UNSUPERVISED DEEP LEARNING NETWORK FOR DEFORMABLE FUNDUS IMAGE REGISTRATION

Giovana Augusta Benvenuto [1] Marilaine Colnago [2] Wallace Casaca [2]

[1]São Paulo State University, Faculty of Science and Technology, Presidente Prudente, Brazil
[2]São Paulo State University, Department of Energy Engineering, Rosana, Brazil

INTRODUCTION AND CONTRIBUTIONS

The process of registering a pair of fundus images, captured at different scales and viewing angles, is of paramount importance to support the diagnosis of diseases and routine eye examinations.

In this paper, we propose an end-to-end unsupervised learning registration framework that unifies Convolutional Neural Network (CNN) and Spatial Transformer Network (STN).

The proposed technique takes advantage of a similarity metric that gauges the difference between the fixed and transformed images, allowing for performing the registration task without any ground-truth data.

In summary, the main contributions of this paper are:

- A fully end-to-end framework for performing retina image registration using Deep Learning techniques.
- A neural network architecture that learns the registration task without using any ground-truth data or artificially created benchmark features.
- A functional and effective registration method capable of operating with distinct classes of *fundus* image pairs.
- Once our network is fully trained, it can achieve one-shot registrations by just providing the desired pair of *fundus* images

THE PROPOSED FRAMEWORK

The proposed framework combines the neural network architecture U-Net [10] with the learning scheme recently proposed by Vos et al [2], where a Convolutional Neural Network estimates a set of matching points from the images, used by a Spatial Transformer Networks [6] to generate a deformation field which leads to the definitive bilinear interpolation.

First, the target *fundus* images go through a segmentation step that captures and highlights their main structures, such as blood vessels and ocular shape. Such a task is performed by applying the so-called *Isotropic Undecimated Wavelet Transform* (IUWT) [1], which in essence computes and takes the transformation coefficients of the images to binarize I_{Ref} and I_{Mov}.

Our deep learning pipeline relies on the U-Net architecture and the steps as follows:

1. The network gets as input the pair of concatenated segmentations, passing them to a block of convolutional layers.
2. Two *downsample* blocks, composed of two convolutional layers and a *max pooling* layer, resample the images by halving their resolution while increasing the number of analyzed features per block. The subsequent block of layers (*upsample* process) is formed by a deconvolution layer and two convolutional layers. Each convolutional layer is followed by the ReLU activation function, and a *Batch Normalization* scheme.
3. The outputs of each level from the *downsample* block are concatenated with the entry of the corresponding level in the *upsample* block.
4. The last layer, which is formed by two kernels, applies a linear activation function so as to generate the grid of points corresponding to the dimensions of the input images.

The output generated by the CNN, which gives a deformation grid, serves as input to the STN so that a bilinear interpolation is computed for aligning the images.

Once the image pair is properly registered, the loss function is then calculated via the Normalized Cross-Correlation (NCC) metric (1), which gauges the overlap among both fixed and processed images without the need for ground-truth data:

$$NCC(x,y)=\frac{\sum_{i=0}^{m}\sum_{j=0}^{n}T_{i,j}R_{i,j}}{\sqrt{\left(\sum_{i=0}^{m}\sum_{j=0}^{n}T_{i,j}^{2}\right)\left(\sum_{i=0}^{m}\sum_{j=0}^{n}R_{i,j}^{2}\right)}}, \tag{1}$$

where $T_{i,j} = t(x+i, y+j) - \bar{t}_{x,y}$, $R_{i,j} = r(i,j) - \bar{r}$, $t(i,j)$ and $r(i,j)$ are the pixel values at (i,j) of the matching and reference images, B_{Trans} and B_{Ref}, respectively, and $\bar{r}$ and $\bar{t}$ are the average pixel values w.r.t. B_{Ref}, and B_{Trans} [8].

The learning pipeline is optimized until convergence by the ADAM algorithm [7].

The process of applying the deformable transformation to the moving image may eventually cause the presence of noise, especially if the images are very distinct from each other. To circumvent this, we add a post-processing step to filter out the noise via Connected Component Analysis (CCA) [3].

DATA SETS, METRICS AND ASSESSED METHODS

Our framework was implemented in Python language (packages *OpenCV*, *Tensorflow*, and *Keras*).

To run the experiments, we took the well-established FIRE database (*Fundus Image Registration Dataset*) [5].

Our network architecture was trained using images of 512×512 pixels, from category S of FIRE database, with eight batches, for 5000 epochs. The tests were accomplished for all the three FIRE image categories in order to inspect how the network would behave when applied to different groups of *fundus* images.

Four very recent image registration methods were taken in our validation analysis: Hernandez-Matas et al. [4], Wang et al. [11], Motta et al [9] and Vos et al. [2], namely here as Rempe, GFEMR, VOTUS, and DIRNet, respectively.

To quantitatively assess the registration results, the following similarity metrics were taken: *Mean Squared Error* (MSE), *Structural Similarity Index Measure* (SSIM), and *Dice Coefficient* (Dice).

RESULTS AND DISCUSSION

QUANTITATIVE EVALUATION

Table 1 summarizes the mean and standard deviation for the registration results produced by each method for the three categories of FIRE database.

By numerically checking the values, our approach was the one that delivered the best scores for all metrics and analyzed data sets.

A particular advantage of our approach is that since other methods were not able to fully register a few image pairs from category P, we pull off these particular cases of failures and compute scores for success cases only. In contrast, our framework was capable to align the image pairs regardless of the category and overlap level.

Table 1. Quantitative analysis of the registration methods.

	Methods	FIRE Dataset		
		A	S	P
MSE (←)	Before	0.0962 (0.0177)	0.0965 (0.0198)	0.1249 (0.0066)
	Proposed	**0.0068 (0.0015)**	**0.0062 (0.0017)**	**0.0121 (0.0027)**
	GFEMR	0.0522 (0.0145)	0.0280 (0.0053)	0.0525 (0.0095)
	Rempe	0.0487 (0.0240)	0.0196 (0.0056)	0.0616 (0.0132)
	VOTUS	0.0525 (0.0229)	0.0189 (0.0052)	0.0514 (0.0119)
	DIRNet	0.0710 (0.0182)	0.0601 (0.0237)	0.1040 (0.0070)
SSIM (→)	Before	0.7307 (0.0421)	0.7237 (0.0457)	0.6510 (0.0177)
	Proposed	**0.9731 (0.0055)**	**0.9749 (0.0068)**	**0.9575 (0.0076)**
	GFEMR	0.8325 (0.0350)	0.8918 (0.0168)	0.8247 (0.0262)
	Rempe	0.8453 (0.0650)	0.9211 (0.0184)	0.8014 (0.0386)
	VOTUS	0.8317 (0.0562)	0.9232 (0.0180)	0.8279 (0.0340)
	DIRNet	0.7852 (0.0459)	0.8099 (0.0611)	0.6816 (0.0173)
Dice (→)	Before	0.2982 (0.1088)	0.3418 (0.1384)	0.1245 (0.0119)
	Proposed	**0.9502 (0.0100)**	**0.9579 (0.0120)**	**0.9103 (0.0238)**
	GFEMR	0.6023 (0.1343)	0.8022 (0.0392)	0.5919 (0.0922)
	Rempe	0.6295 (0.1981)	0.8649 (0.0425)	0.5227 (0.1231)
	VOTUS	0.6105 (0.1802)	0.8702 (0.0388)	0.6149 (0.1004)
	DIRNet	0.4982 (0.1111)	0.6020 (0.1519)	0.2630 (0.0197)

QUALITATIVE EVALUATION

In Figure 1, the aligned images were grouped based on different colorizations. Here, B_{Ref} brings the reference image (in green), while B_{Mov} and B_{Trans} present the moving image before and after the registration (in magenta). The definitive image composition gives the amount of overlap between B_{Ref} and B_{Trans} (in white).

One can observe that our trained model achieves more consistent and pleasant results when compared against other methods, mainly w.r.t. the quality of matching refinement, as depicted by a large amount of white color in the montages.

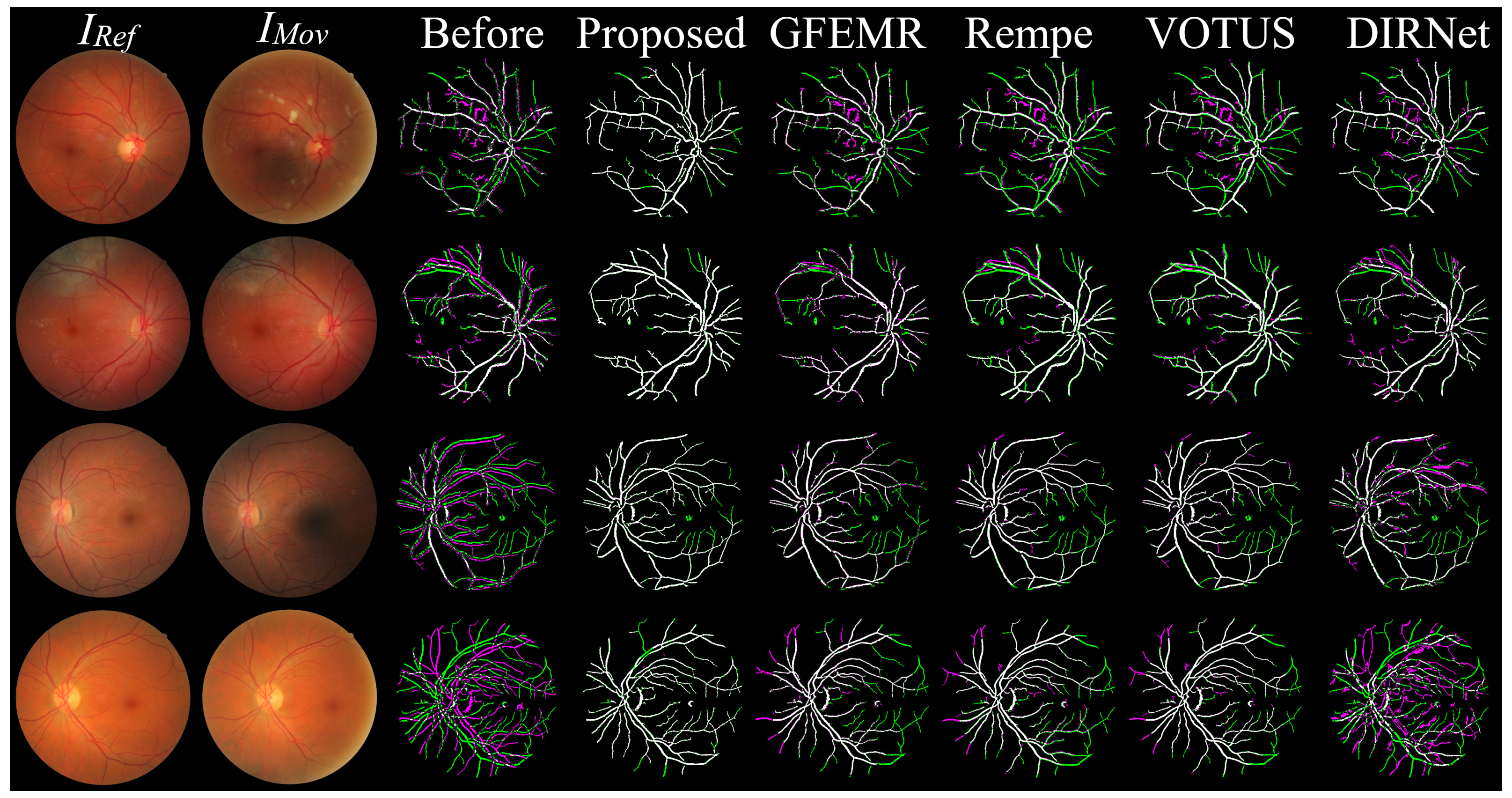

Figure 1. Qualitative comparison between registration results for a pair of images after registration by all the methods.

CONCLUSIONS

This paper proposed an end-to-end framework for deformable registration of retinal images, which was designed to operate in an unsupervised manner so that it does not require any extra mechanism to induce artificially created ground-truth data for training. Once the model is trained, it allows for one-shot registrations by just providing the pair of *fundus* images.

In contrast to other modern image registration methods, our approach produced definitive registrations regardless of the overlap degree and anatomical changes present in the images. As verified by the experiments with three distinct classes of retina images, our framework was able to outperform the others, both in qualitative and quantitative aspects.

In summary, all those properties render the proposed framework a useful and compelling unsupervised registration technique for *fundus* images, achieving a high level of accuracy even in the absence of ground-truth data or large labeled data sets to train a definitive model.

ACKNOWLEDGMENTS

This research was supported by São Paulo Research Foundation (FAPESP), grants #2013/07375-0, #2019/26288-7 and #2021/03328-3, and the National Council for scientific and Technological Development (CNPq).

References

[1] P. Bankhead, C. N. Scholfield, J. G. McGeown, and T. M. Curtis. Fast retinal vessel detection and measurement using wavelets and edge location refinement. *PloS One*, 7(3):e32435, 2012.

[2] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, 2019.

[3] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, and Y. Chao. The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognition*, 70:25–43, 2017.

[4] C. Hernandez-Matas, X. Zabulis, and A.A. Argyros. Rempe: Registration of retinal images through eye modelling and pose estimation. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3362–3373, 2020.

[5] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A.A. Argyros. Fire: Fundus image registration dataset. *Journal for Modeling in Ophthalmology*, 1(4):16–28, 2017. source code: http://www.ics.forth.gr/cvrl/fire/.

[6] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015.

[7] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[8] J. P. Lewis. Fast normalized cross-correlation. *Industrial Light & Magic*, 10, 10 2001.

[9] D. Motta, W. Casaca, and A. Paiva. Vessel optimal transport for automated alignment of retinal fundus images. *IEEE Transactions on Image Processing*, 28(12):6154–6168, 2019.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, page arXiv:1505.04597, May 2015.

[11] J. Wang, J. Chen, H. Xu, S. Zhang, X. Mei, J. Huang, and J. Ma. Gaussian field estimator with manifold regularization for retinal image registration. *Signal Processing*, 157:225–235, 2019.