Automatic Non-rigid Histological Image Registration based on SIFT and RegSnet (ANHIR-SIFT&ReSnet)

h.bai@se18.qmul.ac.uk

**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 data 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 ideas is to measure and correct small, varying discrepancies by deforming one image to match the other.

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 problems, this project combined Scale-invariant feature transform (SIFT) features detection and deep learning techniques to retrieval paired landmarks from images.

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

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

The registration landmarks are retrieved based on paired images. After detected interest points from each image. We expand the landmarks to 16 x 16 image patches. And use these patches to find the real matching landmarks. The basic idea is to make use of the threshold based on k-Nearest Neighbors search and Euclidean distance to filter out feasible paired landmarks. But the performance is not good on the histological images. This registration method filtering is based on the ranking of the output of the convolution neural network (CNN). The CNN is used for determining the pair of the small patches score. The score is added in ranking. The k top of the pairs is the retrieved landmarks of each images.

We also implement a method that directly used CNN to get the landmarks. But 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.

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 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.

**Others researcher:**

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. They(名字) use the deep learning method directly generate the landmarks from paired images. They implement an FCN network that just input paired images into the network. 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.

**Evaluation method:**

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

Target Registration Error is calculated as the Euclidean distance between the retrieved landmarks coordinates and the ground truth landmarks coordinates. And the Euclidean distance need be normalized by the image diagonal rTRE = TRE/(w^2 + h^2)^根号.

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. The Euclidean-distance as a threshold to filter out the pairs as the output. We choose the optimized threshold as the output result to calculate rTRE value.

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).

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

**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:

|  |
| --- |
| clara cell 10 protein |
| prosurfactant protein C |
| hematoxylin and eosin |
| antigen KI-67 |
| platelet endothelial cell adhesion molecule |
| human epidermal growth factor receptor 2 |
| estrogen receptor |
| progesterone receptor |
| cytokeratin |
| 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

|  |  |
| --- | --- |
| Image type |  |
| Lesion tissue |  |
| Lung lobes |  |
| Mammary glands |  |
| COAD | This sub-dataset slices from colon cancer samples |
| Mice Kidney tissue |  |
| Gastric mucosa and gastric adenocarcinoma tissue |  |
| Breast tissue |  |
| Kidney tissue |  |

The more details about the images like the microscope model can be find at(那老哥的链接).

**Our method detail:**

Training:

Expand each corresponding landmark point 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) and train with un-matching points with expectation output is false (0).

Registration processing:

This registration method combined local interest points detection and global pair points matching discriminate methods. This method will use SIFT interest point detector to get local interest point from each image separately. Then we use the Brute-Force matcher to get the paired points based on the points SIFT detected. Then we expand each point to 16 \* 16 small image regions based on the around pixels. The pair sets of regions use the trained CNN model to discriminate each pair points is matching. Then return a pair of landmarks.

Train algorithms

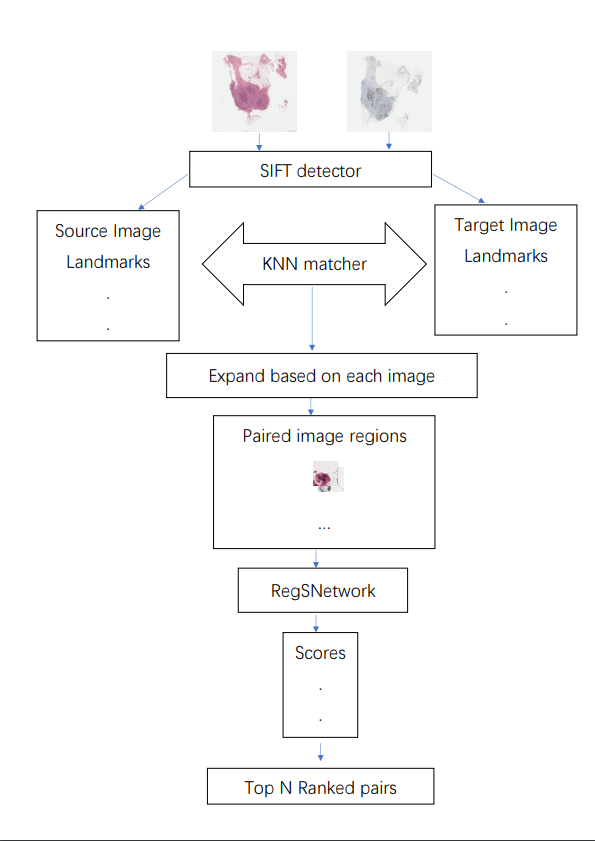
For each pair images in training images set

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

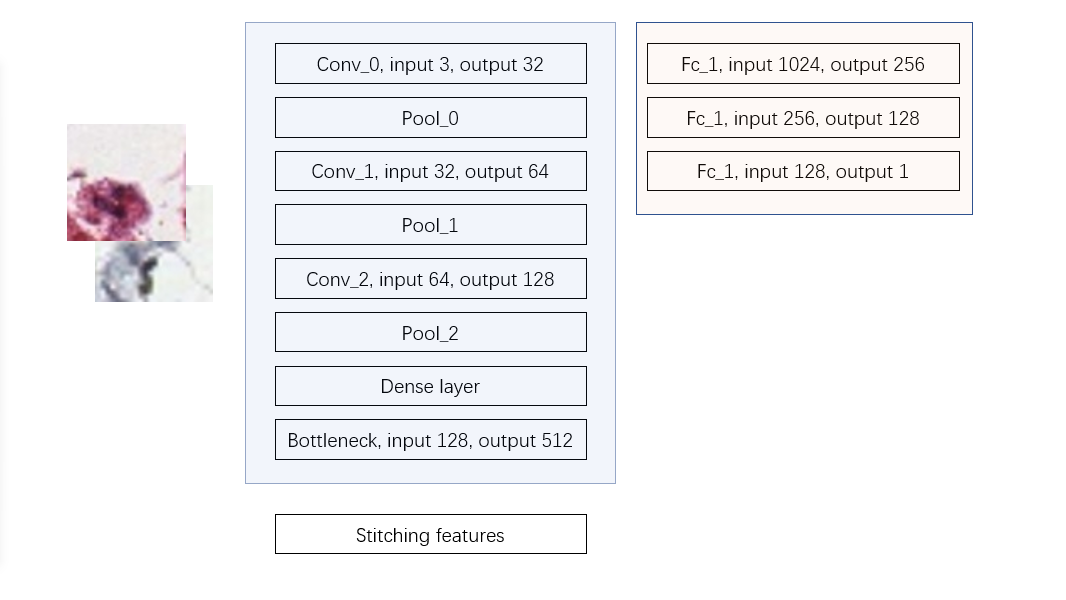
Record the landmarks

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 CNN 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 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. However, this method can also expand our training data set and reduce overfitting and underfitting.

Network

The goal of the network is to identify the paired image and give them a score. The network analysis the paired image features and give each paired image a score. The ranking is based on score which output from the network.

The CNN 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 and 1. The output number used for distinguishing the pair of small image region can satisfied the registration landmarks or not. Binary Cross Entropy (BCE loss) between the target and the output loss function is selected to use in training because the output is only one number. The optimizer is Stochastic gradient descent (SGD).



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 added 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.

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.

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) |
| SIFT | 0.6205 |
| DCGANs | 1.7907 |
| CNN | 0.7814 |
| Our method | 0.0054 |

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.

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.

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.