Ultrasound Video Analysis for Skill Level Assessment in FAST Ultrasound

Robert E. Tyrrell and Matthew S. Holden*

School of Computer Science, Carleton University, Ottawa, Canada

matthew.holden@carleton.ca

Ultrasound Video Analysis for Skill Level Assessment in FAST Ultrasound

Abstract

FAST ultrasound is a medical procedure to assess for free fluid following physical trauma. FAST images can often be difficult to interpret and requires operators to be properly trained. Traditionally, skill is assessed by direct observation from experts, which is expensive and error prone. This project aims to use deep learning to provide automated skills assessment for FAST exams.

Modified I3D networks, a type of modern neural network with a focus on action-based items, were retrained for this purpose. First, a network to identify the skill level of the users from all the ultrasound videos was trained using FAST videos of each vital region divided by novice, intermediate and expert users. Following this, 4 networks corresponding to skill level identification in each region were trained using the previously constructed model. The model's performance was evaluated using k-fold cross-validation.

Results found a testing accuracy of 82.6% for skills assessment using the modified I3D networks.

These results are an improvement over the previous results for skill level evaluation, implying potential use of an I3D network for evaluating skill level from ultrasound video in the future with the proper finetuning.

Keywords: diagnostic ultrasound, skills assessment, deep learning

1. Introduction

Focused assessment with sonography in trauma (FAST) ultrasound plays a crucial role in diagnosis of free fluids within vital regions of the body due to trauma. This is because FAST is safe, with a lack of radiation exposure, allowing for more extensive use than other diagnostic methods. FAST is also inexpensive and can give accurate results within real time, with general

sensitivities ranging between 85% and 95% (Bloom & Gibbons, 2019).

The ultrasound images generated by FAST, however, can be difficult to read (AIUM, 2014). As such, operators must be properly trained for this purpose. This can be a very lengthy and costly process due to the time required for checklist-based assessment by expert preceptors (Ann Emerg Med, 2017). Thus, methods for automated assessment of skill in FAST ultrasound could have added value in training.

Prior works have investigated the use of kinematic data to assess skill level automatically in diagnostic ultrasound. One study found that experts had shorter ultrasound probe path lengths and fewer discreet movements than novices (Ziesmann et al, 2015). Another showed intermediate level users with some prior training scanned more points of interest that novice users, in addition to having shorter path lengths and taking less time (Bell et al, 2017). Recent work confirms that experienced users exhibit shorter times, smaller path lengths, fewer hand motions, and a smaller working volume than beginners (Zago et al, 2019). Overall, prior work shows that skill level can be determined based on summary statistics derived from motion data.

There has also been considerable prior work on deep learning in medical image processing for skills assessment. One such work investigated the utility of retraining Convolutional Neural Networks (CNNs) for Transoesophageal Echocardiography (TEE) graded by experts on a criteria-based checklist for skills assessment (Mazomenos et al, 2018). This was found to have high accuracy in predicting the evaluators' scores. Other work has addressed using a specific type of 3D ConvNet, Inflated 3D ConvNets, in surgical skill assessment (Funke et al, 2019) from videos of minimally invasive surgeries on the open-source JIGSAWS dataset (Gao et al 2017). The retrained network showed high accuracies for predicting skill levels, at approximately 95-100%.

Based on its success, we investigated Two-Stream Inflated 3D ConvNets (I3D), a modern 3D ConvNet incorporating RGB and optical flow information from video, for skills assessment in FAST (Carreira & Zisserman, 2017). The optical flow component of the network is intended to provide greater benefit for classifying action-based items involving movement. This method has been shown to provide superior results in task identification on the open-source UCF-101, HMDB-51, and Kinetics datasets (Carreira & Zisserman, 2017). It is hypothesized that the I3D network might provide an effective method for assessing skill in FAST ultrasound examinations.

The purpose of this study is to develop a method for assessment of skill level in FAST ultrasound examinations from ultrasound videos using the I3D network. We hypothesize that an automated method based solely on ultrasound video will reduce the burden on experts' time, increase standardization across evaluations, and be easy to deploy.

2. Methodology

2.1 Architecture

We used the I3D model proposed by Carreira and Zisserman (Carreira & Zisserman, 2017). This model consists of two separate 3D ConvNets: one trained on RGB data and one trained on optical flow data (FIGURE 1).

The idea is that the optical flow data itself is invaluable in helping to identify the movements taken in a video, thereby providing indication of what is occurring within the video. In this specific case, it is hypothesized that flow data is a boon in determining skill level involved in FAST, as skill level is largely defined by each user's actions. For classification, the videos would then be passed through both the RGB and flow sub-networks and the results averaged to increase accuracy in predictions.

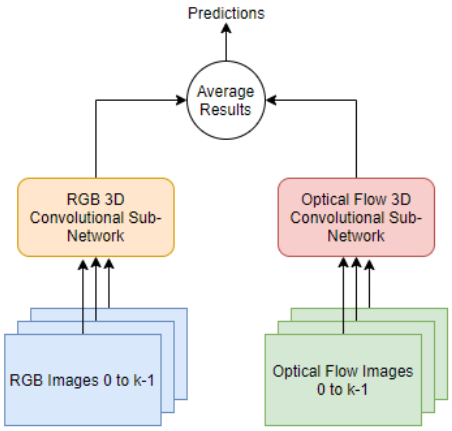

FIGURE 1 - Model of the I3D network's architecture. The original model can be found at: *https://arxiv.org/pdf/1705.07750.pdf* (Carreira & Zisserman, 2017).

The I3D network is pretrained on the Kinetics dataset. We retrained an initial I3D network for skills assessment using FAST ultrasound videos from all regions of the scan (i.e. Right Upper Quadrant, Left Upper Quadrant, Heart, and Pelvis). Subsequently, we recreated four region-specific versions of this network (each with an RGB and flow subnetwork) and finetuned each one using FAST ultrasound videos specific to each region of the scan. At test time, the appropriate region-specific network is chosen to classify the operator's skill level from a FAST ultrasound video.

2.2 Dataset

The dataset used in this experiment is the same as the FAST ultrasound dataset described above (Bell et al, 2017), and consists of multiple ultrasound video scans of each of the vital regions from a FAST ultrasound exam: right upper quadrant (RUQ), left upper quadrant (LUQ), heart, pelvis (FIGURE 2). The scans were performed by Novices, Intermediates, and Experts on a single healthy volunteer in a simulation-based training environment. Novices are users with little experience in FAST ultrasound but who have completed a structured didactic curriculum within

their education. Intermediates are users who have performed at least 50 supervised FAST examinations according to Canadian Point of Care Ultrasound Society guidelines. Experts are attending physicians who have previously completed a fellowship in point-of-care ultrasound.

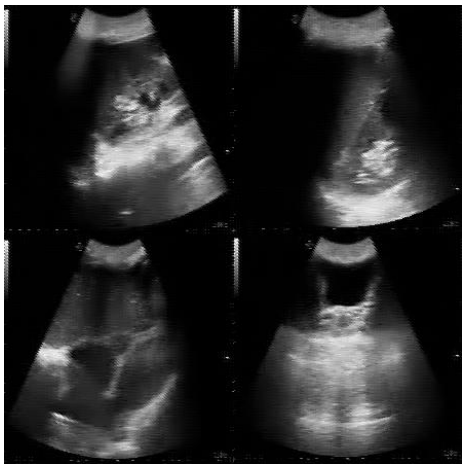

FIGURE 2 - Clockwise from the top left images, the Right Upper Quadrant, Left Upper Quadrant, Pelvis and Heart Region.

This dataset contains 14 novices, 15 intermediates, and 3 experts. For each user in this dataset, there are at least four videos, one pertaining to each region of the FAST examination, with several additional scans for specific regions also existing. This meant there was a grand total of 132 full videos of FAST scans of varying length.

2.3 Experimental Setup

This dataset was then divided up into 3 folds with no repeated entries for training, validation, and testing using the user-out method of k-fold cross-validation (Ahmidi et al, 2019). A maximum of 3 folds could be used, as there were only 3 expert users to divide amongst the folds. Two folds contained 5 novices, 5 intermediates, and 1 expert; one contained only 4 novices, 5 intermediates, and 1 expert This cross-validation protocol ensures every possible combination of

the 3 folds serves as the training, validation and testing sets.

In this case, the validation set only served as a tool for determining which set of checkpoints to use for testing in the end (i.e. early stopping). That is to say, the validation set was used to determine the checkpoints yielding the highest accuracy, provided an accuracy of 85% could be surpassed within the training set.

The main network was trained using scans from all regions. Subsequently, the dataset was divided into 4 smaller datasets, each corresponding to a specific region, and the region-specific networks were finetuned separately. This was done to maximize accuracy in skill level identification for each region.

2.4 Data Pre-Processing

Each video was cropped to remove additional information found along the edges of each video (e.g. scan date, ultrasound parameters, etc.) and to a size of 224x224. RGB videos were created from the greyscale ultrasound videos by taking each RGB component to be equal to the greyscale intensity. Optical flow videos were created for each scan by the TV-L1 optical flow algorithm in OpenCV (Zach, Pock & Bischof, 2007).

2.5 Data Augmentation

Data augmentation primarily took the form of window slicing. This was done to remedy the small size of the dataset in use, providing additional input for more accurate training, and reducing training time by greatly reducing video size. Towards this end, each video was split into up to 10 non-overlapping segments of minimum 25 frames in length. Videos with fewer than 250 frames were split into fewer non-overlapping segments. One snippet of 25 consecutive frames was taken randomly from each of these segments for end use, such that overlap was impossible.

2.6 Systems Used

We used a pretrained version of the I3D network from DeepMind (DeepMind, 2019). For finetuning, we used the I3D Finetune repository of USTC Video Understanding's GitHub account was employed (Hu & Zhou, 2018). The networks were trained and tested on Compute Canada's Graham cluster

3. Results

TABLE 1 depicts the accuracies, skill level-specific sensitivities, macro-sensitivities, and macro-sensitivities without the expert class for each region and for all regions combined. As can be seen, the average accuracy of every region with regards to skill level identification as run through the test set is approximately 82.6%. The other performance measures are included to offer insight into the effectiveness with regards to specific regions, skill level evaluations, and to highlight the negative effect the poor expert sensitivities had on macro-sensitivity.

	All	RUQ	LUQ	Heart	Pelvis
Accuracy	**0.826**	**0.797**	**0.891**	**0.813**	**0.803**
Novice Sensitivity	0.866	0.821	0.964	0.821	0.857
Intermediate Sensitivity	0.967	0.933	1.000	0.967	0.967
Expert Sensitivity	0.000	0.000	0.000	0.000	0.000
Macro-Sensitivity	**0.611**	**0.585**	**0.655**	**0.596**	**0.608**
Macro-Sensitivity w/o Experts	**0.916**	**0.877**	**0.982**	**0.894**	**0.912**

TABLE 1 - Performance Measures for skill level classification in each region of FAST using the main and region-specific networks.

FIGURE 3 shows the confusion matrices for each region and the combined results when all are tallied. Under the 3-fold cross-validation protocol, each video appears in the test fold twice; thus, the reported classification for each video is an average classification over all times the video appears in the test fold.

All	Novice	Intermediate	Expert
Novice	48.5	7.5	0
Intermediate	2	58	0
Expert	9	4	0

RUQ	Novice	Intermediate	Expert
Novice	11.5	2.5	0
Intermediate	1	14	0
Expert	2.5	0.5	0

LUQ	Novice	Intermediate	Expert
Novice	13.5	0.5	0
Intermediate	0	15	0
Expert	2	1	0

Heart	Novice	Intermediate	Expert
Novice	11.5	2.5	0
Intermediate	0.5	14.5	0
Expert	2	1	0

Pelvis	Novice	Intermediate	Expert
Novice	12	2	0
Intermediate	0.5	14.5	0
Expert	2.5	1.5	0

FIGURE 3 - Confusion matrices for skill level classification in each region of FAST. Rows correspond to the true class labels; columns correspond to the predicted class labels. Results are averaged across folds.

The above results in FIGURE 3 and TABLE 1 were computed using a plurality vote run over all the snippets for a video following testing. That is to say that the skill level predicted by the largest number of snippets within a video was decided as the overall prediction of said video.

4. Discussion

The results of this experiment are promising. Skill level evaluation accuracies of 79.7%, 89.1%, 81.3%, 80.3% were found for right upper quadrant, left upper quadrant, heart, and pelvis regions respectively. A combined accuracy of predicting skill level across all videos regardless of region, however, was found to be 82.6%. We observe a considerably lower macro-sensitivity for each region, with the macro-sensitivity for all video predictions being 61.1%. This is likely because only one expert is in the training set at any time. We have thus also reported the macro-sensitivities of the novice and intermediate predictions alone, which was 91.6% for all video predictions. This shows the negative effect the lack of expert predictions had on macro-

sensitivity.

These results compare favourably to a separate experiment performed on the same original data to determine skill level using expert-defined summary statistics from motion data (Holden et al., 2015) and a random forest classifier with the same cross-validation scheme. The performance measures as well as a confusion matrix of these results can be found below in TABLE 2.

	Skill Level Identification
Accuracy	**0.700**
Novice Sensitivity	0.696
Intermediate Sensitivity	0.767
Expert Sensitivity	0.417
Macro-Sensitivity	**0.627**
Macro-Sensitivity w/o Experts	**0.732**

	Novice	Intermediate	Expert
Novice	39	16	1
Intermediate	10	46	4
Expert	0	7	5

TABLE 2 - Performance measures and confusion matrix of the same data using a random forest classifier on expert-defined summary statistics.

For further comparisons, we also report results from just the flow and RGB sub-networks on their own rather than the combined I3D network. This showcases the added value that the full I3D architecture had in prediction.

RGB	All	RUQ	LUQ	Heart	Pelvis	**Flow**	All	RUQ	LUQ	Heart	Pelvis
Accuracy	**0.717**	**0.781**	**0.734**	**0.641**	**0.712**	**Accuracy**	**0.616**	**0.641**	**0.719**	**0.531**	**0.594**
Novice Sensitivity	0.839	0.821	0.786	0.821	0.929	**Novice Sensitivity**	0.464	0.429	0.679	0.357	0.393
Intermediate Sensitivity	0.725	0.867	0.767	0.567	0.700	**Intermediate Sensitivity**	0.883	0.933	0.900	0.800	0.900
Expert Sensitivity	0.154	0.167	0.333	0.167	0.000	**Expert Sensitivity**	0.039	0.167	0.000	0.000	0.000
Macro-Sensitivity	**0.606**	**0.618**	**0.629**	**0.518**	**0.543**	**Macro-Sensitivity**	**0.462**	**0.510**	**0.562**	**0.386**	**0.431**
Macro-Sensitivity w/o Experts	**0.782**	**0.844**	**0.776**	**0.694**	**0.814**	**Macro-Sensitivity w/o Experts**	**0.674**	**0.681**	**0.843**	**0.579**	**0.647**

TABLE 3 – RGB and flow performance measures for skill level classification in each region. Full video performance measures were computed via use of a plurality vote.

One important limitation to this work is overfitting of the training data. We found 0% sensitivity for expert identification within the I3D network. The underrepresentation of expert

data prevents generalization across expert users within the I3D network, with predictions being skewed in favour of the more likely outcomes. However, it is worth noting that differentiation between novices and intermediates is a more difficult task than novices and experts due to more closely related skill levels, and itself has value in providing trainees feedback. The imbalance is made worse using snippets considering that the expert videos tend to be shorter. The mean duration for novices, intermediates and experts was 766 frames, 516 frames, 154 frames respectively. Thus, while novice and intermediate videos can typically be split into 10 segments, expert videos on average produce only 6 segments. This furthers the class imbalance between the three, an issue that can only be corrected by increasing the number of expert snippets in use. This would require overlap of expert snippets or reducing the size of expert snippets which we hypothesize would reduce the accuracy of the flow subnetworks.

Another limitation lies in the use of a plurality vote at the end of testing for the purpose of predicting skill level of full videos using snippets. This could be less accurate than using a consensus layer (e.g. Wang et al. 2018, Funke et al, 2019), which incorporates the aggregation over snippets of the same video into the network. A plurality vote was nonetheless chosen as it allowed considerably reduced training time. There is also notably, the issue of the limited use of the validation set to tune hyperparameters for creating the optimal network. As such, hyperparameters have largely been left intact from their original state, making it more likely that issues such as overtraining have occurred. For example, epoch size and training time are rather large for a dataset so small, and dropout, momentum, and learning rate remain untouched. There is also the issue pertaining to the low accuracy of the flow networks in comparison to the RGB networks. This itself likely stems from the small snippet size of 25 frames, which makes it difficult to judge the actions being taken in such a short amount of time.

From a clinical perspective, there remain two primary limitations. First, it is unclear how skills assessment within the simulated environment of our dataset will translate into a clinical setting. In particular, the experiments were performed on a single healthy volunteer. Furthermore, we used appointment status and previous training as our ground-truth for skill. It may be more appropriate to used checklist-based assessment (Ann Emerg Med, 2017) as our ground-truth, due to fluctuation in individuals' performances.

We propose several avenues of future work to overcome these limitations and generalize this experiment to a real setting. It is suggested that a larger dataset, with more expert scans included, be used in order to better train the network to recognize all levels of skill. It is also suggested that larger snippets be used in order to improve the flow prediction accuracy. We also propose that integrating a consensus layer in our network (Wang et al. 2018), rather than a plurality vote, to classify videos based on snippets could improve performance at the cost of additional training time. Finally, it is suggested that further experimentation with the finetuning of hyperparameters be done to prevent over-correction thanks to a lack of alterations in that regards within this current experiment.

The outlook for skills assessment in FAST ultrasound remains positive. Notably, the accuracy of the I3D networks surpasses not only those presented by the Random Forest classifier, but those of the both the flow and RGB sub-networks that make up each network. This fact implies the possible benefits of using the I3D framework as a machine learning tool for identification of skill level with regards to FAST Ultrasound examinations.

5. Conclusion

This work shows promise for the use of the I3D framework in evaluating the skill level of trainees in FAST ultrasound. The reported results outperform those from previous attempts,

including those of the used 3D sub-networks. In fact, the primary issue displayed with these results existed in the form of a skew of predictions away from the expert class. Future work includes proper hyperparameter finetuning and using a larger dataset with a greater balance between classes. Overall, the I3D framework has a potential future in skill level evaluation for FAST ultrasound scans, and results are favourable enough to warrant further investigation into the issue.

Acknowledgements

This research was enabled in part by support provided by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca).

References

Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD. 2017. A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. IEEE Trans Biomed Eng. 2017. 64(9):2025-2041

American Institute of Ultrasound in Medicine; American College of Emergency Physicians. 2014. AIUM practice guideline for the performance of the focused assessment with sonography for trauma (FAST) examination. *J Ultrasound Med*. 2014;33(11):2047-2056.

Ann Emerg Med. 2017. Ultrasound Guidelines: Emergency, Point-of-Care and Clinical Ultrasound Guidelines in Medicine. Annals of Emergency Medicine. 69(5):27-54

Bell CR, McKaigney C, Ross G, Holden M, Fichtinger G, Rang L. 2017. Sonographic Accuracy as a Novel Tool for Point-of-care Ultrasound Competency Assessment, AEM Education and Training. 1(4):316-324

Bloom BA, Gibbons RC. Focused Assessment with Sonography for Trauma (FAST). StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan- https://www.ncbi.nlm.nih.gov/books/NBK470479/

Carreira J, Zisserman A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:6299–6308.

DeepMind. Kinetics I3D Repository. 2019. https://github.com/deepmind/kinetics-i3d

Funke I, Mees ST, Weitz J, Speidel S. 2019. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J CARS* 14:1217–1225

Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, B´ejar B, Yuh DD, et al. 2014. JHU-ISI Gesture and Skill Assessment Working Set (jigsaws): A surgical activity dataset for human motion modeling. MICCAI Workshop: Modeling and Monitoring of Computer Assisted Interventions (M2CAI).

Hu H, Zhou H. TensorFlow code for finetuning I3D model on UCF101. 2018. https://github.com/USTC-Video-Understanding/I3D_Finetune

Holden MS, Ungi T, McKaigney C, Bell C, Rang L, and Fichtinger G. 2015. Objective Evaluation Of Sonographic Skill In Focussed Assessment With Sonography For Trauma Examinations. CARS 2015—Computer Assisted Radiology and Surgery Proceedings of the 29th International Congress and Exhibition Barcelona.

Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS. 2019. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *Int J CARS* 14:1097–1105.

Mazomenos EB, Bansal K, Martin B, Smith A, Wright S, Stoyanov D. 2018. Automated Performance Assessment in Transoesophageal Echocardiography with Convolutional Neural Networks. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. 11073:256-264 Zach C, Pock T, Bischof H. 2007. A Duality Based Approach for Realtime TV-L^1 Optical Flow. Pattern Recognition. DAGM 2007. 4713:214-223

Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L. 2018. Temporal segment networks for action recognition in videos. IEEE Trans Pattern Anal Mach Intell. 41(11):2740-2755

Zago M, Sforza C, Mariani D, Marconi M, Biloslavo A, La Greca A, Kurihara H, Casamassima A, Bozzo S, Caputo F et al. 2019. Educational impact of hand motion analysis in the evaluation of fast examination skills. Eur J Trauma Emerg Surg. 2019

Ziesmann MT, Park J, Unger B, Kirkpatrick A, Vergis A, Pham C, Kirschner D, Logestty S, Gillman LM. 2015. Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. Journal of Trauma and Acute Care Surgery. 79(4):631-637