Augmented Reality Applications for Image-Guided Robotic Interventions using deep learning algorithms

Jenna Seetohul[1], Mahmood Shafiee[1,] Konstantinos Sirlantzis[1]

[1] School of Engineering, University of Kent
{jls56, m.shafiee, k.sirlantzis}@kent.ac.uk

Abstract. A significant breakthrough in the field of surgery has seen the integration of augmented reality (AR) in standard robot operations, allowing anatomical objects to be digitalized and overlaid onto a real-life scenario during or pre-intervention. This paper provides an overview of the methodology used to reconstruct and register laparoscopic head and neck image sequences for an AR tool. Deep learning (DL) algorithms are designed to strategically place fiducial markers or labels in a dataset, hence enabling a virtual tool path to be set up for guiding the end effector of a robot. We introduce a dataset of 271 images of patients from four different clinics in Quebec with a proven history of head-and-neck cancer. We then propose a marker-based registration method for mapping a trajectory during surgery, utilizing an unsupervised neural network for computing the medical image transformations. During the training stage, we use an optimized convolutional neural network (CNN) that warps a set of labels from the moving image in contrast with the related counterparts in the fixed image. To this end, we compare the loss functions between warped moving labels and fixed labels with respect to the ground truth in the method. Next, we propose a UNet architecture where we measure the accuracies in label localization throughout the test sequences relative to the initial output results. Our experiments showed that the UNet outperformed the initial CNN architecture, with optimum performance outcomes in losses being closer to 1.0.

Keywords: Augmented Reality, Image Registration, Path planning, Supervised Learning

1 Introduction

The use of Augmented Reality (AR) in surgery has plummeted over the past decade, with the ability to provide in situ immersive visualization of a surgical scene in the planning stage as well as during the intervention. Since the groundbreaking release of the Microsoft HoloLens in 2016 [1], the way surgeons perform minimally invasive surgeries has changed, eliminating inherent challenges that the narrow port access and lack of depth estimation causes in the operation theatre. In surgical navigation, most anatomical landmarks are generated in high definition within three dimensional

workspaces, from acquired preoperative CT or MRI datasets. During the intervention, the virtual model is registered to the surgical site using fiducial markers, by removing the backend scenes and overlaying a 3D image onto a see-through display [2]. To ensure the safety of the patient and successful final outcomes of surgery, this method of 3D image overlay is ideal for planning in a nonstructured environment. By combining AR with image-guided robotic surgery, the areas of interest in the body can be displayed through a visualization device in real life, improving a surgeon's hand-eye coordination when manipulating the robot end effector. Despite the plethora of studies available in the literature, medical image registration for surgical guidance is still confronted with valuable constraints such as accuracy of label correspondence throughout sequences of images, computational burden on processing units depending on the DL architecture as well as external factors such as signal fluctuations, noise, and acquisition settings.

Our proposed method is an extended framework on the use of two different deep convolutional neural networks to compare the output of an optimized registration procedure of the head and neck data with an appropriate transformation which converges to a zero value. In a threefold process (Fig 1), we aim to map the warped moving labels to the fixed labels in order to earmark the danger zones around the brainstem and spinal cord. We then calculate the dissimilarity between the dynamic and static labels in the CT image sequence using a dice scoring system as well as sum-square-difference (SSD) for intensity-based loss. Finally, we show that by performing a linear transformation such as an affine registration on the network using an alternative DL model such as UNet or probabilistic dense displacements, we can achieve greater accuracy as compared to the existing DeepReg architecture. The output from this experiment can eventually be used for rendering an estimated target trajectory as shown in the control system.

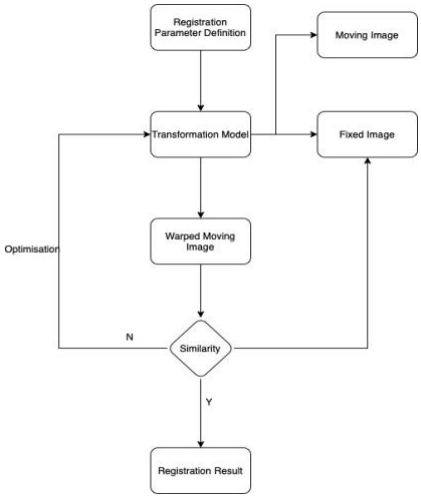


Fig 1: Flowchart of registration procedure

2. Related Work

In this section, we briefly introduce the use of deep learning for medical image registration, as well as the choice of contrasting models after comprehensive study. We then describe the application of such output databases for AR use in surgery.

2.1 Medical Image Registration based on Deep Learning

The innate need for precision in surgical image guidance has seen a dramatic increase in research across the academic community, proposing classic DL algorithms of CT/MRI scans for medical image registration. Ronneberger et al. [3] described the use of conventional neural networks (CNNs) such as the U-Net, where spatial transformations are used to two or more images to a coordinate workspace via an encoder-decoder style network; Qi et al. [4] implemented a modified neural network to extract point clouds from medical scans for semantic segmentation using PointNet architecture, which are then used for AR visualisation. Jaderberg et al. [5] proposed a method of applying STNs during both rigid and deformable transformations using transform feature maps on a grid generator. Sokooti et al. [6] described another method of registration called Displacement Vector Fields (DVF) as ground truth and utilized the RegNet architecture for registering CT images of the chest. This enabled a higher accuracy generation when using alternative real-life datasets, in line with the conventional B-spline methods. De Vos et al [7] proposed an unsupervised end-to-end network using CNNs and STNs to register 2D images of the heart. In line with the dice loss functions for comparing accuracy in training models, authors such as Hering et al. [8] have touched on extending existing algorithms for fixed to dynamic segmentation mapping whilst combining CNN-based square difference loss and similarity scores. Balakrishnan et al. [9] extended the work on Voxelmorph for calculating the Dice score between fixed and warped moving segmented masks. Hansen et al. [10] found that the PDD-Net architecture provided a 15% increase in accuracy during monomodal CT registration using a combination of probabilistic dense displacements and differentiable mean-field regularization.

2.2 Augmented Reality Based on DL Image Registration

The application of AR based technology for surgical guidance has gradually become relevant in clinics. The use of image superposition for pre-planning of complicated surgeries helps surgeons to transfer the reconstructed medical images from the database to the operating room, hence increasing accuracy and reducing surgery periods. Most clinically approved studies use non-invasive fiducial markers displayed through a visor, to track the position of an end-effector with respect to the patient's body using DL algorithms, libraries and software development kits [11]. Jiang et al. [12] used the principle of medical data registration for detecting simple 2D recognizable objects in a workspace using RGB cameras. Ma et al. [13] used preoperative CBCT images for generating a trajectory during dental implant surgery, where the naked-eye 3D reconstructions were superimposed in-situ to form an AR scene around the patient's mandible using matching markers on the patient's body and the matching CT scans. Wang et al. [14] used SDKs as a computing database for tracking 2D and 3D feature

coordinates on medical images and create a calibrated coordinate system between the real scenario and the digital world. Jiang et al. [15] focused on an AR guided navigation platform for dental implant surgery using mesh to point cloud extraction for preoperative image registration, which achieved lower errors and ($p<.05$) for the surgery time.

3. Methodology

In this study, we propose the use of an existing framework based on deep convolutional neural networks (CNN) for CT scan registration of the head and neck. The unsupervised image registration framework consists of two branches as shown in the block diagram below, one for the moving image, M and one for the fixed image, R, each with their associated label. During training, a self-supervised set of labeled data is fed through the neural network, generating a function F' and is resampled to obtain the warped moving image.

3.1 Dataset and Implementation Details

CT image reconstructions of the head and neck were generated using a public dataset from The Cancer Imaging Archive. The DeepReg open-source repository is cloned onto the PC terminal to feed input data of 271 test images, each with 37 slices through supervised network. We perform rigid registration of the dataset first where an image coordinate system is initially mapped onto the other to align tissue deformations. This means that only translation and rotation can be performed for target objects to achieve correspondence. Our experiment involves multi-modal registration of real-time CT scans with preoperative ones which will allow for marker-based planning of a trajectory. We aim to use a displacement vector to project the moving coordinates into the static coordinate space. This transformation is characterized as a combination of vectors which allow for all voxels in a CT image to be equalized in a warping procedure. Generally, the voxels within CT images have a wide range of intensity values across their slices which are calculated using intensity histograms. We use measures such as normalized cross-correlation (NCC), mutual information (MI) and basic sum-square-difference (SSD) to measure the common features between moving and fixed images.

3.2 Evaluation Metrics

It is to be noted that the computationally heavy datasets used for image processing require high GPU processing speeds, which generate complex, inaccurate ground truth transformations and therefore DL algorithms such as weakly supervised methods are more suitable for training them. This enables a pair of corresponding moving and fixed labels to be computed, thus extracting the label dissimilarity during the registration. For this experiment, we compare the prediction array to the mask array with the aim of identifying the positive and negative outcomes as well as the mapping results to calculate a loss function for the regions of interest (ROI). The input data includes a probability map from the model, the mask array containing corresponding ground truths and the base threshold predictions. The base image contains 37 slices, with a dimension

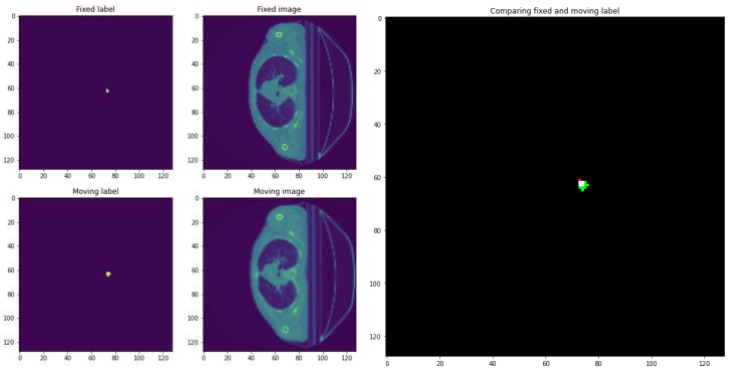
of 128 x 128 pixels, which is the same for the base label. The outputs include a dense deformation ϕ which has an extra index (128,128,2) because at each pixel, we require a direction vector. The output RGB CT images indicate areas of overlap between masks and predictions. We observe that the labels have transformed from the 0^{th} slice at index 1. Upon magnification, a color-coded outcome chart is used to distinguish among true positives from false positives (FP) and false negatives (FN). The intensity based loss between the images shown below is calculated using the mean difference per image tensor with dimensions; f,m φ and is depicted in the equation below :

$$L_{us}(f,m,\varphi) = L_{sim}(f,m \circ \varphi) + \lambda L_{smooth}(\varphi)$$

Where L_{sim} creates differences in appearance, and L_{smooth} creates local spatial variations in φ. In the process of image registration, we perform voxel-wise correspondence between the fixed and moving datasets whereby we may use affine or non-rigid transformations depending on the degrees of freedom. The function below:

$$\mu = \min L (T_{\mu} ; I_{F} , I_{M})$$

describes the optimization problem of registering CT images, where T is the desired spatial transformation which maps M voxels onto F and S is a measure of dissimilarity between the fixed image and the warped moving image. For our experiment, we chose a 3 x 4 affine transformation matrix which is used to visualize the data registration on the fixed images, and then analyze the displacements of consecutive pixels in labels from the test sequence. This means that the straight and parallel lines in the image remain intact but may be translated with a slight change of angle. We found that some of the labels that appeared in the fixed and warped moving images had moved across the slices in the sequence and therefore disappeared from the original moving image. The same process applied for labels in the original moving images disappeared from the fixed and warped moving images, which proved that the warping process was successful.



a) b)

Fig. 2 a) Position of moving label with respect to the fixed label in a 128x128 pixel graph b) A comparison between the positions of the moving label to the fixed label where TP = white, FP = green and FN = red.

3.2.1 Control Experiment using U-Net Module

We focus on the use of supervised learning techniques in order to predict the outcomes of a particular interventional pathway through the brain. In the control experiment, we use a weakly supervised method in order to compare the fixed images to their moving counterpart. We then apply another CNN architecture from the VoxelMorph library, adapted from the UNet, to compare and contrast the accuracy levels in locating labels in the fixed and warped moving segmentations. U-Net is commonly used for image segmentation tasks and provides accurate registration results. It is developed from the FCN network and has multiple features such as enhanced edge detection, minimized information loss, higher background weight amongst others. In this experiment, we describe the network used with an encoder input of size 16 x 32 x 32 x 32 but the framework parameters may vary depending on the requirements. We apply three dimensional, 32-layer convolutions in both the encoder and decoder stages using a kernel size of 3 and a stride of 2, where each convolution is followed by a LeakyReLU layer. The process starts with a downsampling step through different degrees of convolutions, followed by a series of upsampling steps and concatenations to decode the network size after learning from the encoding stages. Successive layers of the decoder operate on thinner spatial scaling which enables accurate CT image alignment, whereby the softmax function activates the pixels and generates a probability map.

3.3 Experiment Results

In this section, we present the results of each experiment and attempt to compare the performance based on certain evaluation metrics.

3.3.1 Image Registration

The results of the prediction test (Fig. 3) are shown below in a warped label simulation whereby we attempt to detect the dissimilarity (SSD) between the fixed label and the moving label by calculating the dice score. In this case, the dice score is 0.517 for 32^{nd} slice of the sequence, where white pixels indicate instances where the model proved that the moving label was in fact located in the same position as the fixed label. The green pixels indicate FP where the moving label was detected in the wrong pixel segment compared to the fixed one and finally, the red FNs indicate a missed segmentation between fixed and moving label. Detecting the image-based loss of each moving tensor or vector compared to the fixed tensor enables us to visualize an average difference between their positions.

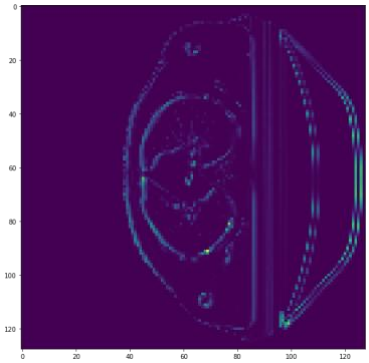


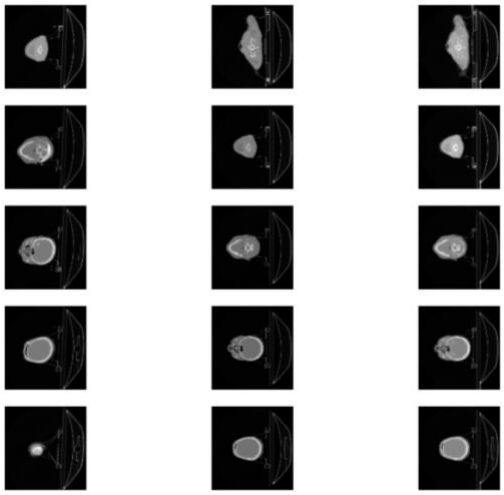
Fig 3: CT image reconstruction after performing an SSD between the warped moving image and the fixed image.

3.3.2 Comparison between U-Net and CNN

In Table 1, we compare the results of the medical image segmentation for the same dataset using both architectures, using metrics such as accuracy, dice scores, SSD and training speeds. The experiment shows clearly that the U-Net and CNN are both suitable for medical image registration. We observed from Table 1 that the U-net network performs better than the original CNN with a dice score of 0.621 on the 32nd slice of the sequence, which was an increase of 10%. Figure 4 shows the difference in intensities and contrasts of the moving label tracked throughout both experiments i.e., the CNN architecture and the U-Net. It is observed that the label appears in most slices in the UNet but appears to fade away during the CNN training, which means that the UNet outperformed the CNN.

Table 1: Results of Image Registration

Heading Level	**U-Net**	**CNN**
Accuracy	0.71	0.65
Training Speed (s)	10	30
F1-score	0.86	0.81
SSD	30450	32342.5
Dice Score	0.75	0.51



a) CNN b) U-Net

Fig. 4. Registration of the moving label after being warped throughout the 32-slice sequence during training of the U-Net versus CNN.

5. Discussion and Conclusions

Deep learning algorithms have been at the core of medical image registration ever since the concept of surgical visualization emerged in the clinical sector. The precision to which surgeons are now able to perform using image overlays and pre-planning marker-based or marker-less trajectories is a steppingstone towards clinical research in the academic community albeit requiring improvement in the medical image quality for operations which involve morphological and volumetric differences, for example, in the resection of the lung in its deflated state using AR may be impractical since 3D reconstructions are made upon inflated lung CT/MRI scans. Medical image registration requires a high amount of accuracy and efficiency, especially when it comes to complicated cases in surgery where minimal invasion and lower operation times are preferred for quicker convalescence. The use of fiducial markers or "labels" for in-situ AR guidance is an evolving technique which can be used to detect, remove, and alter anatomical landmarks precisely.

This paper uses a variant of the U-net network in parallel with a CNN network from the DeepReg tutorial to analyze and compare the efficacy of label registration on a pre-processed and pre-segmented cancer dataset. The optimized CNN architectures are used for detecting the non-invasive markers throughout the sequence and finally, the segmentation results are compared through relevant evaluation criteria. This method is universal, which means that different datasets can be used for analyzing the performance of both neural networks to obtain an efficient registration technique. However, both methods have their flaws since there may be larger datasets whereby the results are easily influenced by the number of training sets. We are continuously optimizing the use of neural networks for image and label registration through various

supervised learning techniques. The performance of the CNNs can be improved by using an image deformation method, hence reducing dice loss of the labels within a sequence, followed by the generated anatomical path for the surgeon's view.

References

1. "Holograms replacing cadavers in training for doctors". (Online) Available at: https://www.theguardian.com/society/2016/nov/17/medical-trainers-look-to-virtual-reality-tech [Accessed on August 10, 2022]
2. Venkatesan, M.; Mohan, H.; Ryan, J.R.; Schürch, C.M.; Nolan, G.P.; Frakes, D.H.; Coskun, A.F. Virtual and augmented realityfor biomedical applications. Cell Rep. Med. 2021, 2, 100348.https://doi.org/10.1016/j.xcrm.2021.100348
3. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
4. Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., & Wang, S. (2019, October). X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In *International conference on medical image computing and computer-assisted intervention* (pp. 247-255). Springer, Cham.
5. Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. *Advances in neural information processing systems*, *28*
6. Sokooti, H., de Vos, B., Berendsen, F., Ghafoorian, M., Yousefi, S., Lelieveldt, B. P., ... & Staring, M. (2019). 3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations. *arXiv preprint arXiv:1908.10235*.
7. De Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., & Išgum, I. (2019). A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, *52*, 128-143.
8. Hering, A., Häger, S., Moltz, J., Lessmann, N., Heldmann, S., & van Ginneken, B. (2021). CNN-based lung CT registration with multiple anatomical constraints. *Medical Image Analysis*, *72*, 102139.
9. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). VoxelMorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, *38*(8), 1788-1800.
10. Hansen, L., & Heinrich, M. P. (2021). GraphRegNet: Deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs. *IEEE Transactions on Medical Imaging*, *40*(9), 2246-2257.
11. Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, *65*(20), 20TR01.
12. Jiang, Z., Yin, F. F., Ge, Y., & Ren, L. (2020). A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. Physics in Medicine & Biology, 65(1), 015011.
13. Ma, L., Jiang, W., Zhang, B., Qu, X., Ning, G., Zhang, X., & Liao, H. (2019). Augmented reality surgical navigation with accurate CBCT-patient registration for dental implant placement. *Medical & biological engineering & computing*, *57*(1), 47-57.
14. Wang, X., Kim, M. J., Love, P. E., & Kang, S. C. (2013). Augmented Reality in built environment: Classification and implications for future research. Automation in construction, 32, 1-13.
15. Jiang, W., Ma, L., Zhang, B., Fan, Y., Qu, X., Zhang, X., & Liao, H. (2018). Evaluation of the 3D Augmented Reality–Guided Intraoperative Positioning of Dental Implants in

Edentulous Mandibular Models. *International Journal of Oral & Maxillofacial Implants*, *33*(6).