Improving Similarity Metric of Multi-modal MR Brain Image Registration Via a Deep Ensemble

Natan Andrade, Fabio A Faria, Fábio A M Cappabianco
GIBIS - Department of Science and Technology - Universidade Federal de São Paulo
Av. Cesare Monsueto Giulio Lattes, 1201 - Eugênio de Melo 12247-014
São José dos Campos/SP, Brazil

***Abstract*—Brain image registration fuses and aligns sets of structural or functional images within individual and population studies. The similarity metric is an image registration component used for detecting the same target region in different images. Multi-modal image registration constitutes one of the greatest challenges in medical imaging as it adds even more variability to the tissue and organ appearance, shape, and positioning. This paper contains two contributions to solve this complex problem: (1) we propose a solution to compute the similarity metric based on a deep ensemble method. It combines multiple traditional and deep similarity metrics into a single improved similarity map; (2) we propose novel evaluation metrics to validate the results. Experiment results in the context of T1- and T2-weighted MR images of the human brain show a major improvement to the state-of-the-art, especially in reducing the false-positive region occurrences.**

I. Introduction

Brain image registration optimizes a geometric transformation of different images into a common space, establishing a known correspondence between them. For this purpose, the registration operates on a transformed image using a static image as its target. Image registration is fundamental for several analytic studies based on structural and functional images, including researches for understanding population tendencies of phenotypes, measuring longitudinal changes of organs and tissues, guiding surgery procedures through images, and correlating an individual's anatomy to a standard atlas space [1]–[3].

Image registration methods are composed of three main components: (1) a transformation model which defines the complexity of possible geometrical changes to the image contents; (2) a similarity metric used to measure how well two images or two image regions (usually defined by patches) are related to each other; and (3) an optimization strategy which consists of the algorithm for searching the best matching between images/patches [4]–[6].

This paper focuses on the second image registration framework component, i.e., the similarity metric. Given a small region of interest or search patch from the transformed image, and convoluting it over the target image, the similarity metric computes the probability of the overlapping regions to contain corresponding contents, generating a similarity map with the same dimensions as the target image as the output. An accurate similarity metric should provide a map with high probability values for the corresponding location between the transformed and target images and low probability values elsewhere.

We may compute the similarity metrics based on manually set markers provided by specialists at well-known locations, or by automatically detecting corresponding regions by their distinctive features. The first procedure is tedious and time-consuming, not practical or feasible for longitudinal studies. The second is more broadly applicable even though the majority of the existing metrics such as the sum of squared differences or sum of absolute differences are not suitable for multi-modal image registration [1], [3], [7]

More recently, deep learning-based metrics became popular in literature [8]–[11]. Still, these approaches are not reliable since they have very low specificity in the target task, as we show in Section IV.

Multi-modal image registration – i.e. involving images of different modalities or acquisition protocols – is challenging due to the distinct appearance of tissues and organs, sometimes not even present in one of the images. For instance, human brain image registration displays a challenging factor because of the enormous shape variability in the cortical gyri and sulci regions. There is also a lack of standard procedure to evaluate registration results as previous works differ in their used validation measures and methodology [3], [9], [12], [13].

In this paper, we propose a solution to compute the similarity metric based on a deep ensemble method trained with both classic and deep similarity metrics. It identifies the precise matching between multi-modal human brain magnetic resonance imaging (MRI) patches of the same subject. We also introduce two novel evaluation metrics to compare the results of the proposed and other state-of-the-art similarity metrics. In Section II, we present the related works of the literature. Section III describes our proposed solution. Section IV contains our evaluation metric and experiments, and in Section VI we state the conclusions and future works.

II. Related Works

The similarity metric in the context of medical imaging registration is a function that identifies corresponding pixels between the transformed and target images. In the case of multimodal registration, the metric takes into account the relative location of the pixels and the region contents in their adjacency [2], [14].

In the literature, there are two classic similarity metrics that are widely used for multi-modal registration: the Mutual Information (MI) [15], [16] (Equation 1) and the Correlation

Ratio (CR) [17] (Equation 2). Other classic similarity metrics are found in [18]. The higher the values of MI and CR, the more correlated are the corresponding pixels x and y of two images X and Y, respectively.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (1)$$

$$\eta(Y|X) = \frac{Var[E(Y|X)]}{Var(Y)} \quad (2)$$

One more classic evaluation metric was proposed in [12]. The authors present an algorithm derived from max-margin output learning to assimilate a linear function similarity metric and perform a multi-modal rigid registration. It selects high information and contrast patches for the registration process. They evaluated the quality of the registration based on the Target Registration Error (TRE) which is the Euclidean Distance (ED) between the transformed and the target images.

Since the advent of deep learning, several solutions have been proposed to improve the accuracy of the classical similarity metrics. In [19], the authors propose the Spatial Transformer Network (STN) which allows the spatial manipulation of input images. It was used in the context of image alignment but was also considered by other medical image registration papers [7], [20], [21].

In [8], the authors propose a deep neural network that takes the transformed and target image patches as input and outputs a probability value for the patches being representative of the same region of interest. They used this probability as a similarity metric. They tested the proposed algorithm on a group of computed tomography (CT) and magnetic resonance (MR) images and computed the proximity of the corresponding patch concerning other regions of the images with a similarity ranking of the patches.

In [9], the authors use a Convolutional Neural Network (CNN) to compute a similarity metric and distinguish between aligned and misaligned multi-modal images, similarly to [12]. They evaluated results using Jaccard and Dice metrics of overlapping transformed and target image patches, with a slight improvement over MI.

In [10], the authors propose using a convolutional stacked auto-encoder to reduce the dimensionality of the image and select its most relevant intrinsic deep feature representations in image patches. Comparison with classical registration methods used Dice metric.

In [22], the authors use a larger image dataset with data augmentation in a semi-supervised fashion to train a CNN. They measure the accuracy by computing the Euclidean distance between image poses.

The work in [23] proposes utilizing a similar method as in [9] to register $T1$-MR and ultrasound images. They also use the Euclidean distance to validate the results.

We could notice that one of the main goals of employing deep learning in the context of image registration is to estimate a measurement of similarity between different images, and in most cases, using patches. The deep neural network similarity metrics have shown a great potential for registering images from the same modality, but results are far from satisfactory for multi-modal images [7].

Even though some previous works explore the usage of different similarity metrics for medical image registration, they do not compare the proposed metrics with other learning-based methods, but almost exclusively with classical metrics such as mutual information. This creates a lack of correct understanding of how good the results are concerning the state-of-the-art methods. Another issue related to their validation is a very high number of false-positive occurrences, which severely impacts the quality of the registration process.

Therefore, in this work, we cover all these issues, proposing a deep ensemble method, which reduces the false-positive region occurrences as compared to different traditional and deep similarity metrics. We present two novel evaluation metrics that allow objective quantification of false positive and false negative errors [24]. Finally, in our experiments, we compare several state-of-the-art deep learning-based algorithms with our proposed solution.

III. Deep Ensemble Method

Figure 1 and 2 show our solution based on a deep ensemble method in a sample application of multi-modal $T1$ and $T2$ brain MR image registration. We divide the multi-modal MR brain dataset ($T1$ and $T2$) into three sets (training, validation, and test) so that they do not share images from the same individuals. We crop each image into smaller square regions and create five sets of image patches: two sets for training deep learning-based metrics (Tra_M, Val_M); two sets for training the proposed deep ensemble method (Tra_E, Val_E); and a test set Tes for the final evaluation.

For training the pipeline, the corresponding patch location in the target image also must be provided as an input. The first step consists of preprocessing the images by having isometric pixels, skull stripping the brain, and normalizing the image intensities. This way, the training and testing of the method become more reliable [25], [26].

Tra_M set is used as input patches of the N different deep learning-based similarity metrics (M_i, where $i = \{1, 2, 3, \ldots, N\}$), resulting in $N \times |Tra_M|$ different similarity maps. We compute a cross-entropy (CELoss) function as loss function between the patch binary masks $p(x)$ (ground truth) and the activation values of the last layer $q(x)$ of the metrics as following on Equation 3.

$$CELoss(p,q) = - \sum_{x \in Tra_M} p(x) \log q(x), \quad (3)$$

We apply the Val_M set to optimize the parameters of the N metrics throughout the learning process and to select the most suitable $N^* \subset N$ similarity maps for the next step. We manually selected and tested a few combinations of the similarity maps for this paper due to the intensive computational power required by the experiments.

On deep ensemble training process, each of the patches from $|Tra_E|$ passes through N^* learned metrics resulting in N^*

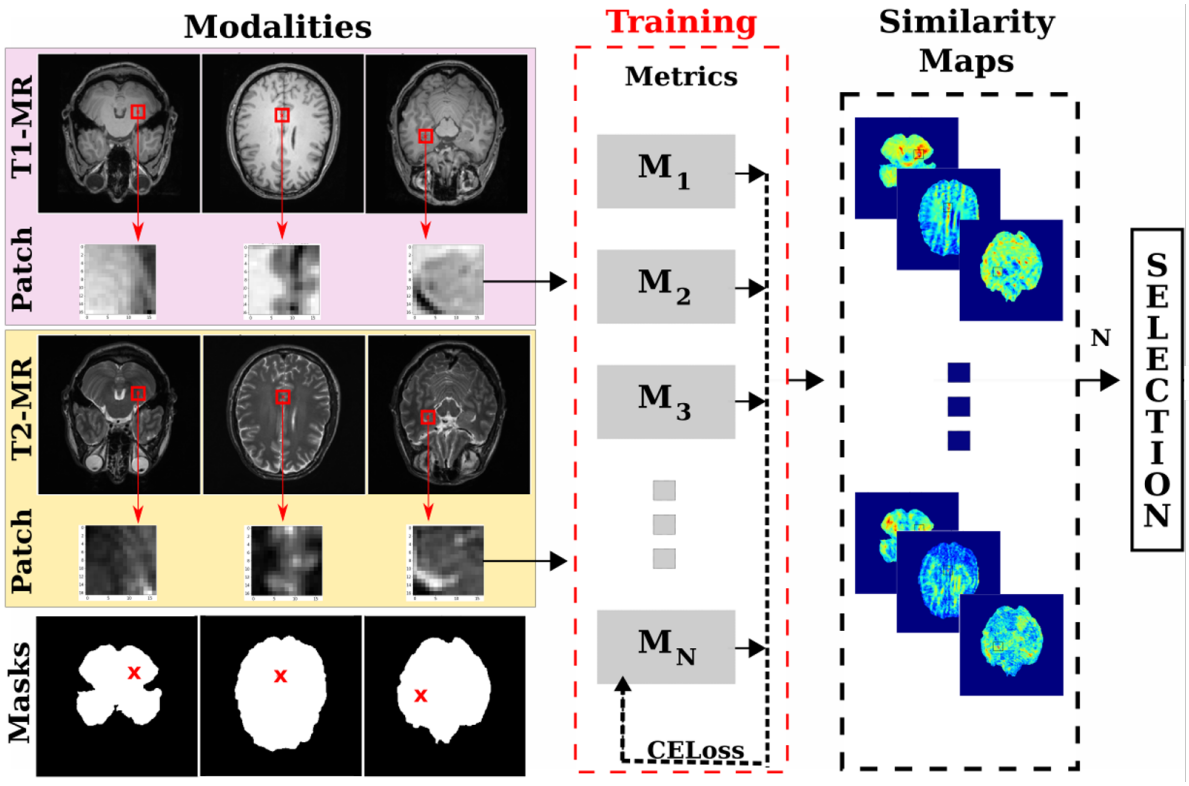

Fig. 1. Individual similarity metric training step, based on multimodal image patches, generating similarity maps for the ensemble.

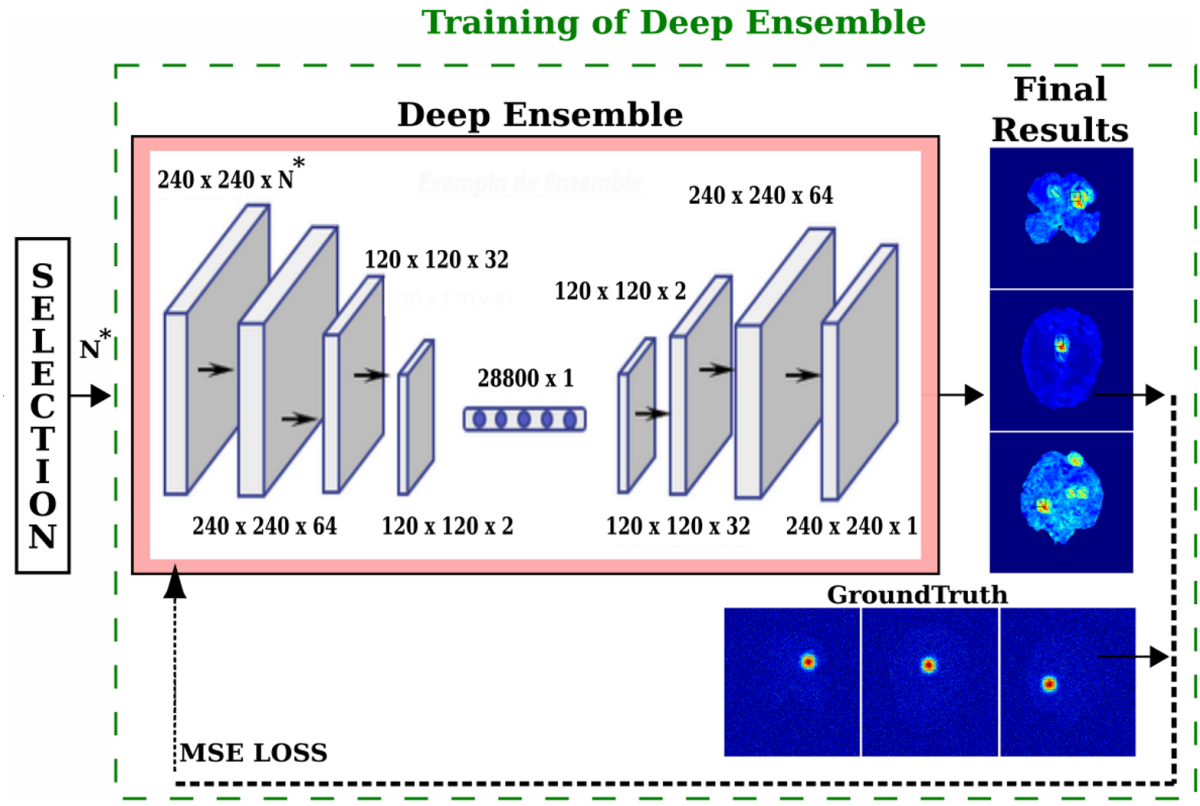

Fig. 2. Deep ensemble training, based on classical and learning-based similarity maps. The final result is an improved similarity map for the region of interest.

similarity maps of dimension 240×240 (one map per metric), which they compose a joint similarity map of dimension $240 \times 240 \times N^*$. We use this new set called Tra_E as an input for a neural network like autoencoder architecture (proposed deep ensemble method).

In the final step, the ensemble method creates $|Tra_E|$ combined maps of dimension $240 \times 240 \times 1$. Even in this process, we run the reconstruction task using the mean score error (MSE) function between the ensemble outputs (Y') and the expected response masks (Y) (Equation 4). Figure 3 shows an example of a response mask in our experiments.

$$MSE(Y, Y') = \frac{1}{n} \sum_{i=1}^{n} |Y - Y'|^2, \quad (4)$$

Finally, we evaluate the deep ensemble method performance over the target task on the test set (Tes).

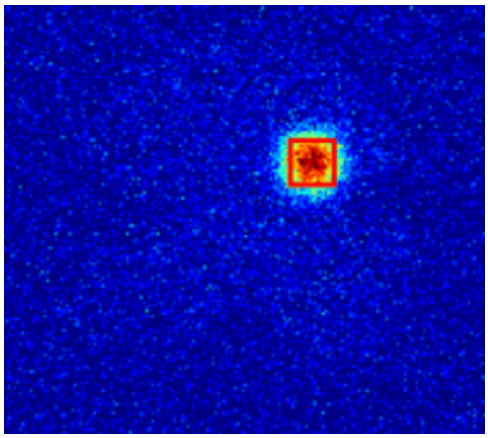

Fig. 3. An example of response mask defined by a Gaussian kernel function used as ground-truth to train the proposed deep ensemble method.

IV. Experimental Methodology

A. Building the Dataset

We have used a set of 100 individuals from IXI database[1]. We performed five pre-processing procedures for each image: (1) The normalization of intensities and spatial; (2) The alignment between $T1$ and $T2$ through the FSL framework and its *Brain Extraction Tool* (BET) [27], [28]; (3) The removal of non-brain regions (skull stripping) to avoid that patches without any information confuse the similarity metrics throughout the training process; (4) The generation and erosion of the brain region of the image (10 pixels) to avoid selecting background regions; and (5) The limitation of 5 points and 10 slices per individual.

For training the deep learning-based metrics, we created $500,000$ corresponding pairs of patches (T1 and T2) and separated them into 70% training $(350,000)$, 30% validation $(150,000)$, according to the individuals of each set. We define these pairs of patches containing corresponding areas of the individual as the positive cases in Tra_M and Val_M. We shuffled the positive cases to generate the negative cases of non-corresponding pairs of T1 and T2 patches, creating $500,000$ new patches for training, and doubling the number of patches in a total of 1 million patches ($|Tra_M| = 700,000$ and $|Val_M| = 300,000$). Again, we carefully avoided mixing individuals of different sets.

During the training of the deep ensemble method, we selected other $52,000$ pairs of patches ($T1$ and $T2$) inside the brains in axial slices and also divided them in training ($|Tra_E| = 70\%$) and Validation ($|Val_E| = 30\%$) sets. All patches from the same subject belong to the same set.

Finally, we created the test set (Tes) with 336 patches from 10 new individuals not present in the set of 100 individuals previously used on sets. We performed experiments with patches of different dimensions $b \times b$ pixels, where $b = \{13, 17, 21\}$, and we cropped them from the same centralized region.

We run all deep learning-based metrics and the proposed ensemble method on an Intel(R) Xeon(R) dual E5-2630 CPU with 64GB of RAM and a Titan V graphics card.

B. Training Setup of the Deep Learning Metrics

We tested several variations of deep learning architectures during the training process of the deep learning similarity metrics, and we describe here the ones that achieved the best results. Among the ones we tested in our pipeline are: the Deep Neural Networks (DNN) proposed in [8], the CNN inspired on the work proposed in [10], the CapsNet based on [29], and an architecture using Spatial Transformer Network layers proposed in [19]. All these architectures have been trained with square patches of $b \times b$, where, $b = \{13, 17, 21\}$ pixels. With exception to STN-based architecture, which we trained with patches of 17×17 pixels in three distinct structures (encoder, decoder, and encoder-decoder).

We trained the DNN for $10,000$ epochs using binary cross-entropy as the loss function, Adadelta optimizer, and batch size equal to 256.

We modified the structure of the CNN proposed in [10] by adding Max-pooling and Batch Normalization layer, which improved its results. Furthermore, we also added the last layers presented in [8] to obtain a continuous similarity metric. We trained it through $1,000$ epochs, using binary cross-entropy as the loss function, Adadelta optimizer, and batch size equal to 256.

We modified CapsNet to obtain a relevant and continuous similarity map as it was not used for that purpose previously. Therefore, we added a separate layer for each imaging modality at the beginning and another layer before merging them for learning the most relevant features. After the output of the capsules, we also included a layer to output the similarity metric as in [8]. We trained it through 500 epochs, using binary cross-entropy as the loss function, Adadelta optimizer, and batch size equal to 256.

We have applied STN within an autoencoder (AE) composed of 12 convolutional layers. It has 5 STN layers and ends with 4 layers for the similarity metric computation. We

[1]http://brain-development.org/ixi-dataset/

have implemented the network in three different ways: (1) the STN layers within the encoder layers (AE-E); (2) the STN layers within the decoder layers (AE-D); and (3) two sets of STN layers within both the encoder and the decoder (AE-ED). Our motivation comes from successful applications of the STN to improve the registration process by convolutional networks [19].

We chose CapsNet and STN structures due to their ability in detecting similarity between geometric transformed images. Even images from the same individual in multimodal imagens may be shifted or rotated. Also, as the goal was to implement deep ensembles, we focosed on small structures with complementary behavior that could be treined in a reasonable time, with the potential to improve the final results.

C. Training Setup of the Deep Ensemble Method

To train the deep ensemble, we used the Mean Squared Error (MSE) metric over a generated response mask based on the target patch position. As we wanted the output similarity map to be accurate, not rejecting results that are very close to the perfect match, the response mask has the highest accuracy score of 1.0 in the perfect match position, i.e., the center of the patch and decreases in a Gaussian kernel function (see Figure 3) with the size of the patch. The entire skull-stripped brain also has a small accuracy value of 0.15 so that the ensemble could learn that all expected matches should occur inside the brain. We also added a Gaussian noise over the image to avoid getting stuck during the ensemble training procedure.

D. Evaluation Metric

We employed two of the evaluation metrics recommended by the methods described in Section II: a distance-based and a region-based evaluation metric. The Euclidean distance metric in Equation 5 computes the distance between the center of the ground-truth patch (G) and the point with the highest probability in the similarity map (P), where G_x, G_y, P_x, and P_y are the x and y coordinates of the center of G and P.

$$ED = \sqrt{(G_x^2 - P_x^2) + (G_y^2 - P_y^2)} \tag{5}$$

We also used the Jaccard Coefficient (JC) in Equation 6 as a region-based evaluation metrics. It is computed based on the number of true-positive (TP), false-positive (FP), false-negative (FN) overlapping areas of the ground-truth patch (G) and the regions (R_i) with probability greater than a threshold value T in the similarity map. Note that a method scores very poorly even if it outputs a single high probability region R with a significant overlapping area with G. As the patch size is always the same, FN and TP are complementary values concerning it in this application.

$$JC = \frac{TP}{TP + FP + FN} \tag{6}$$

One problem of ED is that it disregards the number, size, and distribution of high probability regions in the similarity map. JC, on the other hand, does not allow distinguishing the difference between over-segmentation and under-segmentation. JC, TP, FP, and FN values are expressed in terms of pixels while for image registration, it is more important to evaluate results in terms of the number of correct and incorrect labeled regions.

For a similarity metric to be successfully used in a registration procedure (1) R_i and G should overlap with high probability, close to 1.0, (2) R_i should not be too large in comparison with G even if it completely overlaps with G, (3) there should not be other non-overlapping regions of R_i with high probability, and (4) the metric should be easily interpretable. In this sense, as another contribution of this paper, we propose two novels evaluation metrics: the true-positive overlap rate ($TPOR$) and the false-positive overlap rate ($FPOR$), described in Algorithm 1.

The $TPOR$ is either 0 or 1, being 1 a score for when the correct patch is detected. $FPOR$ is a non-negative value. The lower the value, the better is the score. We compute them as follows: given the similarity map with the probabilities assigned to each pixel, we apply a hysteresis threshold (in our case with a low threshold of 0.8 and a high threshold of 0.9), so that we binarize the similarity map with the highest probability pixels set to 1 and the other pixels to 0. Then, for each connected component C in the binary map:

Algorithm 1: Algorithm to compute TPOR and FPOR metrics.

$TPOR \leftarrow 0$;
$FPOR \leftarrow 0$;
for ***each*** $|R_i|$ **do**
  if $|R_i \cup G| \geq 0.1G$ *and* $|R_i| < 2|G|$ **then**
    $TPOR = 1$;
  else
    $FPOR = FPOR + \lceil |R_i|/|G| \rceil$;

The advantage of the proposed method is that using $TPOR$, we can identify if the proposed method detects the correct patch region and by looking at FPOR we can see if the method is outputting false-positive regions with high probability. We chose the hysteresis parameters and the values of 0.1 and 2 manually, but small variations of these values do not affect the comparative results of similarity metric outputs significantly. Figure 4 shows an illustration of FPOR and TPOR application in a synthetic image.

V. Results and Discussion

In our experiments, we have performed experiments with the proposed method in a variety of combinations of classic and deep similarity metrics. We present here a quantitative and a qualitative evaluation of the results.

A. Quantitative Analysis

Table I shows the average values of the experiments for five different similarity metrics (CapsNet, CNN, DNN, MI, and

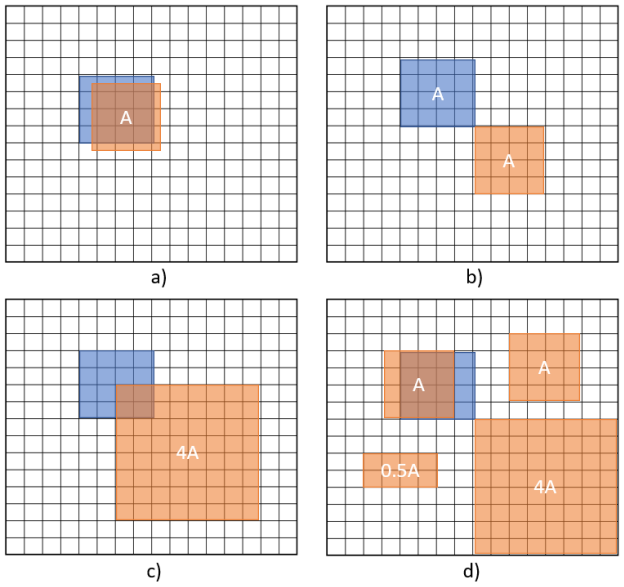

Fig. 4. Illustration of TPOR and FPOR evaluation metrics, where the blue square has area A and represents the target patch $|G|$. The orange patches are the high probability connected component regions $|R_i|$ in the output similarity map of a given method. (a) The transformation and target image patches match in size and location (FPOR = 0 and TPOR = 1). (b) The transformation and target image patches match in size but not in location (FPOR = 1 and TPOR = 0). (c) Transformation and target image patches match in location, but not is size. Target patch is more than double the size of target patch (FPOR = 4 and TPOR = 0). (d) Transformation and target image patches match in size and location, but there are several other FP results (FPOR = 6 and TPOR = 1).

AE) and two best combinations of the proposed deep ensemble methods (ESB5 and ESB3) using IXI dataset. ESB3 means the deep ensemble method using three similarity maps as input (CNN-21, AE-ED, and MI-21), while ESB5 uses five different similarity maps (CNN-17, CNN-21, AE-ED, MI-17, and MI-21).

Bold values are the best scores concerning each evaluation metric. It is clear that ESB3 achieved the best results in terms of FP and $FPOR$ values by far. An $FPOR$ of 0.1 means that ESB3 on average detects only one false-positive region according to the criteria described in Section IV-D for every 10 patches. At the same time, it has the second-best result for $TPOR$, detecting the correct patch region in 92% of the cases. Note that ESB3 also has the highest JC and ED values, indicating that its correct patches are the best positioned on average.

On the other hand, even though AE-ED detects the majority of G with high probability, its $TPOR$ score is much lower, implying that AE-ED regions are numerous and/or large. It also detects on average other 13 false-positive regions, according to AE-ED. Surprisingly, MI-21 has better accuracy in terms of $FPOR$ than all other metrics, including all deep learning-based ones. CNN-21 is also one of the best solutions, having the highest $TPOR$, but detecting seven times more false-positive regions as compared to ESB3. CapsNet was the metric with higher FP and $FPOR$ values.

TABLE I

EVALUATION RESULTS OVER IXI DATASET FOR ALL SIMILARITY METRICS AND OUR PROPOSED DEEP ENSEMBLE. THE USE OF $TPOR$ AND $FPOR$ EVALUATION METRICS IS A CONTRIBUTION TO THIS WORK. GREATER **TP**, **JC**, **TPOR** AND LOWER **FP**, **ED**, **FPOR** ARE BETTER.

Method	TP	FP	JC	ED	TPOR	FPOR
Classic Metric Baseline						
MI-13	31,8	171,5	0,06	41,7	0,65	2,17
MI-17	33,8	91,5	0.08	27.2	0.78	1,11
MI-21	36.4	68.7	0.09	20.4	0.84	0.65
Deep Learning-based Metrics						
CapsNet-13	134.1	2955.5	0.05	50.3	0.86	31.27
CapsNet-17	131.3	2299.6	0.07	49.0	0.89	23.95
CapsNet-21	128.8	1745.8	0.08	47.1	0.89	15.96
CNN-13	130.1	536.0	0.18	35.4	0.86	4.03
CNN-17	123.9	251.2	0.22	23.8	0.93	1.79
CNN-21	103.7	96.8	0.23	14.3	**0.97**	0.71
DNN-13	147.9	2442.1	0.08	45.2	0.41	6.80
DNN-17	66,1	143,3	0,14	30,3	0,79	0,90
DNN-21	44,2	623,4	0,05	57,1	0,34	1,15
AE-ED	**165.2**	2353.1	0.09	49.9	0.63	13.39
AE-D	50.3	233.5	0.09	43.6	0.64	2.01
AE-E	52.9	274.6	0.09	45.1	0.63	2.22
Our Deep Ensemble Method						
ESB3	107.3	**36.0**	**0.28**	**11.8**	0.92	**0.10**
ESB5	94.5	85.9	0.23	23.1	0.78	0.57

Another interesting finding is related to the size of the patch, which considerably influenced the results. In our experiments, 21x21 pixel patches achieved the best results (except for DNN which performed better with a 17x17 patch). Increasing the patch size parameter probably helped CNN, CapsNet, and MI similarity metrics to improve locating the region of interest with more relevant information. Note that our conclusions differ from the ones stated in [8] for a DNN and is aligned with the works in [10], [11] which present more robust structures using 28x28 or 32x32 pixel patches. This is probably because the authors in [8] used the ranking-based metric, giving little concern to the false-positive results.

ESB3 was better than ESB5 according to Table I. ESB3 and ESB5 employ CNN, AE and MI maps as their inputs. ESB5 adds extra variations of CNN and MI with different parameters. We believe that the 2 extra similarity maps do not bring any relevant information in terms of uncertainty and

matching errors to ESB5 and makes its learning process harder and slower. It is possible, though, that in a scenario with more training images and more training epochs, the extra maps could impact the results positively. Another possibility is that more complex structures made the learning process worse [30].

B. Qualitative Analysis

Figure 5 shows some qualitative results among the best similarity metric (MI-21, CNN-21, and AE-ED) performed in this work as well as the final similarity map combined by deep ensemble method (ESB3).

We observe the evident superiority of the proposed deep ensemble (ESB3) compared to three similarity maps (MI-21, CNN-21, and AE-ED), especially in terms of FP score. The red regions are the ones with higher probability and the blue ones with the lowest to be the target region. Notice that the proposed method eliminates most of the high probability regions existing in the other similarity maps, resulting in a cleaner similarity map, fundamental for an image registration algorithm.

This metodology was not applied to other databases and the testing data was not used during any step of the training and validation processes. Using other human brain datasets for training could also help improving the results.

VI. Conclusions and Future Works

In this paper, we proposed a deep ensemble method to combine different similarity metrics in multi-modal brain image registration tasks between $T1$- and $T2$-image protocols. Our pipeline was validated on IXI dataset and surpassed all classic and deep learning-based approaches, especially concerning the low false-positive (FP) occurrences. We also proposed the use of novel evaluation metrics ($TPOR$ and $FPOR$) that are based on labeling regions of high probability into true-positive (TP) or false-positive (FP) areas, providing a much more clear understanding of the method performances.

We believe that the proposed deep ensemble method has a huge potential of being employed for the registration of multi-modal images by selecting the suitable similarity metrics, achieving more accurate similarity maps, and decrease the time-consuming during the target task.

As future works, we include: a deeper analysis of the patch selection for registration purposes; test other deep neural networks such as GANs, U-Nets, VGGs, and Inceptions architectures, among others; test the proposed pipeline over other image modalities such as computed tomographies and ultrasound; and improve the evaluation metrics by analyzing the best parameter selection; implement an automatic procedure to select the most relevant similarity maps as the ensemble input; add the deep ensemble similarity map computation into a full image registration pipeline; train the deep networks with a larger dataset; and include other brain structures to better specify the similarity regions.

Acknowledgment

The authors thank São Paulo Research Foundation (FAPESP grants #2016/21591-5 and #2018/23908-1) and NVIDIA for donating a Titan V GPU used in the experiments of this work.

References

[1] J. V. Hajnal and D. L. Hill, *Medical image registration*. CRC press, 2001.

[2] B. Glocker, A. Sotiras, N. Komodakis, and N. Paragios, "Deformable medical image registration: setting the state of the art with discrete methods," *Annual review of biomedical engineering*, vol. 13, pp. 219–244, 2011.

[3] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: A review," *arXiv preprint arXiv:1912.12318*, 2019.

[4] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[5] N. Andrade, F. A. Faria, and F. A. M. Cappabianco, "A practical review on medical image registration: From rigid to deep learning based approaches," in *31st Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2018, pp. 463–470.

[6] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu *et al.*, "Niftynet: a deep-learning platform for medical imaging," *Computer methods and programs in biomedicine*, vol. 158, pp. 113–122, 2018.

[7] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, 2020.

[8] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 248–252, 2016.

[9] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 10–18.

[10] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1505–1516, 2016.

[11] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, "Adversarial similarity network for evaluating image alignment in deep learning based registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 739–746.

[12] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N. D. Cahill, and B. Scholkopf, "Learning similarity measure for multi-modal 3d image registration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 186–193.

[13] F. Michel, M. Bronstein, A. Bronstein, and N. Paragios, "Boosted metric learning for 3d multi-modal deformable registration," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011, pp. 1209–1214.

[14] O. Rozinek and J. Mareš, "The duality of similarity and metric spaces," *Applied Sciences*, vol. 11, no. 4, p. 1910, 2021.

[15] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[16] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[17] A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 1998, pp. 1115–1124.

[18] G. Song, J. Han, Y. Zhao, Z. Wang, and H. Du, "A review on medical image registration as an optimization problem," *Current medical imaging reviews*, vol. 13, no. 3, pp. 274–283, 2017.

[19] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

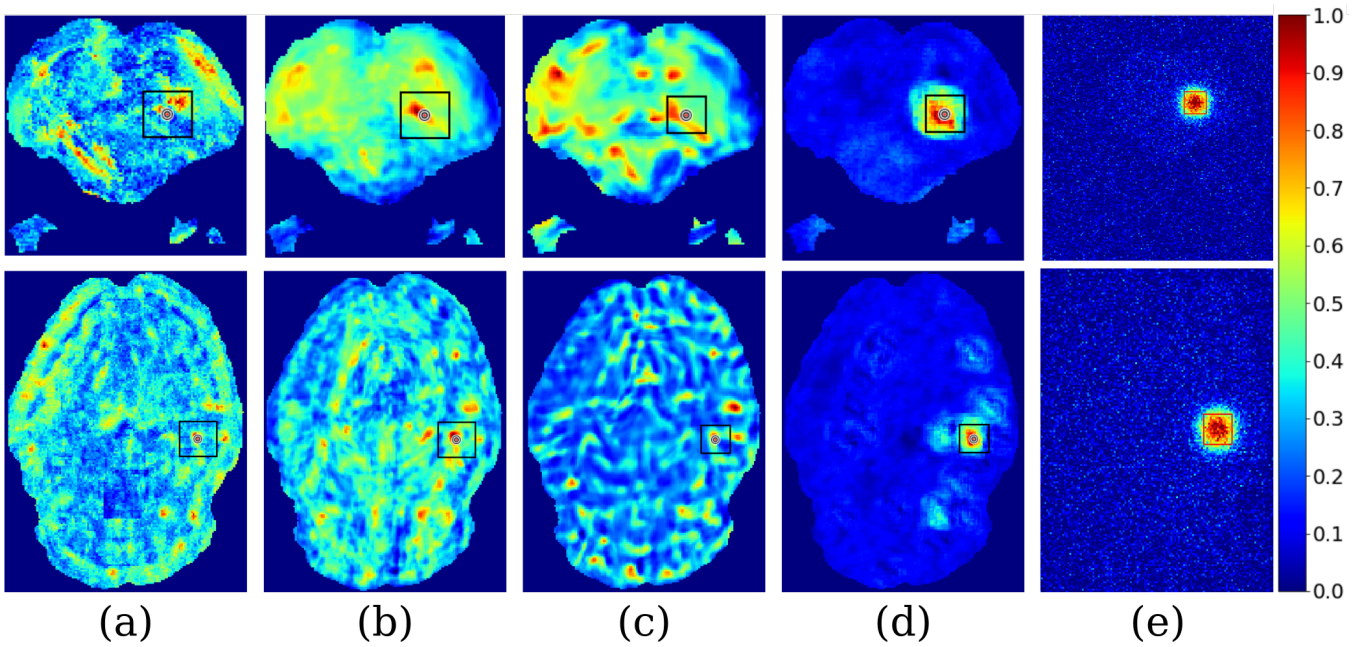

Fig. 5. From left to right are the probability maps (MI-21, CNN-21, AE-ED, and ESB3 methods). The black square and the white arrow indicate the target region.

[20] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 204–212.

[21] E. Ferrante, O. Oktay, B. Glocker, and D. H. Milone, "On the adaptability of unsupervised cnn-based deformable image registration to unseen image domains," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 294–302.

[22] A. Sedghi, J. Luo, A. Mehrtash, S. Pieper, C. M. Tempany, T. Kapur, P. Mousavi, and W. M. Wells III, "Semi-supervised deep metrics for image registration," *arXiv preprint arXiv:1804.01565*, 2018.

[23] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P. A. Pinto, B. J. Wood, and P. Yan, "Learning deep similarity metric for 3d mr–trus image registration," *International journal of computer assisted radiology and surgery*, vol. 14, no. 3, pp. 417–425, 2019.

[24] A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina *et al.*, "Common limitations of image processing metrics: A picture story," *arXiv preprint arXiv:2104.05642*, 2021.

[25] F. A. Cappabianco, P. F. Ribeiro, P. A. De Miranda, and J. K. Udupa, "A general and balanced region-based metric for evaluating medical image segmentation algorithms," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 1525–1529.

[26] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, "Explainable deep cnns for mri-based diagnosis of alzheimer's disease," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[27] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney *et al.*, "Advances in functional and structural mr image analysis and implementation as fsl," *Neuroimage*, vol. 23, pp. S208–S219, 2004.

[28] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[29] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[30] A. Jordao, F. Akio, M. Lie, and W. R. Schwartz, "Stage-wise neural architecture search," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1985–1992.