MedRegNet – Unsupervised Multimodal Retinal-Image Registration with GANs and Ranking Loss

Monty Santarossa\textsuperscript{a}, Ayse Tatli\textsuperscript{b}, Claus von der Burchard\textsuperscript{b}, Lars Schmarje\textsuperscript{a}, Claudius Zelenka\textsuperscript{a}, Stefan Reinhold\textsuperscript{a}, Johann Roider\textsuperscript{b}, and Reinhard Koch\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, Kiel University, Germany
\textsuperscript{b}Department of Ophthalmology, Kiel University, Germany

ABSTRACT

Precise and robust multimodal image registration helps in spatially aligning medical finding from multiple sources for accurate medical diagnosis. We present MedRegNet, a lightweight descriptor module for registration multimodal retinal images. We utilize generative adversary networks (GANs) to learn generator networks which during training can synthesize structurally consistent multimodal image-pairs. From these image-pairs MedRegNet learns to predict stable point descriptors via a relaxed ranking loss. As no dataset-specific incentives are given, MedRegNet learns stable representations by itself. Most importantly, both mono- and multimodal training – including training the GANs – is entirely unsupervised, hence no costly expert annotation is required.

Through evaluation on the publicly available Fundus Image Registration Dataset (FIRE) as well as on our own multimodal dataset containing 340 retinal fundus, autofluorescence and fluorescein angiography image-pairs from 24 patients we show MedRegNet to improve robustness and registration capabilities of classical detector/descriptor algorithms like SIFT, ORB, KAZE and AKAZE. Despite using the same interest points, MedRegNet is matching more points successfully than either baseline even in the monomodal usecase. In the multimodal case, where the classical baselines fail due to large visual difference between the modalities, MedRegNet’s registration performance stays consistent to the monomodal case. Furthermore, MedRegNet shows to be adaptable to point detectors it was not trained on with no or only little cost in performance. MedRegNet can easily be integrated into any feature-based registration pipeline and due to the lack of dataset-specific incentives has the potential to be applied to fields outside retinal imaging.

Keywords: registration, multimodal, retinal images, deep learning, unsupervised, self-supervised, GAN, ranking loss

1. INTRODUCTION

In modern medical imaging, a variety of sensors is utilized to acquire image data. Hence, multimodal medical image registration remains an important field of research, allowing to spatially align findings from multiple modalities for more accurate diagnosis and prognosis. However, the multitude of image acquisition sensors provides several challenges in regards to registration: Spatial differences due to different viewpoints from the same or different sensors result in limited overlap between images. Temporal differences due to progression

![Figure 1: The MedRegNet Descriptor module as used in a typical 4-step feature based registration pipeline](image-url)
or treatment of a disease can see the change or even disappearance of structures (e.g. blood vessels). Finally, multimodality itself can cause vast visual differences between images. In the analysis and treatment of the eye disease central serous chorioretinopathy, where patients are observed over multiple years using a number of different sensors (see Sec. 4.1), we encounter all these challenges, necessitating precise and robust registration.

Such image registration algorithms can be sorted into two broad categories: area- and feature-based. Area-based methods generally aim to maximize between images a similarity metric like mutual information or phase-correlation. This presents a challenge for the multimodal and temporal case. In retinal images especially contrast and intensity are often non-uniform and large homogeneous non-vascular areas in combination with evolving pathologies further degrade performance of area-based approaches.

In contrast, feature-based methods working on matching correspondences between interest points visible in multiple images, are less dependent on similarity between whole image areas and as such appear better suited for multimodal retinal image registration. Typical feature-based registration methods as demonstrated in Fig. 1 work in four steps: 1) Detection of interest points in both images. 2) Description of those points. 3) Matching of corresponding points based on their descriptors. 4) Recovering a transformation between the images using good matches. It is step 2) that is of special importance in the multimodal case, where the same interest points can visually appear drastically different. Such is the case in retinal imaging with different lights and fluorescence (compare Fig. 3 and 9).

In this paper we propose MedRegNet, a descriptor module easily integrable into any 4-step feature-based registration pipeline. By extracting rectangular patches around interest points detected with an off-the-shelf point detector like SIFT, a lightweight CNN learns to predict descriptors via a relaxed ranking loss. Our key contributions are:

- MedRegNet, a lightweight CNN descriptor module which is not given any dataset-specific incentives, allowing it to learn – potentially multimodal – descriptors best suited for the current data by itself
- a training procedure utilizing CycleGANs to generate structurally stable multimodal image pairs, showing that MedRegNet can be trained completely unsupervised
- extensive evaluation on mono- and multimodal retinal datasets for each of the steps Detection, Description and final Registration, showing that MedRegNet on monomodal data improves registration quality and/or robustness over classical point detector/descriptor-algorithms and is able to keep both on multimodal data, where the baselines fail

2. RELATED WORK

We review in this section primarily non-area-based registration approaches on – preferably – multimodal retinal images.

Many older and recent works on retinal images rely on the segmentation or reconstruction of prominent structures such as the vascular tree for registration. Told et al. use CNNs to extract blood vessels from multimodal images, before using elastic frameworks for registration. A similar workflow without deep learning was proposed by Fang and Tang. Recently, Noh et al. utilized transfer learning to extract vessels in FAG sequences and Fundus images for the initial registration step. Similarly, Luo et al. depend on segmentation of the optic disc for the initial coarse registration of CGA and MCSL fundus images. In contrast, MedRegNet does not require any segmentation, which potentially is affected by disease progression and might be costly to annotate.

Other approaches may not require segmentation, but instead other features specific to retinal images. In Lee et al., multimodal registration is based on the patch-wise classification of vascular junctions in the retinal images. In Hervella et al., domain specific landmarks are fundamental for use in a domain-adapted similarity metric. Stewart et al. detect blood vessel centerlines as basis for iteratively expanding bootstrap registration regions. In our work however, MedRegNet does not receive dataset specific or even modality specific incentives.

To sample potentially matchable interest points, MedRegNet utilizes common point detectors like SIFT. Similar approaches are followed by others. Lin and Medioni use SIFT points and features for use in chain
registration of image sequences. Ma et al.\textsuperscript{21} extract SIFT features from the edge map of multimodal retinal images and match them with a gaussian mixture model. In Truong et al.\textsuperscript{22} SIFT is used as a feature descriptor after detecting interest points with a CNN. In MedRegNet, the Detector and Descriptor roles of CNN and SIFT compared to Truong et al.\textsuperscript{22} are reversed, as SIFT features in our case show to be insufficient for multimodal matching.

3. METHOD

The purpose of our registration pipeline is to register a moving image $I$ to a fixed image $I_f$ by estimating a transformation from the first to the latter. For this, our MedRegNet CNN generates descriptor-vectors based on patches extracted around interest points in $I$ and $I_f$. During inference, descriptors from both images are then matched according to their euclidean distance. We explain in this section the architecture of our descriptor model, the relaxed ranking loss we use to train it, the process of generating mono- as well as multimodal training data and finally our complete training procedure.

3.1 Descriptor Model Architecture

MedRegNet’s descriptor is a lightweight CNN which takes as input a single 64×64 px patch cropped from the image with the interest point in the centre. The network itself consists of 3 conv-layers (32, 64, 128 filters with kernel sizes of 7×7, 6×6, 5×5) with ReLU-activation plus 2×2 max-pooling and a final fully connected layer to predict a 1×128 description vector. This descriptor vector is l2-normalized in order to limit the euclidean distance between any two descriptors to $[0, 2]$. In practice, the descriptor network is executed in parallel for all detected interest points in both images.

3.2 Relaxed Ranking Loss

Given a point $p$ in $I$ and its corresponding point $p_{cor}$ in $I_f$, the descriptor model should produce descriptors $v \in [-1, 1]^{128}$ and $v_{cor} \in [-1, 1]^{128}$ such that their euclidean distance $\|v - v_{cor}\|_2$ is smaller than the euclidean distance $\|v - v_i\|_2$ from $v$ to all other descriptors $v_i \in [-1, 1]^{128}$ for points $p_i \neq p_{cor}$ in $I_f$.

Triplet loss approaches\textsuperscript{23–26} try to achieve this by sampling a triple $(v, v_{cor}, v_N)$, where $v_N$ stems from a non-corresponding point in $I_f$, and then apply some strategy to minimize $\|v - v_{cor}\|_2$ while maximizing $\|v - v_N\|_2$. This however often requires expensive sampling strategies like hard negative mining\textsuperscript{27–29} as $v_N$ needs to stem from an interest point similar to the interest point for $v$ in order to make the distinction non-trivial.

Hence, we instead adhere to other works\textsuperscript{11,30,31} and utilize a relaxed ranking loss based on average precision. For this, we consider a single point $p$ in $I$ and $k$ points $p_1, ..., p_k$ in $I_f$. We store correspondences between $p$ and $v_k$ in $I_f$.
and $p_i$ in a binary $k \times 1$-matrix $Y$, where 1 denotes a correspondence and 0 denotes no correspondence. While $p$ could have no corresponding point in $I_f$, during training we assert that $p$ has exactly one, which we call $p_{cor}$. We further calculate all pairwise descriptor-distances $||v - v_i||_2$ for $i \in \{1, ..., k\}$ and store those distances in the $k \times 1$-matrix $X$ as depicted in Fig. 2a.

Since our actual goal, ranking (i.e. sorting) descriptor-pairs $(v, v_i)$ based on their distance, is non-differentiable and as such ill suited for a loss, we instead closely follow Revaud et al.\textsuperscript{11} and make an approximation using histogram binning. This approach is based on two core ideas: 1) Instead of non-differentiable sorting by distance, we approximate an order of all $(v, v_i)$ by soft assignments to ordered bins of a histogram; soft meaning, that $(v, v_i)$ can partially be assigned to multiple bins with the sum of all these soft assignments being 1. 2) Iterating over each bin in the histogram we then calculate approximations for precision and incremental recall, resulting in a quantized average precision score. The earlier $(v, v_{cor})$ appears in the iteration compared to the other $(v, v_i)$, the higher the quantized average precision score is.

First, adapting the approach of Revaud et al.\textsuperscript{11} to our case, we discretize the space $[0, 2]$ of all possible descriptor-pair distances with $M$ bins spaced out equally in intervals of size $\Delta = \frac{2}{M-1}$, such that the position of bin $b_m$ with $m \in 1, ..., M$ is $b_m = (m - 1)\Delta$. Based on its distance $x = ||v - v_i||_2$, each $(v, v_i)$ can then be softly assigned to bin $b_m$ via the function $\delta(x, m) = \max(1 - \frac{|x - b_m|}{\Delta}, 0)$. A visualisation of the soft-assignment function is given in Fig. 2b. By applying $\delta$ to all distances $x \in X$, which we notate as $\delta(X, m)$, we get a $k \times 1$-matrix with soft assignments. Based on these assignments we utilize the approximations of Revaud et al.\textsuperscript{11} for the quantized precision $\hat{P}_m$ and a simplified incremental recall $\hat{r}_m$:

$$\hat{P}_m(X, Y) = \frac{\sum_{m=1}^{M} \delta(X, m')^\top Y}{\sum_{m=1}^{M} \delta(X, m')^\top 1}, \quad \Delta \hat{r}_m(X, Y) = \delta(X, m')^\top Y, \quad (1)$$

with $1 = (1, ..., 1) \in \mathbb{R}^k$. The resulting quantized average precision score $AP_Q$ is:

$$AP_Q(X, Y) = \sum_{m=1}^{M} \hat{P}_m(X, Y) \Delta \hat{r}_m(X, Y). \quad (2)$$

The resulting loss for a single descriptor $v$ is then $L = 1 - AP_Q(X, Y)$, which for the complete loss computation is averaged over multiple descriptors from points in $I$.

3.3 Unsupervised Data Generation

To utilize the relaxed ranking loss, we need to know which interest points correspond in both images and therefore also the transformation between the two input images $I$, $I_f$. In the monomodal case, we can achieve this by loading a single image and applying two different geometric transformations as well as pixel-level augmentations (contrast, brightness, gaussian noise) to synthesize two different images for which we know the registration transformation similar to previous approaches.\textsuperscript{22, 32} For the multimodal case however, we require two structurally equivalent images of different modalities, which is not possible to achieve with simple augmentation. Hence, we use CycleGANs,\textsuperscript{12} which can transfer images from one domain to another while keeping structural information (e.g. position and shape of blood vessels). With this we are able to generate images as in Fig. 3. While real and synthetic images are easily distinguishable in their entirety due to their borders, locally – i.e. on patch-size level – they are evidently similar enough to be used in our training.

Figure 3: Using generative networks (G) to synthesize multimodal image-pairs. MedRegNet requires only local similarity during training. All images are taken from the same patient on the same date.
3.4 Training Procedure
Our complete training setup looks as follows: We load a single image $I$ from our train-dataset and copy it as a second image $I_f$. For multimodal registration with $c$ modalities, $I_f$ with probability $\frac{c-1}{c}$ is transferred into another randomly chosen modality by the corresponding generative CycleGAN model. This ensures that for $I_f$ all modalities are equally likely regardless of $I$’s modality. To both $I$ and $I_f$ we apply pixel-level augmentations as well as geometric augmentations, yielding $I'$ and $I'_f$. We denote the resulting ground truth transformation from $I'$ to $I'_f$ as $H_{gt}$.

After improving contrast on $I'$ and $I'_f$ with CLAHE, an off-the-shelf feature detector is used to detect interest points in both images. From the interest points in $I'$ we filter out unmatchable points, i.e. those points that after transformation with $H_{gt}$ do not lie within an error threshold $\mu_{px}$ to any point in $I'_f$. From the remaining matchable points we keep only $n$. In the rare case that due to the augmentation less than $n$ points should be matchable in $I'$, a new random image of the same modality as $I$ is loaded and we repeat the process. For these $n$ points in both images we extract patches and let the MedRegNet model predict descriptors, which are utilized to calculate the relaxed ranking loss. As neither the GANs nor MedRegNet need any annotations for training, this process is completely unsupervised.

For registration during inference, following Lowe’s method, MedRegNet predicts descriptors for all detected points, which are then matched using Lowe’s ratio test with $r = \frac{|v-v_{f1}|}{|v-v_{f2}|}$ with $v_{f1}, v_{f2}$ being the first and second closest descriptors from $I_f$ to descriptor $v$ from $I$. Only matches with $r \leq \mu_{lowe} \in (0, 1]$ are kept. From inliers determined with RANSAC we recover a homography $H$ for transformation from $I'$ to $I'_f$.

4. EVALUATION

4.1 Datasets
The publicly available FIRE (fundus image registration) inference dataset consists of 134 fundus images-pairs of size 2912x2912 px, divided into categories $A$ (strong anatomical changes due to retinopathy, 14 pairs), $P$ (low overlap, 49 pairs) and $S$ (high overlap, no anatomical changes, 71 pairs). Pairs are annotated by a set of 10 corresponding control points in each image. FIRE has seen wide use to evaluate registration performance on retinal images. We use another public fundus dataset (1000 images) for training.

Our UKSH inference dataset consists of 340 image pairs from 24 patients annotated in the same way as FIRE. Images were acquired at the Universitatsklinikum Schleswig-Holstein. Of the 340 pairs, 117 are monomodal (44 FAF (fundus autofluorescence), 35 FAG (fluorescein angiography), 38 Fundus) and 223 multimodal (77 FAG++FAF, 85 FAG++Fundus, 71 FAG++Fundus). All patients are diagnosed with central serous chorioretinopathy and show slight to substantial structural damage in their retinal images. For each patient, images were acquired from two, preferably far apart dates (min. 8, max. 129, avg. 55 months, std = 34 months), to depict evolving symptoms. For monomodal pairs, corresponding images from both dates are matched. For multimodal pairs, images from the same date are matched. The dataset consists of both left-eye and right-eye images. For FAG, images from the early phase were chosen, where the dye had reached some, but not all blood vessels, to increase registration difficulty. FAG images were taken with a FOV of either 35°, 55° or 102°, FAF with 30°, Fundus with 50°. Images are scaled such that in the image center 200 $\mu m$ correspond to 15 px. For training, 72 single, unannotated FAF, 128 FAG and 38 Fundus images from 7 additional patients are used.

4.2 Metrics
We evaluate the steps Detection, Description and Registration separately. For Matching we use RANSAC with an inlier ratio of 15 px. We refer to any interest point in the moving image $I$ as $p$ and to any interest point in the fixed image $I_f$ as $p_f$.

Detection Metrics: We report as $\#DT$ the average number of possible matches per image-pair, which we define as the minimum of detected points in either image to provide context for the number of correct matches $TP$ (see Description Metrics). A high number of $\#DT$ can, but does not have to imply good spatial coverage of the detected interest point over the image. Hence, we also report MED (mean euclidean distance) to give an indication of how well covered $I$ is with matchable points. A point $p$ from $I$ is matchable if after transformation
with the ground truth transformation $H_{gt}$ it lies within distance $\mu_{px}$ to any point $p_f$. Then MED is the mean euclidean distance of each pixel in $I$ to the nearest matchable point. If zero points should be matchable, MED is set to the diagonal length of the image.

**Descriptor Metrics:** We call TP the absolute number of correct matches for a fixed error threshold $\mu_{px}$. A match $(p, p_f)$ is correct, if $p$ after transformation with $H_{gt}$ lies within distance $\mu_{px}$ to point $p_f$. In accordance with previous works\cite{22,41} we also show matching performance via a Receiver Operating Characteristic (ROC) curve of correct matches as a function of $\mu_{lowe}$. We call the area under this ROC curve rAUC to avoid confusion with the registration metric AUC. A high rAUC-value means that correct matches overall have smaller Lowe’s ratios, hence get matched earlier than incorrect matches, while a high number of correct matches TP can imply better coverage.

**Registration Metrics:** AUC as defined in Hernandez-Matas et al.\cite{35} is the area under curve when plotting the percentage of successfully registered image-pairs in the dataset over a varying error threshold. An image-pair is successfully registered if its median error (MEE) is smaller than the error threshold $\mu_{px}$. MEE is defined as the mean distance between the two corresponding points in each control point pair after registration with the predicted transformation $H$.

We note that MED and rAUC depend on a fixed error threshold $\mu_{px}$ and AUC on a fixed Lowe’s ratio threshold $\mu_{lowe}$. To avoid selection bias, we present MED and rAUC for multiple $\mu_{px}$ and AUC as mAUC and bAUC, where mAUC is the mean AUC value over $\mu_{lowe} \in \{0.5, 0.51, \ldots, 0.99\}$ and bAUC the best (i.e. max) AUC for any $\mu_{lowe}$.

### 4.3 Hyperparameters and Training Setup

Individual MedRegNet models were trained either on a single NVIDIA TITAN X or a single NVIDIA GTX 1070 for a maximum of 500 epochs with a learning rate of 0.001 and the ADAM\cite{42} optimizer. Standard augmentations (scale, rotate, horizontal flipping of both images, perspective transform, brightness, contrast) were applied during training. For filtering non-matchable points in $I$ we use an error threshold of $t = 3$ px. From $I$, at least $n = 30$ interest points must be matchable, otherwise $I$ is discarded and replaced with a random image of the same modality. To speed up training, from $I_f$ a maximum of 1400 randomly chosen interest points are kept. For the ranking loss, $M = 25$ bins were used.

For training the CycleGAN, we utilized the tensorflow implementation of He\cite{43} without changes, using only images from the train-dataset of UKSH for training and validation. Hence, the UKSH validation set is never seen by the CycleGAN.

### 4.4 Monomodal results on FIRE

We compare MedRegNet against the classical detector/descriptor-algorithms SIFT,\cite{10} ORB,\cite{44} KAZE\cite{45} and AKAZE.\cite{46} From Fig. 4 we see that SIFT and ORB on FIRE detect a significantly higher average number of points than KAZE or AKAZE, i.e. 1104 and 2549 vs. 507 and 418 respectively. MED scores as well as visual

![Figure 4: Detection Results on FIRE](image)

<table>
<thead>
<tr>
<th>method</th>
<th>detector</th>
<th>descriptor</th>
<th>$rAUC_5$ (TP)</th>
<th>$rAUC_{15}$ (TP)</th>
<th>$rAUC_{25}$ (TP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>SIFT</td>
<td></td>
<td>.887 (201)</td>
<td>.893 (280)</td>
<td>.898 (310)</td>
</tr>
<tr>
<td>ORB</td>
<td>ORB</td>
<td></td>
<td>.744 (185)</td>
<td>.742 (305)</td>
<td>.741 (347)</td>
</tr>
<tr>
<td>KAZE</td>
<td>KAZE</td>
<td></td>
<td>.879 (130)</td>
<td>.887 (194)</td>
<td>.893 (230)</td>
</tr>
<tr>
<td>AKAZE</td>
<td>AKAZE</td>
<td></td>
<td>.828 (134)</td>
<td>.875 (185)</td>
<td>.890 (210)</td>
</tr>
<tr>
<td>SIFT</td>
<td>oursSIFT</td>
<td></td>
<td>.895 (236)</td>
<td>.890 (468)</td>
<td>.905 (515)</td>
</tr>
<tr>
<td>ORB</td>
<td>oursSIFT</td>
<td></td>
<td>.690 (481)</td>
<td>.692 (1121)</td>
<td>.701 (1451)</td>
</tr>
<tr>
<td>ORB</td>
<td>oursORB</td>
<td></td>
<td>.631 (475)</td>
<td>.679 (1104)</td>
<td>.695 (1409)</td>
</tr>
<tr>
<td>SIFT</td>
<td>oursORB</td>
<td></td>
<td>.879 (234)</td>
<td>.891 (460)</td>
<td>.892 (597)</td>
</tr>
</tbody>
</table>

Table 1: Description Results on FIRE, where ours\textsubscript{D} is MedRegNet trained with detector $D$
Figure 5: Detector comparison on FIRE. Shown is the image where SIFT detects a number of points closest to its average on FIRE. Note that SIFT and ORB have significantly better coverage than KAZE and AKAZE, which mainly detect points only along the largest blood vessels.

inspections like Fig. 5 suggest that these higher numbers indeed correlate with a better spatial coverage in FIRE’s fundus image pairs. Hence, we choose SIFT and ORB as the detectors to train MedRegNet on.

Regarding Description scores shown in Tab. 1, despite detecting more than twice as many points as SIFT, ORB reaches only similar absolute TP numbers, implying that ORB’s detected points are not easily matched. We note that MedRegNet in all cases significantly improves on TP matching numbers over both SIFT and ORB, doubling or even tripling their numbers for error thresholds ≥ 15 px. We also see that it matters only little on which of these detectors MedRegNet was trained. rAUC scores as well as TP stay almost identical for differently trained models if the same detector is used during inference. This shows the robustness of MedRegNet in learning point descriptions not specific to a single detector.

Final Registration scores on FIRE are depicted in Tab. 2 and Fig. 6. We see that both MedRegNet models using SIFT as detector as well as SIFT itself reach highest mAUC and bAUC scores. Though in this case mean registration performance on FIRE is equal between SIFT and MedRegNet, the aforementioned higher number of available TP matches implies that recovering the transformation with MedRegNet is more robust. This is supported by results on the subset A with strong anatomical changes, where all MedRegNet approaches score at least .08 higher bAUC than SIFT.

Comparing against ORB, we see that in this case the higher number of TP matches when using MedRegNet as descriptor does indeed yield better registration scores. This holds especially true for the more difficult subsets P and A. Again, the performance gap between MedRegNet models trained on different detectors is only very slight, though in this case the model trained on SIFT (i.e. a different detector than used here) scores marginally higher.

Table 2: Registration Results on FIRE and its subsets S, P, A, where oursD is MedRegNet trained with detector D. The plots for each bAUC are shown in Fig. 6.
We want to note that our scores do not reach SOTA like Hernandez-Matas et al.\textsuperscript{39} since we use a homography transformation whereas Hernandez-Matas et al.\textsuperscript{39} model the eye shape for transformation. For example, in Hernandez-Matas et al.\textsuperscript{39} SIFT as detector/descriptor is already achieving .721 bAUC compared .685 in our case. Accordingly, it is the low overlap category $P$ that is most challenging for all our methods. This is to be expected, as images in $P$ see the highest distortions which can not be perfectly reconstructed by a homography (compare Fig. 7).

Table 3: Description results on UKSH. \#DT numbers for each detector are given as context for $TP$. The grid detector consists of placing points in an equally spaced 109x109 grid over the entire image.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{Method} & \textbf{Detector} & \textbf{Descriptor} & \textbf{monomodal} & \textbf{multimodal} \\
\hline
SIFT & SIFT & \textbf{rAUC}_5 (TP) & \textbf{rAUC}_{15} (TP) & \textbf{rAUC}_{25} (TP) \\
\hline
ORB & ORB & .860 (257) & .836 (305) & .812 (325) \\
\hline
KAZE & KAZE & .698 (195) & .674 (260) & .657 (290) \\
\hline
AKAZE & AKAZE & .861 (720) & .847 (935) & .834 (991) \\
\hline
SIFT & oursSIFT & .817 (1509) & .823 (2656) & .814 (2767) \\
\hline
ORB & oursSIFT & .637 (2621) & .684 (3807) & .663 (3894) \\
\hline
KAZE & oursSIFT & .710 (1177) & .750 (1767) & .747 (1822) \\
\hline
grid & oursSIFT & .773 (4255) & .779 (6900) & .774 (7100) \\
\hline
\end{tabular}
\end{table}
4.5 Mono- and Multimodal Results on UKSH

Following our findings on FIRE, for UKSH we train MedRegNet on SIFT points. During evaluation, SIFT, ORB, KAZE as well as grid are used as detectors, where grid is not a true detector, but simply sampling points from an equally spaced $109 \times 109$ grid over the image.

Description results are depicted in Tab. 3. Analyzing the monomodal results first, we see MedRegNet again improving TP numbers over all detector/descriptor baselines with increases ranging from a factor of 1.6 up to a factor of 13.4. These improvements are especially notable against ORB, which despite highest #DP numbers on UKSH matches the least TP. Again this indicates that ORB points on retinal images are not easily matched. Despite, using ORB as a detector MedRegNet increases TP by over an order of magnitude. In all monomodal cases MedRegNet’s additional TP come at a slight cost in rAUC compared to the baselines.

On multimodal data, the classical detector/descriptor baselines as expected achieve only very few correct matches, all averaging less than 100 TP. Yet, both MedRegNet’s TP (in relation to #DT) and rAUC stay similar to results on monomodal data. This showcases MedRegNet’s ability to generate descriptors invariant to modality. This is most strongly supported by the fact that MedRegNet’s rAUC stays consistent even when using points sampled from grid, despite many of those points being located in areas without significant structures and MedRegNet having only seen SIFT points during training.

Registration results for each modality pair are shown in Tab. 4 and Fig. 8. Evidently, MedRegNet’s higher TP numbers compared to the baselines significantly improve registration performance due to better spatial coverage of the matched points. This holds true for all baselines and all modalities with the sole exception of KAZE on monomodal FAF, where performance is equal. In addition, with two exceptions for MedRegNet with ORB points, MedRegNet regardless of the used detector achieves higher bAUC than all baselines.

The baselines fail on all multimodal datasets except FAF ↔ Fundus, where blood vessels in both modalities appear darker than the surrounding tissue. While it would be unfair to expect differently for the baselines, the data also shows that MedRegNet using the same points reaches multimodal registration performance equivalent in bAUC to its performance on monomodal data – which in turn is as good or better than the baseline’s, indicating MedRegNet’s robustness on multimodal data.

5. CONCLUSION

We propose MedRegNet, a lightweight CNN descriptor module for use in feature-based registration pipelines capable of robust multimodal registration. Using CycleGANs to generate structurally stable multimodal image pairs, MedRegNet can be trained completely unsupervised with a ranking loss.

We evaluate performance on mono- and multimodal fundus, fundus autofluorescence and fluorescein angiography image pairs. Results on monomodal data show that MedRegNet compared to the classical detector/descriptor baselines improves TP numbers over all detector/descriptor baselines with increases ranging from a factor of 1.6 up to a factor of 13.4. These improvements are especially notable against ORB, which despite highest #DP numbers on UKSH matches the least TP. Again this indicates that ORB points on retinal images are not easily matched. Despite, using ORB as a detector MedRegNet increases TP by over an order of magnitude. In all monomodal cases MedRegNet’s additional TP come at a slight cost in rAUC compared to the baselines.

On multimodal data, the classical detector/descriptor baselines as expected achieve only very few correct matches, all averaging less than 100 TP. Yet, both MedRegNet’s TP (in relation to #DT) and rAUC stay similar to results on monomodal data. This showcases MedRegNet’s ability to generate descriptors invariant to modality. This is most strongly supported by the fact that MedRegNet’s rAUC stays consistent even when using points sampled from grid, despite many of those points being located in areas without significant structures and MedRegNet having only seen SIFT points during training.

Registration results for each modality pair are shown in Tab. 4 and Fig. 8. Evidently, MedRegNet’s higher TP numbers compared to the baselines significantly improve registration performance due to better spatial coverage of the matched points. This holds true for all baselines and all modalities with the sole exception of KAZE on monomodal FAF, where performance is equal. In addition, with two exceptions for MedRegNet with ORB points, MedRegNet regardless of the used detector achieves higher bAUC than all baselines.

The baselines fail on all multimodal datasets except FAF ↔ Fundus, where blood vessels in both modalities appear darker than the surrounding tissue. While it would be unfair to expect differently for the baselines, the data also shows that MedRegNet using the same points reaches multimodal registration performance equivalent in bAUC to its performance on monomodal data – which in turn is as good or better than the baseline’s, indicating MedRegNet’s robustness on multimodal data.

Table 4: Registration Results on UKSH. The plots for the bAUCs and qualitative results can be seen in Fig. 8 and 9 respectively.
descriptor-algorithms like SIFT it uses as detectors significantly improves registration performance and/or, without negative impact, makes it more robust due to significantly increasing the absolute number of correct matches. On multimodal data, where the classical baselines fail, MedRegNet keeps its stable registration performance. Our evaluation also shows that a MedRegNet trained on one detector can in many cases be utilized on other detectors without significant loss in registration performance.

Still, in the future we plan to implement a specialised detector module and train both detector and descriptor module together. This would allow to learn detecting those points that can best be matched by MedRegNet, further improving registration performance and robustness.
Acknowledgements

We acknowledge funding by the KI-SIGS project (Grant number FKZ 01MK20012E) funded by the Federal Ministry for Economic Affairs and Climate Action (BMWi, Germany). We acknowledge funding of L. Schmarje and S. Reinhold by the ARTEMIS project (Grant number 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF, Germany). We would also like to thank Timo Kepp from the Institute of Medical Informatics, University of Lübeck, for his valuable advice.

REFERENCES


