# Implementation of Framework for Deep Learning Based Multi-Modality Image Registration of Snapshot and Pathology Images

Arnab Mandal
Computer Science and Engineering
Shiv Nadar University
2310110428
am483@snu.edu.in

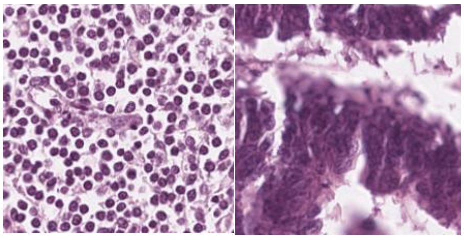
Suryansh Rohil
Computer Science and Engineering
Shiv Nadar University
2310110314
sr738@snu.edu.in

***Abstract*— Multi-modality image registration is an important task in medical imaging because it allows for information from different domains to be correlated. The main challenges in registration tasks involving pathology images come from addressing the considerable amount of deformation present. This work provides an implementation of a framework for deep learning-based multi-modality registration of microscopic pathology images to another imaging modality. The proposed framework is validated on the registration of white light camera snapshot images to pathology hematoxylin-eosin images of the same specimen. A pipeline as specified in the implemented research paper is presented detailing data acquisition, protocol considerations, image dissimilarity and training experiments. A comprehensive analysis is done on the impact of pre-processing, data-augmentation, loss functions and regularizations. Consequently, a robust training configuration capable of performing the desired registration task is found and proven to be useful. Utilizing the proposed approach, we failed to observe a marked increase in the mutual information score, suggesting issues with a lack of resources for proper training to happen.**

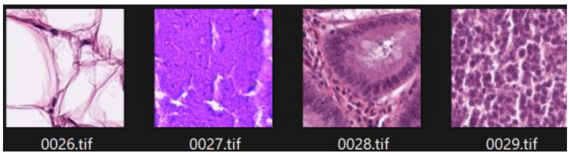
***Keywords—Camera snapshot image, deep learning, deformation, image registration, multi-modality, pathology image.***

## I. Introduction

In Oncologic surgery, pathological analysis is the gold standard for investigating surgical specimens and determining the extent to which excised tissue contains tumor tissue. However, a pathology image is the result of extensive tissue processing that results in a microscopic image of a tissue slice. This tissue processing includes cutting, fixation in formalin, embedding in paraffin, slicing, sectioning, and staining. As a result of these processing steps, the pathology image of a tissue specimen is highly deformed with respect to any other prior imaging done of the same tissue specimen. This makes image registration of diagnostic, pre-surgical, or intra-operative imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, or other experimental technologies to pathology images very challenging.



Classical multi-modal image registration methods can be roughly characterized into two groups: intensity-based methods and feature-based methods. Feature-based methods are based on identifying and matching specific features, such as landmarks, or structures in the images to establish a transformation model. Such an approach, by identifying landmark points, has for instance been done with white-light images and pathology images, and also CT/PET imaging and pathology images. However, a large amount of manual point pairs needs to be selected for reliable registration. Alternatively, intensity-based models use the intensity values of the pixels in the images with the guidance of a cost function to establish a transformation that aligns the images. This approach usually requires extensive pre-processing or a suitable deformation model for registration with pathology images.



In this work, an unsupervised spatial transformer network based on VoxelMorph is used. Although there have been recent advancements in the field of image registration, we chose Voxelmorph because this network architecture, in part, outputs a dense deformation field and does not require additional registration data other than the images to be registered. This is beneficial due to the lack of consistent paired structures or landmarks on the pathology and snapshot images. Despite achieving promising results, several limitations were identified, particularly regarding evaluation, which left room for further improvement. By building on the previous findings, this pipeline first aims to reduce image dissimilarity through several pre-processing steps. Additionally, various training experiments have been conducted to better understand the capabilities of the chosen network architecture and to investigate the effect of a set of key parameters on network performance. These key training experiments involved different modes of data augmentation, loss functions, and regularization factors. No other modifications were made to the neural network architecture. Furthermore, the significance of choosing the right evaluation metrics for such a task when the ground truth is not available has been considered. The effect of pre-processing, as well as the effect of the selected key parameters, was evaluated using both image comparison metrics and clinically established evaluation metrics.

Overall, our primary contributions are summarized as follows:

- Collection of a dataset consisting of paired snapshots and HE images of prostatectomy specimens.
- Development of a pipeline for multi-modal image registration utilizing an unsupervised deep learning model, specifically addressing the significant deformations and dissimilarities between microscopic and macroscopic images.
- Introduction of a new loss function and evaluation metrics.
- Comprehensive analysis of the impact of selecting preprocessing steps, loss functions, and evaluation metrics in applications of multi-modal image registration.

## II. Materials, Methods and Results

### A. *Data Acquisition and Data set*

The original datasets used have not been made available by the publishers of the paper that is being implemented. Accordingly, a synthetic dataset was generated, relying on Colon Cancer histology Images. The NCT-CRC-HE-7K Dataset was used for this purpose. It fits all the original conditions as were specified in the paper, and snapshot images were generated as per the available requirements as well. This was the best suitable alternative that was readily available open-source.

As HE images were already given in the specified dataset, we generated and stimulated snapshot images by

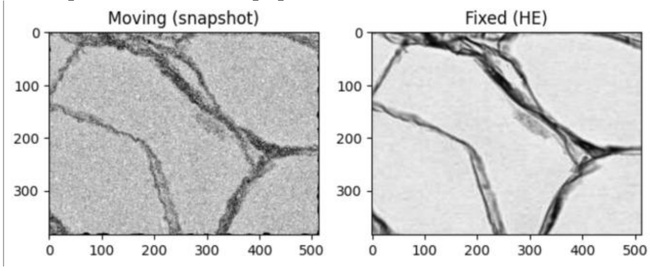
- Applying color distortion
- Adding noise/blurring
- Applying Elastic deformations

### B. *Data Pre-processing*

The introduced pipeline aims to register macroscopic prostate snapshots to microscopic HE images and thus correct for any changes and deformation the prostate tissue undergoes as part of pathology processing. To ensure the network focuses on correcting deformations due to histopathology processing, several pre-processing steps were carried:

- ○ Rotation
- ○ Scaling
- ○ Translation
- ○ Filling Holes

As implemented in the paper.



### C. *Network Architecture*

The network architecture used in this work is an unsupervised neural network based on VoxelMorph as shown in Fig. 3. The network takes as input the concatenation of both a moving image (m) and a fixed image (f), resulting in a two-channel 2D image. The moving image is the one which will be deformed to match the fixed image. The first part of this network is a U-Net based convolutional neural network (CNN). This U-Net consists of ten 2D convolutional layers, where all layers have 32 channels per layer except the last convolutional layer which has 16 channels. Each convolutional layer operates with a convolution kernel size of 3 × 3 with a stride of 1 and zero padding around the edges of the image. Each of the convolutional layers is followed by a Leaky Rectified Linear Unit (ReLU) activation function with a negative slope parameter of 0.2. The encoder part of the U-Net consists of four layers with 2 × 2 max-pooling, consequently, the input is down-sampled to half its spatial dimensions in each dimension after each layer. The decoder part of the U-Net consists of the remaining six layers with 2 × 2 up-sampling and skip connections from the same-sized encoder layers. The main idea is that the output of this U-Net based network is a registration field φ, which, once applied to the moving image, yields the moved and registered image. By including the spatial transformer layer in the network architecture, it becomes possible to use the U-Net based CNN output as a registration field. In this registration field, each pixel contains a displacement vector. The spatial transformer layer then applies this output to the moving image by linear interpolation, which allows for differentiability of the spatial transformer layer. Consequently, a registration field is learned by the U-Net based CNN when it's trained end-to-end, and the final output of the complete architecture is a moved and registered image. Note that the spatial transformer layer does not contain any learnable parameters.

### D. *Training Experiments and Loss functions*

The Model was trained on 800 synthetic pairs with a 80/20 Training-Test split using a batch size of 1 and TensorFlow's Adam Optimizer.

The loss function was a weighted sum of:

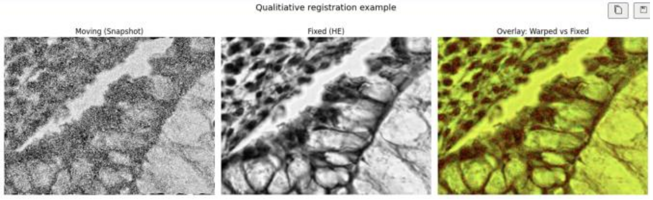
- Mean squared Error
- Mutual Information
- Gradient Regulation with Lambda=2

Weights were empirically chosen as [1.0, 0.5, 2.0] respectively. The model was trained for 5 epochs on CPU hardware, with each epoch taking approximately 5 minutes.

### E. *Evaluation*

Training was monitored qualitatively using red-green overlay visualizations. More yellow-tinted patches implied successful registration.

Alongside this, MI-scores were monitored and compared, before and after registration. Due to limited computation resources, we were unable to observe consistent qualitative improvement in alignment in most test samples.



### F. Discussion

Several aspects of the original published paper prevented us from creating an honest-to-source recreation of it and its implementation.

The lack of an open-source dataset, such as the one used by the authors of the publication, forced us to stimulate a synthetic dataset. While we adhered to the original requirements, the synthetic snapshots failed to replicate real world variability in modality, deformation or acquisition noise. Further, the lack of anatomical annotations or segmentation masks restricted our evaluation to similarity metrics like MI and quantitative overlays.

Another significant obstacle to the successful training of the utilized neural network was the lack of computing resources available to us. We were forced to rely on a CPU and thus were restricted to training for very few epochs. This directly resulted in poorer results than what was expected and observed in the source paper.

Despite these constraints, the pipeline provided a valuable prototype for registration learning under stimulated conditions.

## III. Conclusion

The results show that the pipeline presented in this work is capable of multi-modal registration of camera snapshot images to HE images. This pipeline includes pre-processing of images using rotation, resizing, image translation, inpainting, and grayscale conversion followed by a deep-learning-based registration model. The results demonstrate that the best-performing network can be trained without data augmentation using the MSE + NoBG MI loss function with a regularization parameter of 2.

## Acknowledgement

The authors would like to thank our supervising professor, Saurabh Shigwan for allowing us the opportunity to learn from him, and enabling us to complete our implementation of the paper. We thank Shiv Nadar University for fostering our education, and all open-source data sites for enabling us to find datasets that were viable for implementation.

## References

1. Schoop, R. A. L., et al. "Framework for Deep Learning Based Multi-Modality Image Registration of Snapshot and Pathology Images."

2. Balakrishnan, G., et al. "VoxelMorph: A learning framework for deformable medical image registration."