



Automatic cortical target point localisation in MRI for transcranial magnetic stimulation via a multi-resolution convolutional neural network

John S H Baxter, Quoc Anh A Bui, Ehouarn Maguet, Stéphane Croci, Antoine Delmas, Jean-Pascal Lefaucheur, Luc Bredoux, Pierre Jannin

► To cite this version:

John S H Baxter, Quoc Anh A Bui, Ehouarn Maguet, Stéphane Croci, Antoine Delmas, et al.. Automatic cortical target point localisation in MRI for transcranial magnetic stimulation via a multi-resolution convolutional neural network. International Journal of Computer Assisted Radiology and Surgery, 2021, 16 (7), pp.1077-1087. 10.1007/s11548-021-02386-1 . hal-03283129

HAL Id: hal-03283129

<https://univ-rennes.hal.science/hal-03283129>

Submitted on 9 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Automatic Cortical Target Point Localisation in MRI for Transcranial Magnetic Stimulation via a Multi-Resolution Convolutional Neural Network

John S.H. Baxter* · Quoc Anh Bui ·
Ehouarn Maguet · Stéphane Croci ·
Antoine Delmas · Jean-Pascal Lefaucheur
 · Luc Bredoux · Pierre Jannin

Received: date / Accepted: date

Abstract Purpose: Transcranial Magnetic Stimulation (TMS) is a growing therapy for a variety of psychiatric and neurological disorders that arise from or are modulated by cortical regions of the brain represented by singular 3D target points. These target points are often determined manually with assistance from a pre-operative T1-weighted MRI, although there is growing interest in automatic target point localisation using an atlas. However, both approaches can be time-consuming which has an effect on the clinical workflow and the latter does not take into account patient variability such as the varying number of cortical gyri where these targets are located. **Methods:** This paper proposes a multi-resolution convolutional neural network for point localisation in MR images for *a priori* defined points in increasingly finely resolved versions of the input image. This approach is both fast and highly memory efficient, allowing it to run in high-throughput centres, and has the capability of distinguishing between patients with high levels of anatomical variability. **Results:** Preliminary experiments have found the accuracy of this network to be 7.26 ± 5.30 mm, compared to 9.39 ± 4.63 mm for deformable registration and 6.94 ± 5.10 mm for a human expert. For most treatment points, the human expert and proposed CNN statistically significantly outperform registration, but neither statistically significantly outperforms the other, suggesting that the proposed network has human-level performance. **Conclusions:** The

J.S.H. Baxter is the corresponding author. Email: jbaxter@univ-rennes1.fr

J.S.H. Baxter, Q.A. Bui, E. Maguet & P. Jannin
Laboratoire Traitement du Signal et de l'Image (LTSI - INSERM UMR 1099), Université de Rennes 1, Rennes, France

S. Croci, A. Delmas, & L. Bredoux
SYNEIKA, Rennes, France

J.-P. Lefaucheur
ENT Team, EA4391, Faculty of Medicine, Paris Est Créteil University, Créteil, France
Clinical Neurophysiology Unit, Department of Physiology, Henri Mondor Hospital, Assistance Publique – Hôpitaux de Paris, Crêteil, France

human-level performance of this network indicates that it can improve TMS planning by automatically localising target points in seconds, avoiding more time-consuming registration or manual point localisation processes. This is particularly beneficial for out-of-hospital centres with limited computational resources where TMS is increasingly being administered.

Keywords Transcranial magnetic stimulation, deep learning, convolutional neural networks

1 Introduction

Transcranial Magnetic Stimulation (TMS) is a increasingly used for a variety of psychiatric and neurological disorders, including its established use in the treatment of neuropathic pain and major depression [12] and for motor stroke rehabilitation [12, 17], but also emerging potential use in the alleviation of early Alzheimer's disease [5] and vascular dementia [13]. TMS involves the stimulation of particular functional networks in the cortex through the use of brief intense magnetic fields which are applied by a coil placed on the patient's head proximal to the region of interest. For TMS in specific disorders, such as drug-resistant chronic pain, an initial localisation of the stimulation site can be highly beneficial as the ideal regions to stimulate can sometimes be determined by the patient's symptomatology [9]. For chronic pain, these targeted cortical regions are located in the primary motor cortex (M1) corresponding to the anatomical regions in which the patient feels pain on the contralateral side [9]. The pain alleviating effects of TMS can be short-lasting, often on the order of days or weeks following stimulation. However, multiple follow-up sessions can result in long-term pain alleviation but their frequency means that streamlining procedures for TMS planning is important [12].

TMS planning involves selecting and localising a collection of target points for calibration and for treatment. Calibration target points help determine the stimulation parameters by finding the minimum intensity threshold for which stimulation results in a motor response and thus allow for the stimulation parameters to be adjusted to suit that particular patient, avoiding over- or under-stimulation. Treatment targets are the cortical regions that will be stimulated in order to alleviate the patient's individual disorder. Given the more subjective nature of the disorders being treated, the precise locations of treatment targets may be harder to identify by the human operator, especially as there is often a delay between the stimulation and improvement in the patients symptoms [9]. There is some recent literature regarding identifying the effect site of TMS stimulation using the parameters of said stimulation. However, these approaches are not suitable for pre-operative planning in which said stimulation has yet to be performed [19].

The first method for identifying these points was to initially find the calibration point through trial-and-error (as stimulating this point produced observable muscle contractions) and then navigate the coil a pre-specified distance along the scalp. Another early method avoiding this manual measure-

ment used a fitted cap as markers on this cap allowed for somewhat more accurate and repeatable targeting. However, both *unguided* methods tended to have targeting errors on the order of 1-4 cm. [16]

Although the use of a standard T1-weighted MRI is highly beneficial for the identification of target points and has resulted in improved TMS outcomes [10], MRI-based navigation remains sparsely used for various reasons, among which is the lack of neuroanatomical skill of most operators. The current workflow for MRI-based TMS planning involves the use of these images, which are skull-stripped and then visualised, allowing for target points to be identified on the patient's cortical surface. Recent approaches have attempted to automate this process by deformably registering this image to an atlas containing pre-identified calibration and treatment target points which are then projected onto the visualisation of the cortical surface. However, the process of registration takes several minutes which can be problematic for high-throughput centres. In addition, it is susceptible to error resulting from topological differences between the atlas image and the particular patient due to the high patient variability in the number and position of cortical gyri. Circumventing these image processing steps would be highly beneficial to the TMS workflow.

The TMS workflow is particularly interesting as an increasing proportion of interventions are not performed in hospital, but in specialised out-of-hospital centres. These centres can be more decentralised, facilitating patient access to TMS, but have fewer computational resources and no additional imaging capacity, placing particular constraints on TMS planning software. Firstly, it must be fast as the image processing is often done after the patient has entered the clinic and any delay due to computer processing delays the intervention. Secondly, it must be immediately adaptable to having images from different scanners as patients may be arriving from different hospital centres.

Recent advances in deep learning provide a possible solution and convolutional neural networks (CNNs) have shown great promise in automatically identifying [11] and localising [14] structures in natural images. Although particular affordances must be made to accommodate fully volumetric images due to their size and the inherent memory limitations of CNNs [6], deep learning does offer a new approach to point localisation in structural MRI that may be more robust to patient anatomical variability. In addition, CNNs are highly time-efficient compared to the multiple iterative optimisation-based algorithms used by deformable registration. However, current CNN structures for localising objects are designed for two-dimensional, rather than three-dimensional images. This renders their immediate use difficult as the amount of memory required for storing volumetric activations can easily overwhelm a modern GPU memory even for processing a single volume, let alone batches of volumes. In addition, many of these frameworks are designed for also classifying objects, which involves an additional layer of complexity not required for TMS planning in which a fixed number of points with consistent identities are desired.

Contributions

The purpose of this paper is to estimate the locations of a set of *a priori* defined stimulation target points to assist in TMS planning directly from the T1 images using convolutional neural networks. By using the T1 images directly, this framework avoids the skull-stripping and deformable registration steps, allowing it to more easily fit into the clinical workflow. This paper also relates the accuracy of these target points to that of human expert performance and to the state-of-the-art deformable registration-based method currently employed. To the best of our knowledge, this system represents the first use of deep learning to perform TMS target point localisation as well as the first of such systems with an accuracy approaching that of a human expert.

2 Methods

2.1 Patient Images

26 patient T_1 -weighted MR images (1mm isotropic resolution) have been collected with annotations of multiple TMS target points. (The majority of images have all target points identified.) As the patient base comes from multiple hospital centres, there is some heterogeneity in MRI manufacturer (database includes Phillips Acheiva, Siemens Verio, and GE Signa HDxt) and protocol (T1 3D N NAV, MPRAGE, and CRANE STANDARD/20). To normalise the differing image intensities, approximate min-max normalisation was employed, using the 5 and 95 percentiles as the minimum and maximum intensity estimates. These images have a common RAI orientation but have been resampled to 256x256x256 voxels in order to facilitate downsampling. The images were annotated with a series of targets points which are listed in Table 1. As not all images had all target points identified, Table 1 also reports the number of patients in which the target was identified by at least one human expert and the total number of times it is identified across the entire database.

The non-motor treatment points (i.e. LOFC, ROFC, LDLPFC, RDLPFC, and LHESCHL) were annotated independently by an expert neurologist or neurosurgeon. (All patients were annotated the same annotator for each point but the different point types had different annotators.) The chronic pain treatment points (i.e. LFACEMC, RFACEMC, LULIMBMC, RULIMBMC, LL-LIMBMC, and RLLIMBMC, noting that the LULIMBMC and RULIMBMC are also used for calibration as stimulating them leads to an observable twitching in the hand) have been annotated in each image by three expert neurologists in order to estimate the expert variability (i.e. the average performance of human experts) on this task. To ensure that the accuracy of the training data is of high quality, an "consensus" point is derived from these three expert annotations. This process involves multiple manual steps including: the visual confirmation that the point data is adequate (i.e. appears in the correct cortex on the correct hemisphere), determining which of the three experts are in agreement (i.e. have selected on the same gyral fold as visually determined

Acronym	Region	Type	P#	T#	C#
LOFC	Left orbitofrontal cortex	Treatment (depression, schizophrenia, anxiety, OCD, etc...)	26	26	
ROFC	Right orbitofrontal cortex		26	26	
LDLPFC	Left dorsolateral pre-frontal cortex		24	24	
RDLPFC	Right dorsolateral pre-frontal cortex		24	24	
LHESCHL	Left Heschl's gyrus (i.e. the transverse gyrus) of the temporal lobe		23	23	
LFACEMC	Facial region of the left primary motor cortex	Treatment (chronic pain) Calibration (LULIMBMC, RULIMBMC)	26	78	25
RFACEMC	Facial region of the right primary motor cortex		26	78	26
LLLIMBMC	Lower limb region of the left primary motor cortex		26	78	24
RLLIMBMC	Lower limb region of the right primary motor cortex		26	78	26
LULIMBMC	Upper limb region of the left primary motor cortex		26	78	25
RULIMBMC	Upper limb region of the right primary motor cortex		26	78	26

Table 1 Points used in our TMS dataset containing both depression and chronic pain patients. The P# column refers to the number of patients in this dataset with at least one annotation of this particular target point, T# refers to the total number of target points, and the C# column refers to the number of patients with consensus annotations. Greyed out cells represent regions that could not be computed, having a single expert annotation.

by an expert neurologist visualising all three points) and finding the centre-of-mass point for the experts with the highest agreement. Although not a strict requirement, the further distance between two experts considered to be in agreement was less than 12mm. Often, all three experts agree, that is, they report similar points for a desired target in an individual patient. However, it is a common occurrence for only two of the three experts to agree and for the consensus point to be determined only by those two experts. In the case of 4 points (across the entire patient dataset), no consensus could be found, that is, all three experts identified clearly different anatomical locations as the target point. The network is trained on these consensus points (when available) for the aforementioned point types, rather than on the multiple expert annotations as the consensus points are intuitively more likely to be correct and given the limited amount of data for training, we did not expect the network to be able to disambiguate erroneous manually defined points. For the other point type, the dataset had only been annotated by one expert and whose points are thus used for training.

2.2 Multi-Resolution Convolutional Neural Network Architecture

Our method uses a deep convolutional neural network inspired by multi-resolution architectures such as U-Nets [15] expanding on our previous work in localising points in volumetric images for the purpose of small region segmentation. [3] This work compared a traditional neural network approach for localising a single point, treated as an image-to-vector regression problem,

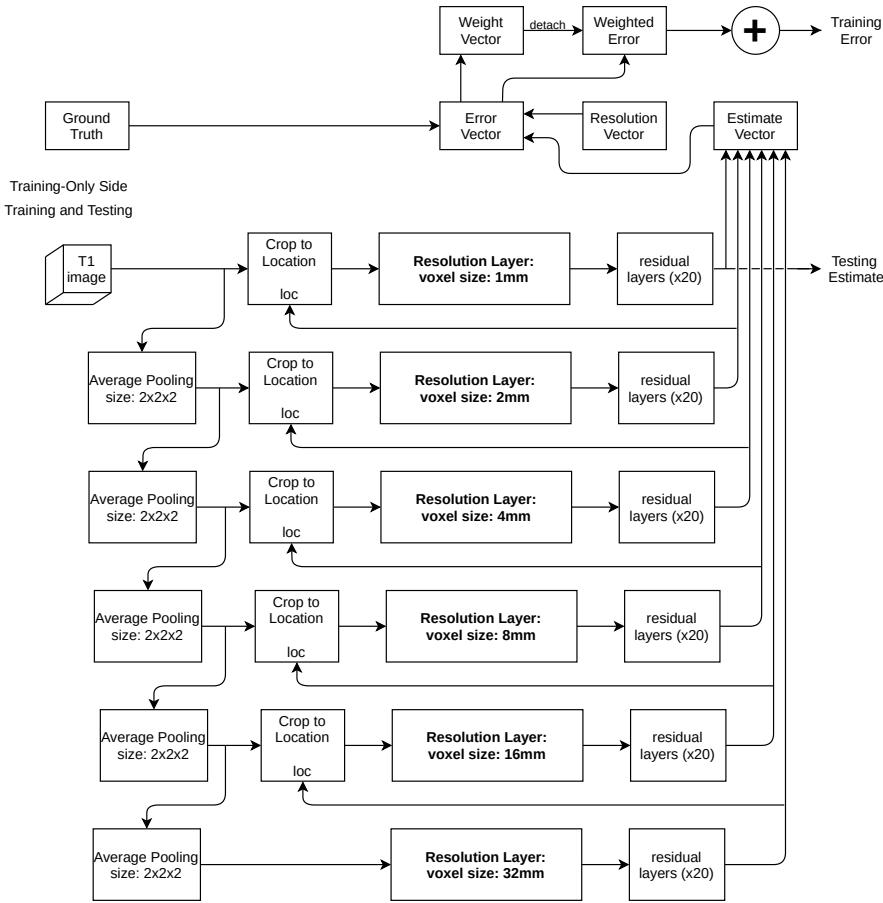


Fig. 1 Network Diagram for Point Localisation with 20 residual layers and 6 ‘single resolution subnetworks’ shown in Figures 2 and 3 respectively.

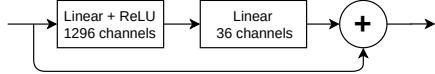


Fig. 2 Residual network layer. All 20 residual layers are used at the end of each resolution subnetwork acting a six residual networks with shared parameters.

with an earlier multi-resolution approach, motivating our network structure. The network architecture is shown in Figure 1. Similarly, there is a sequence of downsampling operations to create the image volumes at each resolution level. (We used average pooling due its small size and fast implementation.)

However, due to the nature of point localisation (distinct from other tasks such as image segmentation) the majority of the image at finer resolutions does not contribute information relevant to the problem at hand. (That is, if the network knows at a coarse resolution that the target point is located

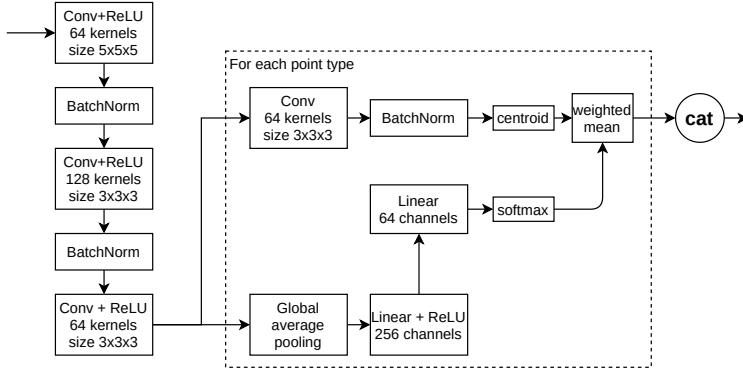


Fig. 3 Architecture for the single resolution subnetwork.

within a particular region, the intensity of voxels outside of said region do not contain information able to improve the network’s accuracy in localising that particular target.) As our problem requires full volumetric images, retaining these non-informative regions represents additional memory and time consumption that could limit its utility. Thus, our network is designed also with a non-differentiable image cropping operation that takes the estimate of the point’s location at a coarse resolution and uses it to crop the image at a finer resolution to a smaller region. This cropped region surrounds the estimated point location, allowing each subsequent layer to update the estimate at a finer without requiring for the entire volume to be stored in memory.

This cropping operation gives us significantly more flexibility in terms of the remaining elements of the network architecture as significantly more GPU memory is available. This allows for the subnetworks that update the centroid locations to be relatively large, each consisting of four convolution layers with a larger number of kernels than in the downsampling operations. These convolutional layers are then interpreted as unnormalised probability maps for a collection of distinct point estimates. To get an updated weighted point estimate, they are passed through a spatial softmax layer and their centroids taken according to the formula:

$$\frac{\sum_i c_i e^{W_i}}{\sum_i e^{W_i}} \quad \text{where} \quad c_i = \frac{\sum_{x \in \Omega} x e^{M_i(x)}}{\sum_{x \in \Omega} e^{M_i(x)}} \quad (1)$$

where i is a channel, x is a location in the image, $M_i(x)$ is the output of the convolution stack for that channel, c_i is the centroid of the resulting probability map, and W_i is the weight of channel i determined by a series of two linear layers, the first of which operates on the flattened image. The architecture of these components is shown in Figure 3.

These are treated as initial estimates for the point locations at this particular resolution layer. However, due to missing or unclear information in the image, it may also be beneficial for these points to be updated given knowledge of the other points. Thus, a series of 20 residual layers is appended to each resolution layer. In order to prevent this from causing a large increase in the

number of parameters, all six resolution layers share the same residual components. This final sub-network also allows for information about the general configuration to be used, similar to a learned statistical shape model. [8]

2.3 Training

Unlike in traditional neural network architectures, it is not feasible to apply a loss function that uses only the final localisation estimate. This is because:

1. The cropping operations are not differentiable with respect to the location they are cropping around meaning that gradients cannot travel from finer resolution layers into coarser ones;
2. Due to the increasing pixel-width, changes in the coarser resolution layers are magnified, i.e., an error of 1 pixel at the coarsest layer is equivalent to an error of 32 pixels at the finest; and
3. Each layer has a limited space of possible estimates, its *field-of-view*, and cannot learn from gradients when the ground truth is outside of that space.

These considerations together require a specialised training structure and loss function. To address the first problem, a separate loss is applied to the estimate from each layer. To address the second, the error for each level can be divided by the resolution. To address this third component, a weighting scheme is used to partition the error towards the layers whose estimates are in the same order of magnitude as their resolution, i.e., the layers that are the best equipped in terms of both resolution and field-of-view. The loss and weighting scheme is described as:

$$L = \sum_l w_l \frac{e_l^2}{r_l^2} \quad \text{where} \quad w_l = \text{sigmoid} \left(-\alpha \left(\frac{\log_2 e_l - \log_2 r_l - 1}{\log_2 f_l} \right) \right) \quad (2)$$

where l is a resolution layer, e_l is the error of that layer's estimate, r_l is that layer's resolution, f_l is that layer's field-of-view measured in pixels. The constant α is used to control how close this weighting is to binary, i.e. how much it allows a particular layer to learn information that is coarser than its optimal resolution but still within its field of view. For our experiments, this value was set to 8. In addition, the weights for each layer are adjusted by its resolution, ensuring gradients for each resolution layer a similar magnitude.

The Stochastic Gradient Descent SGD optimiser was used with a learning rate of 10^{-3} which decays over the number of iterations, i , with a factor of $\frac{1}{1+0.1i}$. All training was performed on an NVIDIA Titan X GPU with 12Gb of memory. A relatively large batch size of 16 full volumetric images was used illustrating the network's capability to conserve GPU memory.

Data augmentation has been implemented in the form of random rotations (std. 10°) and translations (std. 10 mm) which can be easily applied to the point location as well. These augmentations were performed in training time, transforming the image immediately prior to inputting it into the network, rather than saving a database of transformed images.

2.4 Comparative Method

In order to judge the efficacy of the network, it is compared against deformable registration which is showing emerging use in TMS applications. [1] The comparative method uses the SyN toolbox [2] to perform first a rigid registration step, followed by an affine step, then a final non-rigid step. The atlas used is a T1 image of a healthy individual which has been pre-annotated by an expert neurologist. This registration procedure takes on the order of 10-15 minutes to compute per patient compared to 1-2 seconds using the proposed CNN. This registration method has previously been used in research studies performed using the same centre's technology. [4, 7, 18]

2.5 Evaluation Criteria and Methods

In order to evaluate the method, the mean distance error was calculated in a Leave-One-Out cross-validation system. All hyper-parameters were determined prior to performing cross-validation in order to ensure that bias is not introduced through selecting the best model post-cross-validation. The cross-validation is repeated 5 times in order to calculate a dispersion metric, that is, the expected distance between a given prediction and the expected prediction. The dispersion measures how robust the network prediction is to randomness in the network initialisation, which training images are used in the validation set, the order in which training images are presented, random data augmentation etc... The dispersion therefore measures the *precision* of the prediction as opposed to its quality. The dispersion of point p of patient i is:

$$D_{(p,i)} = \frac{1}{N_R} \sum_r \left| P_{(n,p,i)} - \frac{1}{N_R} \sum_n P_{(n,p,i)} \right| \quad (3)$$

where N_R is the number of repetitions (i.e. 5) and $P_{(n,p,i)}$ is the n^{th} network's estimate of point p in patient i . This dispersion also allows for us to intuitively separate the error into the patient/point-specific bias component and a random component that results from non-deterministic aspects in training.

3 Preliminary Results

Preliminary quantitative results are shown in Table 2 with the statistical test results in Table 3. Note that for the non-motor target points only one manual annotation was available. Thus the manual annotation error for these points cannot be estimated. Corresponding qualitative results are shown in Figure 4.

To determine the effect of ensembling and of the residual layers, an experiment was performed in which the LOOCV procedure was repeated five times for six variants of the network. A pair of networks used a shared residual component with another pair using separate residual components for each resolution layer and the last pair having no residual component whatsoever. Each pair had one 'ensemble' that combined the five repetitions, averaging together

the individual network's predictions, while the other item in the pair used the individual network predictions directly. Qualitative results and statistical analysis are given in Tables 4 and 5 respectively.

Point	CNN Error	Disp.	Reg. Error	Expert Var.
LOFC	5.77 ± 3.42 (n = 130)	1.86 ± 2.24 (n = 130)	9.95 ± 4.43 (n = 23)	
ROFC	6.15 ± 3.65 (n = 130)	1.33 ± 0.93 (n = 130)	7.21 ± 4.66 (n = 23)	
LDLPFC	9.08 ± 8.29 (n = 120)	3.24 ± 5.60 (n = 130)	7.44 ± 4.03 (n = 23)	
RDLPFC	7.20 ± 4.30 (n = 120)	2.43 ± 2.49 (n = 130)	8.21 ± 3.34 (n = 23)	
LHESCHL	6.13 ± 3.52 (n = 115)	2.54 ± 1.64 (n = 130)	6.38 ± 3.30 (n = 22)	
LFACEMC	5.54 ± 3.25 (n = 125)	2.29 ± 1.33 (n = 130)	12.93 ± 3.75 (n = 22)	7.12 ± 4.54 (n = 75)
RFACEMC	7.78 ± 5.37 (n = 130)	2.39 ± 1.67 (n = 130)	13.32 ± 3.33 (n = 23)	8.84 ± 5.45 (n = 78)
LLLIMBMC	6.54 ± 5.12 (n = 120)	2.62 ± 2.07 (n = 130)	8.74 ± 5.24 (n = 21)	5.65 ± 3.95 (n = 72)
RLLIMBMC	8.73 ± 7.17 (n = 130)	2.96 ± 2.30 (n = 130)	8.63 ± 5.58 (n = 23)	6.73 ± 6.30 (n = 78)
LULIMBMC	7.82 ± 5.39 (n = 125)	2.58 ± 1.54 (n = 130)	9.51 ± 4.31 (n = 22)	6.85 ± 4.70 (n = 75)
RULIMBMC	8.30 ± 5.01 (n = 130)	2.93 ± 1.82 (n = 130)	10.88 ± 3.99 (n = 23)	6.35 ± 4.79 (n = 78)
MEAN	7.24 ± 5.23 (n = 1475)	2.53 ± 2.39 (n = 1690)	9.39 ± 4.63 (n = 268)	6.94 ± 5.10 (n = 456)

Table 2 Quantitative results (mm) including the error of the proposed method (CNN Error), its dispersion (Disp.), the error of the registration approach (Reg. Error) and the expert variability, i.e. the error between the individual human experts and the consensus point (Expert Var.). The **MEAN** row aggregates together the chronic pain treatment points. Greyed out cells represent regions that could not be computed due to only having a single expert annotation.

Point	CNN vs. Reg.	CNN vs. Expert	Reg. vs. Expert
LOFC	3.74 ** CNN		
ROFC	1.46		
LDLPFC	0.55		
RDLPFC	1.28		
LHESCHL	0.89		
LFACEMC	3.84 ** CNN	2.11	3.68 ** Expert
RFACEMC	3.65 ** CNN	1.64	3.95 ** Expert
LLLIMBMC	2.81* CNN	0.63	2.81* Expert
RLLIMBMC	0.85	1.23	1.55
LULIMBMC	2.66* CNN	0.61	2.85* Expert
RULIMBMC	3.04* CNN	2.12	3.55 ** Expert

Table 3 Results of paired Wilcoxon tests on the mean error (absolute Z-values) for each pair of methods. (Data is paired by patient.) Statistically significant results (after Holm-Bonferroni correction) are shown in **bold** with * meaning $p \leq 5\%$, ** meaning $p \leq 1\%$ and *** meaning $p \leq 0.1\%$. The best performing method in the pair is shown for statistically significant differences.

Point	Ind. Res. No Ensemble	Ind. Res. Ensemble	Shared Res. No Ensemble	Shared Res. Ensemble	No Res. No Ensemble	No Res. Ensemble
LOFC	5.60 ± 3.51	5.14 ± 2.56	5.77 ± 3.42	5.49 ± 2.59	5.97 ± 5.11	5.65 ± 3.21
RORC	6.16 ± 3.94	5.96 ± 3.68	6.15 ± 3.65	6.01 ± 3.57	6.71 ± 4.79	6.45 ± 3.82
LDLPFC	8.90 ± 7.46	8.59 ± 5.87	9.08 ± 8.29	8.81 ± 5.47	8.79 ± 7.25	8.46 ± 5.35
RDLPFC	6.96 ± 3.99	6.30 ± 4.09	7.20 ± 4.30	6.62 ± 3.77	7.22 ± 4.43	6.76 ± 3.67
LHESCHL	6.00 ± 3.30	5.59 ± 2.90	6.13 ± 3.52	5.61 ± 3.35	6.60 ± 3.68	6.15 ± 3.37
LFACEMC	5.79 ± 3.37	5.27 ± 2.81	5.54 ± 3.25	5.08 ± 2.96	5.78 ± 3.37	5.22 ± 2.72
RFACEMC	7.86 ± 5.32	7.46 ± 5.14	7.78 ± 5.37	7.44 ± 5.14	7.66 ± 5.23	7.12 ± 5.05
LLLIMBMC	6.68 ± 5.32	6.25 ± 5.06	6.94 ± 5.12	6.08 ± 4.73	6.77 ± 4.86	6.39 ± 4.42
RLLIMBMC	8.33 ± 7.56	7.76 ± 7.01	8.73 ± 7.17	8.10 ± 7.03	8.62 ± 7.36	7.98 ± 7.43
LULIMBMC	7.79 ± 4.95	7.25 ± 4.81	7.82 ± 5.39	7.48 ± 5.10	7.89 ± 5.45	7.39 ± 5.24
RULIMBMC	8.45 ± 4.81	8.06 ± 4.61	8.30 ± 5.01	7.92 ± 4.46	8.16 ± 4.97	7.71 ± 4.63

Table 4 Quantitative results (mm) of different variants of the proposed method, specifically the presence of the residual layers and of ensembling. (Each ensemble is the average of the models created in the five repetitions of cross-validation.)

Point	Ens. vs No (Ind. Res.)	Ens. vs No (Shared Res.)	Ens. vs No (No Res.)	Ind. vs Sh. (With Ens.)	Ind. vs Sh. (With Ens.)	Sh. vs No (With Ens.)	Ind. vs Sh. (No Ens.)	Ind. vs No (No Ens.)	Sh. vs No (No Ens.)
LOFC	-4.45 ***	-4.45 ***	-4.45 ***	2.10	-2.11	-0.05	1.05	-1.50	-0.13
ROFC	-4.45 ***	-4.45 ***	-4.45 ***	0.85	-1.68	-1.70	-0.31	-1.65	-1.78
LDLPFC	-4.28 **	-4.28 **	-4.28 **	0.8	0.4	0.54	0.40	0.26	0.66
RDLPFC	-4.28 **	-4.28 **	-4.28 **	0.37	-0.99	-0.79	0.49	-1.25	-0.02
LHESCHL	-4.19 **	-4.19 **	-4.19 **	0.40	-1.88	-1.69	0.88	-2.09	-1.12
LFACEMC	-4.36 **	-4.36 **	-4.36 **	-0.73	1.04	-0.08	-1.16	0.63	-0.54
RFACEMC	-4.45 ***	-4.45 ***	-4.45 ***	-0.51	1.36	1.18	-0.59	0.65	0.42
LLLIMBMC	-4.28 **	-4.28 **	-4.28 **	-0.42	-1.05	-1.93	-0.16	-0.70	-2.10
RLLIMBMC	-4.45 ***	-4.45 ***	-4.45 ***	1.13	-0.66	0.44	1.89	-1.17	0.32
LULIMBMC	-4.36 **	-4.36 **	-4.36 **	0.55	-0.08	0.82	0.09	0.00	-0.27
RULIMBMC	-4.45 ***	-4.45 ***	-4.45 ***	-1.2	1.16	0.88	-1.25	1.13	0.44

Table 5 Results of paired Wilcoxon tests on the mean error (Z -values) adjacent methods, i.e. with and without residual layers (Res.) and with or without ensembling (Ens.). (Data is paired by patient.) Statistically significant results after Holm-Bonferroni correction are shown in **bold** with * meaning $p \leq 5\%$, ** meaning $p \leq 1\%$ and *** meaning $p \leq 0.1\%$. (Ind. means individual separate residual subnetworks for each resolution layer, Sh. means a shared residual subnetwork, and Ens. means the use of ensembling.)

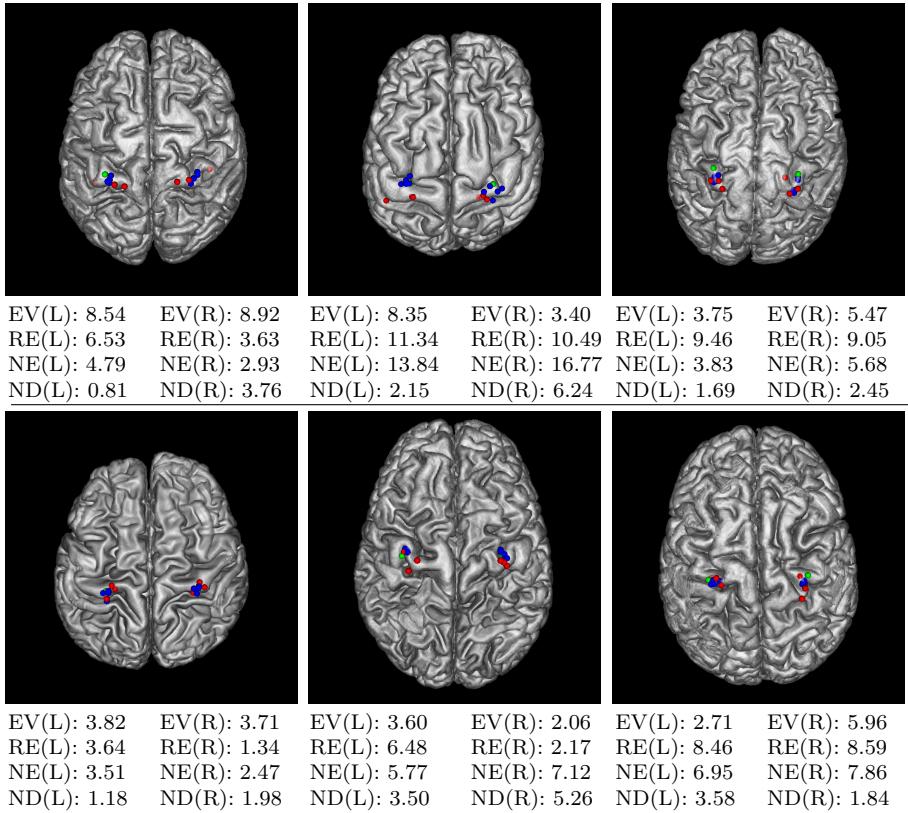


Fig. 4 Results for the left and right lower limb areas of the motor cortex (LLLIMBMC and RLLIMBMC, respectively) of a representative patient. Points selected by a clinical expert are shown in red, the proposed method in blue, and the atlas method in green. The expert variability (EV), registration error (RE), network error (NE) and network dispersion (ND) are given for both sides of the patient shown.

4 Discussion

From our results, we can clearly see that the proposed algorithm borders on human performance as it sometimes outperforms and sometimes underperforms the expert variability but these differences are not statistically significant. The most significant results are in favour of the proposed network for the LFACEMC targets and the human expert for the RULIMBMC targets although both of these reflect a difference in error of approximately 2mm. The similarity between expert variability and the proposed framework's performance is further underlined by the aggregate scores for the chronic pain treatment points which differ by an insignificant 0.30 mm. The registration based approach's error, however, is larger than the expert variability with statistical significance for all but two targets. With respect to the expert users, there is a large degree of variability which is likely a result of the high degree of patient variability and the lack of functional information in the T1w images.

Thus, the use of the proposed network in TMS planning could possibly be considered equivalent to the manual approach, the current standard-of-care, and is a significantly better candidate for automatic target point localisation in TMS than the deformable registration approach. The general improvement in accuracy of the network compared to registration indicates that patient anatomical variability at least has been partially addressed. In addition, due to the deep learning nature of the proposed, it can always be improved through the addition of more annotated images to the training database.

For the non-motor treatment points in which a consensus annotation was not available, the learning and registration methods performed much more similarly. The proposed CNN outperformed registration for five of the six targets (although only with statistical significance for one) whereas registration outperformed the network for only the LDLPFC and not with significance. This indicates that there is a distinct advantage to using consensus points rather than a single observer for this particular machine learning framework. There is a possibility that the singular expert annotations for these points is not fully correct, which renders any comparison between two algorithms using said annotation questionable. However, there is also a possibility that the proposed framework is more sensitive to error in the training dataset although both of these would be difficult to verify without an independent gold standard, which will be discussed in the subsequent section.

We originally hypothesised that certain elements of the network had a non-negligible effect on its performance. However, comparing against small ablations/modifications of the algorithm tells us only that using ensembling improves the accuracy of the methods on the order of 0.1-0.5mm, although this improvement would not affect the statistical significance of comparisons against either the atlas or the human experts. Despite our hypothesis, the effect of the residual layers was negligible, indicating that information about the location of the target points is determined almost entirely from the image, rather than from a statistical understanding of the relationships between the target points' locations.

Limitations

The primary limitation of this work is the method in which the reference target points are collected for both training and evaluating the automatic methods. Our dataset inherently contains a non-negligible level of error due to the possibility that the singular rater for these points is incorrect or that (for targets in which all three raters are available) two raters simultaneously mis-localise the target, resulting in an incorrect consensus being used. Currently, independent and high-accuracy ground-truth target points are not available as the T1 MRI on which these targets are determined does not contain the functional information required to make such a distinction. Previous studies have addressed this issue by defining these points in terms of the point that gives maximal stimulation effect [1] but this can be difficult to ascertain for target points outside of the primary motor cortex. The lack of consensus points in some

scenarios effectively gave the CNN fewer full datasets to use for training that, given the relatively small size of the dataset, could have detrimental effects on its accuracy with respect to the chronic pain treatment targets.

This lack of independent ground truth annotations also has a distinct effect on the interpretation of the network results, specifically a bias in favour of the human experts. For example, a larger number of targets in which all three experts disagreed should intuitively increase the expert variability. However, since no consensus point be identified, there is no reference against which to measure this error, thus decreasing the number of datasets used in the comparison rather than increasing the error. By increasing the number of experts, we would be more likely to find correct consensus points, allowing for a more unbiased comparison between algorithms and human experts.

From a technical perspective, there are some remaining advantages to the registration approach that are not currently implemented by the proposed network. The most important is that new targets can be easily added to the atlas, allowing for new calibration or treatment targets to be incorporated as desired. As TMS is an evolving therapy for a range of neurological disorders, this is likely to happen. To extend a deep learning framework in a similar manner however would involve the annotation of existing images in the training database with these new target points, which requires significantly more time.

Future Work

As suggested in the previous sections, there is still work to be done regarding improving this network and putting it into clinical use. The first is to continue to collect datasets with multiple expert annotations in order to find high quality consensus points to use in training the network. Although the network boarders on human accuracy, it still likely has room to improve.

From a research perspective, we would like to specifically investigate the effects of topological differences to further ensure our hypothesis that this network is more robust to large patient anatomical variations. This would involve the collection of a larger and more varied dataset of individuals that we could then separate into different classes based on common variations (number of gyri, etc...) that have a strong effect on registration performance.

5 Conclusion

This paper presents a multi-resolution convolutional neural neural that is specifically designed for localising points in large volumetric images. The novel aspects of this architecture include the cropping operation which allows for large amounts of GPU memory to be conserved, and the customised loss framework which addresses the non-differentiability of this sampling operation as well as ensuring efficient learning across multiple resolutions.

The proposed network outperforms that of deformable registration which is the state-of-the-art in automatic point localisation for TMS planning and

borders on expert performance which is the current clinical practice. This improvement represents a step towards efficient and fully automatic TMS planning that can be readily used by smaller centres that are increasingly performing these interventions.

Acknowledgements

John S.H. Baxter is supported by the Institut des Neurosciences Cliniques de Rennes (INCR) and the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Post-Doctoral Fellowship (PDF) program. Quoc Anh Bui is funded through a financial support from Région Bretagne. Ehouarn Maguet is also supported INCR. The authors would like to thank X. Morandi, C. Nauczyciel, B. Le Goff, and J.-P. N'Guyen for the annotation of the non-motor treatment points. The authors would also like to thank J.-P. N'Guyen and H. Hodaj for their assistance in annotating the chronic pain treatment points along with J.-P. Lefaucheur.

Declarations

Funding: No funding was received to assist in the preparation of this manuscript.

Conflicts of interest: S. Croci, A. Delmas and L. Bredoux are employees of SYNEIKA. The remaining authors have no financial or non-financial conflicts of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Availability of data and material: Data is not available for this study.

Code availability: Code is currently not made available.

References

1. Ahdab, R., Ayache, S., Brugières, P., Goujon, C., Lefaucheur, J.P.: Comparison of “standard” and “navigated” procedures of tms coil positioning over motor, premotor and prefrontal targets in patients with chronic pain and depression. *Neurophysiologie Clinique/Clinical Neurophysiology* **40**(1), 27–36 (2010)
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* **12**(1), 26–41 (2008)
3. Baxter, J.S.H., Maguet, E., Jannin, P.: Localisation of the subthalamic nucleus in mri via convolutional neural networks for deep brain stimulation planning. In: *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*. vol. 11315, p. 113150M. International Society for Optics and Photonics (2020)

4. Bulteau, S., Beynel, L., Marendaz, C., Dall'Igna, G., Peré, M., Harquel, S., Chauvin, A., Guyader, N., Sauvaget, A., Vanelle, J.M., Polosan, M., Bougerol, T., Brunelin, J., Szekely, D.: Twice-daily neuronavigated intermittent theta burst stimulation for bipolar depression: A randomized sham-controlled pilot study. *Neurophysiologie Clinique* **49**(5), 371–375 (2019)
5. Freitas, C., Mondragón-Llorca, H., Pascual-Leone, A.: Noninvasive brain stimulation in alzheimer's disease: systematic review and perspectives for the future. *Experimental gerontology* **46**(8), 611–627 (2011)
6. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C.: Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging* **37**(8), 1822–1834 (2018)
7. Harika-Germaneau, G., Rachid, F., Chatard, A., Lafay-Chebassier, C., Solinas, M., Thirioux, B., Millet, B., Langbour, N., Jaafari, N.: Continuous theta burst stimulation over the supplementary motor area in refractory obsessive-compulsive disorder treatment: A randomized sham-controlled trial. *Brain stimulation* **12**(6), 1565–1571 (2019)
8. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis* **13**(4), 543–563 (2009)
9. Johnson, S., Summers, J., Pridmore, S.: Changes to somatosensory detection and pain thresholds following high frequency repetitive tms of the motor cortex in individuals suffering from chronic pain. *Pain* **123**(1-2), 187–192 (2006)
10. Kim, W.J., Min, Y.S., Yang, E.J., Paik, N.J.: Neuronavigated vs. conventional repetitive transcranial magnetic stimulation method for virtual lesioning on the b roca's area. *Neuromodulation: Technology at the Neural Interface* **17**(1), 16–21 (2014)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. pp. 1097–1105 (2012)
12. Lefaucheur, J.P., Aleman, A., Baeken, C., Benninger, D.H., Brunelin, J., Di Lazzaro, V., Filipović, S.R., Grefkes, C., Hasan, A., Hummel, F.C., Jääskeläinen, S.K., Kimiskidis, V.K., Koch, G., Langguth, B., Nyffeler, T., Oliviero, A., Padberg, F., Poulet, E., Rossi, S., Rossini, P.M., Rothwell, J.C., Schönfeldt-Lecuona, C., Siebner, H.R., Slotema, C.W., Stagg, C.J., Valls-Sole, J., Ziemann, U., Paulus, W., Garcia-Larrea, L.: Evidence-based guidelines on the therapeutic use of repetitive transcranial magnetic stimulation (rtms): an update (2014–2018). *Clinical neurophysiology* **131**(2), 474–528 (2020)
13. Pennisi, G., Ferri, R., Lanza, G., Cantone, M., Pennisi, M., Puglisi, V., Malaguarnera, G., Bella, R.: Transcranial magnetic stimulation in alzheimer's disease: a neurophysiological marker of cortical hyperexcitability. *Journal of Neural Transmission* **118**(4), 587–598 (2011)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*. pp. 91–99 (2015)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
16. Rusjan, P.M., Barr, M.S., Farzan, F., Arenovich, T., Maller, J.J., Fitzgerald, P.B., Daskalakis, Z.J.: Optimal transcranial magnetic stimulation coil placement for targeting the dorsolateral prefrontal cortex using novel magnetic resonance image-guided neuronavigation. *Human brain mapping* **31**(11), 1643–1652 (2010)
17. Summers, J.J., Kagerer, F.A., Garry, M.I., Hiraga, C.Y., Loftus, A., Cauraugh, J.H.: Bilateral and unilateral movement training on upper limb function in chronic stroke patients: a tms study. *Journal of the neurological sciences* **252**(1), 76–82 (2007)
18. Vignaud, P., Damasceno, C., Poulet, E., Brunelin, J.: Impaired modulation of corticospinal excitability in drug-free patients with major depressive disorder: a theta-burst stimulation study. *Frontiers in human neuroscience* **13**, 72 (2019)
19. Weise, K., Numssen, O., Thielscher, A., Hartwigsen, G., Knösche, T.R.: A novel approach to localize cortical tms effects. *Neuroimage* **209**, 116486 (2020)