GLAMpoints: Greedily Learned Accurate Match points

Prune Truong[1,2*]   Stefanos Apostolopoulos[1]   Agata Mosinska[1]   Samuel Stucky[1]   Carlos Ciller[1]
Sandro De Zanet[1]

[1]RetinAI Medical AG, Switzerland   [2]ETH Zurich, Switzerland

truongp@student.ethz.ch   {stefanos, agata, samuel, carlos, sandro}@retinai.com

Abstract

We introduce a novel CNN-based feature point detector - Greedily Learned Accurate Match Points (GLAMpoints) - learned in a semi-supervised manner. Our detector extracts repeatable, stable interest points with a dense coverage, specifically designed to maximize the correct matching in a specific domain, which is in contrast to conventional techniques that optimize indirect metrics. In this paper, we apply our method on challenging retinal slitlamp images, for which classical detectors yield unsatisfactory results due to low image quality and insufficient amount of low-level features. We show that GLAMpoints significantly outperforms classical detectors as well as state-of-the-art CNN-based methods in matching and registration quality for retinal images.

1. Introduction

Digital fundus images of the human retina are widely used to diagnose variety of eye diseases, such as Diabetic Retinopathy (DR), glaucoma, and Age-related Macular Degeneration (AMD) [42, 52]. For retinal images acquired during the same session and presenting small overlaps, image registration can be used to create mosaics depicting larger areas of the retina. Through image mosaicking, ophthalmologists can display the retina in one large picture, which is helpful during diagnosis and treatment planning. Besides, mosaicking of retinal images taken at different time points has been shown to be important for monitoring the progression or identification of eye diseases. More importantly, fundus image registration has been explored in eye laser treatment for DR. It allows real-time tracking of the vessels during surgical operations to ensure accurate application of the laser on the retina and minimal damage to the healthy tissues.

---

*Work produced during an internship at RetinAI Medical AG
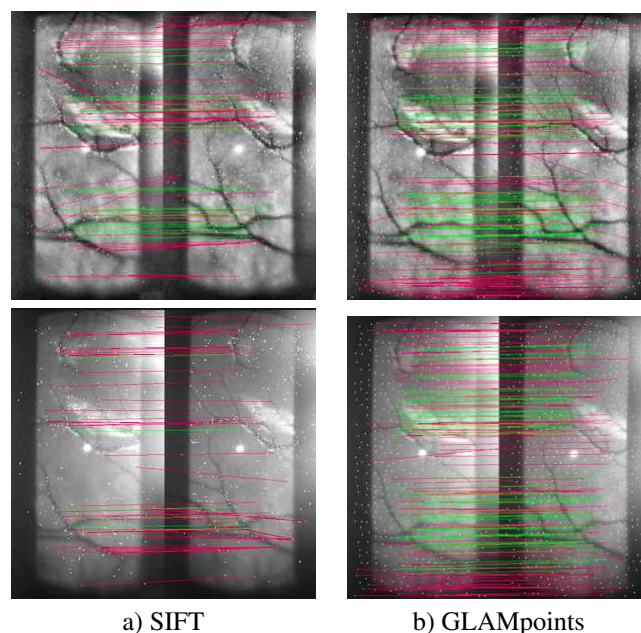

a) SIFT   b) GLAMpoints

Figure 1: Keypoints detected by SIFT and GLAMpoints and resulting matches for a pair of preprocessed (top) and raw (bottom) slitlamp images. Detected points are in white, green matches are true positive, red represents false positive. Our GLAMpoints detector produces more reliable keypoints even without additional pre-processing.

Mosaicking usually relies on extracting repeatable interest points from the images, matching the correspondences and searching for transformations relating them. As a result, the keypoint detection is the first and the most crucial stage of this pipeline, as it conditions all further steps and therefore the success of the registration.

At the same time, classical feature detectors are general-purpose and manually optimized for outdoor, in-focus, low-noise images with sharp edges and corners. They usually fail to work with medical images, which can be distorted,

noisy, have no guarantee of focus and depict soft tissue with no sharp edges (see Figure 3). Traditional methods perform sub-optimally on such images, making more sophisticated optimization necessary at a later step in the registration, such as Random Sampling Consensus (RanSaC) [23], bundle adjustment [44] and Simultaneous Localization and Mapping (SLAM) [18] techniques. In these cases, supervised learning methods for keypoint detection fail or are not applicable, due to missing ground truths for feature points.

In this paper we present a method for learning feature points in a semi-supervised manner. Learned feature detectors were shown to outperform the heuristics-based methods, but they are usually optimized for repeatability, which is a proxy for the matching quality and as a result they may underperform during the final matching. On the contrary, our keypoints - GLAMpoints - are trained for the final matching accuracy and when associated with Scale-Invariant Feature Transform (SIFT) [2] descriptor they outperform state-of-the-art in matching performance and registration quality on retinal images. As shown in Figure 1, GLAMpoints produces significantly more correct matches than SIFT detector.

Registration based on feature points is inherently non-differentiable due to point matching and transformation estimations. We take inspiration from the loss formulation in Reinforcement Learning (RL) using a reward to compute the suitability of the detected keypoints based on the final registration quality. It makes it possible to use the key performance measure, *i.e.* matching power, to directly train a Convolutional Neural Network (CNN). Our contribution is therefore a formulation for keypoint detection that is directly optimized for the final matching performance in an image domain.

The remainder of this paper is organized as follows: we introduce the current state-of-the-art feature detection methods in section 2, our training procedure and loss in section 3, followed by experimental comparison of previous methods in section 4 and conclusion in section 5.

2. Related Work

Existing registration algorithms can be classified as area-based and feature-based approaches. The former typically rely on a similarity metric such as cross-correlation [16], mutual information [38, 31] or phase correlation [29] to compare the intensity patterns of an image pair and estimate the transformation. However, in the case of changes in illumination or small overlapping areas, the application of area-based approaches becomes challenging or infeasible. Conversely, feature-based methods extract corresponding points on pairs of images along with a set of features and search for a transformation that minimizes the distance between the detected key points. Compared with area-based registration techniques, they are more robust to changes of

intensity, scale and rotation and therefore, they are considered more appropriate for problems such as medical image registration.

Typically, feature extraction and matching of two images comprise four steps: detection of interest points, computing feature descriptor for each of them, matching of corresponding keypoints and estimation of a transformation between the images using the matches. As can be seen, the detection step influences every further step and is therefore crucial for a successful registration. It requires a high image coverage and stable key points in low contrasts images.

In the literature, local interest point detectors have been thoroughly studied. SIFT [34] is probably the most well known detector/descriptor in computer vision. It computes corners and blobs on different scales to achieve scale invariance and extracts descriptors using the local gradients. Speeded-Up Robust Features (SURF) [12] is a faster alternative, using Haar filters and integral images, while KAZE [5] exploits non-linear scale space for more accurate keypoint detection.

In the field of fundus imaging, a widely used technique relies on vascular trees and branch point analysis [33, 25]. However, accurate segmentation of the vascular trees is challenging and registration often fails on images with few vessels. Alternative registration techniques are based on matching repeatable local features; Chen *et al.* [15] detected Harris corners [26] on low quality multi-modal retinal images and assigned them a partial intensity invariant feature (Harris-PIIFD) descriptor. They achieved good results on low quality images with an overlapping area greater than 30%, but the method is characterised by low repeatability. Wang *et al.* [47] used SURF features to increase the repeatability and introduced a new method for point matching to reject a large number of outliers, but the success rate drops significantly when the overlapping area diminishes below 50%. Cattin *et al.* [13] also demonstrated that SURF can be efficiently used to create mosaics of retina images even for cases with no discernible vascularisation. However this technique only appeared successful in the case of highly self-similar images. D-saddle detector/descriptor [39] was shown to outperform the previous methods in terms of rate of successful registration on the Fundus Image Registration (FIRE) Dataset [27], enabling the detection of interest points on low quality regions.

Recently, with the advent of deep learning, learned detectors based on CNN architectures were shown to outperform state-of-the-art computer vision detectors [22, 20, 50, 37, 10]. Learned Invariant Feature Transform (LIFT) [50] uses patches to train a fully differentiable deep CNN for interest point detection, orientation estimation and descriptor computation based on supervision from classical Structure from Motion (SfM) systems. SuperPoint [20] introduced a self-supervised framework for training interest point de-

tectors and descriptors. It rises to state-of-the-art homography estimation results on HPatches [11] when compared to SIFT, LIFT and Oriented Fast and Rotated Brief (ORB) [41]. The training procedure is, however, complicated and their self-supervision implies that the network can only find corner points. Altwaijry *et al.* [7] proposed a two-step CNN for matching aerial image patches, which is a particularly challenging task due to ultra-wide baseline. [8] introduced a method to detect keypoint locations on different scales, utilizing high activations in recursive network feature maps. KCNN [21] was shown to emulate hand-crafted detectors by training small networks using keypoints detected by other methods as ground-truth. Local Feature Network (LF-NET) [37] is the closest to our method: a keypoint detector and descriptor is trained end-to-end in a two branch set-up, one being differentiable and feeding on the output of the other non-differentiable branch. However, they optimized their detector for repeatability between image pairs, not taking into account the matching performance.

Truong *et al.* [45] presented an evaluation of SURF, KAZE, ORB, Binary Robust Invariant Scalable Keypoints (BRISK) [32], Fast Retina Keypoint (FREAK) [4], LIFT, SuperPoint and LF-NET both in terms of image matching and registration quality on retinal fundus images. They found that while SuperPoint outperforms all the others relative to the matching performance, LIFT demonstrates the highest results in terms of registration quality, closely followed by KAZE and SIFT. The highlighted issue was that even the best-performing detectors produce feature points which are densely positioned and as a result may be associated with a similar descriptor. This can lead to false matches and thus inaccurate or failed registrations.

Our goal is to tackle this problem by introducing a novel semi-supervised learned method for keypoint detection. Detectors are often optimized for repeatability (such as LF-NET) and not for the quality of the associated matches between image pairs. Our training procedure uses a reward concept akin to RL to extract repeatable, stable interest points with a uniform coverage and is specifically designed to maximize correct matching on a specific domain, as shown for challenging retinal slit lamp images.

3. Methods

Our trained network predicts the location of stable interest points, called GLAMpoints, on a full-sized gray-scale image. In this section, we explain how our training set was produced and our training procedure. As we used standard convolutional network architecture, we only briefly discuss it in the end.

3.1. Dataset

We trained our model on a dataset from the ophthalmic field, namely slit lamp fundus videos, used in laser treatment (examples in Figure 3). In this application, live registration is required for an accurate ablation of the retinal tissue. Our training dataset consists of 1336 images with different resolutions, ranging from 300 px to 700 px by 150 px to 400 px. These images were acquired with multiple cameras and devices to cover large variability of appearances. They come from eye examination of 10 different patients, who were healthy or with diabetic retinopathy.

Let B be the set of base images of size $H \times W$. At every step i, an image pair I_i, I'_i is generated from an original image B_i by applying two separate, randomly sampled homography transforms g_i, g'_i. Images I_i and I'_i are thus related according to the homography $H_{I_i, I'_i} = g'_i * g_i^{-1}$ (see supplementary material). On top of the geometric transformations, standard data augmentation methods are used: gaussian noise, changes of contrast, illumination, gamma, motion blur and the inverse of image. A subset of these appearance transformations is randomly chosen for each image I and I'.

3.2. Training

We define our learned function $f_\theta(I) \rightarrow S$, where S denotes the pixel-wise feature point probability map of size $H \times W$. Lacking a direct ground truth of keypoint locations, a delayed reward can be computed instead. We base this reward on the matching success, computed *after registration*. The training proceeds as follows:

1. Given a pair of images $I \in \mathbb{R}^{H \times W}$ and $I' \in \mathbb{R}^{H \times W}$ related with the ground truth homography $H = H_{I,I'}$, our model provides a score map for each image, $S = f_\theta(I)$ and $S' = f_\theta(I')$.

2. The locations of interest points are extracted on both score maps using standard non-differentiable Non-Max-Supression (NMS), with a window size w.

3. A 128 root-SIFT [9] feature descriptor is computed for each detected keypoint.

4. The keypoints from image I are matched to those of image I' and vice versa using a brute force matcher [1]. Only the matches that are found in both directions are kept.

5. The matches are checked according to the ground truth homography H. A match is defined as true positive if the corresponding keypoint x in image I falls into an ϵ-neighborhood of the point x' in I' after applying H. This is formulated as $\|H * x - x'\| \leq \varepsilon$, where we chose ε as 3 px.
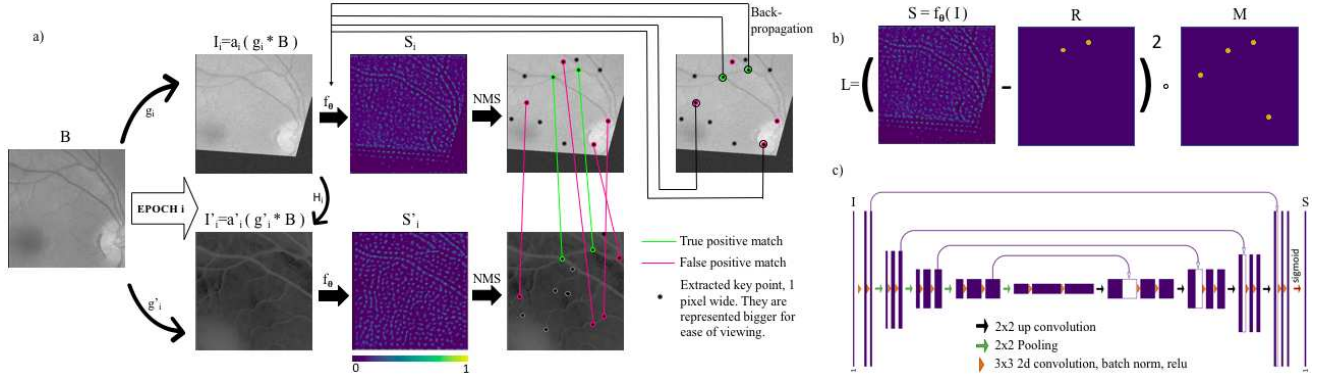
Figure 2: a) Training steps for an image pair I and I' at epoch i for a particular base image B, b) Loss computation. c) Schematic representation of Unet-4.

Let T denote the set of true matches. If a given feature point ends up in the set of true positive points, it gets a positive reward. All other points/pixels are given a reward of 0. Consequently, the reward matrix $R \in \mathbb{R}^{H \times W}$ for a keypoint (x, y) can be defined as follows:

$$R_{x,y} = \begin{cases} 1, & \text{for} \quad (x,y) \in T \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

This leads to the following loss function:

$$L_{\text{simple}}(\theta, I) = \sum (f_\theta(I) - R)^2 \qquad (2)$$

However, a major drawback of this formulation is the large class imbalance between positively rewarded points and null-rewarded ones, where latter prevails by far, especially in the first stages of training. Given a reward R with mostly zero values, the f_θ converges to a zero output. Hard mining has been shown to boost training of descriptors [43]. Negative hard mining on the false positive matches might also enhance performance in our method, but has not been investigated in this work. Instead, to counteract the imbalance, we use sample mining: we select all n true positive points and randomly sample additional n from the set of false positives. We only back-propagate through the $2n$ true positive feature points and mined false positive key points. If there are more true positives than false positives, gradients are backpropagated through all found matches. This mining is mathematically formulated as a mask M, equal to 1 at the locations of the true positive key points and that of the subset of mined feature points, and equal to 0 otherwise. The loss is thus formulated as follows:

$$L(\theta, I) = \frac{\sum (f_\theta(I) - R)^2 \cdot M}{\sum M} \qquad (3)$$

where $\cdot$ denotes the element-wise multiplication.

An overview of the training steps is given in Figure 2. Importantly, only step 1 is differentiable with respect to the loss. We learn directly on a reward which is the result of non differentiable actions, without supervision.

It should be noted that the descriptor we used is the root-SIFT version without rotation invariance. The reason is that it performs better on slitlamp images than root-SIFT detector/descriptor with rotation invariance (see supplementary material for details). The aim of this paper is to investigate the detector only and therefore we used rotation-dependent root-SIFT for consistency.

3.3. Network

A standard 4-level deep Unet [40] with a final sigmoid activation was used to learn f_θ. It comprises of 3x3 convolution blocks with batch normalization and Rectified Linear Unit (ReLU) activations (see Figure 2,c). Since the task of keypoint detection is similar to pixel-wise binary segmentation (class interest point or not), Unet was a promising choice due to its past successes in binary and semantic segmentation tasks.

4. Results

In this section, we describe the testing dataset and the evaluation protocol. We then compare state-of-the-art detectors, quantitatively and qualitatively, to our proposed GLAMpoints.

4.1. Testing datasets

In this study we used the following test datasets:

1. The *slit lamp* dataset: from retinal videos of 3 patients (different from the ones used for training), a random set of 206 frame pairs was selected as testing samples, with size 338 px to 660 px by 190 px to 350 px.
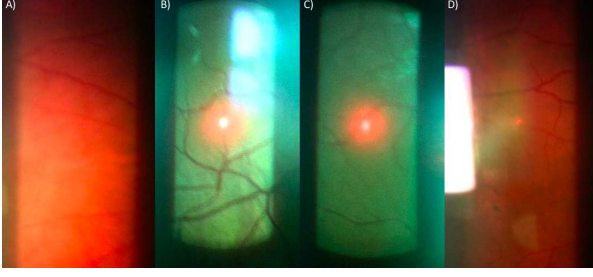
Figure 3: Examples of images from the slit lamp dataset showing challenging conditions for registration. From left to right: low vascularization and over-exposure leading to weak contrasts and lack of corners, motion blur, focus blur, acquisition artifacts and reflections.

Examples are shown in Figure 3. The pairs were selected to have an overlap ranging from 20 to 100%. They are related by affine transformations and rotations up to 15 degrees. Using a dedicated software tool, all pairs of images were manually annotated following common procedures [14] with at least 5 corresponding points, which were then used to estimate the ground truth homographies relating the pairs. As the slit lamp images depict small area of retina, it is justified to apply the planar assumption in generating homographies [13, 24].

2. The *FIRE* dataset [27]: a publicly available retinal image registration dataset with ground truth annotations. It consists of 129 retinal images forming 134 image pairs. The original images of 2912x2912 pixels were down scaled to 15% of their original size, to match the resolution of the training set. Examples of such images are shown in Figure 5.

As a pre-processing step for testing fundus images we isolated the green channel, applied adaptive histogram equalization and a bilateral filter to reduce noise and enhance the appearance of edges as proposed in [19]. The effect of pre-processing can be seen in Figure 1.

Even though the focus of this paper is on the retinal images, we also tested the generalization capabilities of our model by evaluating it on natural images. We used the $Oxford$ [36], EF [53], $Webcam$ [46, 28] and $ViewPoint$ [51] datasets. More details are provided in the supplementary material.

4.2. Evaluation criteria

We evaluated the performance using the following metrics:

1. **Repeatability** describes the percentage of detected points $x \in P$ in image I that are within an ϵ-distance ($\epsilon = 3$) to points $x' \in P'$ in I' after transformation with $H_{I,I'}$:

$$\frac{|\{x\epsilon P, x'\epsilon P' \mid \|H_{I,I'} * x - x'\| < \varepsilon\}|}{|P| + |P'|} \quad (4)$$

2. **Matching performance.** Matches were found using the Nearest Neighbor Distance Ratio (NNDR) strategy, as proposed in [34]: two keypoints are matched if the descriptor distance ratio between the first and the second nearest neighbor is below a certain threshold t. Then, the following metrics were evaluated:

   (a) **AUC**, area under the ROC curve created by varying the value of t computed in line with [17, 49, 48].

   (b) **M.score**, the ratio of correct matches over the total number of keypoints extracted by the detector in the shared viewpoint region [35].

   (c) **Coverage fraction**, measures the coverage of an image by correctly matched key points. A coverage mask was generated from correctly matched key points, each one adding a disk of fixed radius (25px) as in [6].

We computed the homography $\hat{H}$ relating the reference to the transformed image by applying RanSaC algorithm to remove outliers from the detected matches.

3. **Registration success rate.** We furthermore evaluated the registration accuracy achieved after using key points computed by different detectors as in [15, 47]. To do so, we compared the reprojection error of six fixed points of the reference image (denoted as $c_i, i = \{1, .., 6\}$) onto the other. For each image pair for which a homography was found, the quality of the registration was assessed with the median error (**MEE**) and the maximum error (**MAE**) of the distances between corresponding points after transformation.

Using these metrics, we defined different thresholds on MEE and MAE that define "$acceptable$", "$inaccurate$" and "$failed$" registrations. We consider registration "$failed$" if not enough keypoints or matches were found for a homography (minimum 4), if it involves a flip or the estimated scaling component is greater than 4 or smaller than 0.1. We classified the result as "$acceptable$" when $MEE < 10$ and $MAE < 30$ and as "$inaccurate$" otherwise. The values for the thresholds were found empirically by postviewing the results. Using the above definitions, we calculated the **success rate of each class**, equal to the percentage of image pairs for which the registration falls into each category. These metrics are the most important quantitative evaluation criteria of the overall performance in a real-world setting.

4.3. Baselines and implementation details

To evaluate the performance of our GLAMpoints detector associated with root-SIFT descriptor, we compared its matching ability and registration quality against well known detectors and descriptors. Among them, SIFT [2], KAZE [5] and LIFT [50] were shown to perform well on fundus images by Truong *et al.* [45]. Moreover, we compared our method to other CNN-based detectors-descriptors: LF-NET [37] and SuperPoint [20]. We used the authors' implementation of LIFT (pretrained on Picadilly), SuperPoint and LF-NET (pretrained on indoor data, which gives substantially better results on fundus images than the version pretrained on outdoor data) and OpenCV implementation for SIFT and KAZE. A rotation-dependent version of root-SIFT descriptor is used due to its better performance on our test set compared to the rotation invariant version. For the remainder of the paper, SIFT descriptor refers to root-SIFT.

Training of GLAMpoints was performed using Tensorflow [3] with mini-batch size of 5 and the Adam optimizer [30] with learning rate $= 0.001$ and $\beta = (0.9, 0.999)$ for 35 epochs. For each batch we randomly cropped 256×256 patches of the full-resolution image to speed up the computation. GLAMpoints (NMS10) was trained and tested with a NMS window w equal to 10px. It must be noted that other NMS windows can be applied, which obtain similar performance.

4.4. Quantitative results on the *slit lamp* dataset

Table 1 presents the success rate of registration evaluated on the *slit lamp* dataset. Without pre-processing, most detectors show lower performance compared to the pre-processed images, but GLAMpoints performs well even on raw images. While the success rate of acceptable registrations of SIFT, KAZE and SuperPoint drops by 20 to 30% between pre-processed and raw images, GLAMpoints as well as LIFT and LF-NET show only a decrease of 3 to 6%. Besides, LF-NET, LIFT and GLAMpoints detect a steady average number of keypoints (around 485 for pre-processed and 350 non-preprocessed) independently of the pre-processing, whereas the other detectors see a reduction half. In general, GLAMpoints shows the highest performance for both raw and pre-processed images in terms of registration success rate. The robust results of our method indicate that while our detector performs as well or better on good quality images compared to the heuristic-based methods, its performance does not drop on lower quality images.

While SIFT extracts a large number of keypoints (205.69 on average for unprocessed images and 431.03 for pre-processed), they appear in clusters as shown in Figure 5. As a result, even if the repeatability is relatively high, the close positioning of the interest points leads to a large number of rejected matches, as the nearest-neighbours are very

Table 1: Success rates (%) per registration class for each detector on the 206 images of the *slit lamp* dataset. When the original descriptor is not used in association with detector, the descriptor used is indicated in parenthesis.

a) Raw data

	Failed [%]	Inaccurate [%]	Acceptable [%]
SIFT	14.56	63.11	22.33
KAZE	24.27	61.65	14.08
SuperPoint	17.48	48.54	33.98
LIFT	0.0	43.69	56.31
LF-NET	0.0	39.81	60.19
GLAMpoints (SIFT)	0.0	36.41	**63.59**

b) Pre-processed data

Detector	Failed [%]	Inaccurate [%]	Acceptable [%]
ORB	9.71	83.01	7.28
GLAMpoints (ORB)	0.0	88.35	**11.65**
BRISK	16.99	66.02	16.99
GLAMpoints (BRISK)	1.94	75.73	**22.33**
SIFT	1.94	47.75	50.49
KAZE	1.46	54.85	43.69
KAZE (SIFT)	4.37	57.28	38.35
SuperPoint	7.77	51.46	40.78
SuperPoint (SIFT)	6.80	54.37	38.83
LIFT	0.0	39.81	60.19
LF-NET	0.0	36.89	63.11
LF-NET (SIFT)	0.0	40.29	59.71
GLAMpoints (SIFT)	0.0	31.55	**68.45**
Random grid (SIFT)	0.0	62.62	37.38

close to each other. This is evidenced by the low coverage fraction, $M.score$ and AUC (Figure 4). With a similar value of repeatability, our approach extracts interest points widely spread and trained for their matching ability (highest *coverage fraction*), resulting in more true positive matches (second highest *M.score* and *AUC*), as shown in Figure 4.

LF-NET, similar to SIFT, shows high repeatability, which can be explained by its training strategy, which preferred repeatability over accurate matching objective. However, its *M.score* and *AUC* are in the bottom part of the ranking (Figure 4). While the performance of LF-NET may increase if it was trained on fundus images, its training procedure requires images pairs with their relative pose and corresponding depth maps, which would be extremely difficult - if not impossible - to obtain for fundus images.

It is worth noting that SuperPoint scored the highest $M.score$ and AUC but in this case the metrics are artificially inflated because very few keypoints are detected (35,88 and 59,21 on average for raw and pre-processed images respectively). This translates to relatively small coverage fraction and one of the lowest repeatability, leading to few possible correct matches.

As part of an ablation study, we trained GLAMpoints with different descriptors (Table 1, top). While it performs best with the SIFT descriptor, the results show that for every considered descriptor (SIFT, ORB, BRISK), it improves upon the corresponding original detector.

a) Non pre-processed data   b) Pre-processed data

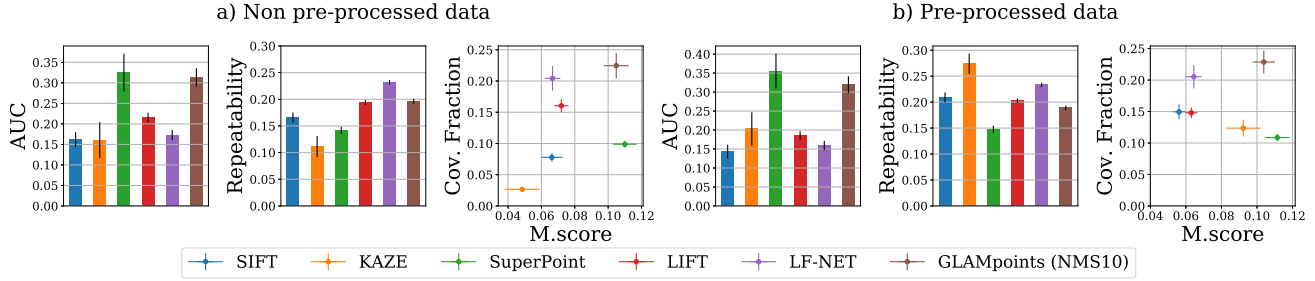SIFT   KAZE   SuperPoint   LIFT   LF-NET   GLAMpoints (NMS10)

Figure 4: Summary of detector/descriptor performance metrics evaluated over 206 pairs of the *slit lamp* dataset.

Table 2: Success rates (%) of each detector on *FIRE*.

	Failed [%]	Inaccurate [%]	Acceptable [%]
SIFT	2.24	36.57	61.19
KAZE	14.18	58.21	27.61
SuperPoint	0.0	13.43	86.57
LIFT	0.0	10.45	89.55
LF-NET	0.0	38.06	61.94
GLAMpoints (OURS)	0.0	5.22	**94.78**

To benchmark the detection results we used the descriptors that were developed/trained jointly with the given detector and thus can be considered as optimal. For instance in [50] the combination of the LIFT/LIFT detector/descriptor outperformed LIFT/SIFT. For completeness, we present the registration results of baseline detectors combined with root-SIFT descriptor in Table 1. As can be seen, it does not improve the result compared to the original descriptor.

Finally, to verify that the performance gain of GLAMpoints does not come solely from the uniform and dense spread of the detected keypoints we computed the success rate for keypoints in a random, uniformly distributed grid (Table 1,bottom), which underperforms in comparison. This shows that our detector predicts not only uniform but also *significant* points.

4.5. Quantitative results on *FIRE* dataset

Table 2 shows the results for success rates of registrations. The presented method outperforms baselines both in terms of success rate and global accuracy of non-failed registrations. As all the images in FIRE dataset present good quality with highly contrasted vascularization, we did not apply pre-processing. We also did not find it necessary to use the available background masks to filter out keypoints detected outside of the retina as generally they were not matched and did not contribute to the final registration.

It is interesting to note the gap of 33.6% in the success rate of acceptable registrations between GLAMpoints and SIFT. As both use the same descriptor, this difference can be only explained by the quality of the detector. As can be seen in Figure 5, SIFT detects a restricted number of keypoints densely positioned solely on the vascular tree and

in the image borders, while GLAMpoints extracts interest points over the entire retina, including challenging areas such as the fovea and avascular zones, leading to a substantial rise in the number of correct matches.

Even though GLAMpoints outperforms all other detectors, LIFT and SuperPoint also present high performance on the *FIRE* dataset. This dataset contains images with well-defined corners on a clearly contrasted vascular tree and LIFT extracts keypoints spread over the entire image, while SuperPoint was trained to detect corners on synthetic primitive shapes. However, as evidenced on the *slit lamp* dataset, the performance of SuperPoint strongly deteriorates on images with less clear features.

4.6. Results on natural images

To further demonstrate a possible extension of our method to other image domains, we computed its predictions on natural images. Note that we used the same GLAMpoints model trained on slit lamp images.

Globally, GLAMpoints reaches a success rate of 75.38% for acceptable registrations, against 85.13% for the best performing detector - SIFT with rotation invariance - and 83.59% for SuperPoint. In terms of AUC, $M.score$ and *coverage fraction* it scores respectively second, second and first best. In contrast, $repeatability$ of GLAMpoints is only second to last after SIFT, KAZE and LF-NET even though it successfully registers more images. This result shows once again that repeatability is not the most adequate metric to measure the performance of a detector. The detailed results can be found in the supplementary material.

Finally, it should be noted that the outdoor images of this dataset are significantly different from medical fundus images and contain much greater variability of structures, which indicates a promising generalization of our model to unseen image domains.

4.7. Qualitative results

In case of slit lamp videos the end goal is to create retinal mosaics. Using 10 videos containing 25 to 558 images, we generated mosaics by registering consecutive frames using keypoints detected by different methods. We calculated

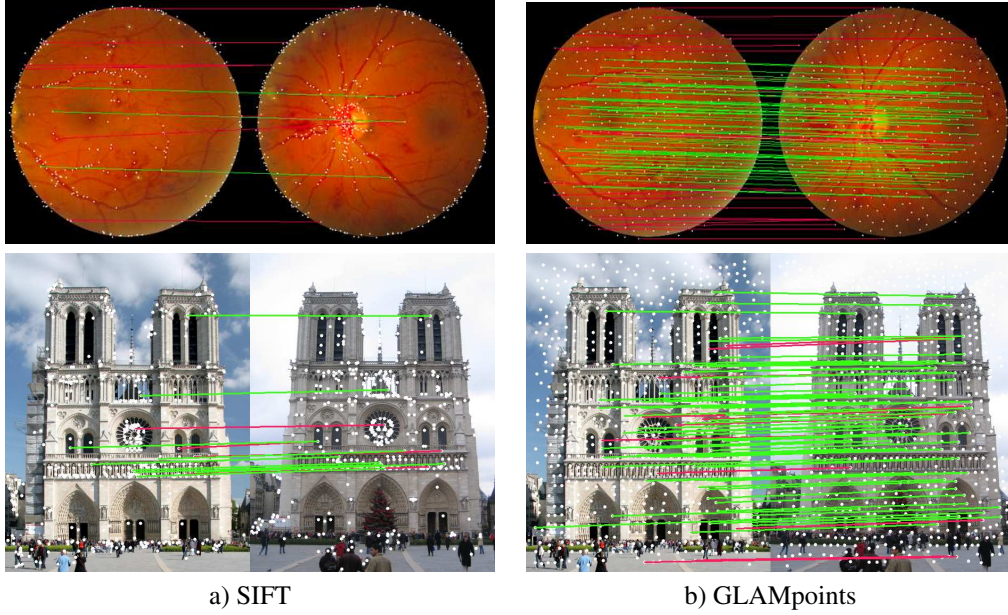a) SIFT   b) GLAMpoints

Figure 5: Interest points detected by a) SIFT and b) GLAMpoints and corresponding matches for a pair of images from the *FIRE* (top) and *Oxford* (bottom). Detected points are in white, green matches correspond to true positive, red to false positive.

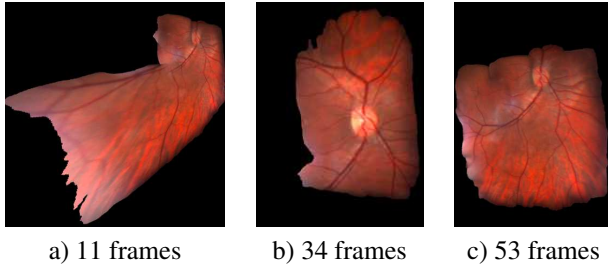a) 11 frames   b) 34 frames   c) 53 frames

Figure 6: Mosaics obtained from registration of consecutive images until failure. a) SIFT, raw images; b) SIFT, pre-processed data; c) GLAMpoints, raw data.

the average number of frames before the registration failed (due to the lack of extracted keypoints or correct matches between a pair of images). Over those 10 videos, the average number of registered frames before failure is 9.98 for GLAMpoints and only 1.04 for SIFT.

Example mosaics are presented in Figure 6. For the same video, SIFT failed after 34 frames when the data was pre-processed and only after 11 frames on the original data. In contrast, GLAMpoints successfully registered 53 consecutive images, without visual errors. The mosaics were created with frame to frame matching with the blending method of [19] and without bundle adjustment.

4.8. Run time

The run time of detection is computed over 84 pairs of images with a resolution of 660px by 350px. The

Table 3: Average detection run time [ms] per image for GLAMpoints and SIFT detector.

	GLAMpoints	SIFT
Pre-processing	0.0	16.64 $\pm$ 0.93
Detection image I	*CNN:* 16.28 $\pm$ 96.86 *NMS:* 11.2 $\pm$ 1.05	28.94 $\pm$ 1.88
Total	27.48 $\pm$ 98.74	45.58 $\pm$ 4.69

GLAMpoints architecture was run on a Nvidia GeForce GTX 1080 GPU while NMS and SIFT used CPU. Mean and standard deviation of run time for GLAMpoints and SIFT are presented in Table 3.

5. Conclusion

In this paper we introduce GLAMpoints - a keypoint detector optimized for matching performance. This is in contrast to other detectors that are optimized for repeatability of keypoints, ignoring their correctness for matching. GLAMpoints detects significantly more keypoints that lead to correct matches even in low textured images, which do not present many features. As a result, no explicit pre-processing of the images is required. We train our detector on generated image pairs avoiding the need for ground truth correspondences. Our method produces state-of-the-art matching and registration results of medical fundus images and our experiments show that it can be further extended to other domains, such as natural images.

References

[1] OpenCV: cv::BFMatcher Class Reference. 3

[2] OpenCV: cv::xfeatures2d::SIFT Class Reference. 2, 6

[3] Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vigas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. 6

[4] Alexandre Alahi, Raphal Ortiz, and Pierre Vandergheynst. FREAK: Fast Retina Keypoint. In *Conference on Computer Vision and Pattern Recognition*, 2012. 3

[5] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J. Davison. KAZE Features. In *European Conference on Computer Vision*, 2012. 2, 6

[6] Javier Aldana-Iuit, Dmytro Mishkin, Ondrej Chum, and Jiri Matas. In the Saddle: Chasing Fast and Repeatable Features. In *International Conference on Pattern Recognition*, pages 675–680, 2016. 5

[7] Hani Altwaijry, Eduard Trulls, Serge Belongie, James Hays, and Pascal Fua. Learning to Match Aerial Images with Deep Attentive Architecture. In *Conference on Computer Vision and Pattern Recognition*, 2016. 3

[8] Hani Altwaijry, Andreas Veit, and Serge Belongie. Learning to Detect and Match Keypoints with Deep Architectures. In *British Machine Vision Conference*, 2016. 3

[9] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012. 3

[10] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors. *CoRR*, abs/1601.05030, 2016. 2

[11] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2017. 3

[12] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006. 2

[13] Philippe C. Cattin, Herbert Bay, Luc Van Gool, and Gábor Székely. Retina Mosaicing Using Local Features. In *Conference on Medical Image Computing and Computer Assisted Intervention*, pages 185–192, 2006. 2, 5

[14] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine. A Partial Intensity Invariant Feature Descriptor for Multimodal Retinal Image Registration. *IEEE Transactions on Biomedical Engineering*, 57(7):1707–1718, 2010. 5

[15] Jian Chen, Jie Tian, Noah Lee, Jian Zheng, Theodore R. Smith, and Andrew F. Laine. A Partial Intensity Invariant Feature Descriptor for Multimodal Retinal Image Registration. *IEEE Transactions on Biomedical Engineering*, 57(7):1707–1718, 2010. 2, 5

[16] Artur V. Cideciyan. Registration of Ocular Fundus Images: an Algorithm Using Cross-correlation of Triple Invariant Image Descriptors. *IEEE Engineering in Medicine and Biology Magazine*, 14(1):52–58, 1995. 2

[17] Anders L. Dahl, Henrik Aanæs, and Kim S. Pedersen. Finding the Best Feature Detector-Descriptor Combination. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 318–325, 2011. 5

[18] Andrew Davison. Real-Time Simultaneous Localisation and Mapping with a Single Camera. In *International Conference on Computer Vision*, 2003. 2

[19] Sandro De Zanet, Tobias Rudolph, Rogerio Richa, Christoph Tappeiner, and Raphael Sznitman. Retinal Slit Lamp Video Mosaicking. *International Journal of Computer Assisted Radiology and Surgery*, 11(6):1035–1041, 2016. 5, 8

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 2, 6

[21] Paolo Di Febbo, Carlo Dal Mutto, Kinh Tieu, and Stefano Mattoccia. KCNN: Extremely-Efficient Hardware Keypoint Detection With a Compact Convolutional Neural Network. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 3

[22] Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT . Technical Report 1405.5769, arXiv, May 2014. 2

[23] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981. 2

[24] Luca Giancardo, Fabrice Meriaudeau, Thomas Karnowski, Tobin Kenneth W., Jr, Enrico Grisan, Paolo Favaro, Alfredo Ruggeri, and Edward Chaum. Textureless Macula Swelling Detection With Multiple Retinal Fundus Images. *IEEE Transactions on Biomedical Engineering*, 58(3):795–799, 2011. 5

[25] Yiliu Hang, Xiaofeng Zhang, Yeqin Shao, Huiqun Wu, and Wei Sun. Retinal Image Registration Based on the Feature of Bifurcation Point. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 2017. 2

[26] Chris Harris and Mike Stephens. A Combined Corner and Edge Detector. In *Fourth Alvey Vision Conference*, 1988. 2

[27] Carlos Hernandez-Matas, Xenophon Zabulis, Areti Triantafyllou, Panagiota Anyfanti, Stella Douma, and Antonis Argyros. FIRE : Fundus Image Registration dataset. *Journal for Modeling in Ophthalmology*, 4:16–28, 2017. 2, 5

[28] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent Temporal Variations in Many Outdoor Scenes. In *Con-

ference on Computer Vision and Pattern Recognition, 2007. 5

[29] Li Ma Jun-Zhou Huang, Tie-Niu Tan and Yun-Hong Wang. Phase Correlation-based Iris Image Registration Model. *Journal of Computer Science and Technology*, 20(3):419–425, 2005. 2

[30] Diederik. P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimisation. In *International Conference on Learning Representations*, 2015. 6

[31] Phil Legg, Paul Rosin, David Marshall, and James Morgan. Improving Accuracy and Efficiency of Mutual Information for Multi-modal Retinal Image Registration using Adaptive Probability Density Estimation. *Computerized Medical Imaging and Graphics*, 37(7-8):597–606, 2013. 2

[32] Stefan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: Binary Robust Invariant Scalable Keypoints. In *International Conference on Computer Vision*, 2011. 3

[33] P. Li, Q. Chen, W. Fan, and S. Yuan. Registration of OCT Fundus Images with Color Fundus Images Based on Invariant Features. In *Cloud Computing and Security*, pages 471–482, 2017. 2

[34] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 20(2):91–110, Nov 2004. 2, 5

[35] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 5

[36] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005. 5

[37] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning Local Features from Images. In *Advances in Neural Information Processing Systems*, pages 6237–6247, 2018. 2, 3, 6

[38] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. Mutual Information Based Registration of Medical Images: A Survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003. 2

[39] Roziana Ramli, Mohd Yamani Idna Idris, Khairunnisa Hasikin, Noor Khairiah A Karim, Ainuddin Wahid Abdul Wahab, Ismail Ahmedy, Fatimah Ahmedy, Nahrizul Adib Kadri, and Hamzah Arof. Feature-Based Retinal Image Registration Using D-Saddle Feature. *Journal of Healthcare Engineering*, 2017:1–15, 10 2017. 2

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015. 4

[41] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An Efficient Alternative to SIFT or SURF. In *International Conference on Computer Vision*, 2011. 3

[42] César A Sánchez-Galeana, Christopher Bowd, Eytan Z. Blumenthal, Parag A. Gokhale, Linda M. Zangwill, and Robert N. Weinreb. Using Optical Imaging Summary Data

to Detect Glaucoma. *Opthamology*, pages 1812–1818, 2001. 1

[43] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Franscesc Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In *International Conference on Computer Vision*, 2015. 4

[44] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, pages 298–372, 2000. 2

[45] Prune Truong, Sandro De Zanet, and Stefanos Apostolopoulos. Comparison of Feature Detectors for Retinal Image Alignment. In *ARVO*, 2019. 3, 6

[46] Yannick Verdie, Kwang Moo Yi, Pascal Fua, and Vincent Lepetit. TILDE: A Temporally Invariant Learned DEtector. *Conference on Computer Vision and Pattern Recognition*, pages 5279–5288, 2015. 5

[47] Gang Wang, Zhicheng Wang, Yufei Chen, and Weidong Zhao. Robust Point Matching Method for Multimodal Retinal Image Registration. *Biomedical Signal Processing and Control*, 19:68–76, 2015. 2, 5

[48] Simon Winder and Matthew Brown. Learning Local Image Descriptors. In *Conference on Computer Vision and Pattern Recognition*, June 2007. 5

[49] Simon Winder, Gang Hua, and Matthew. Brown. Picking the Best DAISY. In *Conference on Computer Vision and Pattern Recognition*, pages 178–185, 2009. 5

[50] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In *European Conference on Computer Vision*, pages 467–483, 2016. 2, 6, 7

[51] Kwang Moo Yi, Yannick Verdie, Pascal Fua, and Vincent Lepetit. Learning to Assign Orientations to Feature Points. In *Conference on Computer Vision and Pattern Recognition*, 2016. 5

[52] Liang Zhou, Mark S. Rzeszotarski, Lawrence J. Singerman, and Jeanne M. Chokreff. The Detection and Quantification of Retinopathy Using Digital Angiograms. *IEEE Transactions on Medical Imaging*, 13(4):619–626, 1994. 1

[53] Larry Zitnick and Krishnan Ramnath. Edge Foci Interest Points. In *International Conference on Computer Vision*, 2011. 5