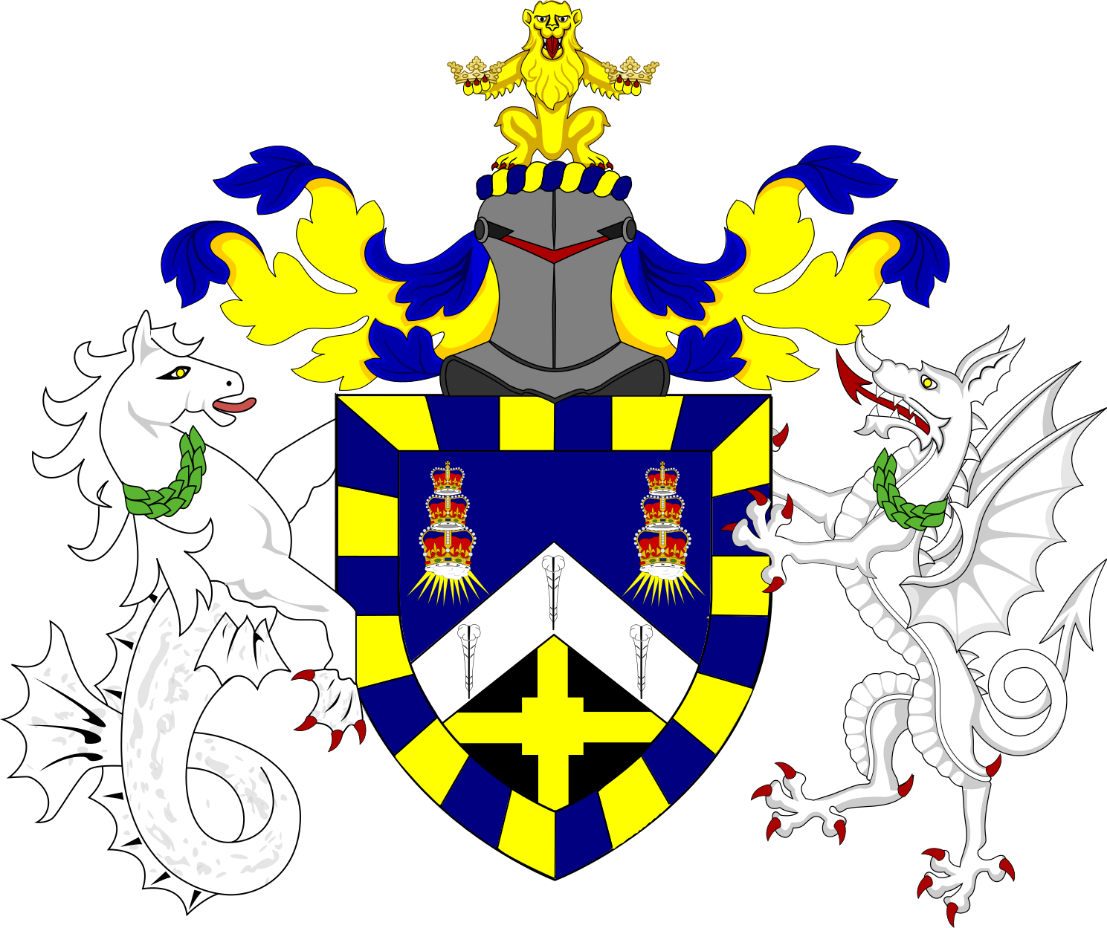
Queen Mary University of London

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



Hao Bai

*Supervisor*: Qianni Zhang

School of Electronic Engineering and Computer Science - Department of Computer Science

August 18, 2019

## 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.

CONTENT

[Automatic Non-rigid Histological Image Registration based on SIFT and RegSnet 1](#_Toc17068963)

[1. Acknowledgement 2](#_Toc17068964)

[2. Abstract 4](#_Toc17068965)

[3. Introduction 5](#_Toc17068966)

[3.1. Background 5](#_Toc17068967)

[3.2. Challenges 7](#_Toc17068968)

[4. literature review: 12](#_Toc17068969)

[5. Dataset: 16](#_Toc17068970)

[5.1. Introduction of the dataset 17](#_Toc17068971)

[5.2. Image content and ground truth content 18](#_Toc17068972)

[6. Requirement and design 19](#_Toc17068973)

[6.1. Requirement 19](#_Toc17068974)

[6.2. Designed approaches 20](#_Toc17068975)

[7. Our approaches details: 25](#_Toc17068976)

[7.1. Our main method: 25](#_Toc17068977)

[7.2. Comparison method 33](#_Toc17068978)

[8. Evaluation: 37](#_Toc17068979)

[9. limitation 45](#_Toc17068980)

[10. Result delivered 46](#_Toc17068981)

[11. Conclusion: 47](#_Toc17068982)

## Abstract

The challenge of the is the registration task is that the approaches need to respond to the changes of the object conditions such as object deformations, rotation, dyes changes and other related changes. The size of the histological image is huge which is up to 50k x 50k pixels and each image’s size is different in our dataset. Therefore, 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 points score. The system includes a ranking function to sort the scores 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 or future optimize the function.

In order to compare performance, we also developed two deep learning architecture based on Hongming Li’s fully convolutional network(need ref) and an unsupervised learning generative adversarial network approach to generate the corresponding landmarks from each pair of images. The performance of the different approaches is evaluated by Target Registration Error (rTRE) which is an accuracy evaluation of the corresponding landmark registration tasks. According to the results during experimenting, our main modular approach is much better than the fully convolutional network approach and generative adversarial network approach especially performed on this histological dataset.

## Introduction

### Background

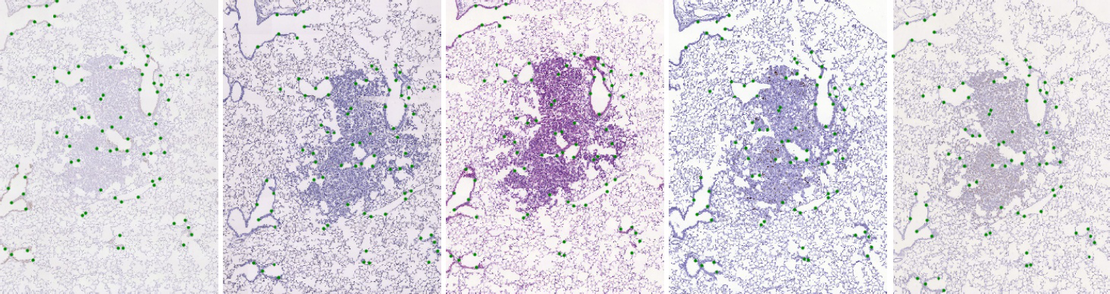
Automatic non-rigid image registration gradually become an important technology in many computer vision tasks(ref). Without registration technology, it is impossible for splicing the overlapping images. The registration technology is widely used in image fusion, change detection and other related areas. This technology is also as an important role in remote sensing application, medical image analysis, military automatic target recognition and many related areas (ref). The image registration overcome the challenge such as image rotation, scale, and skew that are common when overlaying images. The main task for image registration is to get corresponding feature points, image regions or interest points from given image sets. Different from matching function, registration task needs to correspond to the object deformation. In medical image processing tasks, the elastic image registration is a big challenge that the registration approaches 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).

The feature-based registration is widely used in computer vision. The feature-based registration is an approach to get the accurate correspondences between the images. In 1999 (ref), Lowe’s Scale-invariant feature transform research based on feature-based registration proposed a registration method based on local features detectors which improve the performance in both accuracy and speed. 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 point pixel into a small image patches and put image patches into the network. The small patches including the surrounding pixels around the points. The paired patches will be analysis by the network and output a registration score. We rank the output scores and take top N pairs as output.

The Lowe’s Scale-invariant feature transform based registration method can improve the performance to adapt different dataset by replacing the local feature detector method such as Speeded up robust features (SURF)(ref), Oriented FAST and rotated BRIEF (ORB) (ref). However, the traditional method Scale-invariant feature transform (SIFT) descriptor performance is much better than the modern engineered descriptors such as ORB and SURF in our dataset(matchnet ref).

### Challenges

The main challenge of the non-rigid registration of the histological images is that the approaches is required to sensitive to the tissue deformation and retrieval corresponding landmarks. Due to the big histological images size, some traditional methods are time consuming when processing on this dataset. 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 contour or the key points of the tissues.

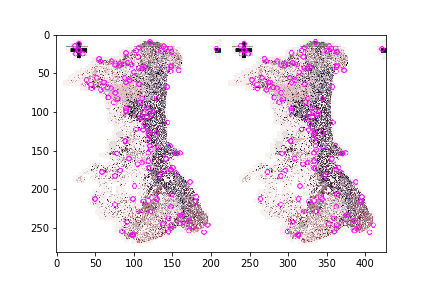


*Some example of the lung lesion tissue, the green points is the ground truth landmarks*

The device is also a considerable problem for the registration approach design. The high quality and large size of the image 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, our main 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 generated from SIFT descriptor, instead of processing the whole images in the network. Our main registration approach proposed a deep leaning approach that measure the registration point score of the small image regions.

There are many deep learning methods generate the landmarks directly. However, there is a problem in deep learning registration method that the images is required to be the same size before putting the image into the convolutional neural network. The multiple sizes images in the dataset is impossible as an input to the neural network 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 convolutional neural network learning and training time and the object deformation. To face the multiple sizes problem, our main 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 multiple size images in the dataset.

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(ref). However, the performance of the Lowe’s SIFT approach is not good as our prediction on our histological images. Our registration output is based on the ranking of the output of the convolution neural network (CNN). We named this network as RegSnet. This network jointly processes the pairs of images and give a mark of each pair. The output of the RegSnet is a registration point score which is the probability of the central coordinate of each pairs of images regions be a pairs registration landmark. This network includes a feature network which jointly process both images features and a metric network which compute the features and generate a registration points probability. In the fourth stage, we rank the scores and take top *K*th pairs. The *K* top of the pairs is the retrieved as the registration landmarks of each images. The amount *K* can be fixed or dynamic. In our approach, the number of output landmarks is fixed. We try to add a network to determine how many landmarks in pairs of images. The result is not as good as our fixed number of pairs of landmarks function. The amount of landmarks generation network needs to perform the pairs of images again. Therefore, the time spending increase. And the training the dataset is too small to get a network model with well learned parameters. So that, we keep the amount of the output landmarks fixed. Each stages of our registration approach handles each part of the registration process clearly. Therefore, all the stages can be changed to other methods or optimized to adapted different kind of dataset. Our main solution for this task is a feasible solution for the low-level devices. These approach needs standard CPU and even no graphics card can also perform the registration task and less than 8GB memory to load the pairs of images and store the small patches.



*A comparison image of our output landmarks and ground truth landmarks (The landmarks is the red round points), left is our generated landmarks, right is the ground truth landmarks*

We also implement a method that directly used convolution neural network to get the landmarks. This approach is inspired by fully convolutional network(ref). This method’s result is not good as our prospected. The reason is might because the total including training and testing dataset only included 215 pairs of images. After training, the network is either underfitting or overfitting. The generated landmarks points include negative number which is meaningless results. We get some ideas on the generative adversarial network (GAN) on another approach. Generative adversarial network uses a discrimination network to determine the generated image is fake or true. We added a discrimination network in training process of the fully convolutional network. However, due to the large amount of the generated landmarks including negative number, the small dataset and the loss function is changed which is not based on the error distance. The results are worse than the fully convolutional network method.

## literature review:

Most of the state of art registration algorithms can be divided into intensity-based approach and feature-based approach. The intensity-based registration approach measure and compare the intensity patterns between the registration images. The feature-based registration approach measures the relationship on the corner, edge, interest points, contours and regions between the registration images. The combination approaches have also been developed.

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. The later local interest point approaches such as Speeded up robust features (SURF) and Oriented FAST and rotated BRIEF (ORB) speed up the SIFT registration method and not lose the accuracy in the nature image registration tasks.

Our RegSnet is influenced by the traditional classification network- VGGNet. VGGNet is a convolutional neural network and is developed by Karen Simonyan et al in 2014 (ref). VGGNet is outperformed in image classification tasks and localization tasks. The network uses smart network layer structure. The network use two 3 \* 3 filter size convolutional layers instead of one 5 \* 5 filter size convolutional layer (two convolutional layers of the 3 \* 3 filters cover 5 \* 5 areas of the original images). There are 25 parameters (5 \* 5 = 25) if using one 5 \* 5 filter size convolutional layers. However, if using two 3 \* 3 filter size convolutional layers, there are only 18 (3 \* 3 + 3 \* 3 = 18) parameters for the network to learn. The number of the parameters reduced by 0.28. Therefore, this architecture’s enables the fewer parameters for the model to learning and converge the faster. The fewer parameters can also reduce overfitting.

Our main approach got some idea from the deep learning in face recognition approaches. The deep leaning approaches doing the face recognition tasks is based on the convolutional neural networks. Some researchers designing a classification convolutional neural network architecture to treat the face recognition tasks. The main aim for the face recognition network is to jointly process the two images and get a similarity score. This approach is widely used in the face recognition tasks and some good network architectures achieved a considerable performance in this task.

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 developed by Hongming Li et al (ref). They implemented a fully convolutional networks (FCN) network that directly generate 3D landmarks from paired of 3D images from the network. The output of the network is X axis coordinates vectors, Y axis coordinates vectors and Z axis coordinates vectors of each input 3D images. This network analysis both images features jointly and generated each image’s 3D landmarks. FCNs is a fast way to generate the result using pretrained deep learning model. However, this 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 testing stage.

There is an end-to-end unsupervised deformable image registration method research by Bob D. de Vos (ref). This research proposed a deep learning approach for deformable image registration. This network including a convolutional neural network regressor, a spatial transformer, and a re-sampler. This method first cut the image into small patches. The convolutional neural network regressor part measures each sub-image similarity. Then use the similarity value of each patches generate a grid of control points. Then send the grid of control points to the spatial transformer to get a full displacement vector field that enables the re-sampler to warp the moving image to the fixed image.

Another function is implemented by Dwarikanath Mahapatra in 2018. Dwarikanath Mahapatra developed a deep learning approach to doing the registration tasks. The deep learning architecture they use is the generative adversarial networks (GANs) which is an amazing idea to implemented on the registration task. This method uses generative adversarial networks to generates the registered image. This approach including a generative network and a discriminative network which is same as traditional generative adversarial networks. The generative network is a convolutional neural network and aims to get the transformed image and deformation field from the inputs which is the floating image and registered (or transformed) image. The discriminative network is to distinguish real or fake based on the generated image and the floating images. This method does not need to do the time-consuming iterative computing to generate the accurately registered images. The registered images and deformation field can directly generate from generative adversarial networks.

However, this method required high memory and good graphics card of the devices to process the high-resolution images. We propose to generate landmark points by this method. The amount of the retrieved landmarks can be dynamic generating and the the result can be more accrue than the fully connected network. However, due to the small training dataset and large image size, this method cannot get a good result in this dataset. The Dwarikanath Mahapatra et al’s generative adversarial networks approach required more data for 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 made 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.

## Dataset:

### Introduction of the 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. The dataset delivered multiple content tissues slices pairs and the ground truth landmarks. This dataset included 215 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:

1. clara cell 10 protein (Cc10)
2. prosurfactant protein C (proSPC)
3. hematoxylin and eosin (H&E)
4. antigen KI-67 (Ki67)
5. platelet endothelial cell adhesion molecule (PECAM-1, also known as CD31)
6. human epidermal growth factor receptor 2 (c-erbB-2/HER-2-neu)
7. estrogen receptor (ER)
8. progesterone receptor (PR)
9. cytokeratin
10. podocin

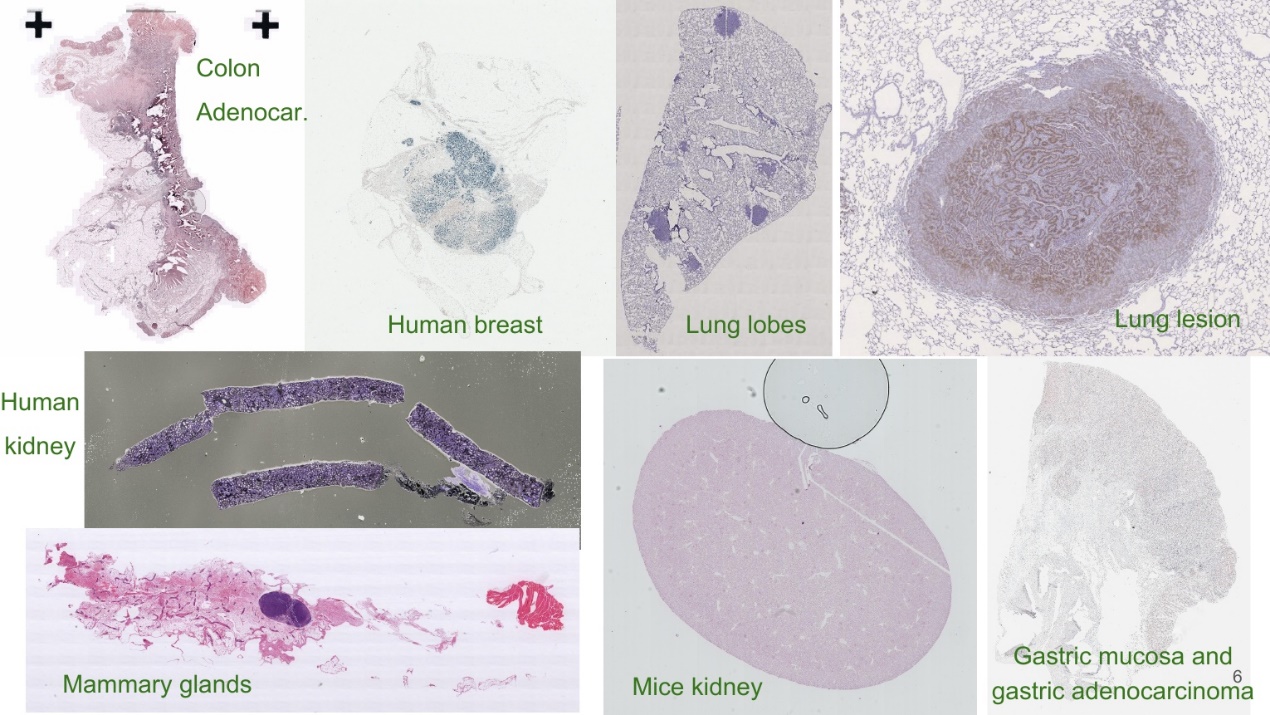
The dataset including 50+ histological sets which is organized by whole slide images of each folders. Each folder is corresponding to the tissues dyed by different stains.

### Image content and ground truth 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 dataset also including a .csv file which including each pairs of images names, corresponding path and landmarks path of the source image and target images. The ground truth landmarks are the X-axis coordinates vectors and Y-axis coordinates vectors. Each pair of the images is corresponding to a pair of landmarks.

The expected output of the dataset is the retrieved landmarks of each pairs of registration images, the *rTRE* score based on the output landmarks set and the ground truth landmarks set and the time spending of each pair.



*Different stained tissues*

## Requirement and design

### Requirement

The goal of the non-rigid image registration task for this histological dataset is to process the paired images and retrieval a pair of corresponding landmarks automatically. The method required to process the paired images at the same time. So that, the generated landmarks are corresponding to both input images. 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 is required to make the task without padding or change the proportion of the object in the images.

### Designed approaches

There are three preliminary implementation methods to deal with the image registration task of this dataset.

#### Our main approaches

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 compared the performance between SIFT, SURF and ORB.

**SIFT 4 stages:**

* 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 key point 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(ref). 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.

#### fully convolutional networks

We got another inspiration from Hongming Li’s fully convolutional networks (FCNs). The Hongming Li’s fully convolutional networks make the registration task on the 3D structural brain magnetic resonance (MR) images and get 3 dimensional coordinates (X-axis coordinates vectors, Y-axis coordinates vectors and Z-axis coordinates vectors) as output landmarks. Our task is to get the 2D landmarks on 2D images. Compared to the Hongming Li’s 3D fully convolutional network. Training the 2D landmark retrieval network might need less data (image pairs) and the time spending to get the optimized network parameters might be shorter. The Hongming Li’s fully convolutional networks (FCNs) including three part of generation which is the X-axis coordinates vector generator, the Y-axis coordinates vectors generator and the Z-axis coordinates vectors generator. Each image is convoluted can return the 3D landmarks. The network including 4 convolutional layers, 2 pooling layers, 2 de convolutional layers and 3 regression layers which corresponding to the output of each axis coordinates vectors. Our idea is to remove the last axis coordinates vector (Z axis coordinates vector) generator sub-network and change the input image size and change the 3D convolutional layers to 2D convolutional layers. The result of the Hongming Li’s fully convolutional networks (FCNs) is not measured by rTRE which is measured by Normalized cross-correlation (NCC) measure. To compare each method performance on our 2D dataset, we will perform rTRE to measure the performance of this protocol which same as our main idea.

#### Generative adversarial network

Another idea is to use unsupervised learning generative adversarial network (GANs). This idea is influenced by generative adversarial network (GANs) can deal with registration image generation tasks. Our generative adversarial network has a generative network and a discriminative network. We use the fully convolutional networks as our generative network part. The network is inspired by Hongming Li’s 3D landmark registration method. The input of this generative network is design as pairs of the images. The output of generative part is a pair of 2D landmarks based on the two input images. The discriminative network input is the output from the generative network which is a pair of 2D landmark vectors. The discriminative network needs to distinguish the pair vectors of the landmarks can be the registration points or not. The output of the discriminative is a value to distinguish true or false. The training process is same as the GANs training. After training the generative adversarial network, the registration processing is using generative network to generate landmark vectors. The discriminative network as a kind of loss function in the training process. However, the optimized function only works in the discriminative network.

#### Designed evaluation method

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.

## Our approaches details:

### Our main method:

We observed that the deep learning method achieved a good performance on the classification or distinguishing task. This kind of the task need less training data and simpler network architecture than other tasks. That can reduce the training and process time and overfitting probability of a dataset. Our dataset included only 215 pairs of registration images and corresponding landmarks. Therefore, performing a complex network to let network learn plenty of parameters or get accurate coordinates from the after-training network is easy to get underfitting. In this situation, we need a simpler network which can also obtain good performance on the registration task and simpler output instead of outputting all the registration landmarks directly. Therefore, we develop a network to solve that problems and enlarged the dataset.

RegSnet is a deep learning architecture. Our RegSnet get some ideas from the deep learning facial recognition approaches (ref). The deep learning facial recognition use the network to generate a similarity score for the related face images. Our 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.

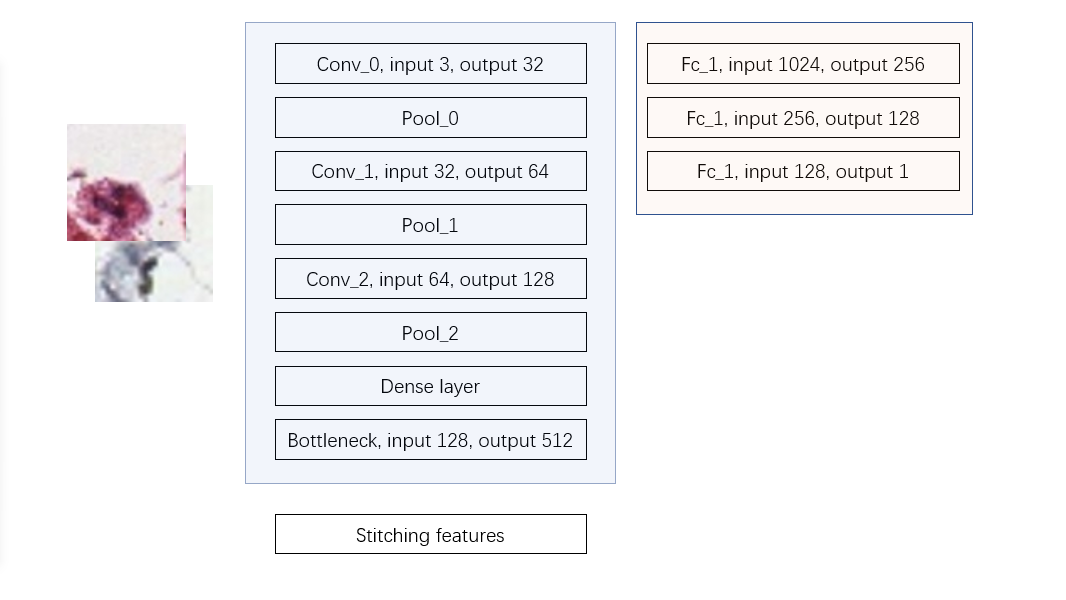
The RegSnet including 3 convolutional layers, 3 pooling layers and 3 fully connected layers. The activation function of each layer expects the last fully connected layer is linear Relu. The last layer activation function is Sigmoid because the expectation output is a number between 0 to 1. The output 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).

Our convolutional layer layout and split got idea from VGGNet which using two 3 X 3 filter size convolutional layers instead of one 5 X 5 filter size convolutional layers to reduce the parameters. By using this approach, the network need fewer parameters to learn. Therefore, the model can converge faster and reduce overfitting issues. If using larger than 16 X 16 patches to determine the registration score, there is a 2\*2 kernel size convolutional layer behind Conv\_2 (not show in current network because we use 16 X 16 patches). In this way, our network reduced 50% parameters (one 4\*4 filter size layer using 16 parameters, two 2\*2 filter size layers using 8 parameters) to learn. This approach reduce help the network converges faster in the network layers construction. Though, we not training the network and we use small patches to train our network so that our network needs to be simple and effective enough. The idea is very useful for the large image process tasks.

#### Network main part

**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 (ref). 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.



*Figure 1: The RegSnet network architecture*

#### Each layer details

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 Algorithms:

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). We minimize the Cross-Entropy loss:

*Formula 1: The Cross-Entropy loss formula*

*is the predicted probability of ith pairs*

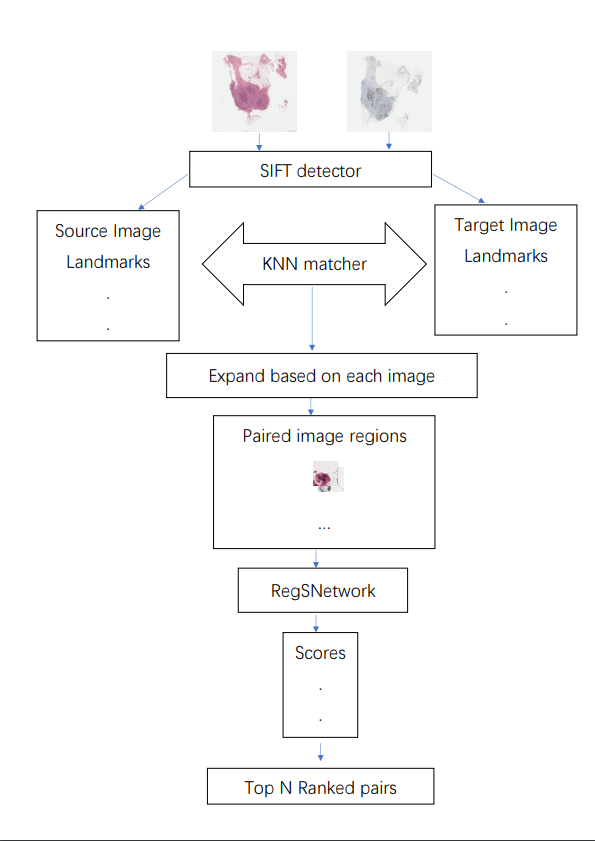
*is the correct classification of ith pairs*

|  |
| --- |
| **Training** **Algorithms:** Get N pairs of images from the dataset |
| **for** i = 0 . . . N-1 **do** |
| Get M paired corresponding ground truth landmarks |
| Expend the landmarks to small image regions |
| **for** j = 0 . . . M-1 **do** |
| **for** q = 0 . . . M-1 **do** |
| **Pair** the jth image in the source image patches and qth image in the target image patches |
| Put the pairs image into the RegSnet and get result |
| **If** j == q **then** |
| Desired output is 1 |
| **else** |
| Desired output is 0 |
| **End if** |
| Calculate Cross-Entropy loss (network output, Desired output) |
| Optimized the network |
| **End for** |
| **End for** |
| **End for** |
| **Return** and **save** model |

#### 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.



*Figure 2: The process of the registration approach* *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 matching 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 cannot 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 \* 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.*

|  |
| --- |
| **Registration Algorithms:** Get N pairs of images from the dataset |
| **for** i = 0 . . . N-1 **do** |
| Process scale-invariant feature transform (SIFT) feature detection to get Rth landmarks in source image and Lth in target image  Expend the landmarks to small image regions |
| Process k-nearest neighbors’ algorithm (KNN) to get M pairs of the landmarks |
| **for** j = 0 . . . M-1 **do** |
| Put the pairs image into the RegSnet and get score |
| **End for** |
| Rank all the scores |
| Take top Kth pairs landmarks as the registration points of each images |
| **End for** |

### Comparison method

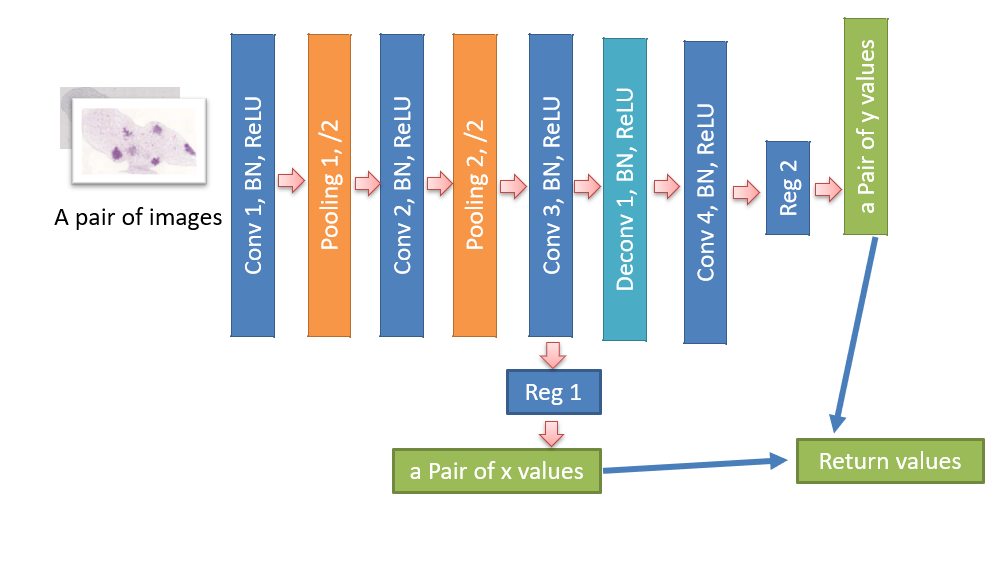
#### Fully convolutional networks:

The fully convolutional network is a deep learning method which directly input the registration image pairs into the network. The output including 4 set of vectors which corresponding to 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. Our Fully convolutional network including 4 convolutional layers, 2 pooling layers, 1 deconvolutional layer and 2 regression layers. We separate the ground truth X-axis coordinates vectors and Y-axis coordinates vectors of the source image and target image. We use the mean square error as our loss function and calculate each vectors’ loss separately. The optimized function is Stochastic gradient descent (SGD). We trained the network using the top 180 pairs of registration images and each images’ ground truth landmarks. And we use the rest of the pairs of images and ground truth landmarks as our test set. Due to the fixed input size of the network, the input pair images need to preprocess to same size (our network the paired input image size is 512 \* 512 RGB images). The number of the output landmarks points is also fixed to 70 pairs of coordinates.

*Formula 2: The m loss formula*

*is the predicted coordinates vector of ith pairs*

*is the correct coordinates vector of ith pairs*

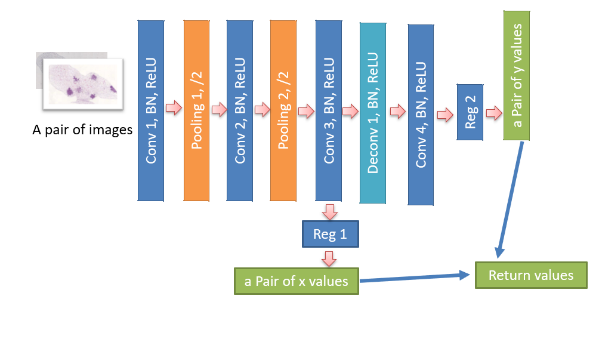


*Figure 3 : Fully convolutional networks architecture, the input is a pair of registration images, the output is 4 coordinate vectors (X-axis coordinates vectors of the source image, Y-axis coordinates vectors of the source image, X-axis coordinates vectors of the target image, Y-axis coordinates vectors of the target image)*

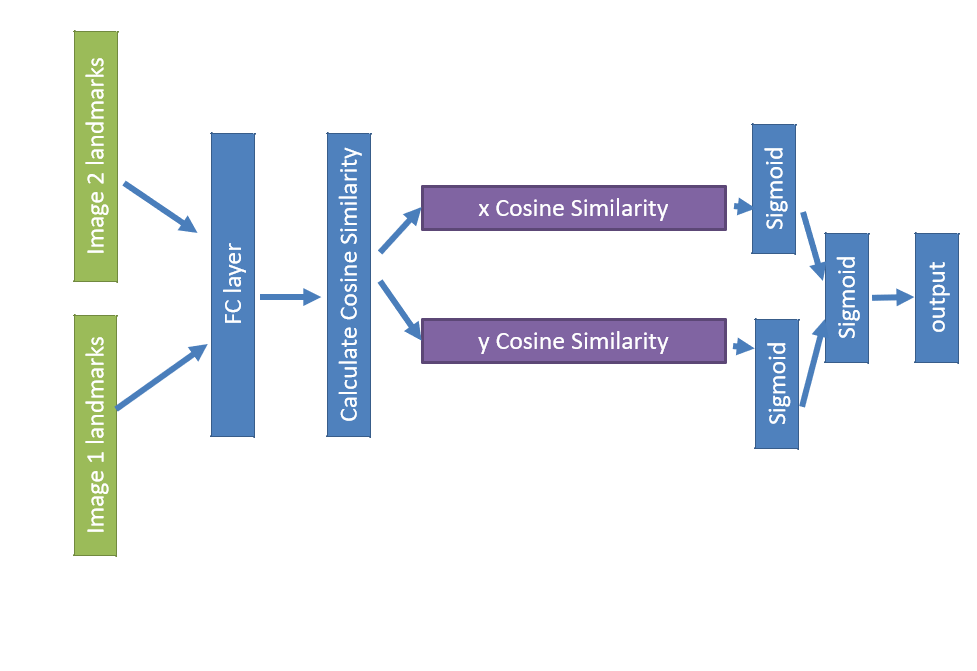
Conv 1, Conv 2, Conv 3, Conv 4 layer is all the convolutional layer. *BN* is the batch normalization function. *ReLU* is the Rectified Linear Unit activation function. The Pooling 1 and Pooling 2 is the pooling layer of the network which reduce the overfitting. Deconv 1 is the deconvolutional layer. The Reg 1 and Reg 2 are the liner regression layer which map the features to the output points. This network jointly processes the images, analysis the features output the X-axis coordinates vectors and Y-axis coordinates vectors for each image.

#### Generative adversarial network (GANs):

Our landmark generative adversarial network including a landmark generator network and a landmark discriminative network in GANs. The landmark generator network architecture is same as the fully convolutional networks. The discriminative network measures the inner relationships between the two landmarks vectors. The discriminative network measures the *COSINE* similarity between the two landmark vectors and generate a score which measure the generated landmarks vectors of the images is good or not. In the GANs approach the input image size is fixed and the output landmark pairs is fixed. Therefore, the input images need to be preprocessed to the same size.



*Figure 4: Fully convolutional networks architecture, the input is a pair of registration images, the output is 4 coordinate vectors (X-axis coordinates vectors of the source image, Y-axis coordinates vectors of the source image, X-axis coordinates vectors of the target image, Y-axis coordinates vectors of the target image)*



## Evaluation:

The performance is evaluated with relative Target Registration Error (*rTRE*) (ref). 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 (*TRE*) 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 the Euclidean distance (*TRE*) need be normalized (divided) by the image diagonal.

*The rTRE formula*

*w: the original image width,*

*h: the original image height*

We also evaluate the registration function processing performance by time spending. We count 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.

**Fully convolutional networks (FCN)**: The fully convolutional network is a deep learning architecture. We get idea of this method from Hongming Li’s fully convolutional network. The Hongming Li’s fully convolutional network make the registration task on 3D images and get the three-dimensional coordinates (X-axis coordinates vectors, Y-axis coordinates vectors and Z-axis coordinates vectors) from the brain images instead of two-dimensional coordinates. Our dataset is 2D histological tissues images and the ground truth is 2D coordinates. Therefore, we changed the Hongming Li’s fully convolutional network to implement our task to get the 2D coordinates (X-axis coordinates vectors, Y-axis coordinates vectors) on 2D histological images. We remain the fully convolutional network architecture before the X-axis coordinates vectors and Y-axis coordinates vectors is got and delete the Z-axis coordinates vectors getting part. The input of the network is paired 2D RGB images. The output is the X-axis coordinates vectors and Y-axis coordinates vectors corresponding to both images. The output of the network is also the final output landmarks of each pair images. The loss function of the fully convolutional network when training is mean square error (MSE error). The optimization function is Stochastic gradient descent (SGD). The accuracy (rTRE) of this method is not as good as our main registration method. This function might predict negative values on X-axis coordinates vectors and Y-axis coordinates vectors. The negative values cause meaningless and inaccuracy of each points.

**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. The landmark generative network we get ideas from Hongming Li’s fully convolutional network. The Hongming Li’s fully convolutional network is a good solution to get paired landmarks from the input pair images. 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.

**Results:** we followed the evaluation protocol and analysis the result as the robustness and accuracy of our approach. The score of the rTRE is related to the accuracy of the generated landmarks and time spending. We measure the mean rTRE of the whole dataset landmarks which generated from 4 optimized function. The deep learning functions (Fully convolutional networks, Generative adversarial network and our main method) training dataset is same which is the top 180 pairs images and corresponding landmarks.

The performance of the method measured by rTRE. The table shows that our main approach result is out performed among these approaches (SIFT, GANs and FCN). The average rTRE is 0.0057252 which is better than our SIFT base line and better than the GANs and FCN methods. The average time taking is similar with other approach. Our baseline method with SIFT matching approach, the average *rTRE* is up to 0.6205 which is the best threshold. The FCN 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 cause the *rTRE* increase. The GANs’ average *rTRE* is greater than 1.0 which is much higher than the FCN approach. The high error rate due to the loss function. In FCN 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 of the networks (Generator and Discriminator). In this task, the dataset is small and easily get underfitting to train a network. If training a dataset many times, the network will be overfitting. This phenomenon occurs between FCN and GANs. To solve the small dataset issues, we change to FCN input from the paired whole images to paired small image regions in our main approach. Therefore, our training dataset is enlarged about 70 times (each image has about 70 ground truth landmarks, so that we can get 70 \* 215 image patches pairs). This approach can reduce overfitting and underfitting. Our method *rTRE* error is 0.0054 which is much lower than the other registration approaches. However, due to the complex of the model, our method might have 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.

The local interested point descriptor performance will heavily influence our function performance. The amount of the extracted points needs neither too many nor too small and the quality of the points is required.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| method | Minimum rTRE | Maximum rTRE | rTRE (after 50 times scale) | Average Time |
| SIFT | 0.5105 | 0.6922 | 0.6205 | 0.0521 |
| GANs | > 1.0 | > 1.0 | > 1.0 | 0.07291 |
| FCN | 0.4916 | 0.9137 | 0.7814 | 0.0432 |
| RegSnet | 0.003954 | 0.00816 | 0.0057252 | 0.058767 |

*The result is 2 times training of the whole training dataset of each approaches’ network. The rTRE shows that 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, number of layers and different local interest point detectors.*

We also observed the potential of the fully convolutional network. The training dataset is only 215 pairs images. The dimensions of the expected output are values that corresponding to the X and Y axis coordinates vectors of each images which is the accuracy value instead of the classification probabilities. That means our dataset is too small for the directly landmarks generation networks. The FCN training needs large amount of training data to let the network learn the parameters. We trained the FCN network use 180 pairs of images for 10 times. The result is going a small amount of converging (the sum of the mean square error loss decreasing) the *rTRE* of the whole dataset also has small amount decreasing after training. But our dataset is too small, if training on same dataset for times, the result will go overfitting. However, the result of the network seems not as good as our main method in this dataset. If the dataset is big enough or other researchers’ pretrained model exist, the increasing spaces of the fully convolutional network accuracy is considerable. As comparison, We also analysis the result given by generative adversarial networks. We found that the trend of rTRE and the loss function is fluctuating wildly and not decreasing. By comparing the results between the generative adversarial networks’ original results and this registration tasks results, we found that the generative adversarial networks make a good job on the randomness image generation instead of this kind of accurate coordinate point acquisition.

To optimize our main method performance, we also use other local interest point descriptors such as Speeded up robust features (SURF), Oriented FAST and rotated BRIEF (ORB) to get the local candidate registration points. Compare with the modern engineered descriptors, SIFT local interest points descriptors speed is slower than Speeded up robust features (SURF) and Oriented FAST and rotated BRIEF (ORB). However, the modern technologies get small amount of the candidate registration point then the traditional SIFT method. The quality of the getting point is not as good as SIFT that many of the key points is missing after extract from the pairs of the images. The spending improvement is only a part of the registration task. The accuracy of the output landmarks is the main measurement of a registration approach. If the future registration task needs a time driven function, our method can make a tradeoff between the accuracy and time spending that our mrthod can speed up the processing time by change from the slower SIFT method to the faster method like SURF or ORB method. Therefore, in this task we choose the SIFT in the main method of this registration task solution.

## limitation

The registration task is to find the relationship between the content of 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 detected points instead of process two images relation jointly at the beginning. In this way, the function might loss some useful registration points. However, the local detector is a useful way to shorten processing the time and lower the devices requirements.

Another issue is the amount of the generated landmarks is fixed. We try to learn the number of the generated 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 215 which might cause underfitting or overfitting of the trained network. In that case, fixed landmarks are a compromised method.

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

There might exits a better architecture of our RegSnet network to get more accuracy registration score. Optimize the RegSnet architecture is a good way to optimize this method that can obtain an accuracy ranking to get the accuracy landmarks.

.

Due to the complex stage of our method, there will be many local minimums in the optimizing stages. Our optimized approaches performance might not a globe minimum optimizing approaches. Therefore, changing or adding components of each stages is a good way to optimizing the functions.

## Result delivered

We delivered the registration landmarks set which our main registration method generated. Each pairs of the registration landmarks set are corresponding to the given registration image pairs. The rTRE score as the accuracy measure between the generated landmarks and ground truth landmarks need to be calculated for each pair of the images. We also measure the time spending when each pair of landmarks generating. We plot the source image landmarks path, target image landmark path, each pair rTRE score and the time spending for each pair in the result.csv file. To have an intuitive feeling of the comparison with ground truth landmarks and the generated landmarks, we plot each landmarks on the images and splice the two images. We also add the spliced image paths of the source images and target images in the result.csv file. The training code of each approaches is included in the supporting material. The trained model for the RegSnet is included.

## Conclusion:

Image registration gradually become a core part of many medical image processing and general computer vision task. We face the main registration challenge is to get the corresponding landmarks for 2D histological images when the image conditions change such as the object deformations, rotation, dyes changes and other related changes. Sometime this image conditions changes occurs at the same time. The other big challenge is the huge size of the histological and large amount of details information’s in the high-resolution images. The traditional and state of art models process the image might need more time and might get huge amount of the landmarks. These challenges of the histological image registration dataset influence the performance of the registration approaches apply in this dataset. To deal with the problem, we proposed a modular solution for the automatic non-rigid histological image registration tasks and a deep learning architecture which can jointly determine the paired images or sub-images registration score. The experimental results are based on the multiple content of 2D histological tissues slices images shows that our modular protocol could obtain a good landmarks registration performance between image registration accuracy and calculation speed.

Ref

(SURF)H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006.

(SIFT)D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

(ORB)E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, 2011.

(matchnet)Xufeng Han, T. Leung, Y. Jia, R. Sukthankar and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 3279-3286.

(Hongming Li’s fully convolutional network) H. Li and Y. Fan, "Non-rigid image registration using self-supervised fully convolutional networks without training data," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, 2018, pp. 1075-1078.

(GAN)

(VGG)Simonyan Karen and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," In *Computer Vision and Pattern Recognition*, 2014.

(Alexnet) Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks, " In *NIPS*, 2012.

(基础)