Unsupervised Non-Deformable Retina Images Registration Using Neural Network

Jérémy Baffou, Noah Kaltenrieder and Théo Damiani
EPF Lausanne, Switzerland

***Abstract*—This project has been part of the Machine Learning course at EPFL by Nicolas Flammarion and Martin Jaggi. It will focus on the broad topic of image registration. A solution will be developed using an unsupervised neural network on pairs of retina images. Allowing only rigid transformations for the main part, the network will calculate the distance between a moving image and another considered as fixed on which we want to align. The transformation matrix found by the network will be then applied to the moving image. Finally, an index will indicate the level of confidence to quantify the similarity between these two images and to know if they come from the same person.**

I. Introduction

A. Task Description

The general task of image registration consists of performing transformations on different datasets to align them into one coordinate system. It is especially useful to see the evolution over time of tissues. In the context of this project, we had retina images that we needed to register together patient wise and following a certain set of rules. We had to register images from the same eye and coming from the same centeredness. Furthermore we were asked to register images coming from different scanner, thus having different sizes and features. This is called multimodal registration as we are dealing with different image modalities [1]. We were also constrained to rigid registration, which means that no deformations of the image were allowed in order to preserve any numerical measure on the image. The pipeline that we developed could now be of use for data analysts to compute metrics such as the curvature of blood vessels and their evolution in time, which is useful to detect diseases.

B. Related Work

Image registration in the biomedical field is an old topic due to its importance in tasks such as atlas construction or medical diagnosis. Previous work was mainly based on what is called feature-based or intensity-based methods. They consist of an optimization problem where, given two images, the goal is to find a transformation between them that minimizes a cost function based on image features or pixel/voxel intensity respectively.

But in the past few years, a growing interest for deep learning based method has emerged. Especially in the field of deformable image registration, where a deformation field is learned [2]. This field gives a mapping pixel-wise from a given image, said fixed, to another one called moved. The task is becoming so common that public packages for such learning methods have appeared. We gave particular attention to one of them called Voxelmorph which, at first, seemed to fit perfectly what we needed.

C. Data

Our data consists of around 5'000 retina images (Fig. 5) captured using various sensors. The data was split in several folders for healthy patient with a single image type, patient with diseases and a single image type and finally, patients with diseases and images captured with different scanners. For every patient we had several images labeled according to some characteristics. The first information indicates which eye we are studying, i.e., the left or right one. Even though it is possible to register a left eye on a right eye, it would lead to poor registration and no scientific interest. The second attribute is based on the centeredness of the image. When the practician takes a picture, it has to center it either on the macula or the optic nerve head (called OHN). Some images in our dataset has really poor quality and thus their centeredness was classified as LQ. We did not use these images for registration as no downstream work could have been possible. Finally, the last characteristic of interest is the image type. We are in a multimodal registration tasks, thus our data is provided by different captors.

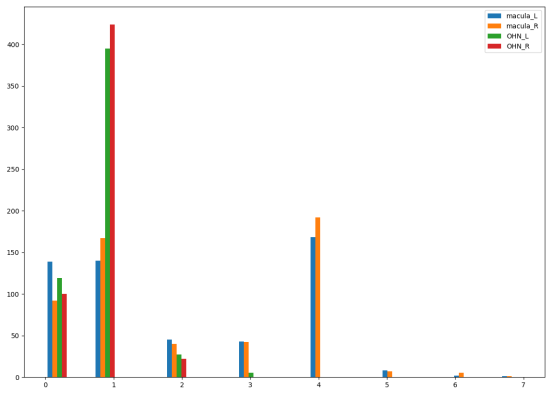

Figure 1. Repartition of the real data.

These images were provided by Mattia Tomasoni, our supervisor from the Foundation Asile des Aveugles and Jules-Gonin Eye Hospital.

The model has been trained on those images but for security and disclosure rules, when data are discussed and presented, they will come from the public STARE project [3] that regroups retina images with diagnostics, annotations, and for some images, blood vessel segmentation labels can be accessed. This is also on this data set that we have trained our prototype as the access to the data for the project is only possible through a secure connection. Thus for prototyping it was easier to work on a local dataset. The STARE dataset consists of around 400 retina images really similar to most of the images on the cluster. Some images also were of poor quality, but we decided to keep them due to the small size of the dataset.

Finally, as described later, we have trained a model to do vessel segmentation as a preprocessing step. This model has been trained on the DRIVE dataset [4]. It consists of 40 retina images with labels indicating where are located the blood vessels.

II. Models and Methods

This report focuses on learning-based registration methods. For this purpose, the Voxelmorph framework [5], [6] was first used. It aims at finding and optimizing a transformation field between two input images. One input is considered as fixed and the model will try to find the best transformation to align the other input called "moving" on the fixed one. The two inputs are passed in a Convolutional Neural Network (CNN), with a U-net structure. A loss (MSE or NCC will be discussed later in II-C) between the moving image, on which the transformation is applied, and the fixed image is then calculated and minimized. The CNN gives, therefore, the optimized deformation field and finally, it is applied to the moving input to output the registered image.

Despite the high accuracy obtained, the output image is deformed. Indeed, the deformation found by the network is pixel or voxel-wise and allows affine transformations. Thus the thickness or angle of some blood vessels is not preserved. This is the main issue because the integrity of the image has to be conserved so data analysis can still be possible after the registration.

Even if it was possible to smooth the flow of the deformation field, the framework was rapidly left over. Thereby, a deep neural network, allowing only rigid transformations, has been developed following the work of two papers [7], [8].

A. Image Preprocessing

As a first-processing step, every image is rescaled and padded with black pixels in order to obtain a square image whose side length is a power of 2. Furthermore the pixels intensities are mapped to $[0, 1]$, which gives a more convenient input to our deep network.

During the prototyping of the model, the STARE dataset was used. As it contains roughly 400 images, an augmentation step has been performed. Random rotation and translation matrices were applied to the original dataset. The tasks given were to align images for each patient, but left eyes, right eyes, and macula, optic nerve heads centered must be treated separately. This means that for some patients, only one left eye and macula centered image was available (see Fig. 5 to have the distribution of images per type per patient). As the model needs at least two images, the dataset of the hospital has been also augmented in the following way. If we had only a single image, we augmented it to have a pair to register. But to avoid bias toward only synthetic registration for some patients, we also randomly augmented some images even though there were other valid images to be aligned with.

However, the model still did not perform well enough. It treated retina images as white uniform circles with black padding and was just able to align them with this "circle". So they were nicely centered but in it, the vessels were not registered properly. It had strong difficulties for rotation.

During our model design, we have seen that the model was working really well on the MNIST dataset [9], partly because the pixel intensity is either 0 or 1, thus the losses penalizes well even small deviations compared to the homogeneous retinal background. Thus we decided to extract the blood vessels of our images, and use these binary masks as the inputs of our registration model. We thus added a segmentation network for blood vessels to the preprocessing pipeline. We relied on an existing work [10], which uses a simple supervised deep network with a U-net structure. Trained on the DRIVE dataset [4] and a small fraction of the STARE dataset [3], it was still able to perform quite reasonably. Additionally, a method provided by our supervisor has also been used as an alternative to segment the vessels. It relies on standard image processing methods but provides some reliable masks.

B. Deep Network Structure

The model, presented in Fig. 2, needs to learn 3 parameters. To construct a rigid transformation, a horizontal and a vertical translation associated with an angle for the rotation is required. The model output this three parameters dx, dy, and θ, so the transformation matrix can be computed:

$$\begin{pmatrix} cos\theta & -sin\theta & dx \\ sin\theta & cos\theta & dy \end{pmatrix}$$

.

As the model is trained on the vessel segmented images, the network needs to output these three parameters and not the registered image as it is the case in the Voxelmorph framework. So it is possible to recover the transformation matrix from the model and apply it to the original, not the segmented moving image which is the binary mask of the blood vessels. To apply transformation matrix on tensors, the work [11] which performed affine deformation, has been adapted to the rigid case.

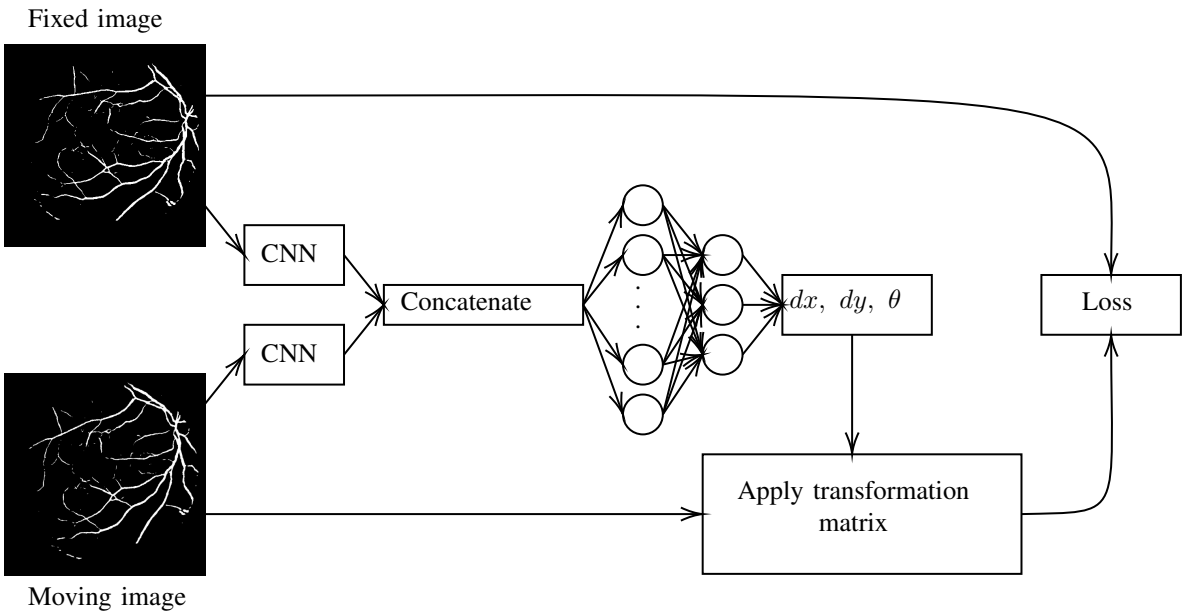

Figure 2. Structure of the Proposed Model

For the specification of the CNNs, the model has been adapted from the work of the cited article [7] Both inputs are passed first in a CNN to extract the best features of the images. Each CNN is built by two layers. A first convolutional 2D (Conv2D(64, 3) expressed in tensorflow.keras.layers function) is followed by another Conv2D(64, 3) but with skip connection on the output of the first one. The outputs of those two CNNs are concatenated, flattened, and go through one dropout layer (Dropout(rate=0.5)), then there are two dense fully connected layers with an output size of 64 (Dense(64)). Finally, the last dense layer of size 3 will render the final three parameters needed to build the transformation matrix. The activation function used at the end of each layer (except after the concatenate, dropout and flatten layer) is the Rectified Linear Unit (ReLU).

C. Loss Function

To quantify the difference between the two input images, several losses have been tested. First of all, the simple Mean Square Error was utilized. Then the regularized MSE loss of paper 2 was also tried, with I_y the fixed image, I_x the moving image, and I_{xy} the registered image:

$$Loss = L_{xy} + L_{yx} + \lambda L_{REGU}$$

$$L_{xy} = \frac{1}{N} \sum_{(i,j)\in\Theta} (I_y(i,j) - I_{xy}(i,j))^2$$

$$L_{REGU} = \frac{1}{3}(\theta^2 + dx^2 + dy^2)$$

But the best result was found using the Normalized Cross-Correlation NCC as loss:

$$NCC = \frac{1}{N} \sum_{(i,j)} \frac{I_y(i,j) - I_y(Mean)}{I_y(Std)} \cdot \frac{I_{xy}(i,j) - I_{xy}(Mean)}{I_{xy}(Std)}$$

Indeed, the NCC is more robust as it works with the cross-correlation between images and not pixel-wise like MSE. Besides, the normalization between the range $[-1; 1]$ helps to output a meaningful index of confidence between the two inputs after the registration.

Also, an interesting loss to test would be the MSE regularized but not on all the pixels of the image but only for the pixel which constitutes the vessels in the segmentation. The loss would be more robust to the black padding added and more accurate as it would not be taken into account.

III. Results

In this section, some results are presented.

Figure 3. Superposition of the fixed image in white and the moving one in blue

Figure 4. Superposition of the same fixed image and with the registered image in red

Figure 5. Template from left to right: Registered, Fixed, Moving

IV. Discussion

The main part of our work has been developed on the STARE data set because the cluster where the images were located had no GPU support, making the training really cumbersome. Furthermore, for such a short-term project, the SIB (Swiss Institute of Bioinformatics) did not give us strong access to the cluster. We had to interact with it through a web interface which was unstable and slow. Nevertheless, we still were able to per a training on the hospital data set and show that our work generalizes well to their images.

Some improvement could be made on the data augmentation. We have only performed rigid transformations, but we could have added some noise to avoid the possibility of a perfect superposition that is unrealistic in most applications. The problem is that we are acting on binary masks, thus it was hard to have a realistic noise which would have added and deleted some vessels in a meaning full way, not purely random.

The work in [1] seemed to be the more promising, as it was the closest one we found on a complete work on multimodal images. But the architecture proposed did not work in our case, maybe because the network is too deep. Unfortunately, we do not have the expertise to properly design or modify cleverly such a complex deep network architecture. But the pipeline described in [1] is probably the way to go. We could have a classifier at the beginning, which would detect the image modality, then pass it as input to the regression part of the deep net. It would allow having slight changes based on the modality. In our case the classifier could even based on the file name as it already gives the image modality.

For multimodal registration, there is an issue as the images do not have the same scale. Some are focused on a tiny portion of the eye and others are more global. We thought about changing slightly the architecture of the model to add another output. This new parameter could be used to rescale the moving image, which would have to be the smallest. Then the network would have to learn how to rescale it so it matches the scale of the static image, and then learn the rigid transform. In practice it would imply to regress four outputs instead of three. And add to the actual transformation matrix M, a rescaled identity matrix : λ*I, where λ is the new output.

Finally, due to cluster and confidentiality issues, we are unable to show results on the real data set. And the results obtained on STARE are encouraging for the next steps but as most of the images are patients with severe diseases and often of low quality, it is hard to have reliable results on this data set.

Acknowledgements

We would like to thank specially Mattia Tomasoni for his advices and help and his availability during the project. We also wanted to thank Sarath Chandra (sarathchandra.knv31@gmail.com), the author of the affine-2d notebook, that gives us the authorization to use his code.

References

[1] K. T. Islam, S. Wijewickrema, and S. O'Leary, "A deep learning based framework for the registration of three dimensional multi-modal medical images of the head," *Scientific Reports*, vol. 11, no. 1, Jan. 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-81044-7

[2] A. V. Dalca, M. Rakic, J. Guttag, and M. R. Sabuncu, "Learning conditional deformable templates with convolutional networks," *NeurIPS: Neural Information Processing Systems*, 2019.

[3] M. H. Goldbaum, "Structured analysis of the retina," U.S. National Institutes of Health, 1975, access to 400 raw images and vessel segmentation labels. [Online]. Available: https://cecas.clemson.edu/~ahoover/stare/

[4] "Drive: Digital retinal images for vessel extraction." [Online]. Available: https://drive.grand-challenge.org/

[5] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE TMI: Transactions on Medical Imaging*, vol. 38, pp. 1788–1800, 2019.

[6] ——, "An unsupervised learning model for deformable medical image registration," *CVPR: Computer Vision and Pattern Recognition*, pp. 9252–9260, 2018.

[7] J. M. Sloan., K. A. Goatman., and J. P. Siebert., "Learning rigid image registration - utilizing convolutional neural networks for medical image registration," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING,*, INSTICC. SciTePress, 2018, pp. 89–99.

[8] H. Liu, Y. Chi, J. Mao, X. Wu, Z. Liu, Y. Xu, G. Xu, and W. Huang, "End to end unsupervised rigid medical image registration by using convolutional neural networks," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2021, pp. 4064–4067.

[9] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[10] N. Thomar, "Unet segmentation in keras tensorflow," https://github.com/nikhilroxtomar/UNet-Segmentation-in-Keras-TensorFlow, 2020.

[11] S. Chandra, "Deep-learning-based 2d and 3d affine registration," 2020. [Online]. Available: https://medium.com/@sarathchandra.knv31/deep-learning-based-2d-and-3d-affine-registration-da73df8d2f24