A Deep Step Pattern Representation for Multimodal Retinal Image Registration

Jimmy Addison Lee[1*], Peng Liu[2*], Jun Cheng[1,3†], Huazhu Fu[4]
[1] Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, China
[2] Big Data Research Center at University of Electronic Science and Technology of China, China
[3] UBTech Research, China
[4] Inception Institute of Artificial Intelligence, UAE

Abstract

This paper presents a novel feature-based method that is built upon a convolutional neural network (CNN) to learn the deep representation for multimodal retinal image registration. We coined the algorithm deep step patterns, in short DeepSPa. Most existing deep learning based methods require a set of manually labeled training data with known corresponding spatial transformations, which limits the size of training datasets. By contrast, our method is fully automatic and scale well to different image modalities with no human intervention. We generate feature classes from simple step patterns within patches of connecting edges formed by vascular junctions in multiple retinal imaging modalities. We leverage CNN to learn and optimize the input patches to be used for image registration. Spatial transformations are estimated based on the output possibility of the fully connected layer of CNN for a pair of images. One of the key advantages of the proposed algorithm is its robustness to non-linear intensity changes, which widely exist on retinal images due to the difference of acquisition modalities. We validate our algorithm on extensive challenging datasets comprising poor quality multimodal retinal images which are adversely affected by pathologies (diseases), speckle noise and low resolutions. The experimental results demonstrate the robustness and accuracy over state-of-the-art multimodal image registration algorithms.

1. Introduction

Image registration has been an important element in the fields of computer vision, pattern recognition, and medical image analysis. It aims to align two or more images into the same coordinate system to receive a comprehensive understanding. In the field of ophthalmology, it is used for assisting ophthalmologists to diagnose diseases and make treatment planning. There are three main retinal registration groups [9, 29, 36], which are monomodal registration, temporal registration, and multimodal registration. The first group aligns monomodal retinal images captured by the same sensor (*e.g.* fundus camera) at different viewpoints during a single session with a patient to form a single mosaic view of the retina. The second group aligns temporal retinal images taken weeks, months or years apart to reveal disease progression. The third group aligns multimodal retinal images captured by different sensors (*e.g.* fundus camera and fluorescein angiography) to obtain a more complete detail of the subject. In this paper, we focus on multimodal registration in the third group.

While medical image registration has been an active research area for more than two decades [23], fully automatic and robust multimodal image registration remains a challenging task. Among different imaging modalities, the intensity differences are non-linear, and a lot of times the images obtained from the clinics and hospitals are adversely affected by pathologies and/or noise. Fig. 1 shows some examples for retinal images. The fluorescein angiographic (FA) images in Fig. 1(b), 1(d) and 1(f) are acquired after injecting of fluorescein dye into the bloodstream. The dye highlights the blood vessels in the back of the eye causing the intensity of the angiograms to vary substantially and appear different from the color fundus images in Fig. 1(a), 1(c) and 1(e). The blood vessels are generally brighter than the background tissues in the FA images but darker in the color fundus images, and are sometimes being obscured due to the effects of diseases. The optical coherence tomography (OCT) fundus image in Fig. 1(h) is constructed by integration of the 3D tomogram along depth to provide a view similar to traditional en-face imaging modalities such as color fundus images. Due to the nature of imaging with coherent light, OCT is susceptible to coherent noise or also called speckle noise [28], which effectively causes significant degradation in spatial resolution and quality.

Inspired by its success in computer vision, we propose to learn pattern patches among different retinal imaging

*Equal contribution
†Corresponding author: chengjun@nimte.ac.cn

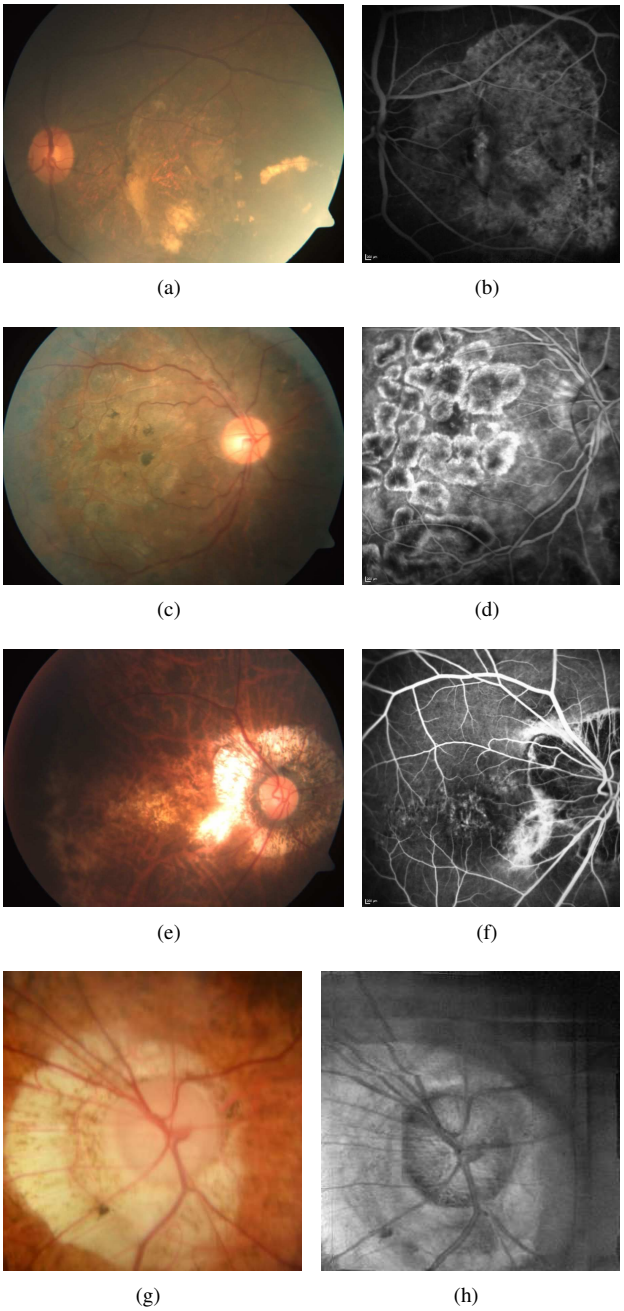


Figure 1. Poor quality multimodal retinal image pairs which are affected by pathologies (diseases) and/or speckle noise. Each row presents a pair, where (a), (c), (e) and (g) are the color fundus images, (b), (d) and (f) are the FA images, and (h) is the OCT fundus image.

modalities based on a CNN. These pattern patches are step pattern representations of the connecting edges from the edge maps which are mainly formed by vascular junctions. The problem is modeled as a classification task, where the goal is to discriminate between informative pattern patches upon observed multiple imaging modalities and use them for multimodal image registration. Optimization is done by rejecting pattern patches with low classification confidences, which hence improves accuracy. The proposed method is fully automatic and no ground truth manual labeling is required. To the best of our knowledge, this is the first time that CNN has been used in the context of multimodal retinal image registration. Since more and more imaging modalities are being developed to better identify eye retina abnormalities, the development of a multimodal image registration method that scales well to new modalities or new image applications with little to no human intervention would have a significant impact on the medical image analysis community.

The rest of the paper is organized as follows. Section 2 discusses related work. In section 3, our methodology is presented. Experiments and results follow in section 4. Finally, we conclude the paper in section 5.

2. Related Work

Although CNN has achieved state-of-the-art performance in image classification and image segmentation [14, 21], there are very few work addressing image registration using CNN [20]. Work on CNN applied to multimodal retinal image registration are even as few as none that we have come across. Recently, a convolutional stacked auto-encoder (CAE) [32] has been proposed to extract features from magnetic resonance imaging (MRI) volumes. It is subsequently combined with a conventional sparse, feature-driven registration algorithm. A CNN based regression approach [26] is introduced to solve rigid 2D or 3D registration for device tracking from 2D X-ray images. The registration method is supervised and not fully automatic. A learnable module called spatial transformer network (STN) [11], in conjunction with a CNN, is used to learn prediction models and has demonstrated on registering non-medical imaging. The STN requires many labeled training samples. A serial-section electron microscopy (ssEM) [34] image registration method is later presented with combination of CAE and STN to generate a deformation map for image alignment via backpropagation of the network. It follows by a feature-based image similarity measure which is learned from the training images by the auto-encoder. Although the above-mentioned algorithms give us some insight into the direction of utilizing CNN for image registration, they are either not applied to medical imaging, or multimodality, or disease cases.

The current state-of-the-art in multimodal retinal image registration is based on feature-based approaches which do not work directly with image intensity values. Information represented by the features is on a higher level. This is suitable for multimodal applications where intensity changes are expected or multi-sensor analysis is demanded [36]. In

fact, feature-based approaches can be further subdivided into two classes: vessel-based and feature descriptor-based. Vessel-based methods typically involve detecting bifurcation points (Y-shape features) [6] by extracting every three dark or bright vessels from circular boundaries of candidate bifurcation locations. The bifurcation points are matched using either local maximization of mutual information [6], or angle-based invariants [35]. There are several other similar methods [3, 15, 29] that utilize vessel bifurcations for image registration. However, although bifurations are invariant to intensity variations, their localizations are imprecise [31]. Moreover, detecting bifurcation points is challenging in poor quality or unhealthy images [8, 16].

Feature descriptor-based methods which do not rely on vasculature have shown to be more suitable for multimodal retinal image registration [5, 8, 16]. A generalized dual-bootstrap iterative closest point (GDB-ICP) [33] uses the popular scale-invariant feature transform (SIFT) [22] with alignment method driven by corner points and face points. An improved version, called edge-driven DB-ICP (ED-DB-ICP) [30], is introduced by enriching SIFT with shape context using edge points. However, it is not robust to scale changes, disease cases, and noise [5, 8, 16]. To overcome the non-linear intensity variations, gradient mirroring method [13] combines opposite gradient directions of SIFT features. However, distinctiveness is compromised due to the reduced dimension of SIFT. To circumvent the problem, a partial intensity invariant feature descriptor (PIIFD)) [4] is presented to achieve higher distinctiveness by combining constrained gradient orientations between 0 to π linearly. It follows by performing a rotation to address the multimodal issue of gradient orientations of corresponding points in opposite directions. Harris-PIIFD [5] is then proposed using PIIFD to describe surrounding fixed size regions of Harris corners [10]. However, the limitations of Harris corners are the non-uniform distribution [8, 17], and the poor repeatability rate due to scale changes or disease cases [8, 16]. A later approach replaces Harris method with uniform robust SIFT (UR-SIFT) [8] for more stable and distinctive features. It has also proven to be more robust to scale changes [16], but it still does not perform well on multimodal retinal images with diseases [16]. Recently, a low-dimensional step pattern analysis (LoSPA) algorithm targets on multimodal retinal images with diseases, and has shown to outperform GDB-ICP, ED-DB-ICP, Harris-PIIFD and UR-SIFT-PIIFD. It uses many customized patterns to describe the vascular junctions and the patterns are able to handle non-linear intensity variations well. However, the multiple patterns are complicated and difficult to implement. Furthermore, the registration success rate for disease cases is still falling below 80% in the paper.

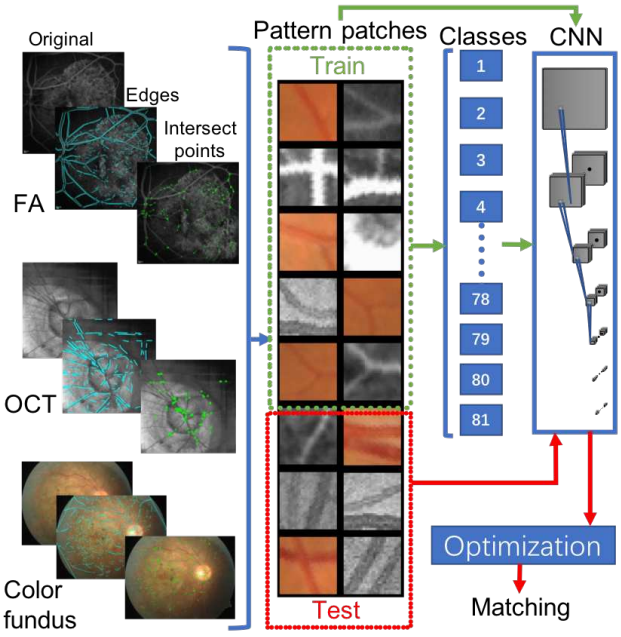


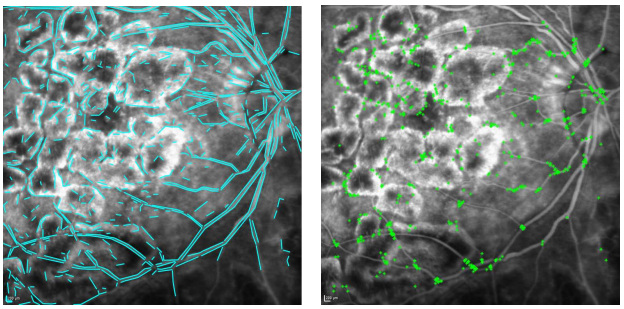
Figure 2. DeepSPa's framework which shows multiple imaging modalities going through the process of extracting feature points, forming into pattern patches, dividing into train and test datasets, all the way to passing through CNN for learning, and optimization during test.

3. Methodology

The DeepSPa's framework is illustrated in Fig. 2. We find intersection points between edges which are mainly formed by vascular junctions in the retinal images. Pattern patches surrounding the intersection points are extracted and sorted into classes according to their pixel patch patterns. CNN is used to learn the patches to be used for matching. During testing, optimization is done by rejecting unreliable pattern patches with low classification confidences among the classes. Regardless of the imaging modalities, the above steps are the same since the pattern patches are sorted according to their patterns and not their modalities. In this paper, we apply our approach to FA, OCT fundus and color fundus images. However, it can be applied to other or more imaging modalities. The following sections will describe each of these parts in more detail.

3.1. Feature extraction

The extraction of distinctive yet repetitive features can be challenging as corresponding feature pairs have to be made across different sensor modalities. Instead of feeding the entire retinal images into our neural network, we exploit the approach of detecting intersection points [17], which are based on connecting edges from the edge maps of the images. Fig. 3 shows an example of extracted edges and their intersection points. Edges are extracted using a strip fitting



(a) (b)

Figure 3. The extracted edges of a FA image are shown in (a), and the intersection points formed by the connecting edges are shown in (b).

algorithm [18]. Many of the connecting edges are situated at the vascular junctions and have shown to achieve high repeatability [17]. The other key advantages of this approach are the insensitivity to non-linear intensity variations across different modalities, and the unnecessary extraction of major or full vascular structure for it to work. It has also shown to outperform various feature detection methods such as the difference of gaussians (DoG) [22] used by SIFT and Harris corner detection [10], in terms of repeatability and stability [17]. However, there are also two common issues encountered. One of them is the missing or fracturing of edges due to the well-known fragmentation of edge maps of real images. Another common issue is the detecting of isolated or insignificant edges which are mainly noise. To circumvent the issues, post-processing steps are applied to fix (concatenate) fragmented edges with end points and angles of close proximity. Two edges e_i and e_j are concatenated if the following condition is satisfied,

$$\theta^{sm}_{e_i e_j} > \tau_\theta \; and \; \min\left(\|P_{k,i}, P_{k,j}\|_2\right) < \tau_{dist}, \quad \forall k \in \{1,2\}, \tag{1}$$

where $\theta^{sm}_{e_i e_j}$ denotes the smaller internal angle between e_i and e_j, and τ_θ represents the allowed internal angle threshold between them. $P_{k,i}$ is the k^{th} end point of e_i, and τ_{dist} denotes the allowed distance threshold between the shortest end-to-end points of e_i and e_j. Subsequently, short and isolated edges (*e.g.* < 5 pixels) are removed.

3.2. Step Pattern Representation

We extract small patches surrounding the intersection points and sort them according to their step patterns. This work focuses on the intensity change patterns rather than the intensity change values as corresponding images of different modalities often do not correlate well due to non-linear intensity changes. We first rotate each patch relative to a mutual orientation in order to achieve rotation invariance, where the center of rotation is at the intersection point $c^{e_{i,j}}$ of two edges e_i and e_j. The angle-to-rotate $\theta_{c^{e_{i,j}}}$ is given by:

$$\min(\theta^x_{e_i}, \theta^x_{e_j}) + [\delta](\max(\theta^x_{e_i}, \theta^x_{e_j}) - \min(\theta^x_{e_i}, \theta^x_{e_j})), \tag{2}$$

where $\theta^x_{e_k}$ denotes the angle from e_k to the positive x-axis, and $[.]$ is a binary indicator function. δ is the inequality derived as:

$$\max(\theta^x_{e_i}, \theta^x_{e_j}) - \min(\theta^x_{e_i}, \theta^x_{e_j}) > 180^\circ. \tag{3}$$

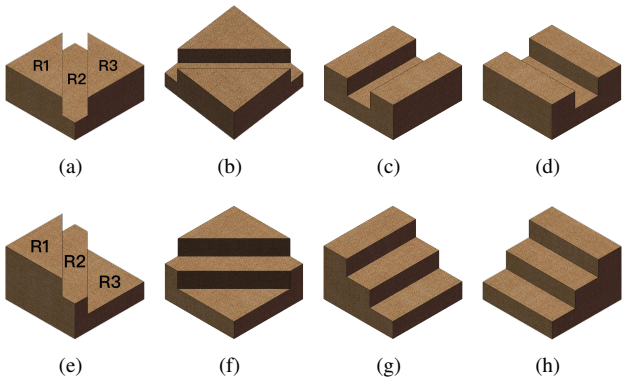
After rotation, a 27×27 local window $\mathbf{W}^{rot}_{c^{e_{i,j}}}$ centered at $c^{e_{i,j}}$ is extracted from the rotated image. We then determine whether $\mathbf{W}^{rot}_{c^{e_{i,j}}}$ comprises the step patterns as shown in Fig. 4. As the name implies, these patterns come in step forms where the higher step regions indicate higher average intensity values in those regions. They are formed using two parallel lines separating each square patch into three equal-sized regions. The regions can be in patterns of two or three level steps, *e.g.* Fig. 4(a) and 4(e) are the same but the former has two level steps and the latter has three. It is apparent that $\mathbf{W}^{rot}_{c^{e_{i,j}}}$ can only comprises either one of the two step patterns and not both at the same time. However, it can comprise one or more other step patterns. This gives a number of possible combination of classes as:

$$\sum_{k=0}^{4} \binom{4}{k} \cdot 2^{(4-k)}, \tag{4}$$

which therefore summed up to a total of 81 classes.

Taking Fig. 4(a) as an example, the number of pixels in $\mathbf{R}_1$, $\mathbf{R}_2$ and $\mathbf{R}_3$ are equal. The average intensity value $I^{\mathbf{R}_k}_{avg}$ in $\mathbf{R}_k$ is formulated as:

$$\frac{1}{N} \sum_{(x,y) \in \mathbf{R}_k} \mathbf{W}^{rot}_{c^{e_{i,j}}}(x,y), \quad \forall k \in \{1,2,3\}, \tag{5}$$



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4. 8 step patterns where (a)-(d) are two level step patterns and (e)-(h) are three level step patterns.

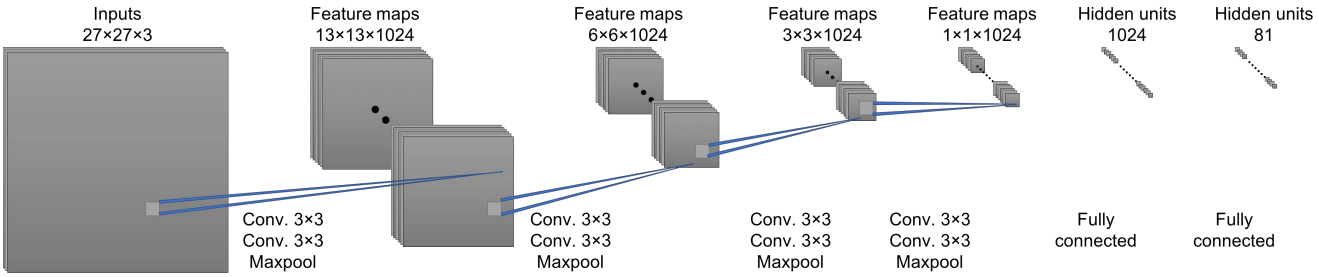


Figure 5. Architecture of our CNN model.

where N is the number of pixels in $\mathbf{R}_k$. The output $out_1^{\mathbf{W}_{c^{ei,j}}^{rot}}$ of $\mathbf{W}_{c^{ei,j}}^{rot}$ for this step pattern can be described as follows:

$$\left[I_{avg}^{\mathbf{R}_1} - I_{avg}^{\mathbf{R}_2} > \tau_{sm}\right] \cdot \left[I_{avg}^{\mathbf{R}_3} - I_{avg}^{\mathbf{R}_2} > \tau_{sm}\right], \quad (6)$$

where $out_1^{\mathbf{W}_{c^{ei,j}}^{rot}}$ is a binary value to indicate whether $\mathbf{W}_{c^{ei,j}}^{rot}$ comprises this step pattern, and τ_{sm} is a small position integer value to avoid noise. In order to cope with contrast reversal problem in different image modalities, e.g. some dark vessels become bright, the step pattern is reversible. Hence, another output $out_2^{\mathbf{W}_{c^{ei,j}}^{rot}}$ can be mathematically revised as:

$$\left[I_{avg}^{\mathbf{R_2}} - I_{avg}^{\mathbf{R_1}} > \tau_{sm}\right] \cdot \left[I_{avg}^{\mathbf{R_2}} - I_{avg}^{\mathbf{R_3}} > \tau_{sm}\right]. \quad (7)$$

The final equation to describe $\mathbf{W}_{c^{ei,j}}^{rot}$ for the step pattern shown in Fig. 4(a) is given by:

$$out_3^{\mathbf{W}_{c^{ei,j}}^{rot}} = out_1^{\mathbf{W}_{c^{ei,j}}^{rot}} + out_2^{\mathbf{W}_{c^{ei,j}}^{rot}}, \quad (8)$$

where $out_3^{\mathbf{W}_{c^{ei,j}}^{rot}}$ is still a binary value. The rest of the patterns in Fig. 4 can also be computed similarly by applying Eq. (5-8). There will be 8 binary values (as a vector) as there are 8 step patterns, and 81 possible combination classes as described in Eq. (4).

3.3. CNN

The most successful type of models for image analysis to date are CNNs. CNNs contain many layers that transform their input with convolution filters of a small extend. In this paper, we use CNNs to learn, classify and optimize the pattern patches extracted from the original images of multiple modalities. Our CNN network architecture is illustrated in Fig. 5. The 27×27 RGB pattern patches are fed to the network. The network consists of a series of convolutional, ReLU and max pooling layers. The output of the network is a classification possibility of 81 classes via a softmax layer.

In order to recognize the pattern patches extracted from all image modalities, we train a multi-modality classification model with multiple modal mixed data. A classification confidence threshold τ_{cl} is used to filter the pattern patches and improve classification accuracy. For an input feature, there are 81 classification confidence output (each between 0.0 and 1.0, summed to 1.0) for 81 classes after the softmax layer. When the max classification confidence is higher than τ_{cl}, the input feature is considered as reliable and reserved to be used in the subsequent steps, else it will be rejected. A higher τ_{cl} means more stringent requirements for the pattern patches. It leads to an increase in the accuracy of the classification but a reduction in the number of remaining features. A good balance will be an optimal τ_{cl} for good classification accuracy while keeping as many features as possible. We experimentally test this in section 4.

3.4. Feature matching and validation

After classification and optimization using CNN, we find corresponding pairs between two sets of classified DeepSPa features by Euclidean distance, using the k-dimensional data structure and search algorithm [2]. The algorithm identifies the k closest neighbors of features in high-dimensional spaces, where we set k to 4 in this paper.

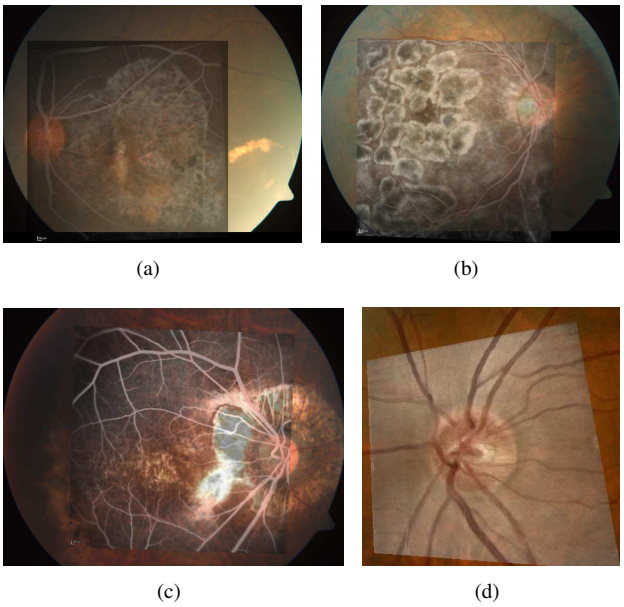
Next, we validate the corresponding pairs in a global transformation function between the two images, regardless of their image modalities. Random sample consensus (RANSAC) [7] with affine transformation setting is applied to the corresponding pairs. Incorrect pairs are excluded using this method.

3.5. Transformation function

Various types of transformation functions can be applied to register retinal images. The most common ones are the linear conformal [25], affine [12] and second-order polynomial [10, 27] models. Linear conformal is the simplest model which only requires two corresponding pairs. As the number of corresponding pairs obtained by our proposed algorithm is usually sufficient and well distributed on the surface of the retina, we can either use the higher order models such as the affine model (three corresponding pairs) or the second-order polynomial model (six pairs of points). We favor affine model over second-order polynomial model as we did not see any significant differences between the re-

sults among the two models in our experiments.

When the transformation function has been applied on the floating retinal image, we simply superpose the transformed retinal image on the fixed retinal image to produce a retinal mosaic. The mosaic image results of images in Fig. 1 are shown in Fig. 6.



(a) (b) (c) (d)

Figure 6. Mosaic results of the proposed DeepSPa algorithm for multimodal image pairs shown in Fig. 1(a) and 1(b), Fig. 1(c) and 1(d), Fig. 1(e) and 1(f), and Fig. 1(g) and 1(h) respectively.

4. Experiments and Results

In this section, we evaluate our proposed DeepSPa algorithm on three parts: CNN settings, rotation invariance and scale change. We also compare our registration performance with state-of-the-art algorithms.

4.1. Datasets

In our experiments, we perform tests mainly on three retinal image modalities: OCT fundus, color fundus and FA images. For CNN evaluation, 67,240 DeepSPa features are being extracted from 200 OCT fundus images ranging in size from 304×304 to 513×385. 196,212 DeepSPa features are being extracted from 292 color fundus images ranging in size from 410×410 to 2588×1958, and 100,125 DeepSPa features are being extracted from 194 FA images ranging in size from 720×576 to 768×768. The total number of DeepSPa features being extracted is 363,577.

The other evaluations and comparative tests are demonstrated on three multimodal retinal image datasets. The datasets are described as follows.

Color fundus-FA (mild-to-moderate retinal diseases) The first dataset [1] is publicly available comprising both color fundus and corresponding FA images of 30 patients with diabetic retinopathy. We classify this dataset as patients with mild-to-moderate retinal diseases in this paper. Both the color fundus and the FA images in the dataset have the same resolution of 720×576.

Color fundus-FA (severe retinal diseases) The second dataset was provided by a local hospital, comprising color fundus and corresponding FA images of 120 anonymous patients with symptoms of severe macular edema and staphyloma which require retinal photocoagulation or photodynamic therapy. The doctors described this dataset as one of the most challenging ones compared to other retinal abnormality cases. We classify this dataset as patients with severe retinal diseases in this paper. Some examples of the image pairs are shown in Fig. 1. The color fundus images range in size from 2588×1958, and the FA images range in size from 768×768. The largest scaling factor in the two datasets is 1.8, however most of these clinical data are of very small scale difference of below 1.5. The largest rotation angle is $30°$.

Color fundus-OCT (speckle noise & low resolution) The third dataset was also provided by a local hospital. It comprises 80 pairs of color fundus and corresponding OCT fundus images. The OCT fundus images are adversely affected by speckle noise. We classify this dataset as images with speckle noise and of low resolution in this paper. Fig. 1(g) and 1(h) show one of these image pairs. The color fundus photographs were acquired with a TRC-NW8 non-mydriatic fundus camera and the 3D OCT data were obtained from a Topcon DRI OCT-1 machine with a size of $992 \times 512 \times 256$ voxels. The OCT fundus images were formed by intensity averaging along A-scans. The color fundus images range in size from 410×410 to 1016×675, and the resized low resolution OCT fundus images range in size from 304×304 to 513×385.

4.2. Robustness test results

This part evaluates on the CNN settings and the robustness of DeepSPa algorithm to rotation invariance and scale insensitivity.

CNN test We divide the datasets described in section 4.1 into training sets and testing sets. Each training set comprises of 80% randomly picked features and the other 20% features are used as testing set. We train three single modality classification models which comprise of OCT fundus, color fundus and FA features individually. We also train a multi-modality classification model with all three modal mixed data. We apply different classification confidence

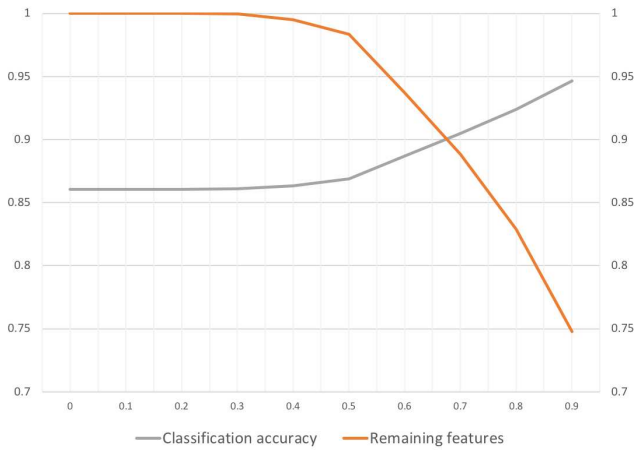


Figure 7. CNN test results with different classification confidence threshold τ_{cl}. The two lines intersect at 0.68.

Table 1. Classification accuracies of different modality models without thresholding, on three datasets. Top score in each is indicated in bold type.

Model	Test Accuracy on Datasets (in %)			
	OCT	FA	Color fundus	Total
OCT	59.03	50.46	19.15	36.84
FA	72.02	85.42	28.43	59.19
Color fundus	53.47	49.50	79.36	63.10
Multimodal	**80.95**	**87.01**	**85.52**	**86.04**

threshold τ_{cl} to filter the pattern patches and the detailed test results are shown in Fig. 7. From the chart, we can see that when the threshold rises, the classification accuracy increases. On the contrary, the remaining features drop when the threshold rises. An optimal τ_{cl} to choose is the intersection point at 0.68, which gives a good balance of 90% in the classification accuracy and remaining features. Tables 1 and 2 compare the classification accuracies between single trained models and multi-modality trained model in three different datasets, without and with thresholding respectively. We can see an increase in accuracy for all models, and it is clear that our multi-modality trained model performs best in all three datasets. We use 0.68 for τ_{cl} throughout the rest of our experiments since it gives the best results.

Rotation invariance test Although the largest rotation angle in our datasets is 30°, we select 20 multimodal image pairs from our datasets and rotate the floating images in the image pairs from 0° to 180° with a 20° step. It should be noted that the reference images are held fixed. We apply DeepSPa algorithm on the reference images and the rotated floating images. All image pairs are successfully registered regardless of the rotation angle, demonstrating that DeepSPa is rotation invariant.

Table 2. Classification accuracies of different modality models with 0.68 threshold applied, on three datasets. Top score in each is indicated in bold type.

Model	Test Accuracy on Datasets (in %)			
	OCT	FA	Color fundus	Total
OCT	70.25	62.23	31.04	53.84
FA	79.13	89.95	31.95	65.61
Color fundus	68.25	56.43	87.28	71.13
Multimodal	**84.59**	**90.10**	**90.76**	**90.11**

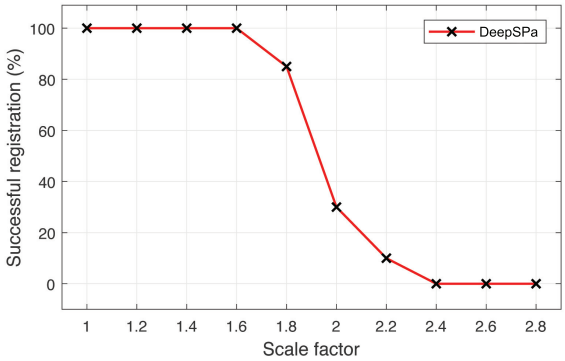


Figure 8. Successful registration relative to scale factor.

Scale change test The largest scaling factor in our datasets is 1.8. Similarly, we select 20 multimodal image pairs from our datasets to perform rescaling. We rescale the floating images with a scaling factor from 1 to 2.8, and apply DeepSPa algorithm on all the images. The registration rates across a range of scale changes are shown in Fig. 8. The results indicate that DeepSPa can provide robust registration when the scale factor is 1.8 and below. It usually fails when the scale factor is above 1.9. However, this is acceptable as clinical data are usually of very small scale differences and are usually less than 1.5 [5, 16].

4.3. Comparative test results

Ground truth We select 10 pairs of corresponding points in every image pair manually to generate ground truth. These points are selected to be distributed uniformly with an accurate localization. The main advantage of this method is that it can handle poor quality retinal images which are adversely affected by diseases and/or noise. For each marked image pair, there will be another team member to verify the correctness of the marked points. The process is time-consuming, but it provides a relatively reliable and fair measurement over all images.

Evaluation criteria As our datasets comprise clinical images of poor quality, e.g. adversely affected by pathologies and/or speckle noise, centerline error measure [3, 29] which measures the median error of the centerline of

Table 3. Multimodal registration results of 7 algorithms on dataset of color fundus-FA (mild-to-moderate retinal diseases). Success rate of registration, average RMSE, average MAE and average MEE are shown. Top score in each is indicated in bold type.

	SIFT	GDB-ICP	ED-DB-ICP	UR-SIFT-PIIFD	Harris-PIIFD	LoSPA	DeepSPa
Success rate (%)	0	10	60	86.67	90	93.33	**96.67**
Average RMSE	N.A.	4.07	2.33	2.97	2.27	1.93	**1.71**
Average MAE	N.A.	7.67	4.06	4.88	3.67	3.35	**2.97**
Average MEE	N.A.	3	1.78	2.54	1.67	1.55	**1.37**

Table 4. Multimodal registration results of 7 algorithms on dataset of color fundus-FA (severe retinal diseases). Success rate of registration, average RMSE, average MAE and average MEE are shown. Top score in each is indicated in bold type.

	SIFT	GDB-ICP	ED-DB-ICP	UR-SIFT-PIIFD	Harris-PIIFD	LoSPA	DeepSPa
Success rate (%)	0	4.17	27.5	35	41.67	79.17	**86.5**
Average RMSE	N.A.	3.81	3.1	4.58	3.93	2.61	**2.13**
Average MAE	N.A.	8.23	6.58	8.86	8.4	6.26	**4.61**
Average MEE	N.A.	3.51	2.81	4.55	3.69	2.34	**2.12**

Table 5. Multimodal registration results of 7 algorithms on dataset of color fundus-OCT (speckle noise & low resolution). Success rate of registration, average RMSE, average MAE and average MEE are shown. Top score in each is indicated in bold type.

	SIFT	GDB-ICP	ED-DB-ICP	UR-SIFT-PIIFD	Harris-PIIFD	LoSPA	DeepSPa
Success rate (%)	0	3.75	22.5	27.5	33.75	63.75	**81.25**
Average RMSE	N.A.	4.21	3.5	4.46	3.97	3.43	**2.62**
Average MAE	N.A.	8.63	6.98	8.98	8.74	6.68	**5.29**
Average MEE	N.A.	3.71	2.95	4.25	3.83	2.72	**2.07**

vasculature is ineffective. In practice, small differences will exist between the coordinates of the transformed points and reference points. Hence, we evaluate the registration accuracy by the root-mean-square-error (RMSE) between 10 pairs of corresponding points in each image pair [8, 16, 19, 24]. A RMSE below 5 pixels is acceptable for clinical purposes [24]. We also report the median error (MEE) [5, 8] and maximal error (MAE) [5, 8] over all corresponding points. For successful registration, we consider the RMSE $<$ 5 pixels [8] in proportion to the image resolution in [8]. In addition, a significant error of MAE $>$ 10 pixels [5] results in a registration failure. We record results for all successful registrations.

Comparison results We run comparative experiments between 7 algorithms: SIFT [22], GDB-ICP [33], ED-DB-ICP [30], UR-SIFT-PIIFD [8], Harris-PIIFD [5], LoSPA, and our DeepSPa. Tables 3, 4 and 5 show the comparison results on the three datasets as described in section 4.1. SIFT algorithm fails to register any image pairs on all three datasets. For less challenging dataset of color fundus-FA (mild-to-moderate retinal diseases), most algorithms pass the 50% success rate mark, with DeepSPa dominating in all scores as shown in Table 3. For datasets of color fundus-FA (severe retinal diseases) and color-fundus-OCT (speckle noise and low resolution) which are more challenging, only LoSPA and DeepSPa algorithms pass the 50% success rate mark, with DeepSPa still unsurpassable in all scores as shown in Tables 4 and 5. DeepSPa is also the only algorithm among the 7 algorithms to achieve above 80% registration success rate on all three datasets. Some registration results of DeepSPa on the three datasets are shown in Fig. 6.

The comparison in this section demonstrates that the deployment of the DeepSPa algorithm to multimodal retinal image registration translates into higher registration accuracy. Although we have demonstrated our algorithm on the color fundus, FA and OCT fundus images, it can be applied to other image modalities such as autofluorescence and en-face OCT images, as well as images from other applications.

5. Conclusion

In this paper, we leverage on the strength of deep neural networks for learning and optimizing our step pattern patches to be used for multimodal retinal image registration. The DeepSPa algorithm is invariant to non-linear intensity changes which is an important requisite for multimodal registration. We have demonstrated DeepSPa on three multimodal retinal image datasets, and results indicate that DeepSPa achieves higher registration accuracy which easily frustrates the other 6 existing algorithms in the experiments. It has also achieved state-of-the-art performance by consistently attaining above 80% registration success rates on all the three datasets, which includes the more challenging ones such as with severe retinal diseases.

References

[1] Shirin H. M. Alipour, Hossein Rabbani, and Mohammad R. Akhlaghi. Diabetic retinopathy grading by digital curvelet

transform. *Comp. and Math. Methods in Med.*, pages 1607–1614, 2012.

[2] Jon L. Bentley. Multidimensional binary search trees used for associative searching. *Comm. ACM*, 18(9):509–517, 1975.

[3] Ali Can, Charles V. Stewart, Badrinath Roysam, and Howard L. Tanenbaum. A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina. *TPAMI*, 24(3):347–364, 2002.

[4] Jian Chen, R. Theodore Smith, Jie Tian, and Andrew F. Laine. A novel registration method for retinal images based on local features. In *Proc. EMBC*, pages 2242–2245, 2008.

[5] Jian Chen, Jie Tian, Noah Lee, Jian Zheng, R. Theodore Smith, and Andrew F. Laine. A partial intensity invariant feature descriptor for multimodal retinal image registration. *TBME*, 57(7):1707–1718, 2010.

[6] Tae E. Choe and Isaac Cohen. Registration of multimodal fluorescein images sequence of the retina. In *Proc. ICCV*, pages 106–113, 2005.

[7] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.

[8] Zeinab R. Ghassabi, Jamshid Shanbehzadeh, Amin Sedaghat, and Emad Fatemizadeh. An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors. *IJIVP*, 2013(25), 2013.

[9] Arthur A. Goshtasby. *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. Wiley-Interscience, 2005.

[10] Christopher G. Harris and Mike Stephens. A combined corner and edge detector. In *Proc. AVC*, pages 147–151, 1988.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proc. NIPS*, volume 2, pages 2017–2025, 2015.

[12] Roger Jagoe, Christopher Blauth, Peter L. Smith, John V. Arnold, Kenneth Taylor, and Richard Wootton. Automatic geometrical registration of fluorescein retinal angiograms. *Comp. and Biomed. Research*, 23(5):403–409, 1990.

[13] Avi Kelman, Michal Sofka, and Charles V. Stewart. Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations. In *Proc. CVPR*, pages 17–22, 2007.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012.

[15] France Laliberté, Langis Gagnon, and Yunlong Sheng. Registration and fusion of retinal images - an evaluation study. *T-MI*, 22(5):661–673, 2003.

[16] Jimmy A. Lee, Jun Cheng, Beng-Hai Lee, Ee-Ping Ong, Guozhen Xu, Damon W.-K. Wong, Jiang Liu, Augustinus Laude, and Tock-Han Lim. A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration. In *Proc. CVPR*, pages 1046–1053, 2015.

[17] Jimmy A. Lee, Beng-Hai Lee, Guozhen Xu, Ee-Ping Ong, Damon W.-K. Wong, Jiang Liu, and Tock-Han Lim. Geometric corner extraction in retinal fundus images. In *Proc. EMBC*, 2014.

[18] Maylor K. Leung and Yee-Hong Yang. Dynamic two-strip algorithm in curve fitting. *Patt. Recog.*, 23(1-2):69–79, 1990.

[19] Jupeng Li, Houjin Chen, Chang Yao, and Xinyuan Zhang. A robust feature-based method for mosaic of the curved human color retinal images. In *Proc. BMEI*, pages 845–849, 2008.

[20] Rui Liao, Shun Miao, Pierre de Tournemire, Sasa Grbic, Ali Kamen, Tommaso Mansi, and Dorin Comaniciu. An artificial agent for robust image registration. In *Proc. AAAI*, pages 4168–4175, 2017.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3431–3440, 2015.

[22] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[23] Primoz Markelj, Dejan Tomaževič, Bostjan Likar, and Franjo Pernuš. A review of 3D/2D registration methods for image-guided interventions. *Medical Image Analysis*, 16(3):642–661, 2012.

[24] George K. Matsopoulos, Pantelis A. Asvestas, Nicolaos A. Mouravliansky, and Konstantinos K. Delibasis. Multimodal registration of retinal images using self organizing maps. *T-MI*, 23(12):1557–1563, 2004.

[25] George K. Matsopoulos, Nicolaos A. Mouravliansky, Konstantinos K. Delibasis, and Konstantina S. Nikita. Automatic retinal image registration scheme using global optimization techniques. *Trans. Info. Tech. Biomed.*, 3(1):47–60, 1999.

[26] Shun Miao, Z. Jane Wang, and Rui Liao. A CNN regression approach for real-time 2D/3D registration. *T-MI*, 35(5):1352–1363, 2016.

[27] Neil Ryan, Conor Heneghan, and Philip de Chazal. Registration of digital retinal images using landmark correspondence by expectation maximization. *IVC*, 22(11):883–898, 2004.

[28] Joseph M. Schmitt, S. H. Xiang, and Kin M. Yung. Speckle in optical coherence tomography. *J. Biomed. Opt.*, 4:95–105, 1999.

[29] Charles V. Stewart, Chia-Ling Tsai, and Badrinath Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *T-MI*, 22(11):1379–1394, 2003.

[30] Chia-Ling Tsai, Chun-Yi Li, Gehua Yang, and Kai-Shung Lin. The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence. *T-MI*, 29(3):636–649, 2010.

[31] Chia-Ling Tsai, Charles V. Stewart, Howard L. Tanenbaum, and Badrinath Roysam. Model-based method for improving the accuracy and repeatability of estimating vascular bifurcations and crossovers from retinal fundus images. *Trans. Info. Tech. Biomed.*, 8(2):122–130, 2004.

[32] Guorong Wu, Minjeong Kim, Qian Wang, Brent C. Munsell, and Dinggang Shen. Scalable high performance image registration framework by unsupervised deep feature representations learning. *TBME*, 63(7):1505–1516.

[33] Gehua Yang, Charles V. Stewart, Michal Sofka, and Chia-Ling Tsai. Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence. *TPAMI*, 23(11):1973–1989, 2007.

[34] Inwan Yoo, David G. C. Hildebrand, Willie F. Tobin, Wei-Chung A. Lee, and Won-Ki Jeong. ssEMnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features. In *Proc. DLMIA workshop, MICCAI*, pages 249–257, 2017.

[35] Frederic Zana and Jean-Claude Klein. A registration algorithm of eye fundus images using a bayesian hough transform. In *Proc. Int'l Conf. Image Process. and Apps.*, volume 2, pages 479–483, 1999.

[36] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *IVC*, 21:977–1000, 2003.