Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Utilizing AI/ML approaches to assess surgical performance and provide virtual coaching
Principal Investigator:	Don Anderson, PhD; Geb Thomas, PhD; Stephen Baek, PhD; Matt Karam, MD
Prepared by (IIAI):	Yanan Liu
Other investigators:	
Date:	

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	January 25, 2022
Funding agency	AHRQ
Grant mechanism	R18 Research Demonstration and Dissemination Grant (resubmission)
If no … Why not? What went wrong?	

Brief summary of accomplished results:

We developed and validated a VGG-16 model to automatically classify whether the image orientation was along the AP/Lat image planes with achieved classification correctness of 0.88. In addition, we developed and validated two U-net models to segment femur bone from each of the two different X-ray image planes (AP/Lat) with average Dice Similarity Coefficient between AI predicted segmentation and manual segmentation of 0.94 for AP images and 0.82 for Lat images.

Research report:

Aims (provided by PI):

AIM 1: Develop a real-time instance segmentation algorithm to automatically label the surgical wire, anatomical landmarks, and specific surgical objectives based on a training set of roughly 1,000 labelled fluoroscopic image sequences collected during surgeries and synthetic training images generated in prior work with our simulator. Algorithms such as Mask R-CNN and U-Net will be explored to achieve real-time computation.

AIM 2: Develop a recurrent neural network (RNN) to recognize simulator-user behaviors and provide machine-generated advice at specific moments during training. The input to the neural net will be hundreds of previous simulator training logs coded with relevant parameters, alongside real-time fluoroscopic Images generated by the simulator and labelled with specific technical shortcomings.

As the project progressed, the team decided to focus on segmenting the femur bone on fluoroscopic images.

Data:

From intraoperative X-ray fluoroscopy imaging, we got 1839 fluoroscopic images with associated manual bone segmentation. They were from two different image planes: 1210 images from the anterior-posterior plane (AP) and 629 images from lateral plane (Lat).

AI/ML Approach:

In this study, a VGG-16 model was implemented for image classification – each image was classified as AP image plane or Lat image plane. Training/validation split was 1422/417. Following the labeling of images as AP or Lat, two U-net models were trained for segmentation of the femur bone in the respective image planes. Training/validation split was 938/272 for AP and 484/145 for Lat.

Experimental methods, validation approach:

Data Preparation

Data preparation or pre-processing is an essential step in any machine learning study. To save computation time, we resized each image to the size of 128x128. In this project, data normalization is an important step which ensures that each input parameter (pixel) has a similar data distribution. This makes convergence faster while training the model. We normalized the image intensity to [0,1] by its maximum value 65535.

Ground truth obtained by manual segmentation

Femur bone segmentation on fluoroscopic images was provided by expert manual segmentation.

Image classification

The VGG-16 is one of the most popular pre-trained models for image classification, its architecture is shown in Fig. 1.

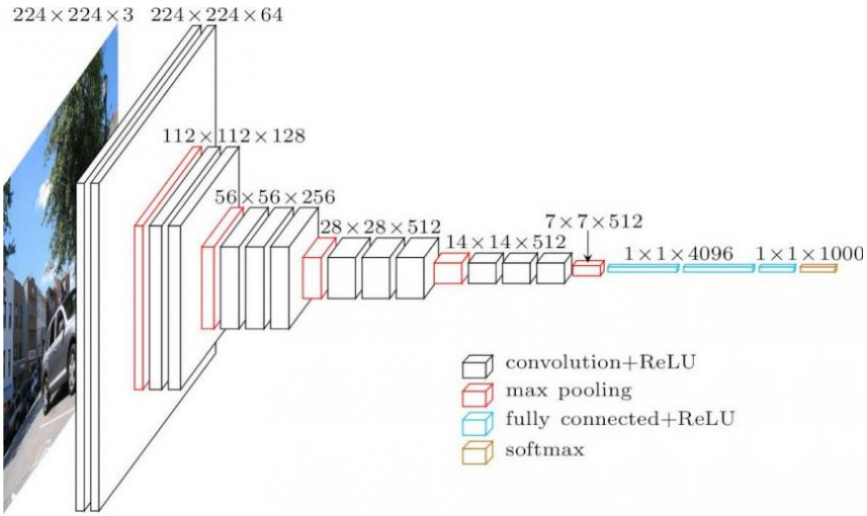

Figure 1. VGG16 architecture [1]

Image segmentation

The Unet is convolutional network architecture for fast and precise segmentation of images. In this project, Unet was implemented with Keras functional API, which makes it extremely easy to experiment with different architectures – see Fig. 2. (https://github.com/zhixuhao/unet).

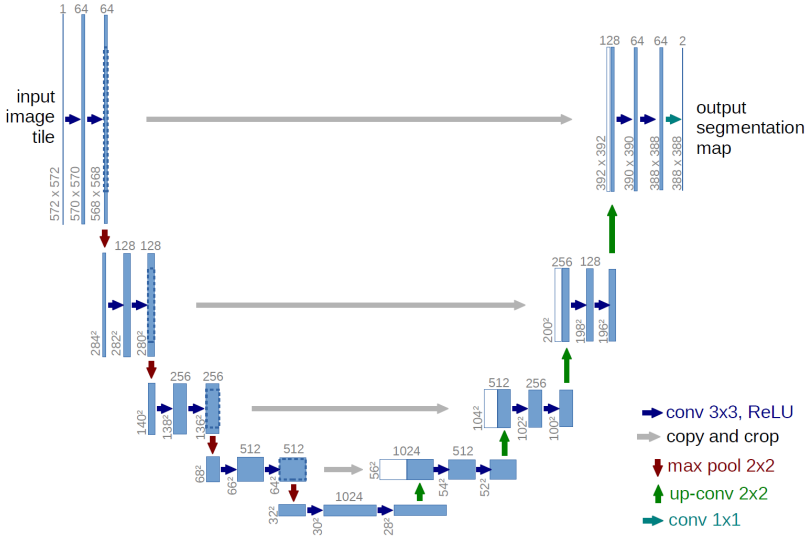

Figure 2. U-net architecture [2]

In the training phase, the input images and their corresponding masks are used to train the U-net, and in the test phase, we give an image as input to generate the corresponding mask as output. Sigmoid activation function makes sure that mask pixels are in [0, 1] range.

Results:

The accuracy of VGG-16 model to automatically classify AP/Lat image planes was 0.88 in the validation group. Averaged Dice Similarity Coefficients between AI-predicted segmentation and manual segmentation were 0.94 for AP images and 0.82 for Lat images.

Since we only had 629 images from the Lat planes, the Dice coefficient can likely be improved after more data become available.

Ideas/aims for future extramural project:

The following text is taken, verbatim, from our recent resubmission of an AHRQ R18 proposal that scored well last cycle but was not selected for funding. The AI/ML capabilities developed as part of this IIAI Pilot Grant provided the background material for this work.

Aim 1: Measure differences in resident OR performance from objective analysis of surgical imagery, and speed up these measurements

The objective of this aim is to measure the precision and efficacy of a set of techniques for objectively measuring intra-operative surgical skill in three orthopedic residency programs. The hypothesis is that *objective skill measurement is consistent with expert subjective analysis, but more sensitive, unbiased, and accurate*. We will test our hypothesis by objectively analyzing imagery and video already collected in surgery. We will compare the variability of the objective assessments with those made by supervising surgeons participating in the surgeries using the statistical approach of generalizability (G) theory.[107-110] G theory is a statistical framework used to model and analyze

measurements. G theory uses generalizability studies to model the composition of assessment scores and decision studies to forecast the reliability of measurements given various conditions (e.g., number of items, number of raters, number of occasions, number of stations, etc.) under which they are obtained.[111] A G study can partition and quantify score variance to provide information regarding both precision (reliability) and accuracy (validity). We will further explore techniques for automating the objective scoring so that it is more convenient.

The rationale for this aim is that the objective skill measures will prove to be move expedient for resident training and assessment while providing greater consistency and precision than the current subjective assessments. When Aim 1 studies are complete, we expect the new measurement techniques to reveal meaningful differences in resident skill consistent with perceptions of supervising surgeons. We also expect them to be sensitive, unbiased, and amenable to becoming a regular part of resident training. Such findings will open the door to more sensitive studies by other researchers on the details of training. This will not only facilitate the study of the training methods but may also ultimately lead to quantifiable skill proficiency levels required for board certification.

…

…

Sub-Aim 1.3 Fully Automated Scoring through Crowd Sourcing and AI. We will continue to advance our automated scoring technology, both for the fluoroscopic image sequences and the arthroscopy video. This will allow us to more rapidly analyze and return results to the residents participating in the study, reduce analysis costs, and help to disseminate our results to the broader surgical community.

We will continue to train CNNs with segmented fluoroscopic images produced by on-line (crowd source) analysis workers through Amazon's Mechanical Turk interface, which pays workers worldwide to perform simple tasks online. In prior work, we provided workers with 1839 fluoroscopic images of femurs, 1210 in the AP plane and 629 in the lateral plane. The workers precisely defined regions of the images that were part of the femoral head, neck, and shaft. Each image was coded thrice, outliers were rejected, and the remaining results were manually inspected. The coded images were then split into a training and validation set, and the training set was used to train a CNN. The VGG-16 model CNN automatically classified the AP/Lat image planes with an accuracy of 0.88 in the validation group. Averaged Dice Similarity Coefficients, a measure of pixel-by-pixel consistency, between AI-predicted segmentation and manual segmentation were 0.94 for AP images and 0.82 for Lat images. We expect to repeat this process for the new surgeries with a larger data set, a total of approximately 50,000 images/year, to automatically define the location of key anatomical features and wire positions for generating IDEA scores from the fluoroscopic data.

We will also explore ways to extend this work towards automatic interpretation of the video analysis of the arthroscopic video sequences. This endeavor will begin with still image analysis of the video using techniques that proved successful for the fluoroscopic analysis. Similar approaches for knee arthroscopy[112] have achieved Dice coefficients of between 0.48 and 0.79 when identifying the femur, tibia, ACL, and meniscus. We expect that augmenting the CNN to account for sequences of successive images will improve its ability to identify structures, and facilitate the timing and scoring procedures devised for Aim 1.2.

Together the elements of Aim 1 will advance from subjective towards objective performance analysis. The analysis techniques that have proven fruitful in the past will be expanded to a broader range of surgeries and move from static surgical images towards video analysis. The order of magnitude increase in experimental scale will allow the analysis to become more detailed and allow us to tease apart more subtle differences in techniques and among learners. Should the analysis of learner behavior prove to be too subtle to reliably detect in the dataset, we can rely on more basic elements of time and

number of images, which are reliably used in simulator training and are quite likely to be evident in surgical data, with appropriate data filtering. Ultimately, we expect our effort to provide techniques that will assess learner performance with greater accuracy and reliability than current subjective approaches. This will provide a more solid foundation on which to advance orthopedic training techniques.

Publications resulting from project:

Not applicable

References:

1. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” ILSVRC, 2014

2. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Med. Image Comput. Comput. Interv. MICCAI 2015, pp. 234–241, 2015.