Computers in Biology and Medicine 140 (2022) 105101

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Color fundus image registration using a learning-based domain-specific landmark detection methodology

David Rivas-Villar [a,b,*], Álvaro S. Hervella [a,b], José Rouco [a,b], Jorge Novo [a,b]

[a] Centro de investigacion CITIC, Universidade da Coruña, 15 071, A Coruña, Spain
[b] Grupo VARPA, Instituto de Investigacion Biomédica de A Coruña (INIBIC), Universidade da Coruña, 15 006, A Coruña, Spain

ARTICLE INFO

Keywords:
Color fundus images
Medical image registration
Medical imaging
Deep learning

ABSTRACT

Medical imaging, and particularly retinal imaging, allows to accurately diagnose many eye pathologies as well as some systemic diseases such as hypertension or diabetes. Registering these images is crucial to correctly compare key structures, not only within patients, but also to contrast data with a model or among a population. Currently, this field is dominated by complex classical methods because the novel deep learning methods cannot compete yet in terms of results and commonly used methods are difficult to adapt to the retinal domain. In this work, we propose a novel method to register color fundus images based on previous works which employed classical approaches to detect domain-specific landmarks. Instead, we propose to use deep learning methods for the detection of these highly-specific domain-related landmarks. Our method uses a neural network to detect the bifurcations and crossovers of the retinal blood vessels, whose arrangement and location are unique to each eye and person. This proposal is the first deep learning feature-based registration method in fundus imaging. These keypoints are matched using a method based on RANSAC (Random Sample Consensus) without the requirement to calculate complex descriptors. Our method was tested using the public FIRE dataset, although the landmark detection network was trained using the DRIVE dataset. Our method provides accurate results, a registration score of 0.657 for the whole FIRE dataset (0.908 for category S, 0.293 for category P and 0.660 for category A). Therefore, our proposal can compete with complex classical methods and beat the deep learning methods in the state of the art.

1. Introduction

Registration consists in aligning a pair of images whose content is coincident (in part), but under different imaging viewpoints. Each image pair consists of a fixed image, which is used as reference, and a moving image that is transformed (or deformed) to match the fixed image, with the same contents appearing in the same locations after the registration. The registration of medical images presents numerous applications in clinical practice, playing an important role in common processing pipelines for medical image analysis [1]. In particular, registration facilitates the simultaneous analysis of several images, allowing the clinicians to draw better conclusions using more data [2]. It also allows the comparison of images that were taken at different time frames, which helps to monitor the progression of a disease and perform longitudinal studies [3]. Additionally, image registration is also useful for aligning the images with a model representing the candidate disease, helping to provide the correct diagnosis and treatment.

The development of automated image registration methods is especially important for computer-aided diagnosis (CAD) systems, that cannot rely on the manual registration of the images. This is due to the time and effort that a clinical expert would have to invest in the manual procedure. The inclusion of automated registration methods in CAD pipelines facilitates the analysis of multiple imaging modalities [4], and even allows the improvement of the quality and resolution of the images [5]. Thus, the availability of accurate registration algorithms can play an important role in the improvement of future CAD systems. Current state of the art in CAD development is dominated by the use of deep learning techniques [6,7]. These techniques have allowed to achieve outstanding performances and robustness in novel and challenging medical image analysis tasks. But, more importantly, deep learning approaches allowed to train systems end-to-end, from the raw data to the expected decisions, without the need of ad-hoc pre-processing or feature engineering. This

* Corresponding author. Centro de investigacion CITIC, Universidade da Coruña, 15 071, A Coruña, Spain.
E-mail addresses: david.rivas.villar@udc.es, david.rivas.villar@udc.es (D. Rivas-Villar), a.suarezh@udc.es (Á.S. Hervella), jrouco@udc.es (J. Rouco), jnovo@udc.es (J. Novo).

https://doi.org/10.1016/j.compbiomed.2021.105101
Received 30 July 2021; Received in revised form 29 November 2021; Accepted 29 November 2021
Available online 3 December 2021

enables flexibility and adaptability of methods to the constantly evolving imaging devices and modalities, along with adjusting to challenging conditions, as those derived from the evolution of pathological lesions. These desirable features of deep learning, in contrast to classical methods, should be also pursued and explored in the development of automatic image registration methods.

The application of image registration methods to retinal image analysis is of special relevance. The eyes are the only organs in the human body that allow non-invasive *in vivo* observation of the blood vessels and neuronal tissue. Furthermore, the retinal imaging techniques, like color fundus retinography, are very common and cost-effective. In clinical practice, these images are often employed to help to diagnose several diseases like Age-Related Macular Degeneration (AMD) or Diabetic Retinopathy (DR), among others. However, many challenges still persist in the registration of color fundus images. For instance, due to the photographic nature of the images, they are subject to multiple variations. These include, but are not limited to, spatial displacements due to movements of the patient or imperfect placement of the machine and the subject, illumination changes, etc. Furthermore, some diseases trigger processes like neovascularization, hemorrhages, drusen or edema that can substantially alter the appearance of the retina. Therefore, morphological changes in the structures of the eyes due to the progression or remission of diseases are also very common.

Existent registration methods can be broadly divided in two groups by considering the type of deformation they apply to the images. These groups are rigid and elastic deformation methods [8,9]. The rigid methods only consider the deformations of rigid bodies under varying imaging views while the elastic models also consider the possible deformations of the imaged objects. Rigid registration methods can use different transformation models that are characterized by their number of parameters and their complexity. An intuitive comparison of the different transformations can be seen in Fig. 1. In increasing order of complexity: rigid body transformations, or Euclidean transformations, only allow 2D translation and rotation (3 Degrees of Freedom, DOF); similarity transformations add isotropic scaling (4 DOF); and affine transformations add shearing (6 DOF). Finally, projective transformations, contrary to previous transformations, do not preserve parallelism as they operate in the projective space (8 DOF) [8,10].

The elastic registration methods can also allow local deformations of the images using additional parameters. These models are appropriate for certain organs or parts of the human body that are subject to size or shape changes. For instance, common movements, or body positioning during the imaging, can alter the shape of some organs, which motivates the wide use of elastic models in the field of medical image registration [9].

As each transformation is more complex and requires more parameters, more point matches among image pairs are needed for a method to accurately create a suitable deformation model. Therefore, the advantage of the simpler models is that, although the produced transformation is simpler, they require less reference points. Elastic models require many more parameters to be optimized (in the order of tenths of thousands [9]) when compared to rigid deformation methods (eight at most). However, using a simpler transformation model than the one required to model the transformation between the images in each registration pair will result in poor performance. The images will be approximately registered, but with a low accuracy due to the lack of power of the used transformation model. Conversely, using excessively complex transformations, allowing for DOFs that are not present in image pairs, results in a more complex optimization process and an increase in the risk of overfitting. Thus, the selection of the transformations is an important step and should be carefully selected for each topic. In terms of retinal fundus images, the expected displacements are due to viewpoint variations, and usually not manifesting tissue deformations, therefore the rigid transformations are commonly the most appropriate. However, state of the art methods use varied transformation functions, depending on the particular goal. Some examples are quadratic models to account for FOV (Field Of View) deformation of the eyes [11], affine transformation [12] or even elastic transformations [13].

Registration approaches can also be classified according to the methods they use to register images. In this regard, there are two groups, classical-methods and the novel deep learning methods. Additionally, classical-methods can be divided into two categories: intensity-based (IBR) and feature-based (FBR) [14].

IBR methods use similarity metrics to maximize the matching between the intensity values of the fixed and moving images being registered. These methods use gradient-descent or similar approaches to optimize the similarity over the considered transformation parameter space. Commonly used metrics include mutual information (MI) [15], normalized cross correlation (NCC) [16], mean squared difference (MSD) [17], etc.

On the other hand, FBR methods use points of interest, called landmarks, to guide the matching between the images to be registered. The overall idea is to find a maximal set of correspondences between the landmarks in both images that uniquely characterize a valid transformation. To ease finding pair-wise landmark correspondences between the images, the landmarks are usually associated to transformation-invariant feature representations. These representations describe the image contents around each landmark. FBR approaches have been extensively used in retinal image registration. Most approaches use broad-domain landmark detectors, like Harris corner detector [18], SIFT [19,20] or SURF [21]. Although the results for these algorithms are accurate, they produce many generic points that require descriptors to facilitate the matching procedure. The main advantage of these methods is that they do not require ground truth or labeled images.

Other works, instead of using broad-domain methods, solve this issue designing domain-specific descriptors [22], although they still rely on generic detection algorithms. Conversely, domain-specific detection methods rely on the extraction of natural landmarks, such as vessel intersections. These methods can greatly reduce the number of detected candidate points. This enables descriptor-less matching [14] as it reduces the complexity of the matching process. This is an advantage as it can reduce the execution time and overall computational complexity. Furthermore, domain-specific interest points are advantageous as they can also be used for multimodal image registration as they are preserved across different imaging methodologies [14,23] whereas generic points are not guaranteed to be present across the different image modalities.

In terms of retinal imaging, feature-based approaches are preferable over intensity-based methods as the relevant patterns for registration are

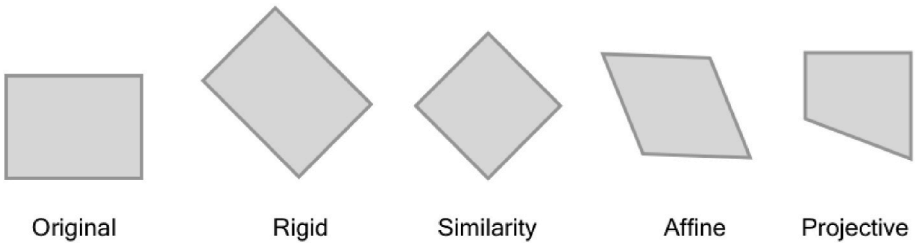

Fig. 1. Appearance of different geometric transformations.

sparse and scattered, and the background of the images is usually homogeneous. Furthermore, these image pairs commonly show the progression of diseases which is more detrimental to intensity-based methods. Therefore, contrary to other medical image registration areas, FBR is preferable over IBR due to the particularities of retinal imaging.

Due to their advantages, novel deep learning methods are becoming more and more widely used, replacing classical methods, specifically in the field of medical image registration. For instance, some recent deep learning works aimed at computing similarity metrics using Convolutional Neural Networks (CNNs) [24]. This allowed to learn the relevant features to estimate the matching between images. Nevertheless, the registration procedure still requires an iterative process to search for the optimal transformation. Other methods can predict the transformation matrix directly by relying on parameter regression [25]. However, there are very few works in the field of retinal image registration, to the best of our knowledge just one [26]. This is due to the specific features of this domain, which prevent the direct adoption of widespread methods. Commonly used approaches are generally intensity-based and thus not adequate and disadvantageous for retinal image registration even if they are successful in other medical areas or if they can be adapted for fundus registration [26].

The specific methodologies for color fundus image registration are currently dominated by classical methods [11,27]. Even if new deep-learning-based pipelines are starting to appear [26], they still cannot compete in performance with the classical proposals. Among these classical approaches, most of them use generic point detectors, domain-specific keypoints or a combination of both. Generic keypoints often require the use of descriptors to properly match those landmark points [18,21]. These descriptors allow to perform direct pairwise matching of keypoints through representation comparison. This allows to reduce the complexity of the matching process. On the other hand, methods that only employ domain-specific interest points usually result in a lower number of specific detections, which can allow to avoid the use of descriptors, while keeping complexity low [14,28]. Retinal image domain-specific landmarks are generally related with the arteriovenous vessel tree. This vascular tree is a complex network of arteries and veins which often intersects and branches. These vessel crossovers and bifurcations are natural characteristic points for ophthalmological images and their relative locations can be used as a biometric pattern due to their uniqueness among individuals [28].

In this work, we propose the use of deep learning to detect representative vascular tree landmarks and use them as reference points for retinal image registration. Particularly, we employ vessel crossovers and bifurcations which have already been successfully exploited as domain-specific landmarks for registration using classical methods [29,30]. Moreover, deep-learning feature-based detectors have not yet been used in retinal image registration despite the fact that this pattern recognition technology usually surpasses the performance of classical methods and that feature-based approaches are the most adequate for this imaging modality. Therefore, combining deep learning methods with feature-based registration is highly desirable. Our method intends to integrate these two approaches in conjunction with domain-specific vascular tree landmarks. Thus, our proposal is the first method of this kind in the state of the art. This way, our method has the combined advantages of these approaches, such as end-to-end training, robustness from deep-learning and the benefits of keypoint-based approaches which are inherent to this domain.

We propose a deep learning method to detect vessel tree landmarks and use them as reference points for retinal image registration in combination with an approach based on RANSAC (Random Sample Consensus). This method is based on proven classical state-of-the-art works [14,28]. The use of domain-related landmarks allows for direct point matching without the computation of complex descriptors, thus, reducing the computational complexity of our approach. This, in turn, can enable fast image registration allowing for easier clinical adoption. Additionally, we propose to use simple similarity transformations to register these images. Similarity transformations assume that the retinal fundus is a plane which is a simplification when compared to state of the art methods which use more complex transformations, with more degrees of freedom [11]. Consequently, our method only requires two matching points per image pair which allows to register images with severe disease progression which can occlude many landmark points, thus facilitating clinical uses. This is an advantage over state-of-the-art methods that require a larger number of matches to produce their transformation models, although it limits the transformation capabilities of our approach.

In summary, our proposal is the first feature-based deep-learning registration method for fundus images. Currently, fundus image registration is dominated by classical FBR methods. However, these classical methods have several disadvantages when compared to novel deep learning approaches. Deep learning methods are highly desirable as they allow to train systems end-to-end, from the raw data to the expected decisions, without the need of ad-hoc pre-processing or feature engineering. This enables flexibility and adaptability of methods to the constantly evolving imaging devices and modalities, along with adjusting to challenging conditions, like those derived from the evolution of pathological lesions. However, current deep learning methods are intensity-based which is not suitable for fundus images, due to the sparse relevant structures and homogenous backgrounds. Although feature-based approaches are preferable there are no deep learning methods of this kind for fundus images. Therefore, using a deep-learning feature-based method is highly desirable. Moreover, our approach proposes to use domain-specific keypoints. These highly specific points can be matched without descriptor computation thus reducing the time complexity of our method.

This manuscript is organized as follows: Section 2 presents the related works, detailing their main features and achievements. Next, Section 3 presents the whole methodology, explaining the different steps, describing the used dataset and the experimental and evaluation approaches. Section 4 describes and discusses the results that were obtained and the main challenges and limitations of this work. Finally, Section 5 includes the conclusions about the proposed method as well as potential future lines of work.f

2. Related work

Nowadays, there are many methods for registering images either medical or non-medical. Specifically, the field of medical image registration has been receiving a lot of attention along the years [1]. In particular, the field of retinal images and especially color retinography registration, has been the topic of several works [31].

Currently, this field is dominated by classical methods, which obtain the best performances [11,27]. For instance, REMPE uses a mix of generic points (SIFT) and domain-specific landmarks (bifurcations) to match the images. This method employs RANSAC to find a first registration approximation and then refines it with Particle Swarm Optimization and with a more complex transformation model. This two-step matching is repeated several times to find the optimal solution, which is selected from all the candidates. Differently, VOTUS [27] creates graphs from the whole arteriovenous tree. These graphs are matched among images using a novel algorithm in conjunction with several classical image features (like Gabor filters or saliency maps). Finally, the transformation is created using DeSAC (Deterministic Sample And Consensus).

Regarding the detection of domain-specific landmarks such as crossovers and bifurcations, classical methods usually consider domain-specific features of the vascular tree. For example, in Ref. [32] filter bank orientations were used to determine the presence of these landmarks. In particular, our proposal is based on previous works [14,28] that also use classical methods to detect these landmarks. Specifically, these works use a creases-based level set extrinsic curvature (LSEC) to segment the blood vessels. From this segmentation, the vessels are

labeled, and their discontinuities fixed. Crossovers and bifurcations are detected using the angles associated with the blood vessels that form them. Finally, a filtering process is used to remove spurious detections. Other more recent works proposed deep learning approaches for their detection [33–35]. These novel methodologies improve the results of classical approaches. However, despite the performance increase and the advantages of deep learning in this domain, none of these methods have been used in image registration.

There are specific deep learning pipelines for the registration of medical images. For instance, some methods can directly predict the transformation for the moving images. These methods create a regression model to recover the transformation matrix parameters and align the images [36]. However, due to the lack of training data it is not always possible to use labeled data, therefore it is common for some methods to use unsupervised learning [37,38]. These methods map the corresponding pair of images to the deformation field that aligns the images [38].

The work of Zou et al. [26] is an adaptation of this type of methods [36,37]. Currently, this is the only deep learning method designed for fundus image registration. This work proposes a deep regression

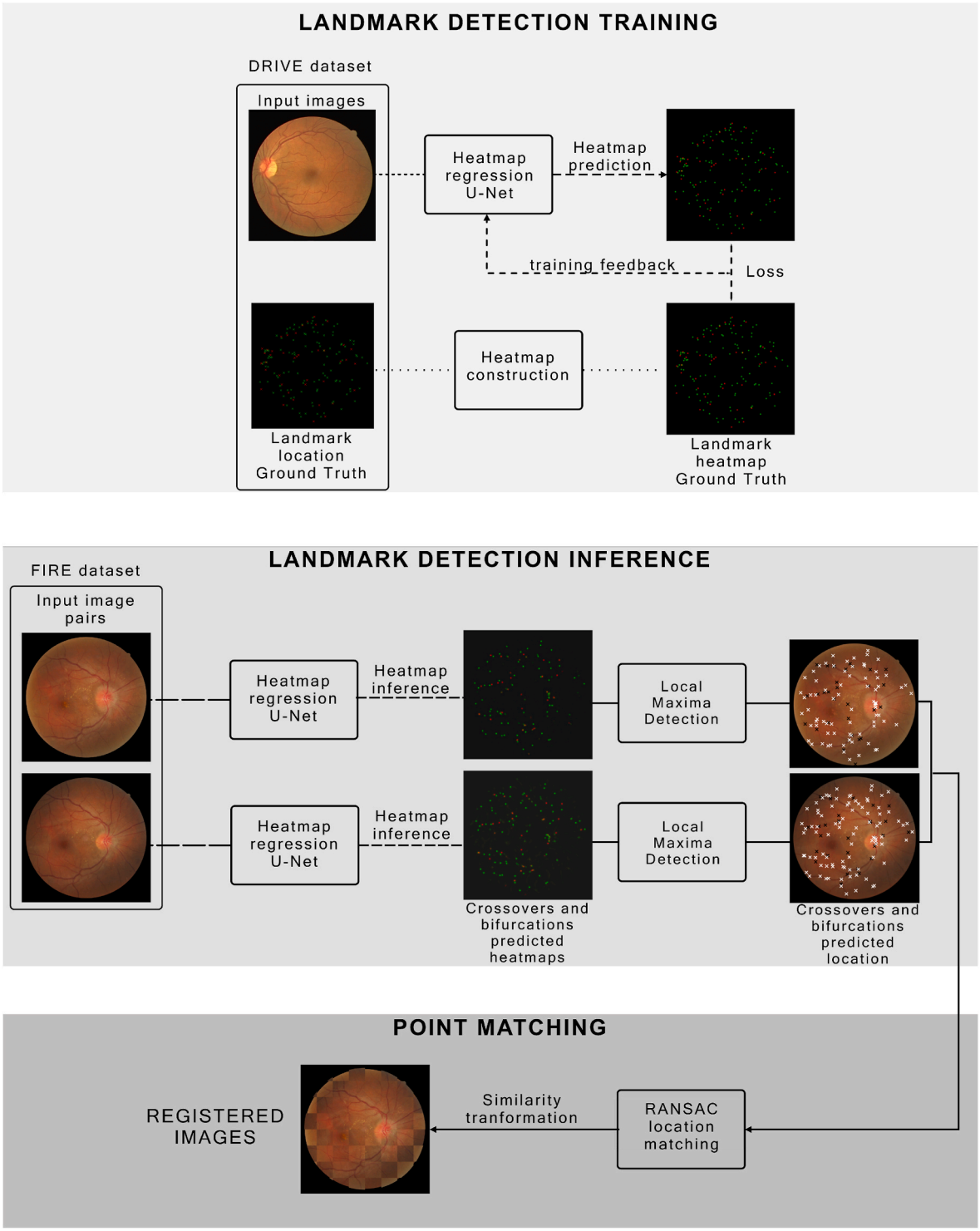

Fig. 2. Overview of the whole methodology.

network (Structure-Driven Regression Network, SDRN), capable of creating deformation fields at different scales. This kind of deep learning-based methods have yet to achieve the performance of the classical ones in fundus images [11,27], contrary to other image registration application fields [38]. This is due to the particular requirements and challenges in the registration process of retinal images, like the preservation of sparse detailed structures (like blood vessels) over relatively uniform backgrounds, the progress of diseases, or the expectedly large displacement transformations. These factors often impede the straightforward adoption of deep learning methods that have been successful in other medical areas.

3. Materials and methods

3.1. Methodology

Our method is schematically described in Fig. 2. Our approach can be separated in several parts:

Landmark detection training: we train a CNN using the DRIVE dataset to detect blood vessels and bifurcations. We transform the landmark location ground truth to heatmaps as depicted in the top-most part of Fig. 2. This has several advantages over using the standard location ground truth. The heatmaps predicted by the network are compared to the ground truth to compute the loss. The overview of this step is detailed in subsection 3.1.1 and the heatmap generation process is specified in subsection 3.1.2. The choice of architecture (U-Net) is elaborated in subsection 3.1.4.

Landmark detection inference: The network is tested using the FIRE dataset. As shown in the middle section of Fig. 2, each image in the pair is given as input to the network, which predicts their landmark heatmaps. These heatmaps are then converted to precise landmark locations using a local maxima detection filter. The local maxima filter is thoroughly described in subsection 3.1.3. The final result of this step is the set of crossovers and bifurcations for each image in the pair.

Point Matching: The crossovers and bifurcations are matched among the images using a RANSAC-based approach. Due to the specific landmarks used, this step does not require descriptor computation. A similarity transformation with 4 DOF is used, therefore only two point matchings are needed to align the images. This step is comprehensively explained in subsection 3.1.5

3.1.1. Landmark detection

The first step of the proposed methodology is to accurately detect the landmarks in each image of the pair using a CNN. The landmark detection requires not only the prediction of each landmark location, but the distinction among the two types of landmarks, crossovers and bifurcations. Furthermore, the number of these landmarks is unknown and variable among the images. Overall, the use of neural networks instead of classical methods to automate this task is desirable as it greatly reduces the time that is employed in feature engineering and is more adaptable to new imaging devices.

This task can be divided in two separate parts, as represented in Fig. 2. Firstly, the landmark detection training which employs the DRIVE dataset as input data. During the training, the landmark location ground truth is transformed to heatmaps. Therefore, the network learns to predict these heatmaps in its training stage. This output is converted back to precise locations in the landmark detection inference task. This is done through a local maxima filter. It should be noted that this second step uses the FIRE dataset to evaluate the performance of the registration methodology. Once the inference heatmaps are transformed to precise locations, these points can be used by a RANSAC matching algorithm to create the suitable transformation models.

3.1.2. Heatmaps

The straightforward way to use a neural network which would simply be predicting the landmark location (binary maps) from a binary ground truth map (the exact pixel-wise location of the landmarks marked as positive class). This is not suitable due to the heavy unbalance between classes i.e., the negative class (background) is much more numerous than the single pixels representing each landmark. On that account, we follow the approach in Ref. [33], which trains the network using heatmaps generated from the binary ground truth. These heatmaps have maximum values where the binary maps had located a landmark, progressively decreasing the values in the surrounding pixels.

The heatmaps are defined as multi-instance as each one represents multiple landmarks. Using heatmaps increases the information that is made available to the network from the original hard or binary labels, which can improve the feedback in the detection step. Furthermore, the use of heatmaps or soft labels is also beneficial in several scenarios as they mitigate potential noise in the original, binary, ground truth since the patterns that make up each landmark can spawn several pixels like in the case of thick blood vessels or in the case of thin vessels with poor contrast. To create these maps, two alternatives are tested, a Gaussian kernel and a Radial Hyperbolic Tangent kernel (Radial Tanh) [33], by convolving the original binary ground truth maps. The saturation distance of the Radial Tanh kernel and the standard deviation of the Gaussian kernel allow to control the region of influence for each landmark, modifying the aspect of the heatmaps.

To distinguish among the two types of landmarks, the prediction is approached as two separate heatmaps, one for each landmark type. Therefore, the network would generate a two output channel, generating two independent heatmaps, one for bifurcations and one for crossovers. However, using two channels penalizes identifying a landmark in the wrong category (i.e., detecting a crossover instead of a bifurcation). Thus, to avoid the network preferring to miss doubtful cases which could penalize it more than simply not predicting a landmark, we employ a third channel. The third channel includes both landmark types, crossovers and bifurcations, encouraging the detection of these landmarks regardless of their type [33]. The network for the multi-instance heatmap regression is trained using the mean squared error (MSE) between the prediction and the target heatmaps as the loss function.

3.1.3. Local maxima detection

The coordinates of each landmark point can be recovered from the predicted heatmaps using a local maxima detection. Specifically, we employ a maximum filter coupled with an intensity threshold, which allows to retrieve only the most salient local maxima. The intensity threshold prevents background noise from being detected as keypoints, preserving only the more relevant detections. This threshold is obtained, for each image, multiplying the maximum value of the network outputs (landmark heatmaps) with a fixed value factor (0.35 in our approach). This can compensate for the particular features of different images and datasets, as the output of the network is linear and thus not bounded. The value of this threshold is selected to maximize the F1-Score in the test set of the DRIVE dataset. Several thresholds were tested in intervals of 0.05 and we found that 0.35 provided the best F1-Score [33].

The maximum filter selects the highest peak in an area. The radius of this area needs to be lower than the minimum expected distance between landmarks of the same type in order to correctly detect two separate maximal points when there are two close candidate points. Mixing crossings and bifurcations does not affect the results as they are predicted in two different channels of the network. A high value for the radius is appropriate to remove spurious peaks due to the low heatmap smoothness. In this work, we select a conservative value of 3 pixels for the radius parameter.

3.1.4. Architecture

The chosen network architecture to detect landmarks is U-Net, specifically the exact architecture described in Ref. [39], a variation of VGG13. This architecture is chosen due to its satisfactory results in previous works related to retinal images [40]. Furthermore, U-Net

obtains the best results in the detection of vessel crossovers and bifurcations in retinal images, demonstrated in the public DRIVE dataset [33]. This combined with its wide use in medical imaging makes this neural network the best candidate for the keypoint detection task. A diagram for the chosen U-Net architecture is shown in Fig. 3. The output function of the network is linear. The total number of trainable parameters is 31,031,875 from a total of 64 base channels or N.

3.1.5. Point matching

Once the landmarks are detected for each image in the pair, they can be used to infer a transformation via a point matching procedure following state of the art methods [14,28]. The chosen method is RANSAC [41], widely used in image registration, even in the top state-of-the-art methods [11]. RANSAC is able to estimate the parameters of a mathematical model, in this case the transformation of the moving image to be aligned with the fixed one, from a set of observed data, in our case, the landmark points. To do so, RANSAC separates the data into inliers which are points that do explain the mathematical model and outliers which are noise and do not fit in the model.

The number of inlier points that are necessary to create a model depends on the complexity of the model [42]. To model the transformations that are required to align the image pairs, we propose to use a simple similarity transformation. This transformation allows for translation (both X and Y axes) as well as rotation and isotropic scaling [8] and thus only requires two matching points. Consequently, our methodology needs at least two crossings or bifurcations per image to correctly register images. Therefore, it can register images even in cases of severe disease progress. As diseases change the aspect of the retina, they can occlude some of the crossovers and bifurcations, thus a low point requirement can be useful in pathological cases. Additionally, the simple transformation model reduces computational complexity and execution time.

Furthermore, our matching method does not require the crossovers and bifurcations to be present in equal (or similar) numbers. The proposed method only matches each point within its type among the two images, which allows for faster execution due to the lower number of checks that are required by the RANSAC algorithm. The only information that is required by the method is the position of the landmark itself, and its type (whether they are a crossover or a bifurcation). Therefore, there is no need for the computation of expensive and sophisticated descriptors for each landmark, a common feat of the state of the art [11,18]. The landmark-detector network is tuned towards high specificity and thus most points should be accurate which should help the RANSAC method to be fast and accurate.

3.2. Datasets

Currently, there is no dataset with both crossovers and bifurcations as well as registration data as labels. The DRIVE dataset [43] contains 40 images with binary crossings and bifurcation labeling (i.e., pixels marked as landmarks). The FIRE [44] dataset contains 134 image pairs, divided into several categories, and contains registration labeling. We propose to learn crossings and bifurcations from the DRIVE dataset [43] and apply this knowledge to the FIRE [44] dataset, which has the appropriate registration labeling to evaluate the method.

The binary labels from the ground truth of the DRIVE dataset corresponding to crossovers and bifurcations are converted to heatmaps. This allows to improve the likelihood of correctly detecting the landmark and reducing the noise. As previously mentioned, this is done by convolving the original binary ground truth maps with two separate alternative kernels: Gaussian kernel and Radial Hyperbolic Tangent.

Regarding the FIRE dataset, it is divided into three separate categories (S, P, A) depending on the image features. Category S contains image pairs with a high degree of overlapping (more than 75%) and have no anatomical differences among them. Similarly, images from category P lack anatomical differences, however, their overlapping is lower (less than 75%). Finally, category A images have large overlaps but have anatomical differences among them, associated to the progression of pathologies. The 134 image pairs are divided in 71 from category S, 49 from P and 14 from A. Representative images from both datasets, FIRE and DRIVE, can be seen in Fig. 4.

Regarding the FIRE dataset, while its control points are indeed blood vessel bifurcations and crossovers they are not suitable as a ground truth to learn how to place these points. The FIRE dataset is designed for registration validation through the evaluation of distances on few landmark positions. These labeled landmarks are only those occupying the overlapping portion between image pairs thus most of all the existent bifurcations and crossovers are missing. Therefore, they are too incomplete to be used as ground truth for learning landmark detection. Instead, for the training stage we use the DRIVE dataset, which has the full ground truth of crossings and bifurcations available. Moreover, we need an independent test dataset to evaluate the registration, which is

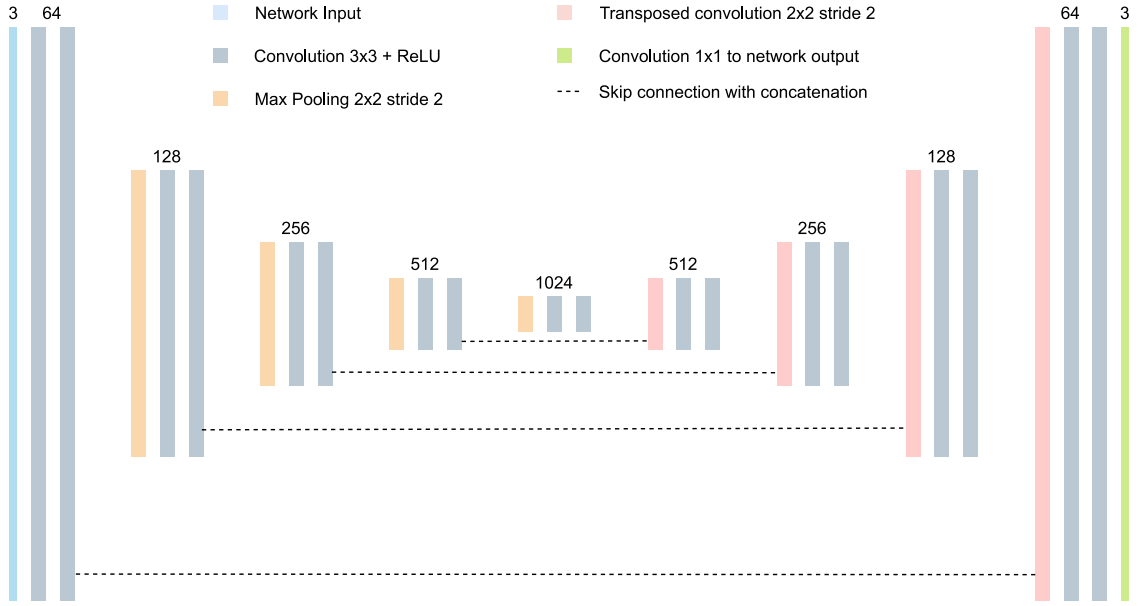

Fig. 3. Diagram of the chosen U-Net network which shows the number of output channels in each convolutional block.

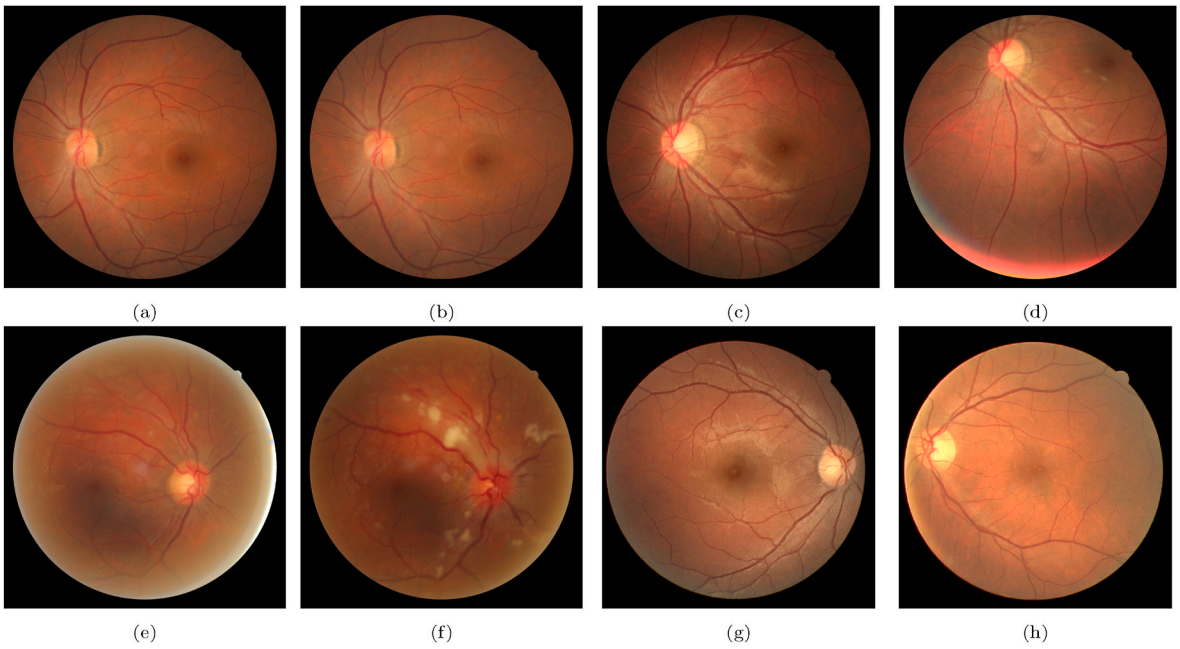

Fig. 4. (a,b) corresponding image pair from Category S, (c,d) from category P and (e,f) from category A, all from the FIRE dataset. (g,h) from the DRIVE dataset.

why the whole FIRE dataset is held out for the test.

The use of different datasets implies certain difficulties, as both datasets have different features. For instance, the DRIVE dataset has an image resolution of 584 × 565 while the FIRE dataset has a resolution of 2912 × 2912. Moreover, all the subjects in the DRIVE dataset are diabetics but only 7 of the 40 images show signs of diabetic retinopathy, an eye pathology, in the specific form of background retinopathy. On the other hand, FIRE contains 134 image pairs which can be divided in several categories depending on the features of the images. In this regard, only category A (14 image pairs) shows cases of retinopathy in the form of increased vessel tortuosity, microaneurysms, cotton-wool spots, etc. This means that, although the source disease is the same, the symptoms vary between datasets adding an additional layer of complexity.

3.3. Experimental details

To train the network, it is initialized following the method by He et al. [45], with a zero-centered normal distribution. The chosen optimizer algorithm is Adam [46] with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the default values proposed by its authors. The network was trained from scratch and hyperparameters were set empirically using the evolution of the validation loss so that they provide stable learning. The learning rate is originally set to $\alpha = 1e-4$ with a patience for the learning rate schedule of 2500 batches before reducing the learning rate. Each batch contains one image and the learning rate is reduced by a factor of 0.1. The training stops after the learning rate reaches $\alpha = 1e-7$ given that there are no significant changes in the validation loss after that point.

To train this network we employ the DRIVE dataset, making use of its standard partition, that is, a specific set of 20 images is reserved for training and the rest for testing. From the test set, 25% of the images were used as the validation set. Furthermore, to avoid overfitting in the training, spatial data augmentation is used. This augmentation consists of random affine transformations that are applied to the original fundus images as well as the ground truth coordinates which are then used to create the target heatmaps. The parameters for said transformations were empirically selected over the training set. Particularly, the images can be rotated, randomly, from −90 to 90°. Each image can also be zoomed from 0.9 × to 1.1 × , also randomly. Furthermore, the images are randomly sheared between −20° and 20°. Color augmentation is also used by randomly changing image components in the HSV color space [47]. The network is trained using the full resolution of the DRIVE images (584 × 565). It should be noted that the keypoint-detection step is completely tested and validated using DRIVE dataset and the method was not tuned in the FIRE dataset.

In this work, we aim at performing a successful registration regardless of the resolution of the images. This is due to the marked differences in image resolution from the dataset used to train the method (DRIVE) with a resolution of 584 × 565 while the FIRE dataset, used to test the method, has a resolution of 2912 × 2912. The main issue is then that the landmark detection network is trained at a fixed image size given by the DRIVE dataset, significantly smaller than the FIRE resolution. In order to overcome the challenge of facing different image resolutions using the same network, we explore several alternatives.

The first alternative that we considered is to upscale the detected points from the resolution in which the landmark detector network operates (DRIVE dataset resolution) to the suitable one. We named this approach point scaling as it directly scales the points. This is a simple approach, and it can incur in some issues if the scaling is not fully accurate. The second alternative consists in upscaling the predicted heatmap and then, calculate the local maxima over the heatmap already in the suitable resolution, the FIRE resolution. We named this second approach heatmap scaling as it scales the heatmap before the calculation of the keypoints. While this approach is more complex, it could prevent some of the inaccuracies that the point matching method may cause.

The proposed methodology is implemented in Python 3 and C++. The U-Net neural network is implemented using PyTorch, an open source Python framework, using CUDA 10.0 and cuDNN 7.5.0. Training, testing and development was performed on a virtual machine with 8 cores from an Intel Xeon Gold 6146 CPU @ 3.20 GHz, a single NVIDIA GRID M60-8Q GPU, with 8 GB of VRAM and 12 GB of RAM.

3.4. Evaluation methodology

In this work, the registration performance is evaluated using two different approaches that are used in the state-of-the-art for the FIRE dataset. First, we use the registration score proposed by Hernandez-Matas et al., authors of the original dataset [44] and state-of-the-art methods [11]. The registration score is based around the idea of

measuring the error between image pairs after the transformation of the moving image, the registration error. This error is calculated regarding the control points which are the ground truth for the FIRE dataset. If the registration error of an image pair is below a threshold, the registration is deemed successful and if it is not, it is unsuccessful. If the registration is successful the image pair is added to the positive cases, and as the error threshold grows more image pairs should fall into the successful category. By plotting the ratio of successfully and unsuccessfully registered images using a 2D graph where the *X* axis corresponds to the value of the error threshold and the *Y* axis to the percentage of successfully registered image pairs, an Area Under Curve (AUC) can be easily computed. The AUC's value as well as the plot serves to easily and robustly compare among the methods. The resulting AUC, specifically defined within specific boundaries (0–100% success ratio in 0–25 pixels) is the final registration score. This registration score is the de-facto standard for the works that employ the FIRE fundus dataset.

Secondly, to facilitate the comparison against all the previous works, we also adopt the evaluation approach that is used in Ref. [26]. In particular, Zou et al. [26] deviate from the standard for the FIRE dataset and use the Root Mean Square Error (RMSE) between the control points in the images, instead of the distance as the error metric for each image pair. Therefore, the final metric is RMSE averaged for the whole set or category of images.

Both metrics will be computed for each of the categories as well as for the whole dataset as specified in other state of the art works as well as the FIRE dataset proposal [11,44].

It should also be noted that, to compare our results to some state of the art works, scaling the images is required. For instance, the work of Hernandez-Matas et al. [11] uses the full resolution of the FIRE dataset, that is, the standard. On the other hand, Zou et al. [26] use the resolution in which their network operates, 256 × 256. This makes direct comparison between methods complex. In order to compare our proposal against all the previous works, we perform our evaluation at different resolutions by scaling the results. This way we can compare our results using both state of the art methods, the ones that use a registration score at 2912 × 2912 and RMSE at 256 × 256.

4. Results and discussion

The results and discussion section is structured as follows: first in subsection 4.1 we explore the alternatives in scaling the results of the landmark detection network (DRIVE dataset resolution) to the operative resolution (FIRE dataset). Next, in subsection 4.2 we detail a comparison among the proposed approach and a previous classical method, since both works share the same point matching method, and therefore, only the landmark detector changes. Subsection 4.3 compares our method with current state of the art works using the same common dataset, FIRE. Finally, in subsection 4.4 we discuss the limitations of our approach.

4.1. Experimental analysis

We conducted the study on the FIRE dataset to evaluate the effectiveness of the proposed method. Furthermore, we compare our method with different state-of-the-art approaches, both classical and deep-learning-based methods.

The keypoint detector network was trained and tested using the DRIVE dataset as this it provides a suitable ground truth for this task. The version trained using the Gaussian kernel obtained 81.22% of precision and 70.80% of recall in the test set. Similarly, the Tanh version 79.63% of precision and 71.58% of recall in the same set. Thus, we can say that neither method holds any advantage above the other.

The results of different variations of the proposed methodology are presented in Table 1, which depicts the registration score, the standard way of evaluating the registration in the FIRE dataset. The results for the different variations of our method are presented in graphical form in Fig. 5. Finally, representative examples of the landmark points detected by the CNN are shown in Fig. 6 and examples of registered images in Fig. 7.

Table 1
Comparison among different variations of the proposed methodology. Metric is registration score (AUC values, higher is better).

Models	Registration Score (AUC)			
	S	P	A	FIRE
Point scaling Gaussian	0.900	**0.298**	0.662	0.655
Point scaling Tanh	0.904	0.296	**0.666**	0.656
Heatmap scaling Gaussian	**0.908**	0.293	0.660	**0.657**
Heatmap scaling Tanh	0.905	0.288	0.658	0.654

The results of the test among the different kernels and scaling methods that are shown in Table 1 demonstrate an adequate performance for every variation of the method. While point scaling methods can incur in some errors due to interpolation in the scaling process, which is mitigated by using heatmap scaling, they still offer better performance in some cases, namely P and A categories. The biggest difference between point scaling and heatmap scaling methods is 0.8% in the case of the Gaussian kernel in category S and in the Tanh kernel in category A, favoring the heatmap scaling method. Therefore, even if both methods are very similar in results and their differences are not noteworthy, for simplicity we choose the heatmap scaling method, as it appears more robust against interpolation artifacts and can, more accurately, obtain the correct landmark position in the higher resolution image. This is due to imperfections in the scaling from the original operating resolution of the landmark detector (the DRIVE image resolution) to the full FIRE resolution. Scaling the heatmap and computing local maxima in the scaled heatmap, instead of the points directly, decreases the error, making the landmarks more reliable. Regarding the kernels that were used, the results indicate that both methods are equivalent. In this case the biggest difference among the kernels is, at most, 0.5% in terms of the chosen kernel, in the particular case of heatmap scaling in the P category. This coincides with the observation in the state of the art [33] which also concludes that both kernels are identical in performance. Therefore, based on the slight advantage in the global FIRE category, even if it is not relevant, we will use heatmap scaling with the Gaussian kernel as our reference to simplify the state-of-the-art comparisons.

Overall, the results of the different kernels and scaling methods are very close to one another, as the AUCs for each method are practically equivalent. This can be appreciated in the registration scores for the experiments in Table 1 and the corresponding graphs that are shown in Fig. 5, where all the curves are very similar. Therefore, for simplicity in the comparison of our work with previous methods and state o the art approaches, we choose the heatmap scaling approach in conjunction with the Gaussian kernel.

4.2. Classical vs. deep learning proposed approach

We compare our learning-based method to our previous classical method [28] in Table 2. As previously explained, this method [28] is very similar to the one we use as it detects crossovers and bifurcations and uses them to estimate a transformation among a pair of images. It shares the point matching algorithm with our proposal, although our version does not need advanced features to complete the matching procedure. That is, both methods employ a similar matching procedure which minimizes the differences of both approaches, reducing it to the keypoint-detection stage. Therefore, this allows a direct comparison between a classical method and a novel deep-learning-based one for the detection of vessel bifurcations and crossovers. This comparison allows to validate the proposed approach in relation to a classical method, highlighting the advantages of the deep learning approach. It should be noted that the chosen classical method has already demonstrated

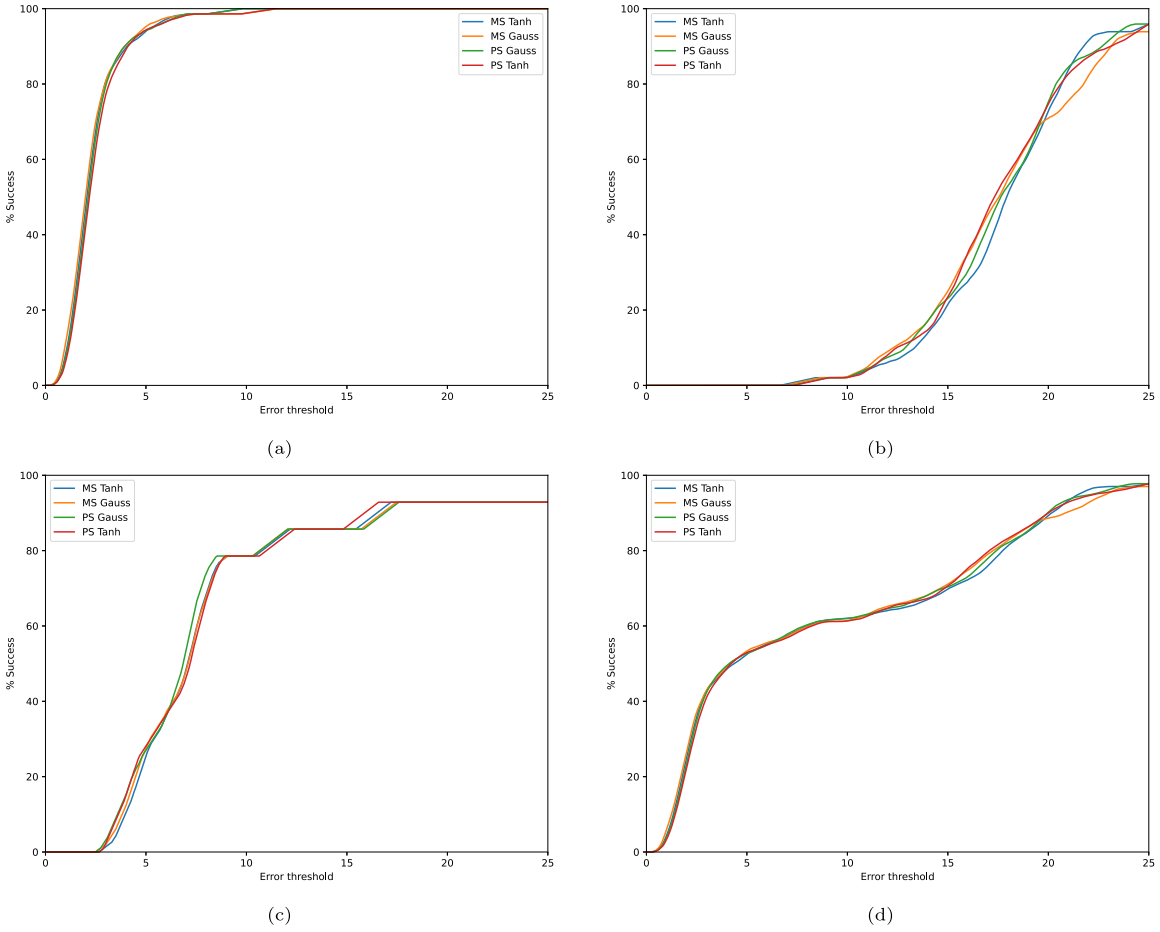

Fig. 5. Evaluation following the standard registration score, depicting the registration success at different error thresholds. Results depicted for (a) Category S, (b) Category P, (c) Category A and (d) the whole FIRE dataset. The curves are smoothed with the Savitzky-Golay Filter [48] to eliminate the edges caused by the low number of images, especially notable in Category A. AUC values are, however, calculated with the original curves.

accurate results in biometry [28] and multimodal registration of fundus images [14]. However, as is common with classical methods, it is specifically tuned towards a resolution of 768 × 584, thus, the FIRE images need to be scaled down to approximately match that resolution. In this case, the FIRE images are converted to 720 × 720 and, later, the transformation matrix is scaled back to full-size so the results can be compared with the rest of methods, including the proposed one. The chosen metric is, again, the registration score.

As shown in Table 2, the results from our method are clearly superior in every category to those of the classical method. Therefore, as the matching method is shared among the proposed method and the classical one, we can concur that the landmarks detected by the network are more accurate. Furthermore, our version does not require the computation of descriptors or the extraction of any feature around each point, contrary to the classical method, thus reducing the amount of computation needed and the overall execution time. A completely fair comparison is complex as the resolutions that the methods use greatly differ (720 × 720 for the classical method and 2912 × 2912 for our proposal). The size of the images plays an important role in the execution time of any algorithm. For instance, the classical method takes an average of 1.5 s to align an image pair in the lower 720 × 720 resolution but this increases to around 15 min per pair when using the full-size resolution. This makes a direct comparison among methods difficult. The proposed method employs approximately 0.65 s per image pair using full-size resolution. Out of that time, it averages around 0.2 s for the U-Net inference and heatmap scaling on both images (keypoint detection) and approximately 0.45 s in the matching process. Moreover, as the proposed method uses deep learning, it has all of its advantages over classical methods. Deep learning can be easily adapted to new images or domains and does not require ad-hoc image pre-processing or to manually tune hyperparameters and use hand-engineered features, like classical methods, which significantly reduces the effort needed to train the system and simplifies the whole pipeline.

Succinctly, our approach obtains better results than our previous works in every category of the FIRE dataset. Therefore, this validates the use of deep learning to detect domain-specific landmarks as the matching method is shared among the two approaches. This means that the deep learning detector produces more accurate keypoints than the previous method since the registration scores improve. Moreover, our proposal is faster than the previous method as it takes 0.65 s to register each pair versus the 1.5 s of the classical method.

4.3. State of the art comparison

The best results for the proposed method and its variations are compared to the state of the art works in Table 3. In this regard, our results are compared to the work of Zou et al. [26] in Table 4 due to the differences in image resolution and evaluation methodology that this paper incurs into. It is important to notice that, in Table 4, the results are evaluated using RMSE, the chosen method by Ref. [26], hence lower values indicate better performance. In contrast, the results depicted in Tables 1 and 3 correspond to the registration score, which indicates better performance at higher values. Additionally, to produce a fair comparison, the registration performance in Table 4 is evaluated at the

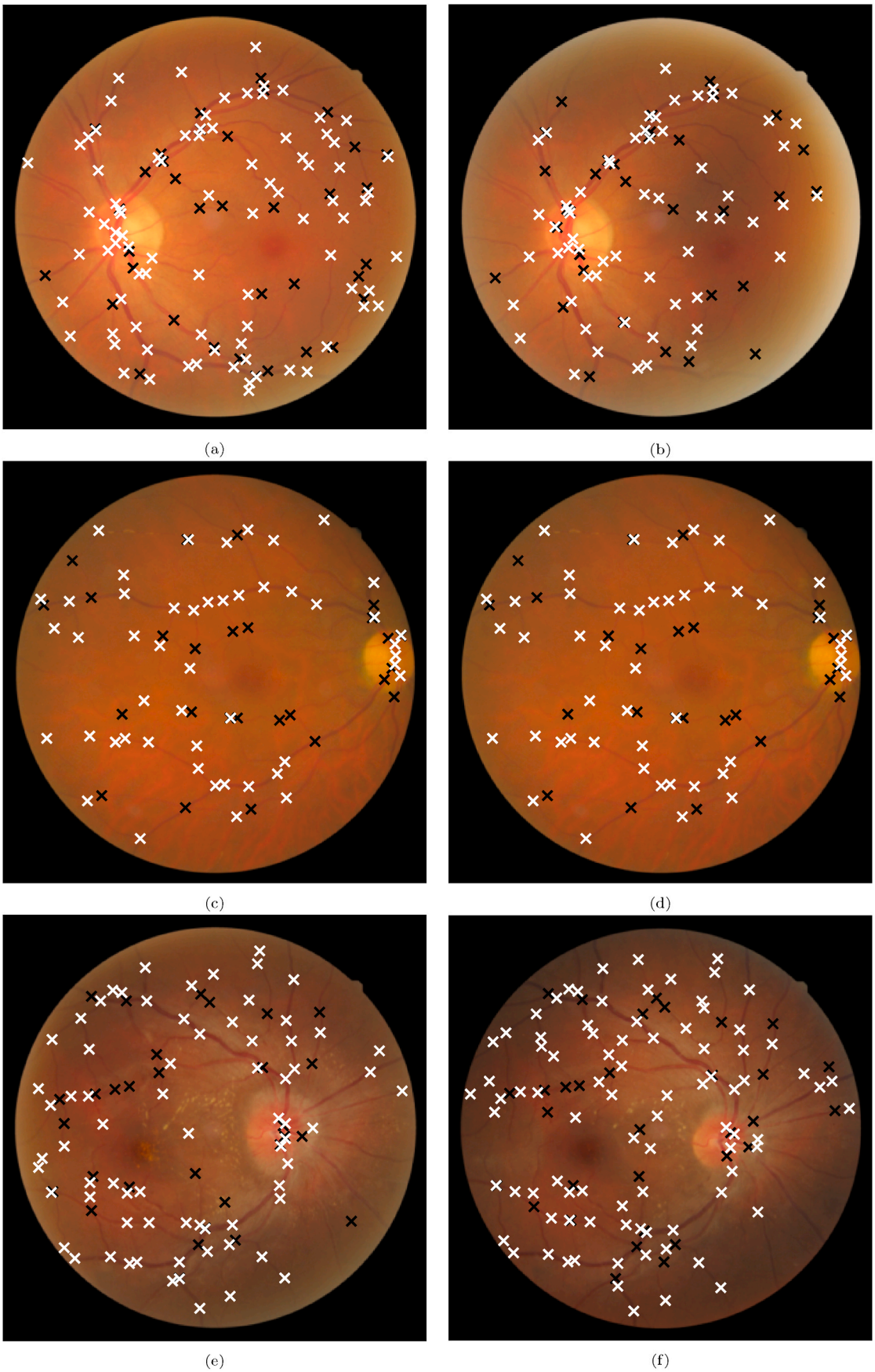

Fig. 6. Detected crossovers (black) and bifurcations (white) for different representative images of each category. Top row corresponds to Category S, the middle row to category P and the bottom row to category A.

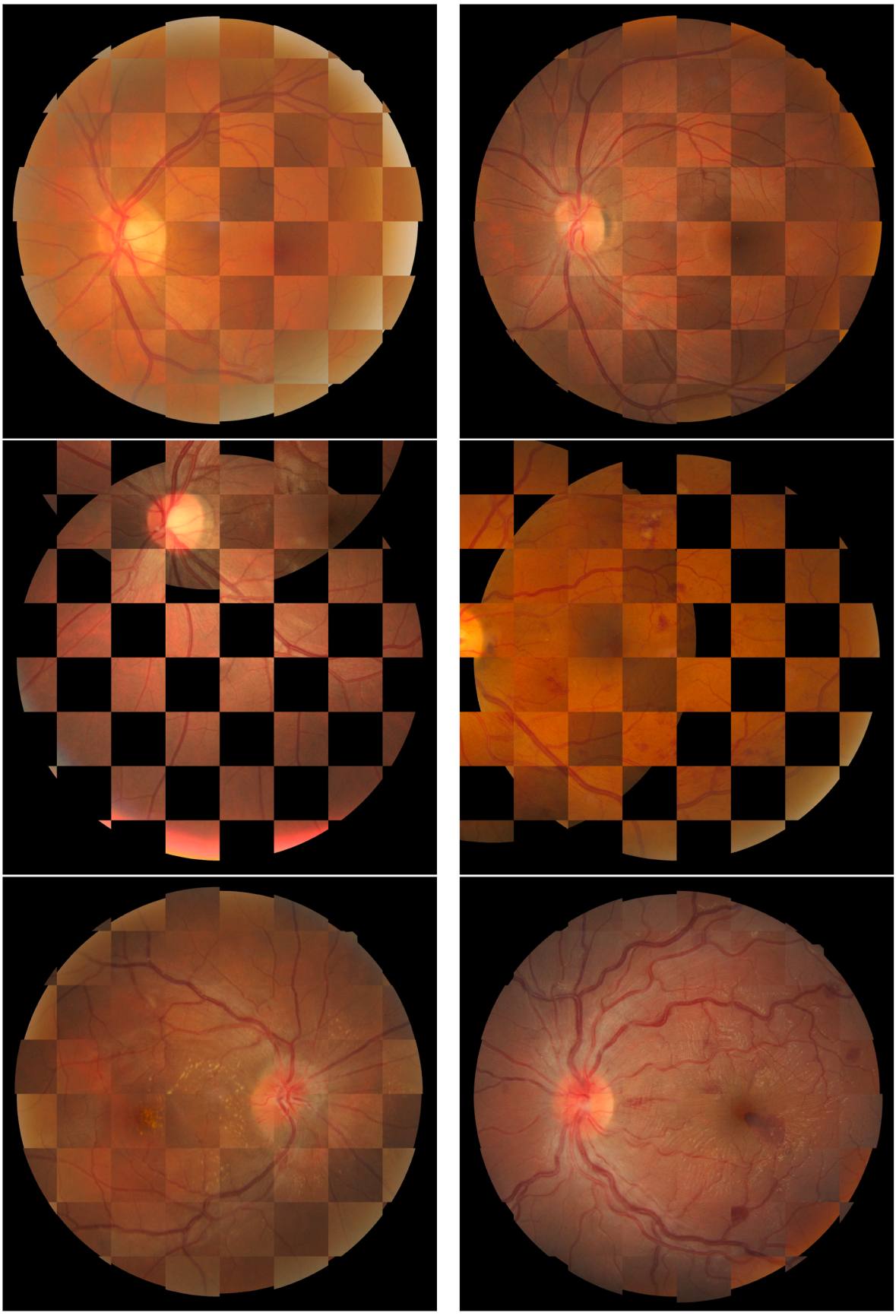

Fig. 7. Representative images from the registered FIRE dataset, top row corresponds to Category S, the middle row to category P and the bottom row to category A.

Table 2
Comparison among the proposed learning-based method and a comparable classical method across all the categories using registration score. Best results, highlighted in bold.

Models	Registration Score (AUC)			
	S	P	A	FIRE
Proposed Method	0.908	0.293	0.660	0.657
Creases feature matching [28]	0.819	0.166	0.550	0.552

same low resolution that is used in the work of Zou et al. [26], i.e., 256 × 256 pixels. This means that most of the pixel accuracy errors in the previously detailed upscaling process are not significant and, furthermore, many details are lost thus bringing the overall errors closer to one another. Moreover, the average used in RMSE is not robust to outliers, which means that even a single failed case, could drastically alter the metric. Therefore, while RMSE is useful to compare our results to some state of the art works, we believe that the registration score proposed with the FIRE dataset is more representative of the actual performance of any registration algorithm as it employs AUCs which are more robust than average as a metric.

In any case, it is clear that our algorithm improves the performance of the work of Zou et al. [26] as it achieves a RMSE of 0.299 for categories A and S as opposed to the state of the art error of 0.915. It is also worth highlighting that our method is capable of registering images in the P category, discarded by this work due to their lack of overlapping. Therefore, our method improves upon purely deep-learning-based methods.

When comparing our method with *ad-hoc* classical methods, highlighted in Table 3, we can see that it surpasses a wide range of methods, especially in the categories A and S. Particularly, in category S our method is only surpassed by REMPE (0.958) [11] and VOTUS (0.934) [27] as it obtains an AUC of 0.908. Similarly in category A, our method ties with REMPE producing an AUC of 0.660. In this case our approach is only surpassed by VOTUS. However, in category P, despite being able to register images unlike [26], the results are not competitive with classical methods. Our approach obtains 0.293 of AUC while the best method, VOTUS, obtains 0.672.

Depicted in Table 3 are the reported execution times for each method. It is not possible to draw fair comparisons from these times as the hardware is not the same. However, we can affirm that, given the execution time of our method relative to the rest of the approaches, a key advantage of our proposal is its speed. Our method is at least an order of magnitude faster than any other approach and two orders of magnitude faster than the best-scoring methods in this dataset. This could be key in real world uses and clinical applications.

To summarize, our approach surpasses current deep learning methods, and it can compete with the classical approaches. Particularly, our method is able to register images in category P unlike previous deep learning proposals, although its results are not competitive with the top state of the art methods. However, in FIRE categories A and S, our proposal obtains comparable results to the best classical methods. Moreover, our method is the fastest in the state of the art, being capable of registering each image pair in under a second, while current state of the art approaches employ minutes.

4.4. Experimental limitations

Overall, we can say that our method is able to compete with the best available registration methods while using a relatively simple pipeline, requiring less parameter tuning, as opposed to the state-of-the-art methods. However, it is clear that our method obtains worse results than the state of the art. Detailing the results for each category, we can see them fluctuate depending on the specific categories. In the category S, which has high degrees of overlapping and no morphological changes, our method is able to place third in the overall state of the art ranking. Similarly, in category A, which has a high degree of overlapping but also has morphological changes due to progress in diseases, our method ties for the second spot in registration score. On the contrary, in category P, our method is not able to compete with the best methods in the state of the art. This particular category is the one with the lowest overlapping among image pairs and, therefore, the highest expected displacement for the moving image. The results for this category negatively impact the score for the whole dataset.

The cause for this decrease in performance in category P is the lack of degrees of freedom in the transformation used. As we employ a RANSAC without budget, the best available transformation (given the detected keypoints) is always found, therefore this is not a limitation in the matching process. Furthermore, despite the cross-dataset application of our keypoint-detector U-Net, the spatial position of the landmarks is accurate, as evidenced by the results in the other categories. Finally, we have observed that the registration results in category P are not completely wrong, following a qualitative assessment, but they are not finely accurate. Thus, we can conclude that the lack of degrees of freedom is the cause for the diminished performance in the P category

Table 4
Comparison of the proposed method and SDRN [26]. Results measured in RMSE at 256 × 256, the operating resolution of SDRN. Best results highlighted in bold.

Models	Registration Error (RMSE)				
	S	P	A	A & S	FIRE
SDRN [26]	–	–	–	0.915	–
Proposed Method	**0.203**	**1.656**	**0.782**	**0.299**	**0.795**

Table 3
Results for the different state of the art methods and our proposals. All the SOTA results extracted from Ref. [11]. Results measured in AUC with the standard FIRE method, higher is better. Best overall results highlighted in bold, results for the proposed method highlighted in italics. The results are ordered according to their performance in the overall FIRE dataset. Execution time measured in seconds. * Indicates execution time extracted from Ref. [27] and † from Ref. [11].

Name	Registration Score (AUC)				Execution Time	Transformation model
	S	**P**	**A**	**FIRE**		
VOTUS [27]	0.934	**0.672**	**0.681**	**0.812**	106*	Quadratic
REMPE [11]	**0.958**	0.542	0.660	0.773	198†	Ellipsoid eye model
GFEMR [13]	0.812	0.607	0.474	0.702	10*	Elastic
SIFT + WGTM [49]	0.837	0.544	0.407	0.685	–	Quadratic
Proposed Method	*0.908*	*0.293*	*0.660*	*0.657*	0.65	Similarity
GDB-ICP [50]	0.814	0.303	0.303	0.576	19*	Quadratic
Harris-PIIFD [51]	0.900	0.090	0.443	0.553	13*	Polynomial
ED-DB-ICP [20]	0.604	0.441	0.497	0.553	44*	Affine
SURF + WGTM [52]	0.835	0.061	0.069	0.472	–	Quadratic
RIR-BS [12]	0.772	0.004 9	0.124	0.440	–	Projective
EyeSLAM [53]	0.308	0.224	0.269	0.273	7*	Rigid
ATS-RGM [54]	0.369	0.000	0.147	0.211	–	Elastic

and, therefore, a limitation of our approach. This is evidenced by the comparison of the different transformation models used in the state of the art methods, shown in Table 3. The use of similarity transformations assumes that the retinal fundus is a plane so that, in images with large deformations (like category P), the difference with reality may be more noticeable. However, in image pairs with smaller deformations (like categories A and S) the inaccuracies are negligible so that the advantages that a simpler transformation offers (lower point requirement, less computational complexity, etc.) are more relevant.

Our proposal uses one of the simplest transformations out of all the methods. This has advantages, like the low point requirement, which could allow to register images with severe pathology progression. However, the main disadvantage is the lack of power to align images with low overlapping and high expected transformations.

5. Conclusion and future work

In this work, we propose a learning-based pipeline, combining deep learning for domain-specific landmark detection and a classical method for point matching. This pipeline is based around proven state-of-the-art methods and provides accurate results.

Our proposal is the first method in the state of the art to use deep learning in a FBR method, detecting domain-specific landmarks and using them to register image pairs. FBR methods are advantageous due to the appearance of retinal images. These images contain sparse relevant structures so that landmark-based methods are preferable over intensity-based approaches. Furthermore, detecting domain-specific landmarks allows descriptor-less matching, due to the lower number of detections, which in turn reduces execution time. Similarly, deep learning methods are more robust than classical methods and require no ad-hoc pre-processing or feature engineering. Moreover, they can be easily adapted to new imaging devices and imaging conditions. Therefore, the combination of deep learning and FBR methods is highly desirable yet still unexplored in previous works. The best available methods are classical methods and the current deep learning approaches are intensity-based. Thus, our proposal is a novel approach to register fundus images.

We propose to use a CNN to detect representative landmarks specific to this domain of images. These landmarks can then be matched without expensive descriptor computation using RANSAC, which produces the desired transformation to register the images.

We use a deep neural network to detect blood vessel crossovers and bifurcations specific to this domain of images. These natural landmarks are unique for each person and, therefore, are very reliable for fundus image registration. Currently, there are no datasets containing landmarks and registration data, therefore we use two separate datasets. We train the landmark detector network using the DRIVE dataset, with ground truth for crossovers and bifurcations. We test the proposed method using the FIRE dataset, which contains registration labeling. This allows to evaluate the performance of the proposed method even if no suitable dataset exists. However, it carries several intrinsic complications like changes in the image resolution, the pathologies present in the images, etc. which our method is able to overcome. The chosen CNN for the landmark detection task is the U-Net. We convert the DRIVE ground truth from a point-based location labeling to a heatmap which improves the performance of the network. Therefore, the network predicts these heatmaps over the FIRE dataset on inference. The heatmaps are converted to the precise location of the landmarks using a local maxima filter.

The landmarks are matched among image pairs using a RANSAC-based classical point matching method. The RANSAC confined to similarity transformations which are able to transform the images with very limited information, just two point matchings. Therefore, our method is resistant to image pairs with severe pathology progress, which can occlude the blood vessels. Furthermore, our method does not require the expensive computation of advanced descriptors for each landmark point as we use domain specific landmarks. Due to the lower number of detections when compared to generic point detectors, it is computationally viable to use RANSAC to test every possible landmark match combination. Therefore, the only information needed by this method is the coordinates and the type of the landmark, differentiating between crossovers and bifurcations. This allows for less computation than state of the art methods. In this regard, our proposal is the fastest in the state of the art by orders of magnitude, as our method takes less than a second to register each image pair while the competing state-of-the-art methods take minutes.

Finally, the experimental results for the proposed method are satisfactory. We validate our approach comparing it with a previous work of ours based on classical methods that detect the same landmarks. As our method improves upon the previous one, and as both share the same matching mechanism, we can affirm that detecting crossovers and bifurcations using deep learning outperforms classical approaches. Furthermore, the proposed method is able to improve the results from purely deep-learning-based state of the art methodologies. Moreover, our method can register the images in the FIRE dataset disregarding their category, unlike the current deep learning method. Therefore, this validates the robustness of our system. In this regard, the proposed similarity transformation is limiting in the P category due to its lack of degrees of freedom when compared with the large, expected transformations. Overall, our proposal shows to be competitive with the best state of the art approaches which are complex classical methods, namely VOTUS and REMPE. Even if the results of our proposal are lacking in P category (0.293 AUC), our method produces accurate results in categories S (0.908 AUC) and A (0.606 AUC).

As future work, we will test transformations with more degrees of freedom than the similarity transformation used in this proposal. The higher order transformations may accommodate the larger displacements of the category P and thus improve the overall results. Moreover, our approach can be extended to detect and describe the crossings and bifurcations. Furthermore, the proposal can also be adapted for the registration of multimodal retinal images, such as retinography-angiography registration. In this regard, domain adaption techniques would be worth exploring to mitigate the lack of annotated multimodal datasets. Finally, another possibility is to develop novel network architectures capable of obtaining similar results to U-Net but more efficiently. This could reduce computational cost, making widespread clinical use even more accessible.

Declaration of competing interest

None declared.

Acknowledgements

This research was funded by Instituto de Salud Carlos III, Government of Spain, DTS18/00 136 research project; Ministerio de Ciencia e Innovación y Universidades, Government of Spain, RTI2018-095 894-B-I00 research project; Consellería de Cultura, Educación e Universidade, Xunta de Galicia through the predoctoral grant contract ref. ED481A 2021/147 and Grupos de Referencia Competitiva, grant ref. ED431C 2020/24; CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%). The funding institutions had no involvement in the study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication. Funding for open access charge: Universidade da Coruña/CISUG.

References

[1] M.A. Viergever, J.A. Maintz, S. Klein, K. Murphy, M. Staring, J.P. Pluim, A survey of medical image registration – under review, Med. Image Anal. 33 (2016) 140–144, 20th anniversary of the Medical Image Analysis journal (MedIA).

[2] J. Hajnal, D. Hill, D. Hawkes, *Medical image Registration*, Ser. Biomedical Engineering Series. Plus 0.5em Minus 0, CRC Press, 4emBoca Raton, FL, 01 2001.

[3] H. Narasimha-Iyer, A. Can, B. Roysam, H.L. Tanenbaum, A. Majerovics, Integrated analysis of vascular and nonvascular changes from color retinal fundus image sequences, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 54 (8) (2007) 1436–1445.

[4] M.S. Miri, M.D. Abràmoff, Y.H. Kwon, M.K. Garvin, Multimodal registration of sd-oct volumes and fundus photographs using histograms of oriented gradients, Biomed. Opt Express 7 (12) (Dec 2016) 5252–5267.

[5] C. Hernandez-Matas, X. Zabulis, Super resolution for fundoscopy based on 3d image registration. " in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 6332–6338.

[6] H.-P. Chan, L.M. Hadjiiski, R.K. Samala, Computer-aided diagnosis in the era of deep learning, Med. Phys. 47 (5) (2020) e218–e227 [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13764.

[7] N. Asiri, M. Hussain, F. Al Adel, N. Alzaidi, Deep learning based computer-aided diagnosis systems for diabetic retinopathy: a survey, Artif. Intell. Med. vol. 99 (2019) 101701 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718307607.

[8] R. Szeliski, Computer Vision: Algorithms and Applications, first ed., Springer-Verlag, Heidelberg, 2010 plus 0.5em minus 0.4emBerlin.

[9] M. Wang, L. Pengcheng, A review of deformation models in medical image registration, J. Med. Biol. Eng. 39 (2018) 4.

[10] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, second ed., Cambridge University Press, 2003 plus 0.5em minus 0.4emUSA.

[11] C. Hernandez-Matas, X. Zabulis, A.A. Argyros, " Rempe, Registration of retinal images through eye modelling and pose estimation, IEEE J. Biomed. Health Inf. 24 (12) (2020) 3362–3373.

[12] L. Chen, Y. Xiang, Y. Chen, X. Zhang, "Retinal Image Registration Using Bifurcation Structures," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2169–2172.

[13] J. Wang, J. Chen, H. Xu, S. Zhang, X. Mei, J. Huang, J. Ma, Gaussian field estimator with manifold regularization for retinal image registration, Signal Process. 157 (2019) 225–235.

[14] Álvaro S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement, Procedia Comput. Sci. 126 (2018) 97–104, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

[15] J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, Image registration by maximization of combined mutual information and gradient information, in: S.L. Delp, A.M. DiGoia, B. Jaramaz (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2000, Plus 0.5em Minus 0.4emBerlin, Springer Berlin Heidelberg, Heidelberg, 2000, pp. 452–461.

[16] G. Balakrishnan, A. Zhao, M.R. Sabuncu, A.V. Dalca, J. Guttag, An unsupervised learning model for deformable medical image registration, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9252–9260.

[17] J. Kybic, M. Unser, Fast parametric elastic image registration, IEEE Trans. Image Process. 12 (11) (2003) 1427–1442.

[18] J. Chen, J. Tian, N. Lee, J. Zheng, R.T. Smith, A.F. Laine, A partial intensity invariant feature descriptor for multimodal retinal image registration, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 57 (7) (2010) 1707–1718.

[19] G. Yang, C.V. Stewart, M. Sofka, C. Tsai, Registration of challenging image pairs: initialization, estimation, and decision, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 1973–1989.

[20] C. Tsai, C. Li, G. Yang, K. Lin, The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence, IEEE Trans. Med. Imag. 29 (3) (2010) 636–649.

[21] G. Wang, Z. Wang, Y. Chen, W. Zhao, Robust point matching method for multimodal retinal image registration, Biomed. Signal Process Control 19 (2015) 68–76.

[22] J. Chen, J. Tian, N. Lee, J. Zheng, R.T. Smith, A.F. Laine, A partial intensity invariant feature descriptor for multimodal retinal image registration, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 57 (7) (2010) 1707–1718.

[23] F. Laliberte, L. Gagnon, Yunlong Sheng, Registration and fusion of retinal images-an evaluation study, IEEE Trans. Med. Imag. 22 (5) (2003) 661–673.

[24] X. Cheng, L. Zhang, Y. Zheng, Deep similarity learning for multimodal medical images, Comput. Methods Biomech. Biomed. Eng.: Imag. Visual. 6 (3) (2018) 248–252.

[25] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: a survey, Mach. Vis. Appl. 31 (1) (Jan 2020) 8.

[26] B. Zou, Z. He, R. Zhao, C. Zhu, W. Liao, S. Li, Non-rigid retinal image registration using an unsupervised structure-driven regression network, Neurocomputing 404 (2020) 14–25.

[27] D. Motta, W. Casaca, A. Paiva, Vessel optimal transport for automated alignment of retinal fundus images, IEEE Trans. Image Process. 28 (12) (2019) 6154–6168.

[28] M. Ortega, M.G. Penedo, J. Rouco, N. Barreira, M.J. Carreira, Retinal verification using a feature points-based biometric pattern, EURASIP J. Appl. Signal Process. 1 (2009) 235746. Mar 2009.

[29] C.A. Lupaşcu, D. Tegolo, F. Bellavia, C. Valenti, Semi-automatic registration of retinal images based on line matching approach, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 453–456.

[30] R. Kolar, V. Harabis, J. Odstrcilik, Hybrid retinal image registration using phase correlation, Imag. Sci. J. 61 (4) (2013) 369–384.

[31] S.K. Saha, D. Xiao, A. Bhuiyan, T.Y. Wong, Y. Kanagasingam, Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: a review, Biomed. Signal Process Control 47 (2019) 288–302.

[32] S. Abbasi-Sureshjani, I. Smit-Ockeloen, E. Bekkers, B. Dashtbozorg, B.t.H. Romeny, Automatic detection of vascular bifurcations and crossings in retinal images using orientation scores, in: 2016 IEEE 13th International Symposium on Biomedical Imaging, ISBI), 2016, pp. 189–192.

[33] Álvaro S. Hervella, J. Rouco, J. Novo, M.G. Penedo, M. Ortega, Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images, Comput. Methods Progr. Biomed. 186 (2020) 105201.

[34] F. Uslu, A.A. Bharath, A multi-task network to detect junctions in retinal vasculature, in: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), " in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Plus 0.5em Minus 0.4emCham, Springer International Publishing, 2018, pp. 92–100.

[35] H. Pratt, B.M. Williams, J.Y. Ku, C. Vas, E. McCann, B. Al-Bander, Y. Zhao, F. Coenen, Y. Zheng, Automatic detection and distinction of retinal vessel bifurcations and crossings in colour fundus photography, J. Imag. 4 (1) (2018).

[36] S. Miao, Z.J. Wang, R. Liao, A cnn regression approach for real-time 2d/3d registration, IEEE Trans. Med. Imag. 35 (5) (2016) 1352–1363.

[37] B.D. de Vos, F.F. Berendsen, M.A. Viergever, H. Sokooti, M. Staring, I. Išgum, A deep learning framework for unsupervised affine and deformable image registration, Med. Image Anal. 52 (2019) 128–143.

[38] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, " Voxelmorph, A learning framework for deformable medical image registration, IEEE Trans. Med. Imag. 38 (8) (2019) 1788–1800.

[39] O. Ronneberger, P. Fischer, T. Brox, " U-Net, Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI), 2015.

[40] Álvaro S. Hervella, J. Rouco, J. Novo, M. Ortega, Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction, Appl. Soft Comput. 91 (2020) 106210.

[41] M. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (1981) 381–395.

[42] D.A. Forsyth, J. Ponce, *Computer vision: a Modern Approach*. Plus 0.5em Minus 0, 4emPearson, 2012.

[43] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, B. van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imag. 23 (4) (2004) 501–509.

[44] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, A. Argyros, Fire: fundus image registration dataset, J. Model. Opthalmol. (1) (2017).

[45] K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification,, in: 2015 IEEE International Conference on Computer Vision, (ICCV), 2015, pp. 1026–1034.

[46] D. Kingma, J. Ba, Adam: a method for stochastic optimization, Int. Conf. Learn. Represent. (2014) 12.

[47] P. Liskowski, K. Krawiec, Segmenting retinal blood vessels with deep neural networks, IEEE Trans. Med. Imag. 35 (11) (2016) 2369–2380.

[48] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627–1639.

[49] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (Nov 2004) 91–110.

[50] G. Yang, C.V. Stewart, M. Sofka, C. Tsai, Registration of challenging image pairs: initialization, estimation, and decision, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 1973–1989.

[51] J. Chen, J. Tian, N. Lee, J. Zheng, R.T. Smith, A.F. Laine, A partial intensity invariant feature descriptor for multimodal retinal image registration, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 57 (7) (2010) 1707–1718.

[52] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Understand. 110 (3) (2008) 346–359 (similarity Matching in Computer Vision and Multimedia).

[53] D. Braun, S. Yang, J.N. Martel, C.N. Riviere, B.C. Becker, Eyeslam: real-time simultaneous localization and mapping of retinal vessels during intraocular microsurgery, Int. J. Med. Robot. Comput. Assist. Surg. 14 (1) (2018) e1848.

[54] E. Serradell, M.A. Pinheiro, R. Sznitman, J. Kybic, F. Moreno-Noguer, P. Fua, Non-rigid graph registration using active testing search, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 625–638.