneous localisation of the anatomy of interest and robust contour extraction (Figure 2).
C. Hough voting with CNN

We introduce a robust segmentation approach that is scalable to multiple regions and implicitly encodes shape priors. This method employs a Hough-voting strategy to perform anatomy localisation and a database containing segmentation patches to retrieve the contour of the anatomy. Instead of relying only on categorical predictions produced by the CNNs we also make use of features extracted from their intermediate layers, in particular from the second-last fully connected one. Several authors [1], [23], [24] have reported that these features (sometimes also called descriptors) can be used for tasks such as image retrieval by mapping images to the feature space and identifying their neighbours. These findings are employed at the core of our voting strategy.

To keep our notation as simple and understandable as possible we describe, in the following, our approach for single region segmentation.

During training, we make use of the dataset of training volumes \( V_j \) with \( j \in \{1...J\} \), and respective binary segmentation volumes \( S_j \) with \( j \in \{1...J\} \). We collect patches from both foreground and background area and train a CNN. As a result, we obtain the parameters \( \theta \) that define the network. The CNN not only differentiates patches belonging to foreground and background through classification, but also associates each patch with its label \( l \in \{0,1\} \). This allows for the network to perform localisation and segmentation. We exhaustively collect a dataset \( T = \{p_1...p_N\} \) of either 2D, 2.5D or 3D patches from the locations \( X = \{x_1...x_N\} \) of the foreground region of each of the training volumes \( V_j \), and we use the CNN to obtain the features \( f_i \) introduced before. Our goal is to create a database storing triples consisting of a feature vector \( f_i \), a vote \( v_i \) and a segmentation patch \( s_i \).

The vote \( v_i \) is a displacement vector joining the voxel \( x_i \), where the i-th patch was collected from, and the position anatomy centroid \( c_j \) in the training volume \( V_j \):

\[
 v_i = x_i - c_j; \quad c_j = \frac{1}{|F_g|} \sum_{x_i \in F_g} x_i
\]

where \( F_g \) is the set of all the voxels belonging to foreground. The binary segmentation patches assume values 1 or 0 respectively for foreground and background area since they are collected from the positions \( x_i \) of the binary annotation volumes \( S_j \).

During testing, in order to segment a previously unseen volume \( I \), we make use of both the trained CNN and the database established before. We first obtain the classification label for each voxel \( x_i \) by processing the relative patch \( p_i \) through the CNN, which delivers also the features \( f_i \) for all the patches being classified as foreground. Each of such features is compared to those contained in the database in order to retrieve the \( K \) closest entries using Euclidean distance as criterion. This \( K \)-nearest neighbour search (\( K \)-nn) [44] is performed computing Euclidean distances \( d_i^{1...K} \) between features, as previously done in (1) for image retrieval.

Once the neighbours are identified, their votes \( v_i^{1...K} \) and associated segmentation patches \( s_i^{1...K} \) from the database, are employed to respectively perform localisation and segmentation. The votes are weighted by the reciprocal of the Euclidean...
distance computed during $K$-nn search $w_{1...K} = \frac{1}{\sum_{i=1}^{K}}$ and contribute to a vote-map at positions

$$\hat{v}_k^i = x_i + v_k^i; \forall k \in \{1...K\}$$

We repeat the steps described above for each of the patches that were classified as foreground (Figure 2). Since the region of interest occurs only once in each volume, we smooth the final vote map and retrieve the region centroid by finding the location $c$ where the maximal value of the vote map is reached (Figure 2). Smoothing reduces the possibility of small localisation mistakes due to “noise” in the vote map around the position where its maximum occurs.

The region of interest can now be segmented by re-projecting the votes $v_k^i$ to the locations $x_i$, where they have been originated from. Not all the votes should be re-projected, since a relevant portion of them is erroneous, i.e. did not contribute to the vote-map anywhere close to the estimated anatomy location. Only those that contributed to the vote-map within a certain range $r$ from the predicted centroid are taken into consideration and are actually allowed to contribute to the final segmentation contour with their own segmentation patch $s_k^i$. The segmentation patches $s_k^i$ are centred at the location $x_i$, weighted by $w_k^i$ and accumulated in the segmentation map $S$ (Figure 2d). Assuming that the segmentations patches $s_k^i$ have been extended to an infinite spatial extent by zero-padding, we can write:

$$\hat{S}(x) = \sum_{x_i} \sum_{k=1}^{K} Ind(\hat{v}_k^i, c) \ w_k^i \ s_k^i(x - x_i)$$

$$Ind(a, b) = \begin{cases} 
1 & \|a - b\| < r \\
0 & \|a - b\| \geq r
\end{cases}$$

In this sense the segmentation patches $s_k^i$ can be seen as basis functions $s_k^i(x)$, which take binary values, that need to be scaled and re-centred at appropriate locations in order to produce the desired effect in the segmentation map. Once the segmentation map $S$ is normalised to take only values comprised between 0 and 1, it is thresholded and the final contour is obtained.

The approach is schematically summarised in Figure 2. Extending this method to multiple regions requires little effort. In our implementation we treated each region independently by creating region-specific databases as well as dedicated vote-maps and segmentations. The memory requirements of this approach can be decreased by retrieving the segmentation patches directly from the volumes $S_{1...J}$ instead of storing them in the database. In this case, the database contains coordinates that are used to fetch contour portions from the $S_{1...J}$.

D. Efficient patch-wise evaluation through CNN

When dealing with images or volumes, patches are extracted in a sliding-window fashion and processed through a CNN. This approach is inefficient due to the high amount of redundant computations that need to be performed for neighbouring patches. In case no padding is used within the convolutional layers, the whole volume can be convolved with the respective kernels in one pass, instead of treating each patch separately, while achieving the same result. The same holds true for pooling layers whose pooling windows can be arranged to process the whole volume at once. However, as soon as fully connected layers are employed, the volume can no longer be processed in one pass due to the fact that the connections of this layer are limited to size of the input patch.

To solve this issue we modify the network structure as proposed by Sermanet et al. in \cite{45} in order to be able to process the whole volume at once, yet retrieving the same results that we would obtain if the data would be processed patch-wise.

IV. EXPERIMENTS AND RESULTS

In this section we show that CNNs not only can be used to robustly segment medical volumes (Figure 3, Figure 4), but they also possess the ability of learning extremely effective features (outputs of upper layers) from the data. Even in ultrasound, where the structures of interest are often not clearly visible or the images are affected by artefacts, CNNs are able to focus on salient information and therefore recognise patterns. We demonstrate the superior performances of our Hough-voting-based segmentation algorithm by evaluating our method on two datasets of US and MRI volumes depicting the human brain. The two modalities provide complementary information, but are inherently different both from the point of view of the challenges they offer and the range of anatomy they can image. We want to additionally highlight that, to the best of our knowledge, this is the first study where segmentation is performed on ultrasound data using CNNs.

A. Datasets and ground-truth definition

Our MRI dataset is composed of MRI volumes of 55 subjects, which were acquired using 3D gradient-echo imaging (magnitude and phase) with an isotropic spatial resolution of 1x1x1 mm. The sequence \cite{46} is designed for quantitative susceptibility mapping (QSM) and sensitivity towards iron deposits, which are a biomarker for movement disorders like Parkinson’s Disease and create visible contrast in relevant
basal ganglia like SN and STN. For our study, we created a novel group atlas using the Advanced Normalisation Toolkit (ANTS) and a state-of-the-art template creation routine [12]. It is important to note that this routine is based on a group-wise optimisation [13] and creates an unbiased, average brain template with particularly high SNR and contrast, since template appearance and shape are simultaneously optimised during template building. The high tissue contrast greatly facilitates manual delineation in the template, especially since certain structures become visible that are otherwise hard or impossible to discern in individual scans. One example is the internal medullary lamina separating the globus pallidus into its external (GPe) and internal (GPI) pallidal segments. Basal ganglia and other deep-brain structures were annotated in two ways. One set of bi-lateral atlas labels (brainstem, n. accumbens, amygdala, caudate, thalamus, hippocampus, pallidum, putamen) were annotated semi-automatically via a shape- and appearance-model segmentation (FSL FIRST [47]) plus manual correction of generated labels (one neuroimaging technician, verified by one expert neurologist). Another set of bi-lateral labels (separation of pallidus into GPe and GPi, midbrain, red nucleus, substantia nigra pars compacta and substantia nigra pars reticula) was annotated in a fully manual manner (neuroimaging technician, verified by expert neurologist) based on visible contrast. As a summary, the list of structures of interest is also visible in Figure 6.

After atlas creation and correction of labels in template space, the ground truth annotation for 55 subjects was generated via label transfer from the atlas through the inverse deformation fields computed previously. Data from 45 subjects were used for training and 10 for testing. In a single slice of five randomly selected test subjects, we additionally obtained two manual segmentations by two experts for computing inter-rater Dice coefficients. We use this to provide a baseline, for better assessment of CNN results as detailed in section IV-E.

The US dataset was acquired transcranially on 34 subjects, with several freehand 3D sweeps recorded through the left and right temporal bone window each. Altogether, 162 volumes were acquired with slight variations in bone window positioning, and reconstructed at 1mm isotropic resolution. For all 162 TCUS volumes, midbrain outlines were annotated in 3D by a single human expert. Inter-rater agreement of the midbrain annotations, in terms of Dice coefficient, has been reported in [19] to be 0.85. CNN training was performed on data from 8 subjects (40 sweeps), and testing on data from 24 previously unseen subjects (114 sweeps) while validation data was acquired from 8 sweeps from 2 subjects. Performing segmentation on more than 100 test volumes is a good indicator of actual clinical applicability of (Hough-)CNN-based segmentation. The experiments show that the method generalises very well on previously unseen data, which is a highly desirable property in clinical settings.

In order to test our approach and to benchmark the capabilities of the proposed CNNs when they are trained with a variable amount of data, we establish, for each dimensionality (2D, 2.5D and 3D) two differently sized training sets in US and three in MRI respectively. For each of the 40 training volumes in US we collect either 2K or 10K patches per volume such that half of the training set depicts the background and the other half the foreground. The resulting training sets have respective sizes of 80K and 400K patches. A validation set containing 5K patches has been established for US using images of subjects that have not been used for training or testing and employed to assess the generalisation capabilities of the models. From the 45 MRI training volumes, we extract either circa 100, 1K or 10K patches per volume per region (including background). The resulting training sets have respective sizes of 135K, 1.35M and 13.5M patches.

B. CNN parameters

We analyse six different network architectures, presented in Table I by training each of them for 15 epochs using Stochastic Gradient Descent (SGD) with mini-batches of 64 or 124 samples, learning rate varying between $10^{-2}$ and $5 \cdot 10^{-3}$ depending on the individual network architecture, momentum 0.9 and weight decay $5 \cdot 10^{-4}$. All our models converged after a few epochs, and often before the seventh epoch.

Each network is analysed three times, with patches capturing the same amount of context from the neighbourhood, but having different dimensionality. That is, our networks process 2D data, 2.5D data and 3D data in order to investigate how the networks respond to the higher amount of information carried by patches in 2.5D and 3D patches compared to 2D. During training we randomly sample patches from annotated volumes and we feed them to the networks along with their ground truth labels. The patches of the 2D dataset are all square and have a size of $31 \times 31$ pixels; the 2.5D dataset is composed of patches having the same size and three channels consisting of 2D patches from the sagittal, coronal and transversal plane.

![Fig. 4. Visual comparison of semantic segmentation results (top two rows) and Hough-CNN results (bottom two rows) on same MRI volumes using the same trained CNN. Coloured regions represent ground truth annotation. Coloured contours represent segmentation outputs. Best viewed in digital format.](image-url)
centred at the same location; the 3D dataset contains cubic patches having size $31 \times 31 \times 31$ voxels.

Some of the parameters supplied to our Hough-CNN algorithm are empirically chosen. Parameters names and respective values are reported in Table II. These parameters remained constant throughout all experiments, both in ultrasound and MRI. All the trainings were performed on Intel i7 quad-core workstations with 32 gigabytes of ram and graphic cards from Nvidia, specifically “Tesla k40” or ”Titan X” (12GB VRAM). All tests were made on a similar workstation equipped with a Nvidia GTX 980 (4 GB VRAM).

### C. Experiments and results in ultrasound

We train our CNNs with different amount of data having different dimensionality, as explained in Section V-A. Each of the six proposed architectures is trained six times (five for 3D) in order to cover all the possible combinations of dimensionalities ($2D$, $2.5D$, $3D$ patches) and amount of data (training set sizes $80K$, $400K$). We test each CNN on 114 ultrasound volumes acquired from subjects whose scans have never been used during training or validation.

Table II shows the average performance in terms of Dice coefficients, mean distances of the estimated contours to the ground truth annotations and failure rates of the proposed Hough-CNN segmentation approach when different CNNs are employed. Since we segment one region per volume, the failure rate represents the percentage of volumes where the region of interest could not be segmented due to wrong localisation (Dice 0). In Figure 5 we provide summary of the performances of each network, when various amounts of training data are used and patches of different dimensionality are supplied. Better networks produce Dice histograms whose higher values are occurring far away from the origin.

Visual examples of ultrasound segmentation results are visible in Figure 5. It is notable that the Hough-CNN segmentation is able to localise and segment the midbrain accurately, regardless of whether the scan was acquired through the left or right bone window. It is also robust to bone window quality and overall visibility of structures, as well as signal-drop regions and blurring.

### D. Experiments and results in MRI

We train each of our networks nine times (eight for 3D) in order to explore all the possible combination of different data dimensionality and size of the training set as explained in Section IV-A. We test each of the models on 10 volumes, using their respective atlas-based annotations for evaluation. We verified, through visual inspection performed by a technician and an expert neurologist, that the annotation appropriately delineate the regions of interest.

Table IV reports the average performance in terms of Dice coefficients, mean distances of the estimated contours to the ground truth annotations and failure rates of the proposed Hough-CNN segmentation approach when different CNNs are employed at its core. The failure rate, in particular, refers to the percentage of regions of the whole training set (total number: $26 \times 10$ regions), that were not segmented correctly by Hough-CNN due to the fact that they could not be correctly localised. The results are clustered by the size of the training set employed to train the model to improve readability and the possibility of making comparisons between CNNs employing data having different dimensionality ($2D$, $2.5D$ and $3D$). From these results we observe that the best performing architecture is “7-5-3”.

In Figure 6 we compare the results achieved by the architecture “7-5-3”, on each of the 26 brain region of interest separately, when different data dimensionalities are used. The bar plot shows the results in terms of Dice coefficient, while the dashed line plot conveys the results in terms of average distance of the estimated contour to ground-truth delineation. We observe that Hough-CNN yields better Dice coefficients when bigger regions and high contrast area are segmented. Small and low contrast regions could be correctly localised but they were in general harder to segment.

Visual examples of MRI segmentation results are visible in Figure 6. It is notable that the Hough-CNN segmentation is able to correct localise and segment multiple structures, despite large anatomical variability, such as cortical atrophy and enlarged lateral ventricles.

### E. Comparison to inter-rater agreement

Since we trained and tested the semantic- and Hough-CNN methods using labels transferred from the atlas, we created two human expert segmentations in the MRI volumes of individuals. We compare these to atlas segmentations on the one
### Table III

Midbrain segmentation results in 114 previously unseen TCUS volumes, using Hough-CNN with variations of architectures (single rows), patch dimensionalities (column blocks) and training set sizes (row blocks). The best result for each architecture (across the data dimensionalities) are highlighted using bold typeface. The best results for each dimensionality (across the architectures) are underlined.

<table>
<thead>
<tr>
<th>Dimensionality →</th>
<th>2D</th>
<th>2.5D</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Averages</td>
<td>Distance (mm)</td>
<td>Failures</td>
</tr>
<tr>
<td>Training set size 80K patches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-3-3-3-3-3-3</td>
<td>0.83</td>
<td>0.92</td>
<td>3%</td>
</tr>
<tr>
<td>3-3-3-3-3-3-3-3-3</td>
<td>0.80</td>
<td>0.93</td>
<td>5%</td>
</tr>
<tr>
<td>5-5-5-5-5-5</td>
<td>0.77</td>
<td>1.07</td>
<td>9%</td>
</tr>
<tr>
<td>7-7-7-7-7-7-7-7-7</td>
<td>0.80</td>
<td>0.96</td>
<td>5%</td>
</tr>
<tr>
<td>9-7-5-3-3-3-3-3</td>
<td>0.79</td>
<td>0.96</td>
<td>7%</td>
</tr>
<tr>
<td>SmallAlex</td>
<td>0.85</td>
<td>0.81</td>
<td>1%</td>
</tr>
</tbody>
</table>

Training set size: 400K patches

<table>
<thead>
<tr>
<th>Dimensionality →</th>
<th>2D</th>
<th>2.5D</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Averages</td>
<td>Distance (mm)</td>
<td>Failures</td>
</tr>
<tr>
<td>3-3-3-3-3-3</td>
<td>0.84</td>
<td>0.90</td>
<td>1%</td>
</tr>
<tr>
<td>3-3-3-3-3-3-3-3-3</td>
<td>0.83</td>
<td>0.91</td>
<td>1%</td>
</tr>
<tr>
<td>5-5-5-5-5-5</td>
<td>0.83</td>
<td>0.94</td>
<td>2%</td>
</tr>
<tr>
<td>7-7-7-7-7-7</td>
<td>0.83</td>
<td>0.94</td>
<td>2%</td>
</tr>
<tr>
<td>9-7-5-3-3-3</td>
<td>0.82</td>
<td>1.01</td>
<td>2%</td>
</tr>
<tr>
<td>SmallAlex</td>
<td>0.83</td>
<td>0.91</td>
<td>3%</td>
</tr>
</tbody>
</table>

Fig. 5. The midbrain segmentation performance of each network on 114 TCUS test volumes, under different training conditions, is summarised through histograms. The horizontal axis is subdivided in Dice bins having a width of 0.05 Dice. The vertical axis represents the number of volumes falling in each Dice bin. Each CNN architecture is depicted with its own colour.
hand, and to CNN segmentations on the other. This helps to create some baseline values for better assessment of the CNN results in MRI data and to put previous results into perspective. Manual segmentations were created separately by two raters, one neuroimaging technician and one neurological resident. It is important to note that creating manual segmentations imposes a significant time effort on human experts, which forced us to restrict them to single slices only, in five randomly selected MRI test volumes, on the 21 out of 26 anatomic regions of the striatal system that were visible in those slices. We compute mean Dice overlap measures for each structure across five volumes, and show a bar-plot representing the distribution of Dice coefficients in Figure 6. We compare five different inter-rater agreements in order to better assess and position the Hough-CNN performance:

1) We compare human inter-rater agreement for blinded experts, i.e. experts segmenting the single slices on MRI data from scratch. This indicates a baseline as to how easy it is to localise the structures, and how visible the boundaries are to human observers. The median Dice coefficient across 21 structures is 0.686 ($\sigma = 0.11$).

2) The inter-rater agreement of two atlas-guided experts has a median Dice coefficient of 0.846 ($\sigma = 0.08$). Here, "atlas-guided" means that human experts started from atlas-segmentation and only adapted them when they saw fit. The improvement in agreement compared to blinded experts is highly significant (Wilcoxon signed-rank test $p < 10^{-5}$). This can be accounted to a much easier localisation and delineation of consistent boundaries with help from the atlas, since it preserves shape context from the template.

3) Atlas-guided experts agree even more closely with the atlas segmentations themselves, with a median Dice of 0.902 ($\sigma = 0.08$). The difference between 2) and 3) is highly significant ($p < 10^{-4}$), i.e. wherever atlas-guided experts decided to deviate from the atlas, they deviated in slightly different directions, leading to less agreement among each other. This also shows that atlas-segmentations can be considered as ground truth ("bronze standard" $^{[14]}$), since experts tend to agree more with atlas segmentations than with each other.

4) Compared to that, Hough-CNN segmentations have a median Dice overlap with the atlas of 0.829 ($\sigma = 0.09$), which is on the order of the inter-rater variability between atlas-guided experts in 2), with significant difference ($p < 0.01$). This shows that the Hough-CNN method is able to preserve both shape and inter-region context of anatomical structures trained from the atlas.

5) Eventually, the two human experts on average have an overlap with Hough-CNN segmentations with a median Dice of 0.771 ($\sigma = 0.11$).

The agreement between Hough-CNN and experts in 5) is lower ($p < 10^{-5}$) than the agreement of Hough-CNN with the atlas in 4). As noted, it might actually be a better choice for Hough-CNN to stay closer to segmentations from the atlas, as even humans tend to agree with the atlas more than with each other as seen in 2) and 3).

V. DISCUSSION

Training of CNNs requires a large amount of data in order to achieve satisfactory voxel-wise classification results and perform semantic segmentation. However, as described in the introduction, obtaining such large annotated datasets is rarely possible in clinical settings. By using a voting-based strategy, it is possible to localise the anatomy of interest with high precision, even when the rate of mis-classified voxels is very high. Additionally, our Hough-CNN approach implicitly enforces shape priors which facilitate segmentations in images where the anatomy of interest is poorly visible.
TABLE IV
AVERAGE SEGMENTATION RESULTS OF 26 STRUCTURES IN 10 MRI TEST VOLUMES, USING HOUGH-CNN WITH VARIATIONS OF ARCHITECTURES (SINGLE ROWS), PATCH DIMENSIONALITIES (COLUMN BLOCKS) AND TRAINING SET SIZES (ROW BLOCKS). THE BEST RESULT FOR EACH ARCHITECTURE (ACROSS THE DATA DIMENSIONALITIES) ARE HIGHLIGHTED BY USING BOLD TYPEFACE. THE BEST RESULTS FOR EACH DIMENSIONALITY (ACROSS THE ARCHITECTURES) ARE UNDERLINED. THE BEST RESULT IS OBTAINED USING THE ARCHITECTURE “7-5-3” AND 3D DATA.

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>2D</th>
<th>2.5D</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averages</td>
<td>Dice [0, 1]</td>
<td>Distance (mm)</td>
<td>Failures</td>
</tr>
<tr>
<td>3-3-3-3-3</td>
<td>0.61</td>
<td>0.52</td>
<td>6%</td>
</tr>
<tr>
<td>3-3-3-3-3-3-3-3</td>
<td>0.61</td>
<td>0.52</td>
<td>8%</td>
</tr>
<tr>
<td>5-5-5-5-5</td>
<td>0.64</td>
<td>0.49</td>
<td>6%</td>
</tr>
<tr>
<td>7-5-3</td>
<td>0.67</td>
<td>0.48</td>
<td>4%</td>
</tr>
<tr>
<td>9-7-5-3-3</td>
<td>0.60</td>
<td>0.52</td>
<td>8%</td>
</tr>
<tr>
<td>SmallAlex</td>
<td>0.61</td>
<td>0.33</td>
<td>3%</td>
</tr>
<tr>
<td>Training set size: 1.35K patches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-3-3-3-3</td>
<td>0.63</td>
<td>0.51</td>
<td>3%</td>
</tr>
<tr>
<td>3-3-3-3-3-3-3-3</td>
<td>0.63</td>
<td>0.51</td>
<td>3%</td>
</tr>
<tr>
<td>5-5-5-5-5</td>
<td>0.64</td>
<td>0.51</td>
<td>3%</td>
</tr>
<tr>
<td>7-5-3</td>
<td>0.68</td>
<td>0.47</td>
<td>2%</td>
</tr>
<tr>
<td>9-7-5-3-3</td>
<td>0.63</td>
<td>0.53</td>
<td>4%</td>
</tr>
<tr>
<td>SmallAlex</td>
<td>0.64</td>
<td>0.51</td>
<td>4%</td>
</tr>
<tr>
<td>Training set size: 1.35M patches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-3-3-3-3</td>
<td>0.64</td>
<td>0.52</td>
<td>3%</td>
</tr>
<tr>
<td>3-3-3-3-3-3-3-3</td>
<td>0.65</td>
<td>0.56</td>
<td>2%</td>
</tr>
<tr>
<td>5-5-5-5-5</td>
<td>0.64</td>
<td>0.51</td>
<td>3%</td>
</tr>
<tr>
<td>7-5-3</td>
<td>0.68</td>
<td>0.49</td>
<td>3%</td>
</tr>
<tr>
<td>9-7-5-3-3</td>
<td>0.63</td>
<td>0.52</td>
<td>5%</td>
</tr>
<tr>
<td>SmallAlex</td>
<td>0.65</td>
<td>0.52</td>
<td>4%</td>
</tr>
</tbody>
</table>

Fig. 8. Single-slice Dice coefficients for 21 structures (averaged across five subjects each). We perform five comparisons which provide a baseline and perspective on Hough-CNN results. We compare inter-rater variability (HE1 vs. HE2, blinded or with atlas guidance) to variability between atlas (ATL) and Hough-CNN (CNN), as well as variability between both experts (Avg. HE1/HE2) vs. CNN.

Furthermore, when using 3D patches, only 1.35M training patches were required to surpass the performance obtained with datasets of 13.5 millions 2D and 2.5D patches, which marks a 90% reduction of required training data. In all three dimensionalities, 2D, 2.5D and 3D, Hough-CNN outperforms voxel-wise segmentation (cf. Figure 7). Similar to related works [30], [33], [34], [35], we thus demonstrate that it may be beneficial to embed CNNs as powerful classifiers into higher-level methods which encode anatomic shape- and appearance priors.

The experiments performed on MRI highlight important aspects of both our CNNs and the modality itself. Most of the brain regions considered in this study (e.g. midbrain, STN, caudate) can be recognised by a human rater by clearly visible contrasts, while the position and boundaries of other, more difficult regions (e.g. GPi, GPe, SNpc, SNpr) can be inferred through anatomical knowledge and neighborhood context. Ultrasound volumes are much more challenging from this point of view. Human midbrain in TCUS can be difficult to discern and human observers can be mislead by artefacts and signal-loss areas having similar shape. The CNNs employed in this study had various architectures and therefore different pattern recognition capabilities. In MRI, where the most part of regions of interest have good contrast while the position of the others can be inferred by the context, the best performing network was “7-5-3”. Although this architecture is the simplest, it delivered best results in all the MRI experiments. In US, which is a challenging modality, the networks that delivered best results were among the most complex. “SmallAlex” and “3-3-3-3-3-3-3-3-3” are deeper and therefore recognise more complex visual content than “7-5-3”.

While we observed a strong performance advantage when segmenting MRI volumes considering 3D data (Table IV), we observed the opposite effect when segmenting ultrasound as shown in the bottom left of Table III. In MRI, processing data in 3D brings additional useful information which improves the performance of both automated methods and human raters, who refer simultaneously to sagittal, coronal and axial views when establishing the ground truth. In US we observed that experts segmenting the ground truth used only the axial plane, since it is the only plane in which the characteristic shape of the midbrain can be recognised. This might indicate that CNNs learned to recognise patterns in a similar way as humans do, and therefore produce best results when they are not supplied with misleading information.

Altogether, using Hough-CNNs, we achieved very robust midbrain segmentation in 3D-TCUS, in a test dataset of more than 20 subjects and 114 volumes, with a large variation...
of 3D sweep geometry, bone window qualities, midbrain appearance, location and orientation. We also segmented 10 previously unseen MRI volumes achieving very high Dice coefficients for large and high-contrasted regions, while some of the smallest and most challenging regions were almost always localised accurately and segmented with sub-voxel mean surface distance.

Compared to our atlas-based segmentation in MRI, Hough-CNN preserves anatomical shapes and context across multiple regions, and on an inter-rater variability basis, it outperforms humans who try to segment the volumes from scratch (cf. section IV-E). Compared to atlas-segmentation, Hough-CNN is faster (30 seconds in US, and 3–4 minutes in MRI on the machine employed for testing) and entirely registration-free, i.e. concerning translation the method is invariant due to the patch-wise classification approach. Concerning rotation, both training and testing data for each individual was provided in scanner coordinates, which shows that the method is also invariant to head rotations present in clinical practice. Even though Hough-CNN is biased to the atlas segmentations after training, the method separates classification through CNN and segmentation through back-projection of surface patches. Thus, a complete exchange of the atlas-segmentations with (very difficult to obtain) manual segmentations can be achieved without re-implementation and without re-training, since only the segmentation patches in the databases need to be updated. Our approach is also very flexible since each intensity patch can be associated to several segmentation patches, collected from the annotations of multiple volumes, during segmentation. Even though we have not performed experiments yet, we are confident that the trained CNN classifier can be also easily extended to other modalities via transfer learning, e.g. from our QSM sequences to T1 or T2. Similarly, we have demonstrated the straightforward application of Hough-CNN to a completely different modality, TCUS, in which registration-dependent methods like atlas-based segmentation would be extremely difficult, if not impossible due to largely missing anatomical and structural context. It is also noteworthy that in this work, we have only used the CNN method for segmentation. However, as other works have demonstrated [2], the learned data representations in the last layers of the CNN can be directly used for classification or regression of disease parameters. This can be interleaved with segmentation, which goes far beyond the capabilities of pure atlas-based methods.

VI. CONCLUSION

In this work, we applied CNNs to medical image segmentation, under the constraints of clinical feasibility both in terms of limited training data and computational resources. We performed an unprecedented study of several CNN parameters, including architectures, patch dimensionality and training set size. We proposed Hough-CNN, a patch-wise multi-atlas method which implicitly encodes priors on anatomic shape and context. The method outperformed voxel-wise semantic segmentation of CNNs in all parameter settings, while using less training data and delivering smooth segmentation contours without the need for post-processing. The method is modality-independent and scalable to multiple regions and harnesses the impressive classification power of CNNs and Deep Learning for application in clinical settings.

ACKNOWLEDGMENT

This study was funded by the Lüneburg Heritage and Deutsche Forschungsgesellschaft (DFG) Grant BO 1895/4-1. We gratefully acknowledge the support of NVIDIA Corporation in donating a “Tesla K40” GPU for this study.

REFERENCES


Fausto Milletari received his B.E. from the university of Pisa and his M.Sc. from the Technical University of Munich. Since October 2013 he is pursuing a Ph.D. in informatics at the chair for Computer Aided Medical Procedures (CAMP) of the Technical University of Munich. His research has been focused on problems from both medical image analysis and computer vision domain. He has authored several contributions applying machine learning methods to perform anatomy detection, segmentation and visual tracking.

Seyed-Ahmad Ahmadi graduated with a double M.Sc. degree in Electrical and Computer Engineering from the Technical University of Munich (TUM) and the Georgia Institute of Technology in 2008. In his Ph.D. studies, supervised by Prof. Dr. Nassir Navab at TUM, he worked on novel image segmentation methods in transcranial 3D ultrasound towards early-detection of Parkinson’s disease. Since 2013 he is a Post-Doctoral researcher at the department of neurology, Klinikum Grosshadern, Ludwig-Maximilian University (LMU). His research focus is on multi-modal image registration, segmentation and data analysis for improved diagnosis and treatment of neurological movement disorders.