# Automatic Non-rigid Histological Image Registration based on SIFT and RegSnet

Hao Bai

h.bai@se18.qmul.ac.uk

ID: 180223545

**Acknowledgement**

I would like to express my special thanks of gratitude to my supervisors Dr. Qianning Zhang and Yaqi Wang. My supervisors kindly advise and guide me during this research. I came to know about so many knowledge, technique and skills from my supervisors. I am really thankful to them.

**Abstract**

The challenge of the histological image registration task is that the size of each image is huge up to 50k x 50k pixels and each image’s size is different.

We proposed a modular approach to solve the automatic non-rigid registration on histological images issues. Our approach combined the SIFT local features extraction and convolutional neural network. The SIFT feature extract the interested point coordinate based on each image. Then expand the points to a small image region based on surrounding pixels. We developed a convolutional neural network named RegSnet which analysis both image patches and give each pair of image patches a registration score. The system includes a ranking function to sort the score and return top N pairs as the output of the approach. Our approach is highly modular that others can change component to make tradeoff between the accuracy and speed.

In order to compare performance, we also developed a deep learning method and GANs method to generate the landmarks. The evaluation method we use the rTRE which is a accuracy evaluation of the registration tasks. According to the results during experiment, our approach is much better than the deep leaning method and GANs in this dataset.

**Introduction**

Different from matching function, registration is transforming different set of images into one coordinate system. In medical image processing elastic registration need to respond to the object deformation. The deformable registration allowed an un-uniform mapping between images. The state of art method is to measure and correct small, varying discrepancies by deforming one image to match the other (ref).

There are several methods can detect local interest points like SIFT, SURF, ORB and so on. This registration method compares the performance of the SIFT, SURF and ORB local interest point detection method. Though, the SURF and ORB method is faster than SIFT. The quality and quantity of the detected points is not as good as the SIFT detector. (need a form and some picture)

Scale-invariant feature transform (SIFT) is a traditional algorithm which detect and describe local features (interest points) in images. SIFT algorithm can implement matching task based on best-bin-first search algorithm which deal with the nearest neighbor search problem in high dimensional spaces. And filter out the useful pairs based on Euclidean distance. Deep learning techniques is widely implemented in computer vision task. Many researchers implement deep learning techniques on image registration tasks. Due to large image size, small datasets and histological image content issues, this project method combined Scale-invariant feature transform (SIFT) features detection and deep learning techniques to retrieve paired landmarks from images. However, the SIFT method not focus on the content of the original images. The SIFT matching protocol use Euclidean distance between the paired points and a threshold based on Euclidean distance to filter out the useful pairs. Compared to SIFT matching method, our method expands the coordinate into small image patches and put them into the network. The small patches including the surrounding pixels around the points. The paired patches will be analysis by the network the output is a score. We rank the output scores and take top N pairs as output.

(The challenge of the registration of the medical images is that the body might change posture even skeletal arrangement can change)

The main challenge of the non-rigid registration of the histological images is required to sensitive to the human tissue deformation and retrieval useful landmarks. Due to the big histological images size, some traditional methods are time consuming. And the textures in the images not homogeneous, so when make use of the SIFT or other traditional method might get huge amount of the points on the texture area instead of the area required.

The device is also a considerable problem for this challenge. Due to the high quality of the image, the image size become a big issue when processing the image. Limited require of this method is the memory of running device is at least 8gb to ensure the image can be loaded and processed.

To solve the problem, this method will scale each image 50 times before using SIFT to detect local interest point from each image. And focus on the surrounding pixels of each landmark, instead of processing the whole images.

There is a problem in landmarks generation directly by deep learning method that the image is required to be the same size before putting the image into the CNN. The multiple sizes images in the dataset is impossible to import to the CNN unless preprocessing the images into same size. However, the preprocessing might cause loss useful information’s or the object deformation when cutting or shrinking large images. When zooming or padding the images, it is required to process the image into the largest images’ size. That will increase the CNN learning time and the object deformation. To face the multiple sizes problem, our method takes local detector at the beginning. And then, out method expanded the point into same size image regions. That solution is a good way to deal with different input size.

The aim of the function is to get the registration landmarks based on paired images. We propose a modular solution that combined local interested points technique and deep learning method. The first part, our method using SIFT detector to get each image key points. The second stage, after using local interest points detector detected interest points from each image. We expand the landmarks to 16 x 16 image patches. The third stage, we aim to use these patches to filter out the real matching landmark pairs. The basic idea proposed by David G. Lowe in 2004 which based on k-Nearest Neighbors search to match paired landmarks and use a Euclidean distance threshold to filter out feasible paired landmarks. However, the performance is not good on the histological images. Our registration method filtering is based on the ranking of the output of the convolution neural network (CNN). The CNN output determines the pair of the small patches score. In the fourth stage, we rank the scores and take top N pairs. The k top of the pairs is the retrieved landmarks of each images. The amount K can be fixed or dynamic. All the stages can be changed to adapted different dataset.

We also implement a method that directly used CNN to get the landmarks. The result implemented on this dataset is not good. The dataset including only 230 pairs of images. After training, the network is either underfitting or overfitting. And the landmarks points including negative number which is useless. To solve this problem, we get some ideas on the generative adversarial network (GAN). Which use a discrimination network to determine the generated image is fake or true. We added a discrimination network in training process. However, due to the large amount of the generated landmarks including negative number and the small dataset and the loss function is changed to not based on the error distance. The results are worse than the CNN method.

**Related work:**

The traditional method using SIFT to match the two images which developed by David G. Lowe in 2004. The SIFT detector can get local features of each images. This approach using SIFT detector to get local interested point and then using the Brute-Force matching (KNN) to get the paired points. Then using Euclidean-distance to filter out the points which above the threshold value. SIFT approach robust to get local scale invariant features and doesn’t need to resize the paired images to the same size. However, the SIFT matching function not based on the content of paired images which might lose useful information or can not distinguish between useful and less useful landmarks.

Deep learning method is a modern approach in image processing task. Many researchers use this method to deal with the registration task. There is a FCNs network using deep learning method directly generate the 3D landmarks from paired 3D images. They implement an FCN network that the input is paired of 3D images. After the network processing, the output will separate the X coordinate and Y coordinate of both images. This network analysis both image features and processed to each image’s landmarks. That is a fast way to generate the result. However, we implement the network and find that this method has some limitations. The method cannot suitable for the dynamic size input images. And the result is unstable that some coordinate is negative value which is useless landmarks. The amount of the generated landmarks is fixed in training and using stage.

Another function is implemented by IBM, the method determines the probability of each pixel which can to the registration landmarks. This method put the two images directly into the network and the result is two matrixes with each pixel probability. This method required high memory and good graphics card of the devices to process the images. Though this method, the amount of the retrieved landmarks can be dynamic the result can be more accrue than the above one. However, due to the small dataset, large image size, this method is required more data to training to let the network learn the parameters to reduce underfitting. The input image size of this method needs to preprocess the original images.

MatchNet is classical deep learning method deal with registration tasks. The network consists of a feature network and a metric network. The feature network in MatchNet is influenced by AlexNet which can achieve better performance for object recognition. The feature network in MatchNet including 5 convolution layers and 3 pooling layers. The metric network is make up of 3 fully connect layers to measure the patches matching probability. The input of this network is a pair of image patches, the output is two values from the two units of the 3rd fully connect layer. However, this network is just to measure images or image patches match or not. The registration points of each image are not considered by this method.

There

…More

**Dataset:**

The Histology (CIMA) dataset is offered by Phd.Jiří Borovec. The images and landmarks data are free to use in adapted and share for any purpose. This dataset will be only used in this research and never be used in commercial. This dataset included 480 high-resolution (up to 40x magnification) whole-slide images of different types of tissue (lesions, lung-lobes, mammary-gland). The size of images in dataset can be varied from 15k x 15k to 50k x 50k pixels. The images are acquired by consecutive tissue slices and each slice dyed by different stains. The dyes are showed below:

The dataset including 50+ histological sets which is organized by whole slide images of each folders. Each folder is corresponding to the tissues dyed by different stains.

Image content

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Tissue | Availability | Resolution [µm/pixel] | Avg. size [pixels] |
| lung-lesion\_ | Lung lesion | Public | 0.174 | 18k×15k |
| lung-lobes\_ | Whole mice lung lobes | Public | 1.274 | 11k×6k |
| mammary-glands\_ | Mammary glands | Public | 2.294 | 12k×4k |
| mice-kidney\_ | Mice kidney | Public | 0.227 | 37k×30k |
| COAD\_ | COlon ADenocarcinoma (colon cancer) | Public | 0.468 | 60k×50k |
| gastric\_ | Gastric mucosa and gastric adenocarcinoma tissue fragments | Public | 0.2528 | 60k×75k |
| breast\_ | Human breast | Public | 0.2528 | 65kx60k |
| kidney\_ | Human kidney | Public | 0.2528 | 18kx55k |

The more details about the images like the microscope model can be find at Phd.Jiří Borovec official website.

**Requirement:**

The goal of the non-rigid image registration task is to analysis paired images and retrieval a pair of corresponding landmarks.

The method required to process the paired images at the same time

The size of the images in the dataset is not same. Resizing the images to same size cause the proportion of the objects in the images changing. Cutting images into same size will cause the information loose. Padding the images into same size cause the process time longer. So that the registration method required to doing the task without resizing the images.

Dynamic amount of the output landmarks based on the images (Not enhancement the performance)

**Design:**

There are three preliminary implementation methods to make the registration task.

The main idea is combined the local key points descriptor and global matching method which influenced by the classic matching method. To face the different image size issues, the designed process using local feature detector to get the local interested point to filter out useful landmarks. The input of the local feature detector technique does not require same size images. This method will compare the performance between SIFT, SURF and ORB.

The first stage of SIFT is Scale-space extrema detection. The SIFT descriptor convolved the images with Gaussian filters with different scales and get different Gaussian-blurred images. Maxima and minima point of the Difference of Gaussians (DoG) that occur at multiple scales is the points that will be taken at this stage.

The next stage is key point localization. The previous stage will detect many candidate landmarks. This stage is required to filter out the points that sensitive to noise or poor positioning along the edges. And get accuracy location of each points.

In Orientation assignment step, each key point will be assigned local image gradient directions. This step can achieve invariance to image rotation.

The last step of SIFT is key point descriptor. This step will compute a descriptor vector for each key point. This stage makes each keypoints highly distinctive and partially invariant to the illumination or 3D viewpoint.

After getting each local candidate points from each image, we need to match (registration) points to paires. The designed matching function is k-nearest neighbors’ algorithm (KNN). This technique matches each pair based on minimum Euclidean distance. This part is same as the SIFT matching technique.

In this way, the matched paired landmarks included many useless pairs points and un-matching pairs. The amount of ground truth landmarks for each image is much less than the retrieved pairs. Each image is about 60 to 75 pairs of landmarks. Therefore, we need a tool to filter out the meaningful pairs. The SIFT matching use a threshold based on Euclidean distance. We want to distinguish the matched pairs and un-matched pair based on the content of the images. So, we decide to use deep learning method to distinguish a pair of the landmarks is matched or not and generate a score which measures each image patches registration value. Our basic idea is to expand each paired point into small image patches based on surround pixels and put these patches into a convolutional neural network. The convolutional neural network is to analysis each pair and give each pair a score. The score can be ranked to filter out useful landmarks. The top N pairs can be retrieved as the output landmarks for each image.

The evaluation method is designed to use relative Target Registration Error (rTRE) which measures geometric error between the ground truth and generated landmarks based on target image. This accuracy measure protocol needs to calculate Target registration error (TRE) first which is the distance between ground truth points and the retrieved points after registration. The Target Registration Error (rTRE) is the Target registration error (TRE) divided by the image diagonal length. The final rTRE for the whole retrieved landmarks is the mean value of all Target Registration Error (rTRE) of all the generated and ground truth landmarks of this image. We also evaluate the time spending of the registration task processing on the two images. Each processing time measured from the reading of the two images to the paired landmarks generated. We will take the average time spending of the whole dataset.

Another method is to GANS

**Our method detail:**

Network:

RegSnet is a deep learning architecture. This network jointly learns features and analysis the splice features together. This network including different types of layers that handles features extracting, mapping and analyzing. Our RegSnet network architecture is influenced by MatchNet. MatchNet is a good architecture for jointly processing the paired image patches in a single network.

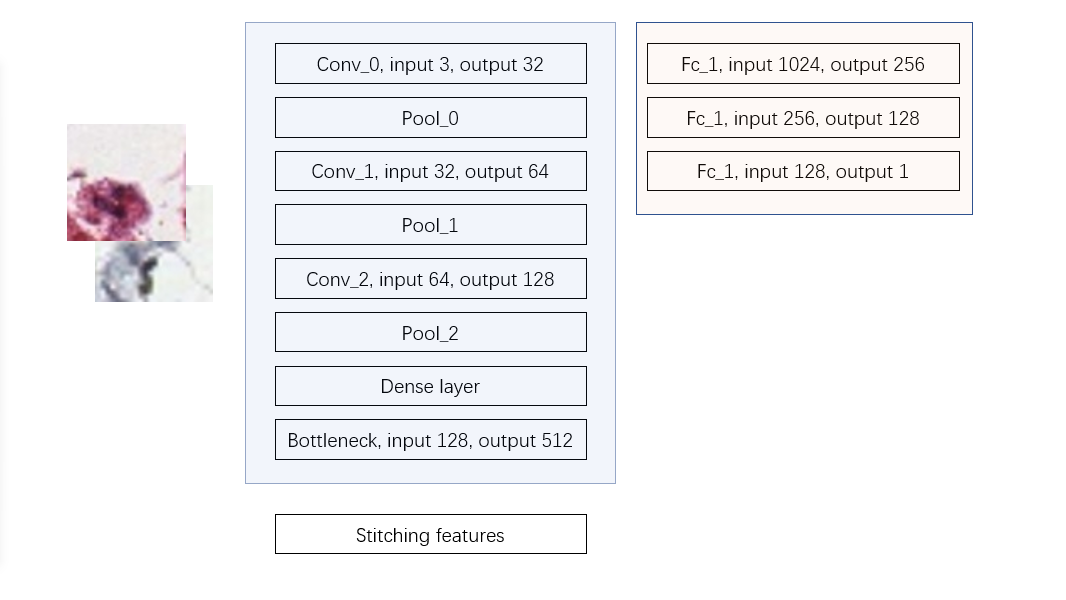
The basic idea of our network is to identify the paired image and give them a registration score. The score represents the paired images can be a registration landmark or not. The output scores will finally send to the ranking function. Therefore, we need modified the MatchNet network among the input images size, the inner layers and the output.

Our network consists a feature network and a metric network.

The RegSnet including 3 convolutional layers, 3 pooling layers and 3 fully connected layers. The activation function of each layer expects the last fully connected layer is linear Relu. The last layer activation function is Sigmoid because the expectation output is a number between 0 to 1. The output scores use for distinguishing paired of small image regions whether can satisfied the registration points requirements or not. Binary Cross Entropy (BCE loss) loss function is selected in training process. The optimizer is Stochastic gradient descent (SGD).

**Feature network**: The feature network inspired by AlexNet which have a good combination of the convolutional layers and pooling layers. The inspired network achieves a good performance on object recognition tasks. In our feature network, the convolutional layers and pooling layers works together to map each image feature separately. The bottleneck layer will stretch each image’s features and splice these features.

**Metric network**: The metric network measures the similarity between the spliced these features using fully connect layers. The activation function of the fully connect layers is ReLu and Softmax which is the last layer activation function. The input of this part is the output of the feature network which is the spliced features of the two images. The output of this network is the score in [0,1] of the two images which is also the output of the whole network. The score is a positive value which represent the probability of the paired images can be a registration point.



conv\_0 is a convolutional layer that the input channel is 3, output channel is 32, kernel size is 2 x 2, stride is 1 x 1. pool\_0 is a pooling layer that kernel size is 2, stride is 2. conv\_1 is a convolutional layer the input channel 32, output channel is 64, kernel size is 2 x 2, stride is 1 x 1. pool\_1 is same as pool\_0. conv\_2 is a convolutional layer that the input channel is 64, output channel is 128, kernel size 2 x 2, stride is 1 x 1. pool\_2 a pooling layer which is same as pool\_1 and pool\_0.

Bottleneck is a linear layer (fully connected layer) that input features is 128, output features is 512. After the Bottle neck layer, the pair images features will be spliced together (512 + 512 = 1024) and send to next layer. fc\_1 is a Linear that input features is all the paired images features (1024), output features is 256. fc\_2 is a Linear layer, the input features is 256, output features is 128. fc\_3 is a Linear layer which input features is 128, output features is 1.

Training:

We expand each corresponding ground truth landmark points to 16 x 16 region pair with surrounded pixels. Then put this pairs of regions into the network and the expectation output is true (1). We also trained with un-matching points with expectation output is false (0). The loss function is Binary Cross Entropy (BCE loss) loss. The optimize function is Stochastic gradient descent (SGD).

Registration processing:

The registration method inspired by classic SIFT descriptor matching which develop David G. Lowe. The classic method uses SIFT descriptor to get interest point. And then, the method uses the KNN matches to get paired key points. Finally, the sift matching method using a Euclidean distance and threshold to filter out the better landmarks. However, the SIFT descriptor matching is a local matching method which not considered the matching feature pair content in each image.

Our registration method combined local interest points detection function and global pair points matching discriminate methods. The local interest point detection reduced large number of useless points of each image. That can reduce the processing time in the paired points distinguishing stage. Our method uses SIFT point detector to get local interest point from each image separately. This part is same as SIFT matching. Then we use the Brute-Force matcher (KNN) to get the paired points based on the points SIFT detected. We expand each point to 16 \* 16 small image regions based on the around pixels. We will put the pair sets of regions to the trained RegSnet network to get a registration point score. We filter out this score and get the top N landmarks as the output of this registration task.

Train algorithms

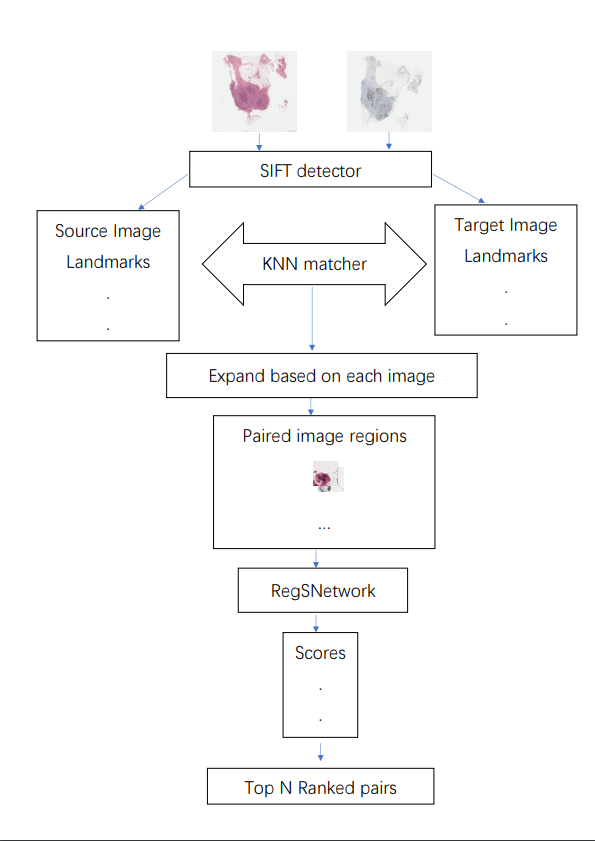
For each pair images in training images set

Using the ground truth pair landmarks and expand the points to 16 x 16 small images

Put pairs of matching small images into the network the expectation output is true (1)

Put pairs of unmatching small images into the network the expectation output is false (0)

Calculate the loss and update the network parameters



Registration algorithms

For each pair images need registration

Using scale-invariant feature transform (SIFT) to detect features in each image

Expand the points to 16 x 16 small images

Put the images to the trained network and record each score

Ranking the score and take the top n pairs

Record the landmarks

After using SIFT detector, the local interested point needs to be matched in to pairs. The matching function is using the KNN search to get pairs. The traditional method is using Euclidean distance as a threshold value to filter out the useful pairs. The performance is not as good as we expected. So, we take RegSnet to filter out the useful pairs. However, simply input the landmark pixels or coordinate into the network can not obtain a good solution to the non-rigid registration due to the deformation of the object in the image might have. We expanded the pixels to 16 x 16 small image regions that surrounded the pixels. And put the paired small regions into the network to get a score. Using small image region for training instead of the whole images can expand our training data set and reduce underfitting of the network and the processing time in the network is reduced.

After getting the ranked coordinates, we need to sort the landmarks from small to large based on the coordinates. The final landmark can be recorded to the csv file and test the performance based on the rTRE method.

**Evaluation:**

The performance is evaluated with relative Target Registration Error (rTRE). The evaluation method is also suggested to measure the performance in the registration task. The Target Registration Error (rTRE) measures the accuracy between retrieved landmarks and the ground truth landmarks. The registration error measuring is related with the images size.

Before calculating the rTRE, the Target Registration Error is required to calculate. This error is calculated as the Euclidean distance between the retrieved landmarks coordinates and the ground truth landmarks coordinates. To calculate the rTRE, all of the the Euclidean distance (TRE) need be normalized (divided) by the image diagonal.

rTRE = TRE/(w^2 + h^2)^2.

We also evaluate the registration function processing performance by time spending. We evaluate the time spending from the protocol reading the images to the protocol generate the landmarks. We take the average time spending of the all paired image registration.

**SIFT baseline**: we use OpenCV’s sift detector to get the sift local features. We followed David G. Lowe’s SIFT matching process as baseline. The baseline process is using SIFT to get each image local features (interest points) and using KNN method to pair points. We use the Euclidean distance and a threshold based on the Euclidean distance to filter out the paired landmark as the output. We choose the optimized threshold as the output result to calculate rTRE value. The optimized Euclidean distance threshold is around 0.7 and the optimized rTRE is 0.6205.

**RegSnet**: We trained RegSnet by expanding the ground truth landmarks to small image regions and input these same size image regions into the networks. The network will analysis paired images at the same time. We splice the features after the bottleneck layer. The expectation output of the real matched paired regions is 1 and the fake pairs is 0. The loss function is Binary Cross Entropy (BCE loss). The optimization function is Stochastic gradient descent (SGD). The architecture of the network and registration method we specified in the network and method details part. We take the after trained network to make our registration task.

**Generative adversarial network (GANs)**: GANs is a deep learning architecture. We developed this method before our main combination protocol. This method is influenced by Goodfellow’s GANs network architecture. This method consists a generative network and a discriminative network. We input the full paired images to the generative network. The expected output of the generative network is the X-axis coordinates vectors of the source image, Y-axis coordinates vectors of the source image, X-axis coordinates vectors of the target image and Y-axis coordinates vectors of the target image. The generated paired landmarks will put into the discriminative network as input to get whether the vectors are matched or not. The loss function of the network is Binary Cross Entropy (BCE loss). The optimization function is Stochastic gradient descent (SGD). Due to the result seems not as good as our main function, the rTRE value and the time spending of this method will used for comparing performance with the main function.

The

Results: we followed the evaluation protocol and analysis the result as the robustness of our approach. The score of the method is related to the accuracy of the generated landmarks and time spending.

#####################

The performance of the method measured by TRE. This result of this method is much better than the directly generate landmarks by CNN. The reason is that the SIFT detectior is already get the landmarks.

|  |  |  |
| --- | --- | --- |
| method | rTRE (after 50 times scale) | Time |
| SIFT | 0.6205 | 0.0621 |
| GANs | 1.7907 | 0.07291 |
| FCN | 0.7814 | 0.0432 |
| Our method | 0.0057252 | 0.058767 |

The result is 2 times training of the whole training dataset each methods rTRE. According to the rTRE, our method has considerable accuracy and reliable output of other methods. The accuracy made a tradeoff between the accuracy and speed. The accuracy can be improved by increasing the filter size, output channels.

Formula:

Discuss:

Our baseline method with SIFT matching approach, the average eTRE is up to 0.6205 which is the best threshold. The CNN network is lower than our base line which is 0.7814. The reason is might be the training data set is too small and the generated landmark including negative number which is useless. The DCGANs’s average eTRE is up to 1.7907 which is higher than the CNN. The high error rate due to the loss function. In CNN the loss function is mean square error. However, the loss function of the GANs is mainly the discriminator network. The discriminator network as the loss function can reduce overfitting but need large dataset to learn parameters in the networks (Generator and Discriminator). In this task, the dataset is small and easily get underfitting. If training the dataset many times, the network will be overfitting. This phenomenon occurs between CNN and GANs. To solve the small dataset issues, we change to CNN input from the paired whole images to paired small image regions. The dataset is enlarged about 70 times. Which reduce overfitting. Our method rTRE error is 0.0054 which is much higher than the others. However, due to the complex of the model, our method has many local minimums that the optimizing might be stuck in. To solve that, we change different component in each part randomly and record the optimized result.

**Experiment**

**limitation**

the registration task is to find the relationship between the two target images. Due to huge image size, our method detected local interested point first, and then deal with the relation between the paired images based on the points. In this way, the function might loss some useful matching points for registration. But the local detector is a useful way to shorten the time.

Another issue is the amount of the generated landmarks is fixed.  We try to learn the number of the landmarks of the paired images by CNN.  But the method is time consuming and unstable. There is no significant increasing of the accuracy of the generated landmarks. Therefore, this function is not a good solution to deal with the dynamic number of landmarks. Another solution is that we put the paired images into the CNN and determine the probably of each pixel. However, our device cannot hold the huge paired images size. Furthermore, training need large dataset for the network to learn parameters. Although the size of each images is huge, the paired images is only 230 which might cause underfitting or overfitting of the trained network. In that case, fixed landmarks is a compromised method.

This method can be modified and optimized by change the local interest point detector function. In this function, the SIFT, ORB and SURF method is comparison in this paper. There might be some local interested point detector more suitable than SIFT to get useful points.

The better CNN network is a good way to optimize this method that can obtain a better rank to get good landmarks. Due to the method is separate to modular functions, our method can be optimized to dynamic landmarks generation just add the dynamic landmarks amount determine function before the ranking. The method will generate the dynamic amount landmarks.

Due to the complex stage of our method, there will be many local minimums. The result might not a globe minimum. Which is a problem when optimize our result.

**Conclusion:**