Research Article | Vol. 12, No. 1 / 1 January 2021 / Biomedical Optics Express 619
Biomedical Optics EXPRESS

Deep-learning based multi-modal retinal image registration for the longitudinal analysis of patients with age-related macular degeneration

THARINDU DE SILVA,[1] EMILY Y. CHEW,[1] NATHAN HOTALING,[2,3] AND CATHERINE A. CUKRAS[1,4]

[1]National Eye Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]National Center for Advancing Translational Science, National Institutes of Health, Bethesda, MD 20892, USA
[3]nathan.hotaling@nih.gov
[4]cukrasc@nei.nih.gov

Abstract: This work reports a deep-learning based registration algorithm that aligns multi-modal retinal images collected from longitudinal clinical studies to achieve accuracy and robustness required for analysis of structural changes in large-scale clinical data. Deep-learning networks that mirror the architecture of conventional feature-point-based registration were evaluated with different networks that solved for registration affine parameters, image patch displacements, and patch displacements within the region of overlap. The ground truth images for deep learning-based approaches were derived from successful conventional feature-based registration. Cross-sectional and longitudinal affine registrations were performed across color fundus photography (CFP), fundus autofluorescence (FAF), and infrared reflectance (IR) image modalities. For mono-modality longitudinal registration, the conventional feature-based registration method achieved mean errors in the range of 39-53 µm (depending on the modality) whereas the deep learning method with region overlap prediction exhibited mean errors in the range 54-59 µm. For cross-sectional multi-modality registration, the conventional method exhibited gross failures with large errors in more than 50% of the cases while the proposed deep-learning method achieved robust performance with no gross failures and mean errors in the range 66-69 µm. Thus, the deep learning-based method achieved superior overall performance across all modalities. The accuracy and robustness reported in this work provide important advances that will facilitate clinical research and enable a detailed study of the progression of retinal diseases such as age-related macular degeneration.

1. Introduction

Retinal diseases such as age-related macular degeneration (AMD) affect ~3 million Americans and the associated vision loss among the ageing population creates a growing healthcare burden [1]. Early diagnosis and prediction of disease progression is important in developing therapeutics to prevent late stage disease and irrecoverable vision loss. In this context, longitudinal clinical studies are critical in understanding the natural history of AMD and other retinal diseases with multi-modal imaging playing an important role in measuring the structural changes that occur in the course of the disease. Color fundus photography (CFP), fundus autofluorescence (FAF), and infrared reflectance (IR) are 2D imaging modalities that are routinely acquired in the clinic. Co-registering these longitudinally acquired, multi-modal images facilitates the characterization of disease progression as well as assists the exploration of imaging biomarkers that precede vision loss. This work aims to develop a registration framework that automatically aligns several 2D imaging modalities of the retina to advance research in large-scale clinical studies.

#408573 https://doi.org/10.1364/BOE.408573
Journal © 2021 Received 28 Aug 2020; revised 29 Oct 2020; accepted 30 Oct 2020; published 23 Dec 2020

Biomedical Optics EXPRESS

Multiple previous work in the context of retinal imaging have developed registration tools using conventional feature-based and intensity-based methods [2,3]. Feature based methods establish point correspondences typically relying upon feature detectors that capture high-contrast vascular structures in retinal images [4–6]. Point correspondences are used to estimate the transformation parameters and solve for alignment [7]. Mapping non-linear intensities in multiple modalities is a challenging ill-posed problem and the success of previous hand-crafted descriptors relies upon several pre-processing and/or fine tuning steps to address the specific challenges of the modalities being analyzed. This requires investing significant effort to optimize and tailor the feature detection and description process [8–10]. Li et al. [9]. found that MIND itself did not confer enough specificity for feature-based matching and instead proposed a method based on mean-phase of histogram of oriented gradients (HOG-MP) in conjunction with an intensity-based refinement using MIND to achieve robust performance in registering color and infrared reflectance images. Ghassabi et al. [10] proposed modifications to existing SIFT and Harris detectors to enhance robustness. Wang et al. [11] had to overcome deficiencies in single detectors and combined multiple detectors and descriptors to eliminate outlier matches and improve accuracy in multi-modal retinal image registration. Even in the presence of an extensive body of work in conventional handcrafted, feature based registration, there are two-fold challenges to achieving accurate and robust performance in a large multi-modal clinical data set using a fully automated algorithm without requiring specific fine-tuning steps to improve robustness in a given modality pair. First, the image content in terms of structure and texture changes over time due to disease and/or confounding conditions such as cataract and other obscurations of the media. Fine tuning hand-crafted feature detectors to account for mismatch and improve robustness resulting from such variable sources of change is a challenge. Second, cross-sectional, multi-modal images have non-linear, heterogeneous mapping in image intensities due to different cameras, fields of view, and image modalities that capture different spectral properties. This necessitates feature detectors and descriptors to specifically account for different properties such as large changes in image scale and/or inhomogeneous mapping of image-intensities across modalities and customize feature-based algorithms to achieve accurate alignment. Thus, having a registration framework that inherently addresses these challenges both across modalities and longitudinally and achieves accurate and robust performance is the driving motivation behind this work.

Deep-learning based approaches have recently shown promise for challenging registration problems and thus could be suitable for addressing the challenges in retinal image registration [12,13]. While there has been previous work on disease/feature detection in diabetic retinopathy, AMD, retinopathy of prematurity (ROP) and image quality assessment using deep-learning [14,15], it has gained recent interest for applications in ocular/retina image registration [16,17]. While deep-learning based registration methods are less common compared to object detection and segmentation algorithms, Lee et al. [16] describe a relevant method to learn feature classes from patches extracted from vascular structures in retinal images. The presented method solves feature selection and matching using CNN models and relies upon basic image processing methods involving several empirically defined parameters for vascular structure identification and affine transformation estimation. While it showed improvements over several conventional feature-based registration methods, the reported success rate is ~80% indicating the need for more robust solutions to meet clinical requirements. This work aims to improve upon such models by handling multiple modalities and facilitating end-to-end learning including feature selection, description, and matching within a single framework. Moreover, machine/deep learning methods have targeted specific elements of the registration framework such as learning registration initialization [18], similarity metric [19], intensity mapping [20], and intermediate images/image synthesis for multi-modality registration [21]. Furthermore, end-to-end learning has been performed targeting the entire registration framework in multiple instances using U-Net like architectures to produce a deformation field, often applying constraints on the type of deformation

Biomedical Optics EXPRESS

[22]. While U-Net like architectures utilize a large number of parameters with up-sampling convolutional layers to compute a non-rigid deformation field, learning based end-to-end affine regression models may be suitable for retinal image registration due to the presence of feature rich vascular structures and the ability to correct for the misalignment caused by the optical camera with an affine transform. While the model can be extended to correct for perspective distortions of the camera and/or any deformation arising from the spherical eye shape, such improvements could yield incremental advantage in accuracy considering the relatively small magnitude of such distortions. On the other hand, an affine transformation that corrects for the gross motion of the patient and camera would compensate for the majority of the misalignment and provide clinically useful level of accuracy as evaluated in this work. There are computer vision-based registration methods previously developed for alignment of natural images using such models [23]. However, the reported alignment accuracy in natural imaging-based problems has been coarse and has not been extensively validated for geometric accuracy. These do not meet the clinical requirements for applications in retina where the anatomical structures need to be aligned at <100 micron precision.

This work explores affine regression models for registration mirroring the conventional feature-based algorithms in a deep learning framework. Direct computation of affine registration parameters from the network [23] could be challenging due to the non-linearity in the affine parameter space. Moreover, deep learning algorithms have been previously identified as difficult to apply to registration problems due to the involvement of non-overlapping physical spaces [24]. In this work, we evaluate modifications to the regression models to solve feature correspondences accurately and robustly. The region overlap between the image pair was incorporated and predicted as a part of the network output to improve the accuracy in estimating the affine transformation using feature correspondences. This extends our preliminary work [25] that compared the performance of existing conventional and deep-learning methods [23] for mono-modality retinal image registration

The current work develops deep-learning based algorithms for fully automated alignment of retinal images to translate to moderate-to-large data sets consisting of longitudinal and cross-sectional multi-modal images. Compared to well-established conventional feature-based registration methods, this approach offers important advances. With conventional methods, optimizing the feature detector and descriptor often requires laborious customization to account for differences in image quality, image characteristics, and heterogenous disease presentation in each modality. Alternatively, the deep learning approach learns a generalized representation for feature detection, description, and matching using example training images without the need to hand-craft each step. We thus develop a single, generalizable model that can register images across time and modality which is capable of handling different image pairings. New retinal image modalities (e.g., near infrared autofluorescence or indocyanogreen angiography) can be incorporated in a straightforward manner within the framework. This is a practical consideration in real-life situations where new image modalities may be incorporated in the course of an ongoing study. The performance of deep learning registration was evaluated using CFP, FAF, and IR image modalities and was compared to a traditional features-based registration method comprised of extracting scale-invariant feature transform (SIFT) features and computing the registration transformation using random sample consensus (RANSAC) algorithm. The methods were validated by computing the target registration error using manually identified corresponding landmarks (i.e., vessel bifurcations) in images.

2. Methods

2.1. Conventional feature-based registration

In a feature-based registration framework, independently detected features from each image are used to construct a correlation matrix using a distance metric that quantifies the similarity

Biomedical Optics EXPRESS

between feature pairs. Feature correspondence is established after eliminating any outlier matches in the feature point sets and the transformation between two corresponding feature point sets is computed. The main steps of the conventional feature-based alignment are shown in Fig. 1.

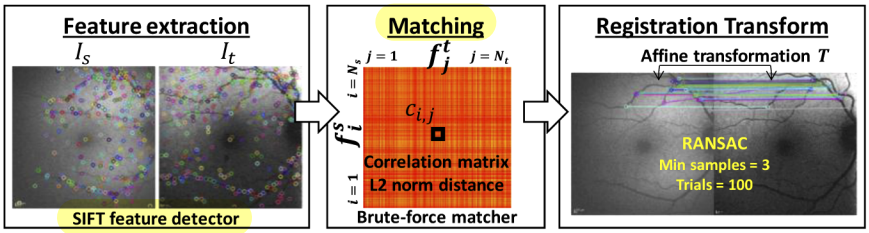

Fig. 1. Image alignment using the conventional method (C_{SIFT}). SIFT features were extracted from each image and correlation matrix was computed between two feature sets using L2 norm metric. Affine transformation was computed using RANSAC to eliminate outlier feature matches.

In our implementation, prior to feature detection, images were preprocessed using bilateral filtering followed by contrast-limited adaptive histogram equalization (CLAHE). In order to select a suitable descriptor, we tested six hand-crafted feature descriptors (i.e., SURF, BRIEF, ORB, local binary patterns, HOG, and MIND) during initial experiments. While these included descriptors specifically designed to handle occlusion (i.e., HOG) and non-linear intensity mappings in multiple modalities (i.e., modality independent neighborhood descriptor - MIND), SIFT [26] was selected to collectively represent the conventional feature based registration using a hand-crafted descriptor due to its superior performance during empirical testing in our data set. Given a source (I_s) and a target image (I_t), the algorithm independently detects feature locations and represent them using L-length feature descriptors (f^s, f^t). The number of features (N_s and N_t) detected in source and target image is variable. The element (i, j) of the ($N_s \times N_t$) dimensional correlation matrix is computed with L2-norm distance of L-length feature vectors according to:

$$c_{ij} = \sum_{k=1}^{L} (f_{ik}^s \cdot f_{jk}^t)^2 \tag{1}$$

Brute force matcher identified the similarity between feature point pairs to solve for correspondence. RANSAC algorithm (minimum samples = 3, number of trials = 100) was applied to compute the affine transformation by eliminating any outlier matches. This implementation of the conventional method is referred to as $\mathbf{C_{SIFT}}$ method in subsequent analysis.

2.2. Deep learning-based registration

2.2.1. Deep-learning architecture mirroring conventional feature-based registration

We investigate the convolutional neural network (CNN)-based architecture resembling the conventional feature-based registration process proposed by Rocco et al. [23] Three main parts in the network, i.e., feature extraction, matching, and computing the transformation, are analogous to the conventional method as shown in Fig. 2. Feature extraction was performed with a Siamese architecture using dual VGG 16 feature extractors. Unlike the conventional method, the number of features does not change for a fixed image size. It is dependent on the VGG layer from which the feature descriptors are extracted. For example, with source and target image size = 240×240, VGG pool4 layer after downsizing the original image four times yields $N_s = N_t = 15 \times 15 = 225$ for every image. A dense correlation matrix (with dimensions $N_s \times N_t$) was constructed with the L2-norm of the feature extractors according to Eq. (1).

Correlation matrix was connected to a regression network which outputs parameters related to the affine transformation to align the image pair. In lieu of brute-force matcher and RANSAC

Biomedical Optics EXPRESS

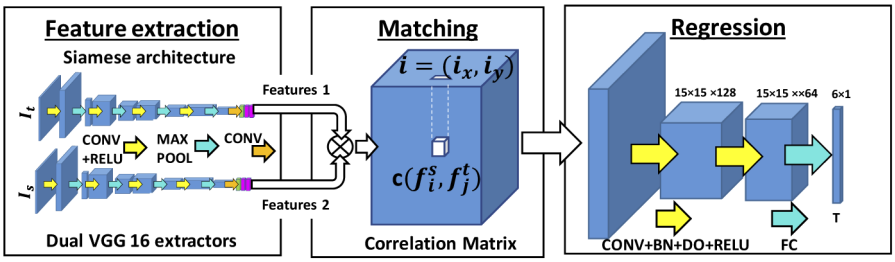

Fig. 2. Feature based registration using deep-learning (DL_{AFFINE}). CNN-based architectures for feature extraction, matching, and computing the registration transform. The network outputs 6-parameter affine transformation. CONV: convolution layer, BN: batch normalization, DO: drop off, RELU: rectified linear unit, FC: fully-connected layer.

in the conventional method, the regression network eliminates outliers, finds the mapping between correspondences, and estimates the non-linear affine registration parameters. Training was thus performed end-to-end using the architecture shown in Fig. 2, where the combination of Siamese feature detectors, correlation matrix, and the regression layer directly output the 6-parameter affine transformation vector for a source and target image pair. The loss function was computed after applying the affine transformation to a spatial grid of points in the image space and computing the grid-point displacement error compared to the grid generated with the ground truth transformation. This method is referred to as **DL_{AFFINE}** in subsequent analysis.

Several characteristics of the regression network are noteworthy when comparing it to the conventional feature-based approach. Firstly, unlike in **C_{SIFT}**, there is no explicit method to identify feature locations. For a given image size, the number of extracted features is fixed and each feature could be considered to represent a patch in the image. For example, with original image size = 200×200 px and $N_s = 10$, the elements in the resulting 10×10 correlation matrix are representative of contiguous 20×20 px patches in the image. Constructing the correlation matrix using contiguous patches results in a dense matrix where features represent all regions of the image (compared to the sparse correlation matrix in the conventional method with a limited set of detected features). Solving patch-to-patch correspondence between images limits registration accuracy compared to establishing point-to-point correspondence (with known detected feature locations) in the conventional method. Increasing N_s can improve the accuracy by decreasing the patch size at the cost of increasing the number of parameters computed by the regression network. For example, $N_s = 200$ (200×200 correlation matrix for 200×200 px image) enables the computation of pixelwise correspondences at the cost of an exponential increase in the number of parameters of the regression network. Alternatively, small N_s limits the precision of solving locations of corresponding features. In Rocco et al. [23], this complexity is buried in the regression network where the precise pixel-wise correspondences need to be extrapolated using the coarse, patch-based correlation matrix to estimate the affine output transformation. While a decrease in accuracy due to patch-wise matching may be acceptable for certain applications in computer vision, it may not yield micron-level precision that is required in retinal imaging applications.

Secondly, allowing the network to train end-to-end demands the estimation of the registration output parameters (6-parameter vector for affine transformation consisting of rotation, translation, shear, and scale) that lie in non-linear space. Rocco et al. [23] introduced a grid loss function that applied the affine transformation parameters to a spatial grid of points and computed the sum of displacement errors at grid locations as the loss, thereby converting errors in the non-linear space to a set of displacement errors. The complexity due to non-linear parameters is hidden in

Biomedical Optics EXPRESS

the regression network and it needs to estimate the non-linear registration parameters without the knowledge of the affine transformation operation that computes the loss.

Thirdly, computing a dense correlation matrix could be suboptimal at non-overlapping regions in the image pair. Some patches in source image have no corresponding region in the target image and vice versa. The regression network needs to learn how to suppress these elements of the correlation matrix comprising of non-overlapping regions. While convolutional neural networks are capable of hiding such complexity, we explored alternative implementations to the regression network that explicitly address these concerns to improve performance.

2.2.2. Regression network to predict patch displacements

Instead of using convolutional layers to directly compute the non-linear affine parameters as in $\mathbf{DL_{AFFINE}}$, we experimented with the regression network predicting relative displacements of image patches. Thus, we train for the corresponding feature displacements by measuring the displacement relative to the center of each patch. The objective of the regression network with this approach is to compute relative displacements for all the patches in the image from the correlation matrix. The loss function is computed as the total Euclidean distance of patch center displacements. Using relative displacements of all the patch centers in the network output, the registration transformation was computed in closed form solution using the methods described in Horn et al. [27]. This method is referred to as $\mathbf{DL_{DISP}}$ in subsequent analysis.

2.2.3. Regression network to predict patch displacements with region overlap

Non-overlapping areas in the source-target image pair cannot effectively contribute to the dense correlation matrix representing contiguous image patches. Thus, estimating the correspondence using only regions of overlap that effectively contributed to the correlation matrix could improve accuracy and efficiency. Thus, we modified the regression network to estimate the region of overlap as a binary prediction for each patch and compute patch displacements only within the overlapping region as shown in Fig. 3. We propose a loss function combining the binary overlap prediction and displacements in x and y directions as following:

$$Loss = Cross\ Entropy\ (O) + \lambda \sum_{i \in O} \nabla a^2 + \nabla b^2 \tag{2}$$

where O denoted the region of overlap between source and target images, ∇a and ∇b represent patch displacement errors in x and y, and weighting parameter λ. The resulting affine transformation is computed using the correspondences within the overlapping region.

We refer to this method as $\mathbf{DL_{OVLP}}$ in subsequent analysis. Figure 4 compares the three approaches to the regression network that predicts different parameters related to the underlying misalignment between the image pair.

2.3. Experiments

The images were acquired as part of an IRB-approved study and consisted of 167 patients who exhibited a range of AMD severities. Both study eyes and fellow eyes were examined using standard imaging methods during study visits. CFP [TRC-50DX retinal camera (Topcon Medical Systems, Oakland, NJ)], FAF [Heidelberg Spectralis (Heidelberg Engineering, Heidelberg, Germany) with excitation wavelength of 488 nm and a barrier filter at 500 nm, 30 degrees] and IR images [Heidelberg Spectralis (Heidelberg Engineering, Heidelberg, Germany) with 795nm, 30 degrees] were acquired at baseline and four years from baseline.

Mono-modality, longitudinal registrations were performed among FAF-FAF, CFP-CFP, IR-IR modalities for images at baseline and 4-year visits. Multi-modality, cross-sectional registrations were performed among CFP-FAF, IR-FAF, CFP-IR modalities at baseline. CFP images were converted to grayscale to be consistent with inherently grayscale FAF, IR modalities. The

Biomedical Optics EXPRESS

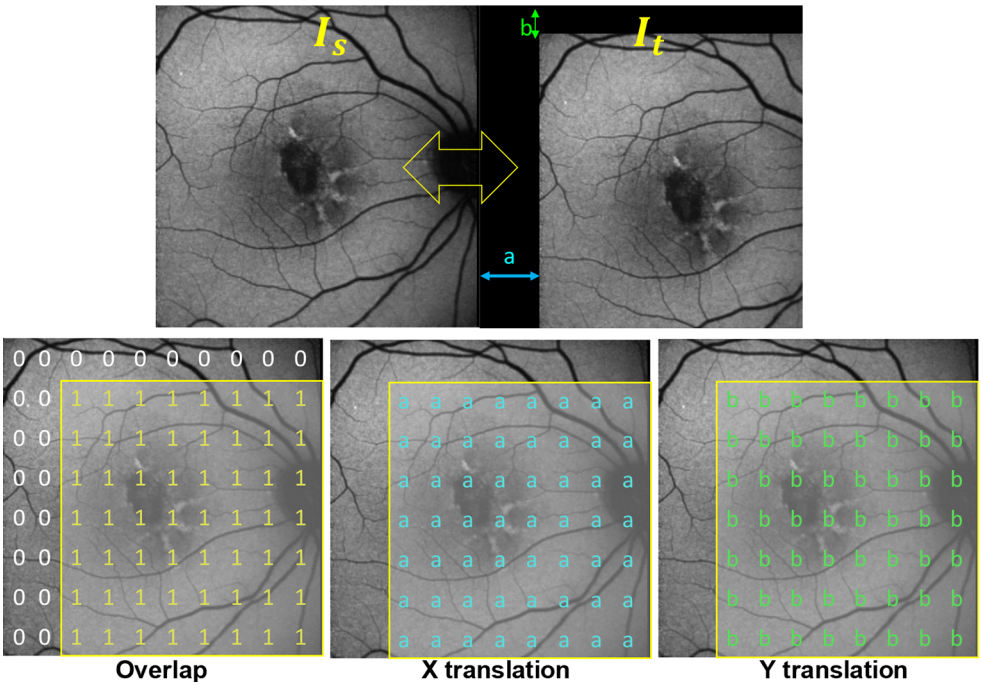

Fig. 3. Overlap and displacement prediction for a toy example with translation (t_x = a, t_y = b). Target image I_t is a translated version of the source image I_s. The three outputs of the network predicts the overlap (binary non-zero response at each overlapping patch) and x and y displacement at each patch relative to patch center. The translation error in this toy example is same for each patch and could be variable if rotation and scaling are involved.

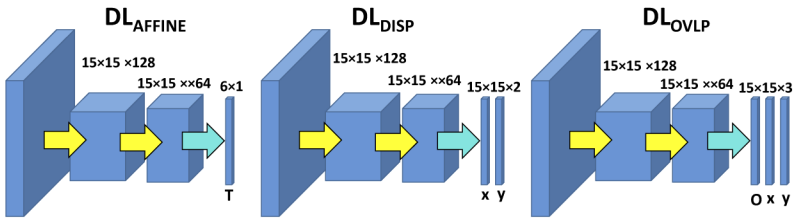

Fig. 4. Different regression networks within the deep-learning architecture evaluated in this work. DL_{AFFINE} method directly outputs the 6 parameter vector for the affine transformation, DL_{DISP} method outputs patch displacements in x and y directions and DL_{OVLP} method output patchwise region overlap in addition to the patch displacements.

Biomedical Optics EXPRESS

registrations were performed using both conventional and deep-learning based methods described in Sections 2.1 and 2.2. The ground truth for training deep learning methods was generated using the output affine transformation of the conventional feature-based method (i.e., C_{SIFT}) after qualitative verification and removal of failures. 1224 image pairs from 145 patients (90%) were included in the training set and 114 image pairs from 19 patients (9%) were included in the test set. 18 image pairs from 5 patients (1%) were used as a validation set to select the best trained model based on optimal validation error. There was no overlap of patients between training and test set images. All images were cropped and resampled to a fixed size 248 px × 248 px (37.4 µm × 37.4 µm pixel spacing) prior to training to enforce constant-size image input to the network.

Data augmentation was implemented using four different methods. First, each image pair was flipped horizontally to generate a synthetic right eye image from the patient's original left eye image and vice versa. Second, the same image pair was input twice to the network alternating each as a source (I_s) and target image (I_t). Third, aligned images were perturbed with a random transformation such that the network learns to solve a different misalignment for the same image pair at different training epochs. The introduced random transformations covered a range that were representative of the maximum misalignments exhibited in clinical images with translation ranging from -1.2 mm to +1.2 mm (± 32 px), rotation ranging from ± 0.1 rad, and scale ranging from ± 0.3. Finally, image pairs for all longitudinal and multi-modal registrations were pooled into a single training set. While increasing the number of images available for training, this enforces the feature extractors to generalize the learning of anatomical structures observed across all image modalities. Training was performed using Adam optimizer with learning rate=0.001, 1st momentum=0.9, 2nd momentum = 0.99, and batch size=16. Pre-trained (on ImageNet as available in PyTorch v1.1) VGG-16 feature extractors were used for initialization of the Siamese architecture. The network consisted of multiple stages with feature extraction, correlation matrix computation, and regression as illustrated in Fig. 2 and the network is trained based on a singular loss function that estimates the error of grid/patch displacements relative to ground truth. To facilitate backpropagation to all stages, consecutive optimization epochs changed between three settings. First, feature extraction stage weights were updated and the remainder of the network weights was frozen. Second, feature regression layer weights were updated and remainder of the network was frozen. In the third setting all weights of the network were unfrozen and updated to collectively train all stages of the network.

Both conventional feature-based and deep learning based registration were validated using ground truth obtained via manually identified anatomically homologous point pairs for all the test images whereas the images used for training deep learning methods was obtained via successful feature based registrations. During manual point identification, 666 conspicuous bifurcations of the vascular structures were annotated across 111 test image pairs. Target registration error (TRE) was computed as the average displacement between homologous point pairs after registration. Fiducial registration error (FRE) was calculated by performing the optimal registration relative to the manually identified points using Horn's method [27]. FRE yields a lower-bound on the expected registration error considering the limitations due to the additional parameters necessary to compute the transform (i.e., any perspective distortion and non-rigidity of the underlying transform results in high FRE) and the variability in manually annotating the point locations (i.e., error in point localization would reflect in high FRE). Fiducial localization error (FLE) was measured as the variability in repeatedly identifying 30 point landmarks manually in a subset of 5 images of the test set.

To evaluate the different implementations of regression networks ($\mathbf{DL_{AFFINE}}$, $\mathbf{DL_{DISP}}$, $\mathbf{DL_{OVLP}}$), each was trained separately, and the accuracy was quantified in the test set. To compare the performance of deep-learning based methods with the conventional method, all test image pairs were also registered using $\mathbf{C_{SIFT}}$ method. To explore the robustness of conventional and deep learning algorithms using the limited set of test images with manually-annotated ground

Biomedical Optics EXPRESS

truth landmarks, test image pairs were registered repeatedly after misaligning with random transformations in a separate experiment. The magnitude of the random transformation was increased systematically to increasingly challenge the registration algorithm with larger initial misalignments. The accuracy of the registration was analyzed as a function of the magnitude of randomly introduced misalignment.

To augment the training data set, we pooled multi-modality imaging together in our experiments. Alternatively, training can be performed separately to solve each of the six registration problems involving different combinations of image modalities (e.g., training a network to solve FAF-CFP registration using only FAF-CFP images in the training set). To evaluate whether the network can benefit from separately solving a specific registration problem, 100 additional training epochs was performed to the deep-learning model by only providing training images relevant to a specific registration.

The number of layers in the feature extraction (N_s), and regression (N_R),) networks of the deep-learning method were important hyperparameters that could affect registration accuracy and performance. Separate networks were trained after varying the output of the extracted features (N_s) within a range 49-961 (corresponding to pool3-pool5 layers in VGG) and the number of regression network layers in a range 1-4 and the accuracy was quantified for each network.

3. Results

3.1. Performance evaluation of conventional and deep-learning registration methods

Figure 5 shows violin plots for registration error distributions comparing the conventional registration method, $\mathbf{C_{SIFT}}$, and different approaches to deep-learning regression networks evaluated in this work (i.e., $\mathbf{DL_{AFFINE}}$, $\mathbf{DL_{DISP}}$, and $\mathbf{DL_{OVLP}}$) for different mono-modality longitudinal and multimodal applications. Table 1 summarizes the mean ± standard deviation (std) error for each registration problem evaluated in this study. When comparing mono-modality, longitudinal registrations among FAF-FAF, CFP-CFP, IR-IR images, conventional feature-based method demonstrated mean errors 39-53 µm (for different modalities), substantially improving misalignment prior to registration that ranged from 246-674 µm. On the other hand, $\mathbf{DL_{AFFINE}}$ method exhibited suboptimal mean errors in the range 105-174 µm, even though it showed an improvement relative to the initial misalignment. $\mathbf{DL_{DISP}}$ and $\mathbf{DL_{AFFINE}}$ methods showed successful and comparable registration performance with mean errors ranging from 53-60 µm and 54-59 µm respectively. While $\mathbf{C_{SIFT}}$ method exhibited the best overall geometric accuracy for mono-modality registration, $\mathbf{DL_{DISP}}$ and $\mathbf{DL_{OVLP}}$ methods showed acceptable registration performance.

Table 1. Mean ± std errors for different registration methods evaluated in the study.

Reg. Error (mean ± sth)	FAF-FAF	CFP-CFP	IR-IR	CFP-FAF	FAF-IR	CFP-IR
Before (µm)	331 ± 242	674 ± 291	246 ± 168	933 ± 213	311 ± 286	938 ± 237
C_{SIFT} (µm)	39 ± 15	53 ± 11	48 ± 10	>5000	>5000	>5000
DL_{AFFINE} (µm)	120 ± 91	105 ± 52	174 ± 118	203 ± 115	398 ± 262	267 ± 157
DL_{DISP} (µm)	53 ± 16	60 ± 18	60 ± 11	71 ± 36	89 ± 66	79 ± 43
DL_{OVLP} (µm)	54 ± 12	54 ± 12	59 ± 12	66 ± 31	69 ± 29	68 ± 20
FRE	37 ± 8	31 ± 16	42 ± 12	47 ± 27	48 ± 17	41 ± 13

For multi-modality, cross-sectional registrations among FAF-CFP, FAF-IR, and CFP-IR modalities, $\mathbf{C_{SIFT}}$ method frequently exhibited gross failures where the solution resulted in an error larger than the initial misalignment in more than 50% of the cases. The failures could occur due to spurious matches resulting from suboptimal image quality and/or insufficient number of reliable features. $\mathbf{DL_{OVLP}}$ method demonstrated the best geometric accuracy with mean

Biomedical Optics EXPRESS

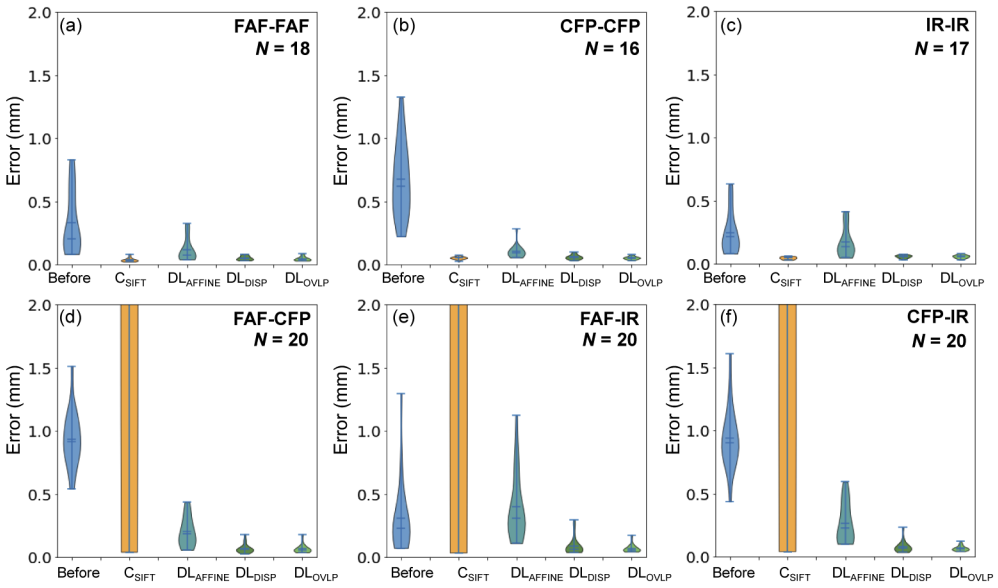

Fig. 5. Comparison of error distributions prior to registration, conventional method, C_{SIFT}, and deep learning approaches with different regression networks DL_{AFFINE}, DL_{DISP}, and DL_{OVLP}. The plots evaluate the performance for mono-modality longitudinal registrations (a) FAF-FAF, (b) CFP-CFP, and (c) IR-IR and multi-modality registrations (d) FAF-CFP, (e) FAF-IR, and (f) CFP-IR. Along the center line of each distribution plot, the mean, median, and error bars representing the data range are shown. The number of test image pairs (N) used to generate error distributions are shown in the top right corner of each plot.

errors ranging from 66-69 μm, substantially improving the initial misalignments that ranged from 311-938 μm. **DL_{AFFINE}** method again exhibited suboptimal errors with means ranging from 203-398 μm whereas **DL_{DISP}** method showed markedly improved performance with errors ranging from 71-89 μm. Whereas conventional method was susceptible to gross failures, **DL_{OVLP}** method yielded the best overall performance across all mono-modality and multi-modality registrations. FRE, (i.e., the registration error after aligning with ground truth fiducials) ranged from 31-48 μm for different registrations as shown in Table 1. Thus, the accuracy of **DL_{OVLP}** falls within a comparable range to the accuracy captured in FRE. The FLE was measured to be 29.7 + 12.8 μm (mean ± std) from repeatedly identifying the same set of landmarks five times and represents the uncertainty associated with the manual feature identification process. Figure 6 shows image registration results for representative cases showing performance of the **DL_{OVLP}** method qualitatively. In mono-modality, longitudinal cases, registration overcame mismatch due to structural changes resulting from disease progression. In multi-modality, cross-sectional cases, registration successfully aligned anatomy overcoming image intensity mismatch in different modalities. Considering the visual quality of the alignment and the magnitude of variability of the validation process captured by FRE/ FLE, **DL_{OVLP}** method demonstrated successful performance across all registrations evaluated in this study. Figure 7 illustrates representative cases for suboptimal registration outputs from both conventional (C_{SIFT}) and deep-learning (DL_{OVLP}) methods. Conventional method usually exhibited conspicuous failures resulting from erroneous feature matching. Deep learning method on the other hand exhibited subtle failures closer to the solution.

Deep learning methods with different regression networks, **DL_{AFFINE}**, **DL_{DISP}**, and **DL_{OVLP}**, showed monotonic improvements over the initial misalignment in multi-modality registrations.

Biomedical Optics EXPRESS

Predicting patch displacements from the regression network in $\mathbf{DL_{DISP}}$ was more accurate than directly predicting the affine transform parameters that lie in a non-linear space in $\mathbf{DL_{AFFINE}}$. The output of the $\mathbf{DL_{DISP}}$ method was displacements of patch centers in x and y directions that

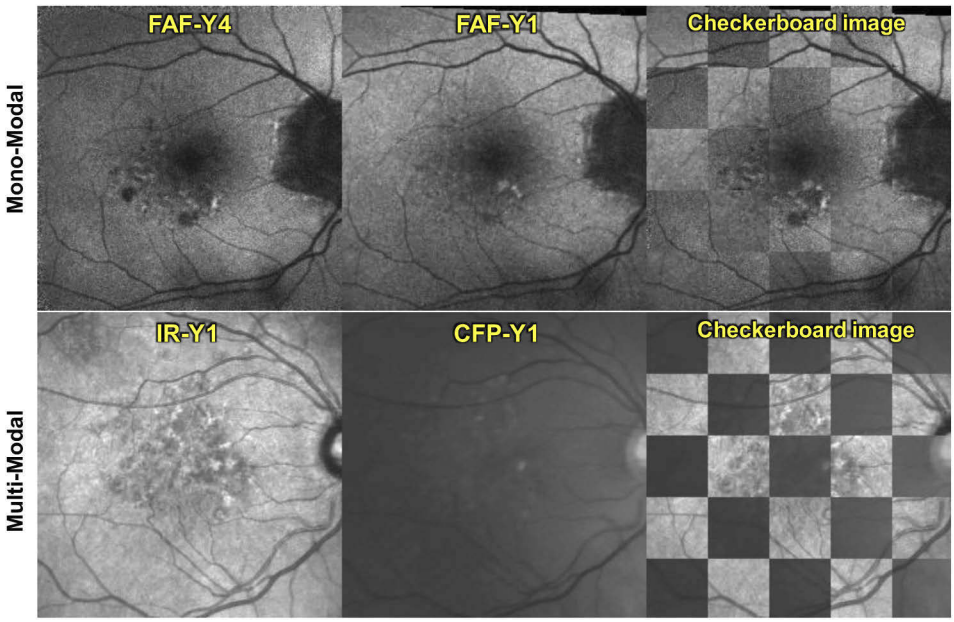

Fig. 6. (Top row) Qualitative assessment of mono-modality (FAF) registration of longitudinal images in the presence of mismatch due to the growth of atrophic regions using DL_{OVLP} method. (Bottom row). Qualitative assessment of multi-modality (IR and CFP) registration of cross-sectional images using DL_{OVLP} method.

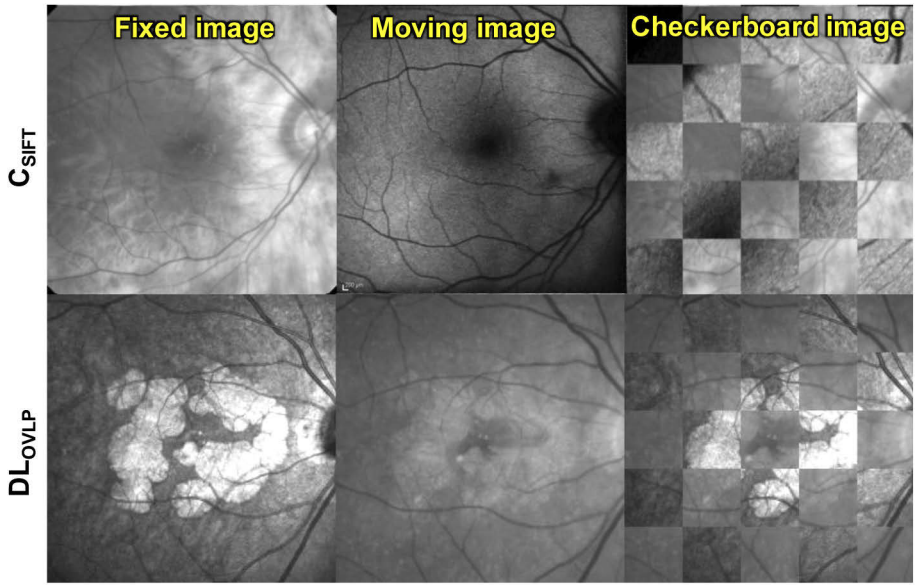

Fig. 7. Suboptimal registration output from C_{SIFT} (top row) and DL_{OVLP} (bottom row) methods.

Biomedical Optics EXPRESS

was used to estimate the affine transformation. The errors of the individual patch displacements in $\mathbf{DL_{DISP}}$ output compared to ground truth were found to be 47 ± 45 µm and 53 ± 64 µm in x and y directions respectively for CFP-FAF registration. The direct outputs of the $\mathbf{DL_{OVLP}}$ method are binary variables indicating the overlap of the patches in addition to the displacement errors in x and y directions as in $\mathbf{DL_{DISP}}$. For the binary classification task of overlap prediction, $\mathbf{DL_{OVLP}}$ method exhibited accuracies with precision $= 0.99 \pm 0.01$, recall 0.97 ± 0.02, and F1 score 0.97 ± 0.01. The displacement prediction task in $\mathbf{DL_{OVLP}}$ method exhibited accuracy of 45 ± 43 µm and 37 ± 28 µm for x and y displacements respectively. Thus, the overlap was successfully predicted using the network, which helped to incrementally improve the accuracy of the registration compared to $\mathbf{DL_{DISP}}$.

Figure 8 shows the convergence of the loss function during training for evaluated regression networks $\mathbf{DL_{AFFINE}}$, $\mathbf{DL_{DISP}}$, and $\mathbf{DL_{OVLP}}$. To facilitate comparison of the loss functions of different networks, the loss was normalized using the training loss value after the first epoch of each network. $\mathbf{DL_{DISP}}$ and $\mathbf{DL_{OVLP}}$ exhibited less fluctuations in training loss as the number of epochs increased with more desirable convergence properties. Some of the training fluctuations are attributable to the changing optimization strategy by freezing different parts of the network in subsequent epochs. However, this strategy empirically yielded better validation error after convergence for all regression networks. While the test loss exhibited more fluctuations compared to training loss, with the increasing number of epochs $\mathbf{DL_{OVLP}}$ showed a rapid decrease in test loss indicating better generalization of the model during training. Thus, $\mathbf{DL_{OVLP}}$ also exhibited desirable convergence during training.

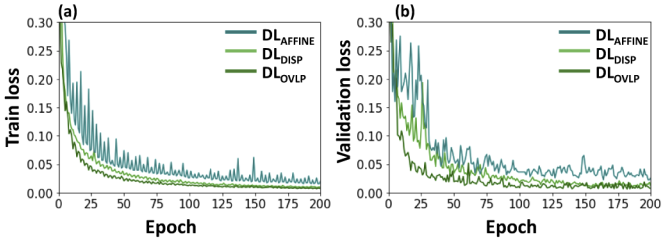

Fig. 8. Convergence properties of the loss function during optimization iterations for (a) training and (b) validation data.

3.2. Performance evaluation in terms of robustness to initialization error

To evaluate the robustness of registrations and estimate the operating range as a function of initialization error, we performed registrations after introducing random transformations for test images in each multi-modality and mono-modality registration problem. Each test image was transformed with randomly generated ranges to increase the initial misalignment and monotonically challenge the registration algorithm. Figure 9 plots the median registration errors with upper quartiles (upward pointing arrow) as a function of the initial misalignment. $\mathbf{DL_{OVLP}}$ method showed consistent registration behavior across all registration problems whereas $\mathbf{C_{SIFT}}$ demonstrated superior performance with large operating ranges for mono-modality registrations and exhibiting majority gross failures in multi-modality registrations. It is also interesting to observe that the $\mathbf{DL_{OVLP}}$ method degraded performance with increasing misalignment. However, it showed robust performance approximately within the range of misalignments that it was trained on and suboptimal accuracy was observed with large initial misalignments >1.5 mm unobserved during training.

Biomedical Optics EXPRESS

3.3. Performance benefit for separately training for each registration

In deep learning based registration, we pooled all image modalities in the training set to learn a generalized model to solve multiple registrations involving multiple modalities using a single trained network. In order to analyze, whether the accuracy of the model was compromised when solving for multiple modalities simultaneously in a single network, we trained the model for 100 additional epochs using only training examples involving CFP-FAF registration. Figure 10 shows the performance comparison with and without this additional training step that guided the model to solve a specific registration problem. Clinically and statistically insignificant improvement was observed with 100 additional epochs of training using only CFP-FAF images showing an error = 63 ± 35 μm compared to an error = 66 ± 37 μm without the additional training step. This indicates the network was able to generalize well without compromising on accuracy when multi-modality images were included in the training set.

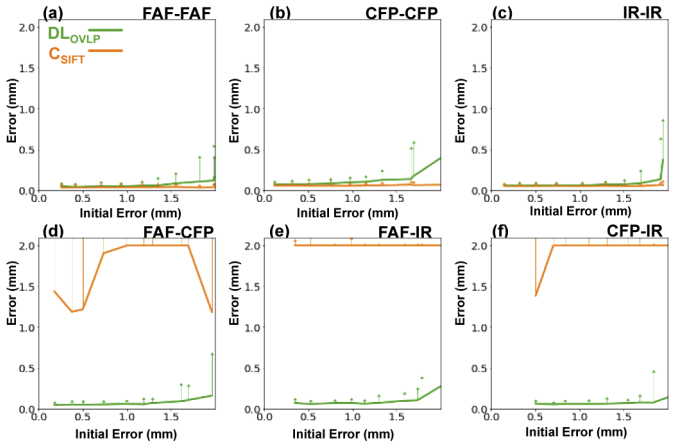

Fig. 9. Median registration error (solid line) and upper quartiles (upward arrow) for different initial errors after randomly perturbing the images to introduce misalignment prior to registration, C_{SIFT} (orange) and DL_{OVLP} (green) are compared for mono-modality registrations of (a) FAF-FAF, (b) CFP-CFP, and (c) IR-IR and multi-modality registrations of (d) CFP-FAF, (e) IR-FAF, and (f) CFP-IR.

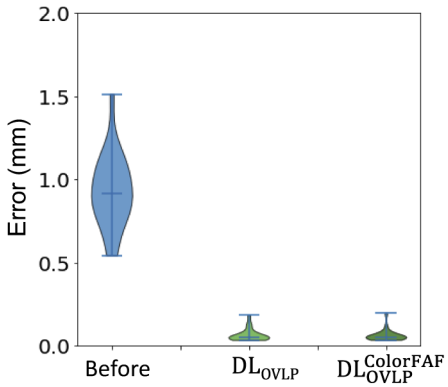

Fig. 10. Comparison of performance of networks that trained on multiple image modalities pooled together (DL_{OVLP}) versus that additionally trained on a specific modality-pair (DL_{OVLP}^{CFPFAF}).

Biomedical Optics EXPRESS

3.4. Hyperparameter sensitivity of the deep-learning architecture

Figure 11 shows registration performance sensitivity to different hyperparameters used in the deep learning network. In terms of the feature extraction network with the Siamese architecture, the performance was compared in Fig. 11(a) after extracting features at different stages of the VGG network. Extracting features at pool4 layer in VGG with $N_s = 225$ yielded the optimal performance in terms of accuracy. There is a tradeoff in selecting the feature extraction layer too shallow or too deep from the VGG network. N_s=961, corresponding to pool3 feature extraction layer in VGG, improved the resolution of the correlation matrix. However, it has a shallow VGG feature extraction network that limits the parameters learned in the feature descriptor. The decrease in performance could be attributed to this decrease in feature descriptor length. VGG pool5 layer with N_s=49 on the other hand possesses a deeper VGG network, but it results in a coarse correlation matrix and increases the patch size of the feature descriptor, thus degrading accuracy relative to N_s=225. In terms of the regression network, the number of layers (NR) was increased from 1 to 4 and the performance was compared in Fig. 11(b). N_R>2 exhibited comparable accuracy without diminishing accuracy with the increasing number of layers. $N_R = 3$ was used in the experiments of this paper considering the computation and memory requirements with increasing number of layers.

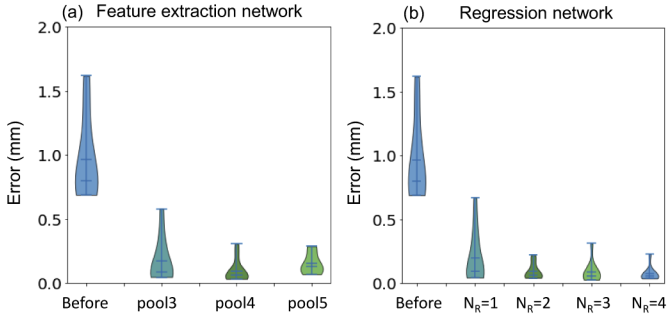

Fig. 11. Evaluating network hyperparameters for feature extraction and regression networks.

4. Discussion and conclusions

The current study evaluated two different approaches to feature-based registration using conventional and deep learning-based methods. The conventional method accurately registered mono-modality, longitudinal images, whereas it exhibited gross failures often driving the solution away from the initial misalignment due to spurious feature matching when registering multi-modality images. The deep learning method that mirrored the architecture from conventional feature detection and matching workflow was able to achieve accurate registration performance after training on successful conventional method registration cases. $\mathbf{DL_{OVLP}}$ method that predicted the region overlap between source and target images in addition to image patch displacement yielded the best overall accuracy and robustness with mean errors ranging from 54-69 μm across all registrations.

The conventional feature-based method relies upon a feature localizer that samples potential candidate locations from the image and a feature descriptor that captures image-intensity variations around a local neighborhood. For mono-modality registration the alignment was achieved accurately by finding robust feature matches from the sampled candidate locations in the image pair. When applying $\mathbf{C_{SIFT}}$ in a cross-modality setting, the feature descriptor needs to be robust to find matches overcoming variable image-intensity mappings across modalities. While previous research [10] aims to develop robust descriptors and image similarity metrics to

Biomedical Optics EXPRESS

achieve robust multi-modality registration, it poses different challenges specific to the involved modalities and often require customization for each registration with a different modality pairing. Deep learning methods could be more desirable in this context, since the laborious descriptor customization process can be replaced by learning a feature representation generalizable across modalities using example images provided during training. The single model that trained on all three modalities evaluated in this work achieved accurate and robust registration indicating the success in generalization of the feature description and matching process involved in multiple registration tasks. Figure 12 shows successful registration performance with two additional modalities [IR autofluorescence (IRAF) and indocyanogreen (ICG) angiography] not encountered during training. The performance can be improved if necessary, by devising a continuous learning approach where training images from additional modalities are incorporated in the model without requiring custom algorithm modifications. Such amenability in solving versatile registration problems within a single framework is desirable in clinical translation.

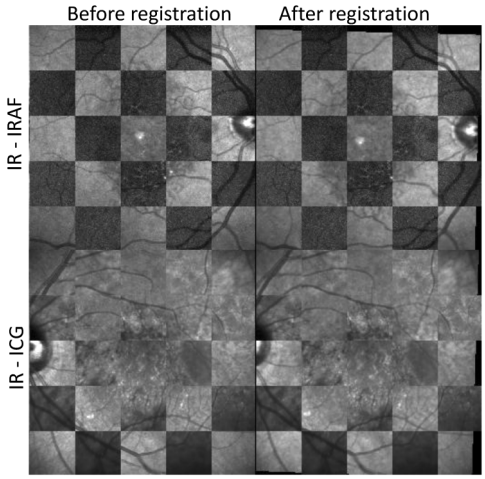

Fig. 12. Successful registration of additional modalities (IRAF, ICGA) not encountered during training

The overall registration error of $\mathbf{DL_{OVLP}}$ was within a reasonable range (17-27 μm) relative to FRE that represents the error after alignment using the manually identified fiducials that provided ground truth to assess different registration methods. FRE captures both the variability of the fiducial localization process and any residual error after aligning with an optimal affine transformation. Multi-modality FREs (41-48 μm) were observed to be slightly larger than those for mono-modality (31-42 μm), which could be attributable to larger perspective distortions in multi-modality images not fully compensated with an affine registration. For mono-modality registration $\mathbf{C_{SIFT}}$ method achieved accuracy comparable to FRE. Since the feature localization process in $\mathbf{C_{SIFT}}$ is more precise and repeatable, the feature localization variability could be lower than the manual process used to generate ground truth fiducials. Thus, the $\mathbf{C_{SIFT}}$ method provided a good alternative to generate a large ground truth data set for training deep learning methods compared to the labor-intensive alternative to ground truth with manually identified fiducials. Regardless of the method (manual or $\mathbf{C_{SIFT}}$) used to generate ground truth training data, its associated variability and error could pose a fundamental limitation to the accuracy attainable with deep learning methods. Thus, errors of successful $\mathbf{C_{SIFT}}$ registrations may provide a lower bound on the expected error that could be realized with a deep-learning method trained on successful $\mathbf{C_{SIFT}}$ data.

Biomedical Optics EXPRESS

We observed monotonic increase in accuracy when comparing the performance of different regression networks evaluated in this study. Computing affine transformation from the regression network directly hides the complexity of the computations involved in estimating the output affine parameters in a set of convolution layers. Specifically, we addressed the complexity of the regression network in estimating regression parameters due to patch-based correlation matrix, non-linearity of the affine transform parameter space, and presence of non-overlapping physical spaces. The improvements observed in $\mathbf{DL_{DISP}}$ and $\mathbf{DL_{OVLP}}$ methods suggest that deep-learning networks can benefit from guiding the computations in convolution layers to achieve the desired output and validating the output of the network architecture with relevant ground truth data.

The use of affine models for end-to-end registration has inherent limitations in correcting for possible additional distortions resulting from the image acquisition process and/or anatomical changes/patient positioning. Perspective or deformable models could further improve the accuracy of aligning anatomy in longitudinally acquired multi modal images. The presented model is amenable to be extended to incorporate deformation by increasing the number of parameters and fitting a deformable model to the resulting patch displacement field. In this work we observed gross misalignment was compensated to a large extent by the affine registration model and the measured error (54-69 µm) was approaching the variability of the manual fiducial identification process (FRE=31-48 µm FLE = ~30 µm). Deformable models could improve the results further by correcting any residual errors subsequent to affine registration as shown qualitatively in [28] in subsequent to rigid registration. However, generating ground truth for training and quantitative validation of such high precision models would be challenges that need to be addressed. Additionally, when selecting the appropriate model the clinical objectives need also be considered, since non-rigid registration can diminish anatomical changes in some cases (e.g., GA growth in longitudinal images).

The performance of the $\mathbf{C_{SIFT}}$ method was less sensitive to the initial misalignment in that it had a wide operating range solving for accurate alignment from large initial misalignments for mono-modality registration whereas it did not achieve robust performance even for small initial misalignments for multi-modality registration. The ability to accurately identify feature locations and find matching features in the image pair allows for robust performance insensitive to the magnitude of the initial misalignment. Having a wide-operating range is a desirable property in feature-based registration compared to other image-based registration algorithms [29,30] that are susceptible to local optima and have limited capture ranges due to the use of iterative local optimization from initialization. Deep learning architecture that resembled the conventional feature detection/matching would ideally learn a generalized model to solve accurate registration irrespective of the initial misalignment. However, it was interesting to observe that $\mathbf{DL_{OVLP}}$ method accuracy degraded for initial misalignment > 1.5 mm which represented larger misalignments that were not encountered during training. This could be due to the regression network only outputting displacements in the range observed from the training data. While this performance meets the clinical requirement in successfully solving practically encountered misalignments in clinical imaging, there are dependencies on the training data that limits the generalizability of the deep learning method that need to be considered accordingly for different clinical application settings.

The experiments performed in this work used the pixel spacing provided by the imaging system to measure geometric errors before and after registration. The accuracy thus measured in the imaging plane may differ from the actual anatomical measurements in the physical space up to the camera magnification factor. We also assume isotropic distribution of errors in the 2D imaging plane, whereas in reality, due to the ellipsoidal shape of the 3D eye anatomy the accuracy could degrade away from the image center line.

The reported accuracy of $\mathbf{DL_{OVLP}}$ method in this work provides an automatic method to align cross-modality and longitudinal images frequently encountered in retinal imaging. This allows

Biomedical Optics EXPRESS

for mapping longitudinal images into a common physical space for qualitative monitoring as well as quantitative measurement of changes. The required accuracy for clinical translation in part depends on the smallest anatomical structures of clinical significance that need to be monitored longitudinally. Considering drusen-like structures that are <100 μm, the accuracy reported in the paper provides an important advancement to support large clinical studies that monitor the longitudinal progression of AMD as well as other retinal diseases.

Funding

National Eye Institute Intramural Research Program, National Institutes of Health (EY000509-10).

Disclosures

The authors declare no conflicts of interest.

References

1. D. S. Friedman, B. J. O'Colmain, B. Muñoz, S. C. Tomany, C. McCarty, P. T. V. M. DeJong, B. Nemesure, P. Mitchell, J. Kempen, and N. Congdon, "Prevalence of Age-Related Macular Degeneration in the United States," Arch. Ophthalmol. (2004).
2. T. Chanwimaluang, G. Fan, and S. R. Fransen, "Hybrid Retinal Image Registration," IEEE Trans. Inform. Technol. Biomed. **10**(1), 129–142 (2006).
3. J. Zheng, J. Tian, K. Deng, X. Dai, X. Zhang, and M. Xu, "Salient Feature Region: A New Method for Retinal Image Registration," IEEE Trans. Inform. Technol. Biomed. **15**(2), 221–232 (2011).
4. C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, and A. A. Argyros, "Retinal image registration under the assumption of a spherical eye," Comput. Med. Imaging Graph. **55**, 95–105 (2017).
5. J. Parekar, P. Porwal, and M. Kokare, "Automatic retinal image registration using fully connected vascular tree," in *2016 International Conference on Signal and Information Processing (IConSIP)* (IEEE, 2016), pp. 1–5.
6. H. Tang, A. Pan, Y. Yang, K. Yang, Y. Luo, S. Zhang, and S. H. Ong, "Retinal Image Registration Based on Robust Non-Rigid Point Matching Method," J. Med. Imaging Heal. Informatics **8**(2), 240–249 (2018).
7. C. V. Stewart, C.-L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," IEEE Trans. Med. Imaging **22**(11), 1379–1394 (2003).
8. R. Ramli, M. Y. I. Idris, K. Hasikin, N. K. A. Karim, A. W. Abdul Wahab, I. Ahmedy, F. Ahmedy, N. A. Kadri, and H. Arof, "Feature-Based Retinal Image Registration Using D-Saddle Feature," J. Healthc. Eng. **2017**, 1–15 (2017).
9. Z. Li, F. Huang, J. Zhang, B. Dashtbozorg, S. Abbasi-Sureshjani, Y. Sun, X. Long, Q. Yu, B. H. ter Romeny, and T. Tan, "Multi-modal and multi-vendor retina image registration," Biomed. Opt. Express **9**(2), 410 (2018).
10. Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, and E. Fatemizadeh, "An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors," Eurasip J. Image Video Process. (2013).
11. G. Wang, Z. Wang, Y. Chen, and W. Zhao, "Robust point matching method for multimodal retinal image registration," Biomed. Signal Proces **19**, 68–76 (2015).
12. 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," Med. Image Anal. (2019).
13. S. Wang, M. Kim, G. Wu, and D. Shen, "Scalable High Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning," in Deep Learning for Medical Image Analysis (2017).
14. F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," npj Digit. Med. (2019).
15. P. Burlina, D. E. Freund, N. Joshi, Y. Wolfson, and N. M. Bressler, "Detection of age-related macular degeneration via deep learning," in *Proceedings - International Symposium on Biomedical Imaging* (2016).
16. J. Lee, P. Liu, J. Cheng, and H. Fu, "A deep step pattern representation for multimodal retinal image registration," in *Proceedings of the IEEE International Conference on Computer Vision* (2019).
17. M. Arikan, A. Sadeghipour, B. Gerendas, R. Told, and U. Schmidt-Erfurt, "Deep Learning Based Multi-modal Registration for Retinal Imaging," in (2019), pp. 75–82.
18. S. Miao, Z. J. Wang, and R. Liao, "A CNN Regression Approach for Real-Time 2D/3D Registration," IEEE Trans. Med. Imaging **35**(5), 1352–1363 (2016).
19. X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **6**(3), 248–252 (2018).
20. A. Roche, X. Pennec, G. Malandain, and N. Ayache, "Rigid registration of 3-D ultrasound with MR images: a new approach combining intensity and gradient information," IEEE Trans. Med. Imaging **20**(10), 1038–1049 (2001).
21. X. Cao, J. Yang, Y. Gao, Q. Wang, and D. Shen, "Region-Adaptive Deformable Registration of CT/MRI Pelvic Images via Learning-Based Image Synthesis," IEEE Trans. Image Process. **27**(7), 3500–3512 (2018).
22. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A Learning Framework for Deformable Medical Image Registration," IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019).

Biomedical Optics EXPRESS

23. I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2017 (2017).
24. X. Cao, J. Fan, P. Dong, S. Ahmad, P.-T. Yap, and D. Shen, "Image registration using machine and deep learning," in *Handbook of Medical Image Computing and Computer Assisted Intervention* (Elsevier, 2020), pp. 319–342.
25. T. S. De Silva, N. Hotaling, E. Y. Chew, and C. Cukras, "Feature-based retinal image registration for longitudinal analysis of patients with age-related macular degeneration," in *Medical Imaging 2020: Image Processing*, B. A. Landman and I. Išgum, eds. (SPIE, 2020), p. 106.
26. M. Grabner, H. Grabner, and H. Bischof, "Fast approximated SIFT," in *Lecture Notes in Computer Science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2006).
27. B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," J. Opt. Soc. Am. A (1987).
28. H. Rabbani, M. J. Allingham, P. S. Mettu, S. W. Cousins, and S. Farsiu, "Fully automatic segmentation of fluorescein leakage in subjects with diabetic macular edema," Investig. Ophthalmol. Vis. Sci. (2015).
29. T. De Silva, D. W. Cool, J. Yuan, C. Romagnoli, J. Samarabandu, A. Fenster, and A. D. Ward, "Robust 2-D-3-D Registration Optimization for Motion Compensation during 3-D TRUS-Guided Biopsy Using Learned Prostate Motion Data," IEEE Trans. Med. Imaging **36**(10), 2010–2020 (2017).
30. T. De Silva, A. Uneri, M. D. Ketcha, S. Reaungamornrat, G. Kleinszig, S. Vogt, N. Aygun, S.-F. Lo, J.-P. Wolinsky, and J. H. Siewerdsen, "3D–2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch," Phys. Med. Biol. **61**(8), 3009–3025 (2016).