Keypoint matching networks for longitudinal fundus image affine registration

Peng, Yunzhen, Chen, Xinjian, Xiang, Dehui, Luo, Gaohui, Cai, Mulin

Yunzhen Peng, Xinjian Chen, Dehui Xiang, Gaohui Luo, Mulin Cai, "Keypoint matching networks for longitudinal fundus image affine registration," Proc. SPIE 11596, Medical Imaging 2021: Image Processing, 115963D (15 February 2021); doi: 10.1117/12.2581041

Event: SPIE Medical Imaging, 2021, Online Only

Keypoint Matching Networks for Longitudinal Fundus Image Affine Registration

Yunzhen Peng[a], Xinjian Chen[ab]， Dehui Xiang[*a], Gaohui Luo[a], Mulin Cai[a]
[a]MIPAV Lab, School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu Province, 215006, China; [b]State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, Jiangsu Province, 215006, China

ABSTRACT

Registration of retinal images is an important technique for facilitating the diagnosis and treatment of many eye diseases. Recent studies have shown that deep learning methods can be used for image registration, which is usually faster than conventional registration methods. However, it is not trivial to obtain ground truth for supervised methods and popular unsupervised methods perform not well for retinal images. Therefore, we present a weakly-supervised learning method for affine registration of fundus image. The framework consists of multiple steps, rigid registration, overlap calculation and affine registration. In addition, we introduce a keypoint matching loss to replace common similarity metrics loss used in unsupervised methods. On a fundus image dataset related to multiple eye diseases, our framework can achieve more accurate registration results than that of state-of-the-art deep learning approaches.

Keywords: affine registration, color fundus image, keypoint matching, longitudinal analysis

1. INTRODUCTION

Color fundus photography is a non-invasive imaging that allows for detection of various eye diseases. In order to accurately compare the evolution of these diseases over time, color fundus images collected from the same patient at different time must be perfectly superimposed. This is what we called "registration", a fundamental preprocessing step for longitudinal analysis. A registration illustration is shown in Fig.1.

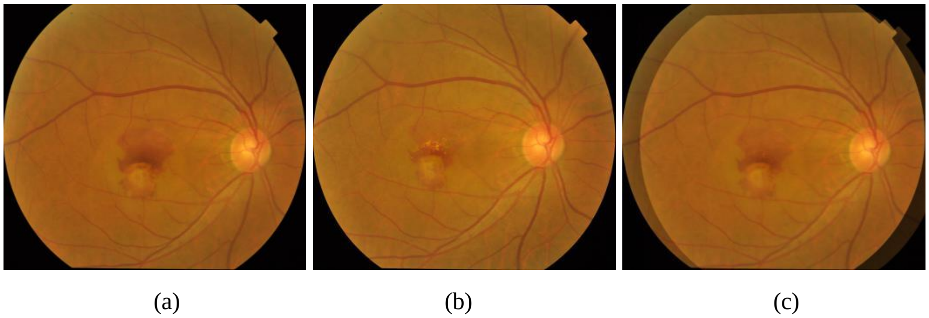

Fig.1. Illustration of fundus image registration. (a) Fixed image. (b) Moving image. (c) Mosaic image.

Traditional retinal image registration can be broadly divided into two groups: feature based and image intensity based. Feature based methods often extract a set of key points from a pair of images, and then match these key points to find the transformation parameters. The key points are usually vascular bifurcation points, cross-over points, etc. Intensity based methods usually find the optimum transformation model through optimizing a certain similarity function. Although image intensity-based registration is widely used in medical image registration, it does not perform very well for retinal images[1].

Recently, many deep learning-based methods have been applied in medical image registration. These methods can be divided into 2 categories: supervised learning methods and unsupervised learning methods. Supervised learning methods

Medical Imaging 2021: Image Processing, edited by Ivana Išgum,
Bennett A. Landman, Proc. of SPIE Vol. 11596, 115963D · © 2021
SPIE · CCC code: 1605-7422/21/$21 · doi: 10.1117/12.2581041

usually require ground truth of registration fields, and are often attained by traditional registration methods[2]. Generating ground truth is time-consuming and tedious. Some papers explore unsupervised strategies that build on the spatial transformer network[3]. Unsupervised methods usually rely on intensity-based similarity measures such as MSE or LNCC[4]. However, there is no robust image similarity measure. Therefore, a weakly-supervised registration framework[5] has been proposed. During training, a convolutional neural network is optimized by output transformation parameters that warp a set of available anatomical labels from the moving image to match their corresponding counterparts in the fixed image. During inference, only unlabeled image pairs are used as the network input.

For retinal image registration, since it's hard to find a proper intensity-based similarity measure, unsupervised learning methods are not suitable. While supervised learning methods need ground truth made by conventional registration methods, which is very time-consuming and unpractical Therefore, the weakly-supervised method maybe is a good choice. The original weakly-supervised registration framework aligns multiple labelled corresponding structures to train the network. However, for the fundus images with pathological changes, it is not easy to segment the accurate blood vessels, especially when the patient is seriously ill.

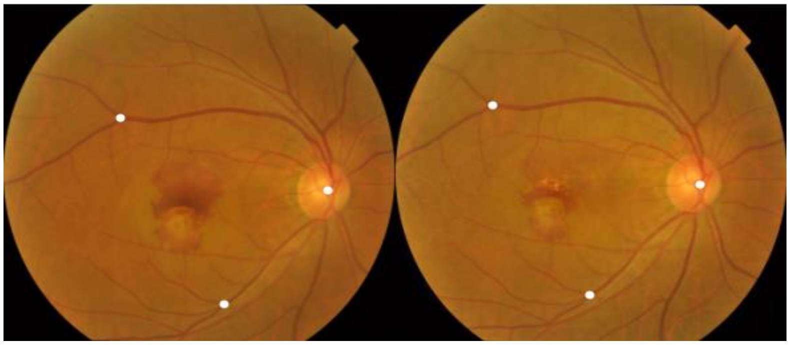

Fig.2. Illustration of key points in a pair of images, where the white dots indicate key points.

In this paper, we propose a key point matching network for fundus image registration, combining the traditional fundus image registration method with the weakly-supervised registration method based on deep learning. Different from weakly-supervised registration method, we replace the labelled segmentation masks with blood vessel bifurcation points, key points namely. In other words, input labels are 12 coordinate values from 3 pair of matched key points rather than binary images. We design the keypoint transform layer to calculate the corresponding key points coordinates in the distorted image, and we take the RMSE between the distorted coordinates and the fixed coordinates as our objective function. In addition, many studies show that a single-step network is hard to registration images with large transformation, and so we propose a two-step affine registration framework for fundus image longitudinal registration.

In summary, our paper makes the following contributions:

- A weakly-supervised registration framework based on key points is proposed. Inspired by conventional fundus image registration methods, labeled segmentation masks in weakly-supervised registration framework are replaced by key points coordinates. RMSE between key points coordinates take the place of the widely used similarity metric based on image density, which is likely to fall into local optima and does not perform very well for retinal images.
- Two-step affine registration framework is proposed for registration of color fundus images related to four eye diseases. To the best of our knowledge, this is the first universal deep learning method for fundus images affine registration related to multiple diseases.
- Our framework can achieve more accurate registration results than that of state-of-the-art approaches on a dataset related to multiple diseases.

2. METHODOLOGY

2.1 Weakly-supervised registration framework based on keypoints

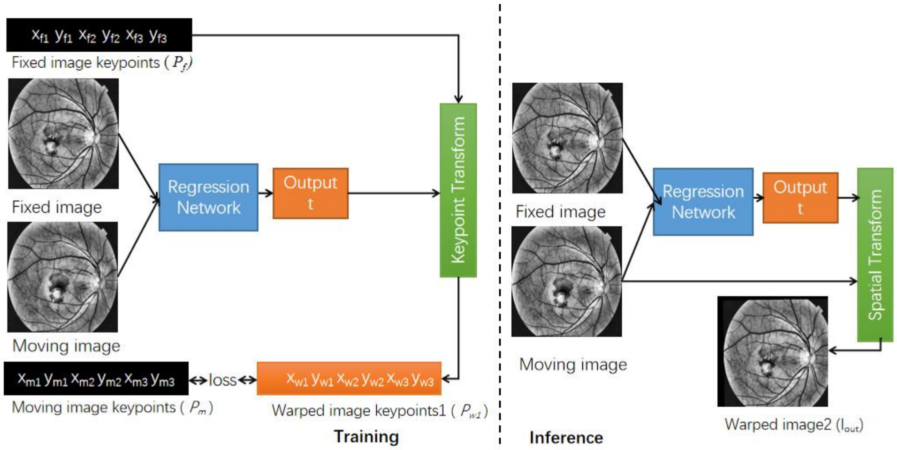
$x_{f1}\ y_{f1}\ x_{f2}\ y_{f2}\ x_{f3}\ y_{f3}$
Fixed image keypoints (P_f)
Fixed image
Moving image
Regression Network
Output t
Keypoint Transform
$x_{m1}\ y_{m1}\ x_{m2}\ y_{m2}\ x_{m3}\ y_{m3}$
loss
$x_{w1}\ y_{w1}\ x_{w2}\ y_{w2}\ x_{w3}\ y_{w3}$
Moving image keypoints (P_m)
Warped image keypoints1 (P_{w1})
Training
Fixed image
Moving image
Regression Network
Output t
Spatial Transform
Warped image2 (I_{out})
Inference

Fig.3. The left part illustrates the training strategy of the proposed registration framework, where the keypoints coordinates are only required in training. The right part illustrates the inference, requiring only the image pair.

Inspired by weakly-supervised registration framework[5], we design the framework as Fig.3 shows. The proposed convolutional neural network aims to predict transformation parameters (i.e. six parameters for affine registration) to align three pairs of matched key points for input image pairs during the training, while the coordinates are not required for inference.

Considering that an affine transform matrix can be concluded from three pairs of matched points, we select three pairs of keypoints as our labels during training. However, it doesn't mean that three pair of key points picked up randomly can work. We choose those key points distributed evenly, and then we will evaluate these key points by openCV tool. Only if these key points could register corresponding images successfully by openCV can we pick them as our labels.

During inference, to perform a spatial transformation of the moving image I_m, the spatial transform layer is adopted. The spatial transform layer is firstly used to calculate a set of sampling points, and then the warped image I_w is generated by image sampling. In the affine case, coordinates of the sample points can be formulated as

$$\begin{pmatrix} x_i^m \\ y_i^m \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \end{bmatrix} \begin{pmatrix} x_i^f \\ y_i^f \\ 1 \end{pmatrix} \tag{1}$$

where (x^f_i, y^f_i) is the fixed coordinate of the regular grid in the warped image, (x^m_i, y^m_i) is the moving coordinate in the moving image that define the sample point, and the matrix is the affine transformation matrix.

Following this work, during training, we assume that if all sample points are just the matched moving coordinates for the fixed coordinates, the registration perform perfectly. We design the keypoint transform layer to implement the formulation (1) and then to calculate sample points. In order to make the corresponding three sample points as close to matched moving points as possible, the loss is defined as Root Mean Square Error (RMSE),

$$loss = \sqrt{\frac{\sum_{i=1}^{3}((x_{mi} - x_{wi})^2 + (y_{mi} - y_{wi})^2}{3}} \tag{2}$$

where (x_{wi}, y_{wi}) corresponds to the sample point mentioned above.

As shown in Fig.4, the regression network is inspired by the globle-net[5]. The input of the network is a concatenated image pair. First, a feature extractor adapted from Resnet18 generates feature maps with 512 channels. Second, global average pooling is applied to generate a feature vector. Finally, we pass the feature vector into two fully connected layers to regress transform parameter.

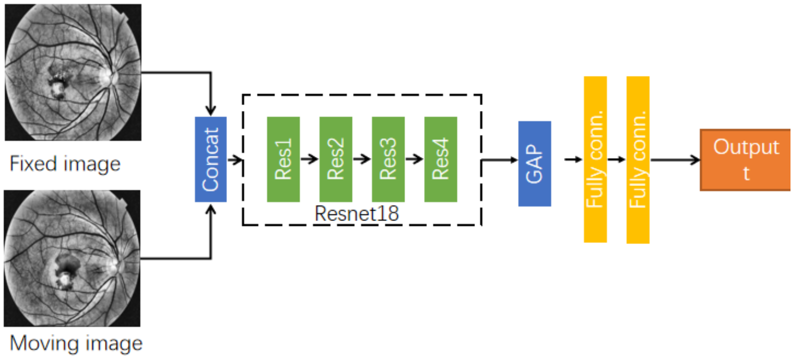

Fig.4. Architecture of the regression network

2.2 Two-step affine registration

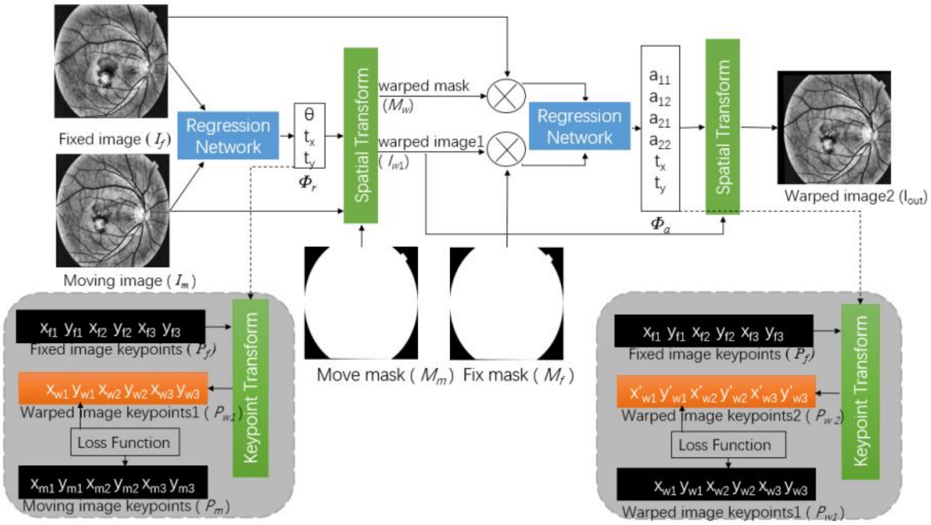
Fixed image (I_f)
Moving image (I_m)
Regression Network
θ
t_x
t_y
Φ_r
Spatial Transform
warped mask
(M_w)
warped image1
(I_{w1})
Regression Network
a_{11}
a_{12}
a_{21}
a_{22}
t_x
t_y
Φ_a
Spatial Transform
Warped image2 (I_{out})
Move mask (M_m)
Fix mask (M_f)
x_{f1} y_{f1} x_{f2} y_{f2} x_{f3} y_{f3}
Fixed image keypoints (P_f)
x_{w1} y_{w1} x_{w2} y_{w2} x_{w3} y_{w3}
Warped image keypoints1 (P_{w1})
Loss Function
x_{m1} y_{m1} x_{m2} y_{m2} x_{m3} y_{m3}
Moving image keypoints (P_m)
Keypoint Transform
x_{f1} y_{f1} x_{f2} y_{f2} x_{f3} y_{f3}
Fixed image keypoints (P_f)
x'_{w1} y'_{w1} x'_{w2} y'_{w2} x'_{w3} y'_{w3}
Warped image keypoints2 (P_{w2})
Loss Function
x_{w1} y_{w1} x_{w2} y_{w2} x_{w3} y_{w3}
Warped image keypoints1 (P_{w1})
Keypoint Transform

Fig.5. Illustration of the training process of the proposed Two-step affine registration framework, where the gray-dotted lines represent data flows only required in training.

The overview of our proposed registration framework is shown in Fig.5. Following conventional hierarchical multi-stage strategy, we divide the registration into two steps: rigid registration and affine registration. In the first step, we take a pair of preprocessed images of size 512×512 as inputs. Let I_f, I_m be these two images. We model a function $f_r(I_f, I_m) = \Phi_r$ using a regression network, Φ_r represents three output parameters: 1 rotation angle and 2 displacement values. The spatial transform layer warps the moving image to the warped image. Let $I_m \circ \Phi_r$ represents I_m warped by Φ_r. Since the warped image and the fixed image are not fully overlapped, which may cause hard divergence when affine network is training. Therefore, in the second step, we calculate the overlap by multiplying the image and mask of the other image, and then send the 2 overlap images to the second regression network to estimate 6 affine transform parameters Φ_a. This can be written as

$$\Phi_a = f_a(I_f \otimes M_w, (I_m \circ \Phi_r) \otimes M_f) \tag{3}$$

where the fixed image’s mask is denoted as M_f, the warped image mask from the first step is denoted as M_w, and $\otimes$ means the pixel multiplication. Eventually, the wholly inference process can be written as

$$I_{out} = I_m \circ \Phi_r \circ \Phi_a \tag{4}$$

3. DATA

The proposed method is evaluated on 492 pair of fundus images obtained from patients suffer from at least one of diseases including diabetic retinopathy (DR), age-related macular degeneration (AMD), central retinal vein occlusion (CRVO) and branch retinal vein occlusion (BRVO). All the original image size is 3046×2572, and we resize them to 512×512. In our experiments, we divide 492 images into 381 training images, 61 validation images and 50 test images.

We emphasize that there is no relationship between the key points during training and that during testing. During training, we labeled three pair of key points, and the network aims to predict transformation parameters that can align those key points. During test, we label ten pairs of key points as ground truth to calculate the measure metric.

4. RESULTS

We implement the registration framework with Keras. For data augmentation, we use ShiftScaleRotate with the package albumentations. In the first step, we set the parameter shift_limit as 0.3, which means the range of shift is [-512*0.3, 512*0.3]. In the second step, we set the parameter shift_limit as 0.03. We also preprocess the original image. We extract the green channel of color fundus image, and resize it into 512*512. For the gray image, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) to augment image. We train the network using RAdam optimizer with batch size 16 and initial learning rate 1e-4.

The registration results are compared to the ground truth according to 2 metrics: Area Under Curve (AUC)[6] and average RMSE. Some examples of registration results are shown in Fig.4. Considering that there are few deep learning methods for fundus image registration. We compare our methods with two other deep learning methods for medical image affine registration[7,8]. In addition, we do the ablation study to demonstrate the two-step affine registration framework.

As Table.1 shows, the weakly-supervised registration framework based on key point can improve the registration accuracy greatly compared with unsupervised methods based on intensity similarity. A single-step affine registration network can’t perform well on images with small overlap, and a two-step affine registration framework can implement from-coarse-to-fine registration.

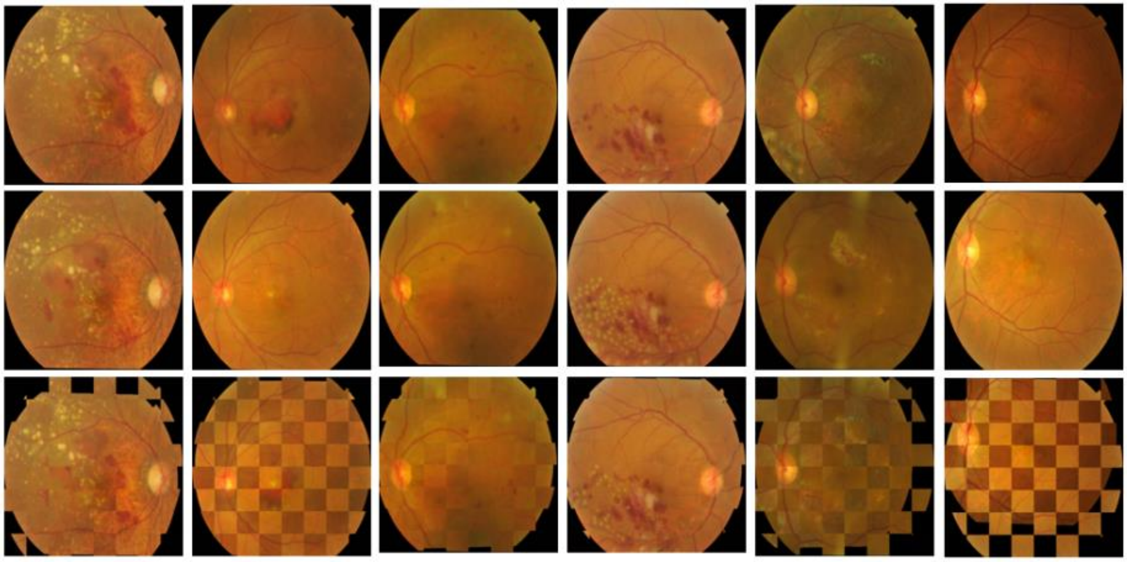

Fig.6. Examples of registration results. The first row shows the fixed image; the second row shows the moving image; the last row shows checkerboard generated by warped image and fixed image.

Table.1. Comparison of registration results for different methods. The first four rows refer to the comparison of loss with other two methods (loss); the last three rows refer to the ablation study of our proposed registration framework. The loss we proposed is used as default.

Methods	AUC	RMSE
RigidNetV1 (LNCC)	0.22	39.2±25.4
RigidNetV2 (MSE)	0.26	37.8±26.7
RigidNetV1	0.69	11.12±8.55
RigidNetV2	0.78	8.03±5.72
AffineNetV2	0.71	10.56±6.93
RigidNetV2+AffineNetV2	0.88	4.32±2.66
RigidNetV2+Overlap+AffineNetV2	0.89	**3.88±1.78**

5. CONCLUSIONS

In this paper, we propose an affine registration framework for fundus images. The framework consists of multiple steps. In order to solve the problem that unsupervised learning methods are not applicable to our dataset, we introduce the keypoint matching loss to replace common similarity metrics, inspired by weakly-supervised registration framework. Our framework can achieve more accurate registration results than that of state-of-the-art approaches.

REFERENCES

[1] Sajib Kumar Saha, Di Xiao, Alauddin Bhuiyan, Tien Y. Wong and Yogesan Kanagasingam, "Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: A review," Biomedical Signal Processing and Control, 2019, 47(JAN.):288-302.

[2] E Chee, Wu and Zhenzhou, "AIRNet: Self-Supervised Affine Registration for 3D Medical Images using Neural Networks," arXiv Prepr, arXiv:1810.02583v2

[3] Jaderberg, Max, Karen Simonyan and Andrew Zisserman, "Spatial transformer networks," in Advances in neural information processing systems, 2015, pp. 2017-2025.

[4] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. and Dalca, A. V., "VoxelMorph: A Learning Framework for Deformable Medical Image Registration," IEEE Transactions on Medical Imaging, 2019:1788-1800.

[5] Hu, Y., Modat, M., Gibson, E., Ghavami, N. and Bonmati, E., "Label-driven weakly-supervised learning for multimodal deformable image registration," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI), 2018, pp. 1070-1074.

[6] Hernandez-Matas, Carlos, Xenophon Zabulis, and Antonis A. Argyros, "An experimental evaluation of the accuracy of keypoints-based retinal image registration," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 377-381.

[7] S. Christodoulidis, M. Sahasrabudhe, M. Vakalopoulou, G. Chassagnon, M.-P. Revel, S. Mougiakakou, and N. Paragios, "Linear and deformable image registration with 3d convolutional neural networks," Image Analysis for Moving Organ, Breast, and Thoracic Images, vol. 11040, pp. 13-22, 2018

[8] de Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M. and Išgum, I，"A deep learning framework for unsupervised affine and deformable image registration," Medical image analysis, vol. 52, pp. 128-143, 2019.