

MULTI-MODAL LEARNING FROM VIDEO, EYE TRACKING, AND PUPILLOMETRY FOR OPERATOR SKILL CHARACTERIZATION IN CLINICAL FETAL ULTRASOUND

Harshita Sharma^{*1}

Lior Drukker²

Aris T. Papageorgiou²

J. Alison Noble¹

¹ Institute of Biomedical Engineering, University of Oxford, Oxford, UK

² Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

ABSTRACT

This paper presents a novel multi-modal learning approach for automated skill characterization of obstetric ultrasound operators using heterogeneous spatio-temporal sensory cues, namely, scan video, eye-tracking data, and pupillometric data, acquired in the clinical environment. We address pertinent challenges such as combining heterogeneous, small-scale and variable-length sequential datasets, to learn deep convolutional neural networks in real-world scenarios. We propose spatial encoding for multi-modal analysis using sonography standard plane images, spatial gaze maps, gaze trajectory images, and pupillary response images. We present and compare five multi-modal learning network architectures using late, intermediate, hybrid, and tensor fusion. We build models for the Heart and the Brain scanning tasks, and performance evaluation suggests that multi-modal learning networks outperform uni-modal networks, with the best-performing model achieving accuracies of 82.4% (Brain task) and 76.4% (Heart task) for the operator skill classification problem.

Index Terms— Multi-modal learning, ultrasound, convolutional neural networks, eye tracking, pupillometry.

1. INTRODUCTION

Obstetric ultrasound scanning is recognized as a highly-skilled task, requiring years to master well. Ultrasound operator skill assessment and characterization can form part of initial training on simulators, but has not been studied extensively in the clinic using objective computer-aided methods. In the emerging parallel field of surgical data science [1], such methods have been proposed for surgical skill assessment and evaluation. Our paper explores a similar perspective under sonography data science, where novel multi-modal deep learning networks are designed to automatically classify operator skills using heterogeneous spatio-temporal sensory cues acquired from routine settings, namely, the scan video recording, eye-tracking data, and pupillometric data, in the context of second-trimester fetal ultrasound scanning. The acquired video appearance provides knowledge of ‘what’ and ‘how well’ the operator captures information; the eye tracking enables the understanding of ‘where’ the operator looks, and pupillometry determines ‘how hard’ the operator concentrates during the scanning. Specifically, the quality of the captured standard plane (SP) in the scan video, deter-

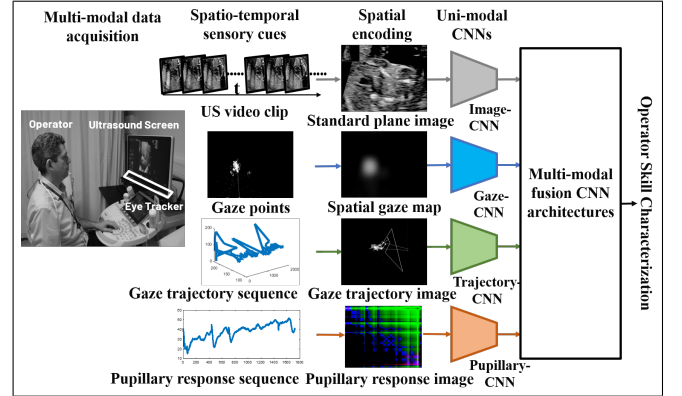


Fig. 1: Overview of the proposed multi-modal learning method for operator skill characterization.

mined by the appearance of certain mandatory anatomical landmarks, is a potential indicator of operator skill. Gaze has been shown to be informative to differentiate the visual expertise and behavior between experts, trainees, and novices in radiology [2], and can be indicative of ultrasound operator skill as well. Pupillometry, the study of eye pupil diameter changes, correlates with cognitive workload [3], and medical professionals with varying skill levels have shown to exhibit distinct pupillary responses, for example, emergency medicine [4] and ultrasonography [5]. Thus, we hypothesize that these novel sensory cues can help discriminate newly qualified and experienced ultrasound imaging operators.

In the literature, the most widely explored modalities for multi-modal learning include images, audio, video, and text [6]. The analysis of heterogeneous multi-modal data acquired in real-world settings, such as that used in our work, presents two unique challenges. Firstly, it is well-known that deep learning models require large-scale data for successful training. However, clinical multi-modal data acquired in specialized acquisition setups is often small-scale. Further, considering gaze and pupillary response as examples of sequential data, pre-trained models in similar domains do not exist, which means deep temporal models (e.g. LSTM, 1D CNN) need to be built from scratch. This may lead to overfitting, or even training failure (as empirically observed for this work). The problem is enhanced by the natural variable length of sequences, also found in our data, which requires length-adjustment such as zero-padding. To address this challenge, we propose spatial encoding of the limited

variable-length sequential data into fixed-sized images, followed by transfer learning on pre-trained image-based CNN models for operator skill characterization. The second challenge is how to combine the sensory cues for end-to-end multi-modal learning. For instance, late fusion has shown success on benchmark datasets (e.g. Kinetics, mini-Sports), but is also associated with overfitting [7]. Late fusion models learn intra-modal dynamics; however, inter-modal interactions can allow information to be exchanged between multi-modal CNN layers. In this work, we implement late fusion to explore intra-modal learning, intermediate fusion to capture inter-modal interactions, and hybrid fusion to combine both benefits. Tensor fusion was introduced to model inter-modal dynamics by explicitly aggregating uni-modal, bi-modal and tri-modal interactions between three modalities [8]. Though tensor fusion is computationally more expensive for four modalities in this work, we investigate two efficient CNN architectures using tensor fusion.

The main contributions of the paper are: 1) We propose a novel and comprehensive multi-modal analysis pipeline using heterogeneous spatio-temporal sensory cues, namely, scan video, eye-tracking data, and pupillometric data, acquired from routine obstetrics ultrasound, for automatic operator skill characterization. 2) We propose methods to encode limited-size and variable-length sequential datasets to enable transfer learning and alleviate problems associated with complex training from scratch. 3) We perform an ablation study of the uni-modal CNN models, and compare five end-to-end multi-modal CNN architectures.

2. METHODS

2.1. Multi-modal Data Acquisition

The multi-modal data came from the Perception Ultrasound by Learning Sonographic Experience (PULSE) study¹. Routine full-length second-trimester ultrasound scan videos were recorded along with synchronized eye tracking of operators [9]. The full-length scan videos were temporally partitioned into variable-length video clips consisting of frames before the first automatically detected freeze until the last sequential freeze frame. The gaze and pupillary response sequences were obtained corresponding to the extracted video clips, using the spatial gaze points (relative x and y coordinates) and pupil diameters with matching timestamps as the output of the eye tracker [10].

We selected 370 scans undertaken by 12 operators for this study. Operators were identified as newly qualified (NQ, 3 operators, ≤ 2 years' experience, 225 scans) or experienced (XP, 9 operators, > 2 years' experience, 145 scans). We selected the 'Brain' and the 'Heart' scanning tasks in the ultrasound scans for further analysis [11], because these tasks require the assessment of the fetal brain using two SPs and

the fetal heart using five SPs, respectively, which can prove challenging for ultrasound operators, thereby allowing differentiation of their skills. These were found to be the most commonly occurring tasks [12, 13], indicating that the operators spent most time on them. A total of 2,309 video clips and sequences (732 Brain, 1,577 Heart) were extracted from the full-length scan videos and eye-tracking data, respectively.

2.2. Spatial Data Encoding

The raw multi-modal data consists of N video clips, and corresponding gaze sequences and pupillary response sequences. The n^{th} video clip with K video frames is $\mathbf{V}_n = [v_n^1, v_n^2, \dots, v_n^K]$, where k^{th} frame is $v_n^k \in \mathbb{R}^{H \times W \times C}$, ($H = 224$, $W = 288$, $C = 3$) after cropping and scaling the relevant imaging area from the ultrasound screen RGB image. From the multi-modal sensory cues, we generate four types of data inputs to train the multi-modal skill characterization models, which are described and spatially encoded as follows.

1. *Standard Plane (SP) Image*: The n^{th} SP image is $S_n = v_n^k$, $S_n \in \mathbb{R}^{H \times W}$, where the frame is randomly extracted from the freeze part of \mathbf{V}_n , as the appearance of the captured image shows negligible variance during the interpretation of the frozen segment. The quality of appearance of the captured SP image is indicative of operator skill.
2. *Spatial Gaze Map*: The n^{th} gaze point sequence is $\mathbf{P}_n = [p_n^1, p_n^2, \dots, p_n^K]$, where $p_n^k \in \mathbb{R}^2$ is the k^{th} gaze point coordinate. Gaze points in \mathbf{P}_n are temporally accumulated to obtain a spatial point map $A_n \in \mathbb{R}^{H \times W}$ and the spatial gaze map $G_n \in \mathbb{R}^{H \times W}$ as defined in Eqn. 2.

$$A_n(x, y) = \begin{cases} 1, & \forall p_n(x, y) \in \mathbf{P}_n. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$G_n(x, y) = A_n(x, y) * K(\sigma) \quad (2)$$

$K(\sigma)$ is a 2D Gaussian Kernel with $\sigma = 1.5^\circ$ visual angle corresponding to the foveal spread of the human eye [10]. G_n encodes spatial aspects of operator gaze in the cumulative dwell time, such as locations of visual attention, number of regions of interest, and spatial dispersion of gaze.

3. *Gaze Trajectory Image*: From \mathbf{P}_n , a weighted line graph with edges $\mathbf{L}_n = [l_n^1, l_n^2, \dots, l_n^{K-1}]$ is generated, where l_n^k represents an edge between successive gaze points p_n^k and p_n^{k+1} . Edge l_n^k is weighted by $w_n^k \in [0, 1]$ given by $w_n^k = 1/(K - k)$. The corresponding gaze trajectory image $T_n \in \mathbb{R}^{H \times W}$ consists of all edges \mathbf{L}_n with weighted gray levels. The weights give higher importance to gaze points traversed relatively later in time, when the operator is interpreting frozen frames. The image encodes spatio-temporal gaze information such as the scan-path pattern and the fraction of cumulative gaze that was fixated (ratio of closer to dispersed edges).
4. *Pupillary Response Image*: The n^{th} pupil diameter sequence is $\mathbf{D}_n = [d_n^1, d_n^2, \dots, d_n^K]$ where $d_n^k \in \mathbb{R}$ is the k^{th} mean pupil diameter of left and right eye. The sequences are first pre-processed according to bespoke

¹Project PULSE, funded by the European Research Council (grant ERC ADG-2015 694581) <https://www.eng.ox.ac.uk/pulse>

guidelines [14] to filter noisy and artifact samples. Then, a task-evoked pupillary response (TEPR) sequence $\mathbf{E}_n = [e_n^1, e_n^2, \dots, e_n^K]$ is computed from the processed \mathbf{D}_n [3].

$$e_n^k = \frac{(d_n^k - d_r)}{d_r} \quad (3)$$

where d_r represents a rest pupil diameter, which is the minimum pupil diameter for a given scan. The TEPR sequence \mathbf{E}_n is encoded into image $R_n \in \mathbb{R}^{H \times W \times C}$ using summation and difference Gramian Angular Fields (GAF) and Markov Transition Field (MTF) [15] as the RGB channels, where the GAF encodes static information, while MTF depicts the dynamics embedded in the raw time-series.

Hence, the resulting n^{th} input data instance is a 4-tuple $I_n = \{S_n, G_n, T_n, R_n\}$. A ResNet-18 CNN architecture [16] was selected for transfer learning *via* fine-tuning of pre-trained models to perform operator skill characterization, as it provides a good balance between computational complexity and classification accuracy. The resulting ensemble of uni-modal CNNs for J data sources is $\mathbf{M} = \{M_j\}_{j=1}^J$.

2.3. Multi-modal Learning

Five multi-modal fusion CNN architectures are explored to learn skill characterization models. Consider feature extractor f_j^X corresponding to layer X of a uni-modal CNN M_j . The five multi-modal CNN architectures are defined as below.

1. *Late fusion CNN (LF-CNN)*: Here, X is the ‘pool5’ (average pool) layer of ResNet-18 CNN. The LF-CNN model consists of J uni-modal feature extractors followed by a fusion layer $f_{fusion}^{X+1} = c(\{f_j^X\}_{j=1}^J)$, where $c()$ represents concatenation operation. Resulting feature vector $F_{fusion}^{X+1} \in \mathbb{R}^{1 \times 1 \times 2048}$ is input to dropout, fully connected and softmax layers. This CNN architecture provides intra-modal interactions before a late fusion during the training.
2. *Intermediate fusion CNN (IF-CNN)*: Here, X is the ‘res3b-relu’ layer (last layer of third convolutional block) of ResNet-18 CNN. The IF-CNN model consists of uni-modal feature extractors $\{f_j^X\}_{j=1}^J$ followed by a fusion layer $f_{fusion}^{X+1} = c_D(\{f_j^X\}_{j=1}^J)$, where $c_D()$ is a depth-concatenation operation. Resulting intermediate feature vector $F_{fusion}^{X+1} \in \mathbb{R}^{28 \times 28 \times 512}$ is input to lightweight randomly initialized CNN layers including a convolution layer ($7 \times 7 \times 512$), batch-normalization and ReLU, followed by global average pooling (GAP), dropout, fully connected and softmax layers. This CNN architecture offers inter-modal interactions during training due to an intermediate feature fusion and depth-concatenation.
3. *Hybrid fusion CNN (HF-CNN)*: In this CNN, fusion layers are first obtained for late and intermediate fusion as $f_{late-fusion}^{X_1+1}$ and $f_{inter-fusion}^{X_2+1}$ where X_1 and X_2 represent CNN layers ‘pool5’ and ‘GAP’ of the two fusion CNN models described above. The following fusion layer is $f_{fusion}^{X_{(1,2)}+1} = c(f_{late-fusion}^{X_1+1}, f_{inter-fusion}^{X_2+1})$. Resulting

hybrid feature vector $F_{fusion}^{X_{(1,2)}+1} \in \mathbb{R}^{1 \times 1 \times 2560}$ is input to dropout, fully connected and softmax layers. This CNN architecture combines the benefits of intra-modal and inter-modal interactions during training.

4. *4-way Tensor fusion CNN (TF-4M-CNN)*: Here, X is the ‘pool5’ layer of ResNet-18 CNN. The TF-4M-CNN model consists of J uni-modal feature extractors, each followed by randomly initialized uni-modal fully connected layers of 32 neurons, and a 4-input fusion layer $f_{fusion}^{X+2} = c_{4T}(\{f_j^{X+1}\}_{j=1}^J)$, where $c_{4T}()$ represents a 4-way tensor fusion operation. The tensor fusion layer computes the outer product between the individual representations [8], and captures uni-modal, bi-modal, tri-modal and quadri-modal dynamics. Resulting 4-D feature tensor $F_{fusion}^{X+2} \in \mathbb{R}^{33 \times 33 \times 33 \times 33}$ is input to GAP, dropout, fully connected and softmax layers.
5. *3-way Tensor fusion CNN (TF-3M-CNN)*: Here, X is the ‘pool5’ layer of ResNet-18 CNN. The TF-3M-CNN model consists of J uni-modal feature extractors, each followed by randomly initialized uni-modal fully connected layers of 16 neurons, and four 3-input tensor fusion layers, each combining three uni-modal inputs, given by $f_{fusion}^{X+2} = c_{3T}(\{f_j^{X+1}\}_{j \in \{1,2,\dots,J-1\}})$, where $c_{3T}()$ represents a 3-way tensor fusion operation. This setting leads to four 3-D tensor cubes, similar to the interactions in [8], representing uni-modal, bi-modal and tri-modal interactions between three modalities. Resulting 3-D feature tensors $F_{fusion}^{X+2} \in \mathbb{R}^{17 \times 17 \times 17}$ are each input to four fully connected layers (16 neurons), followed by concatenation, dropout, fully connected and softmax layers.

3. EXPERIMENTS AND RESULTS

The proposed models were evaluated through five-fold cross-validation experiments for the Brain and the Heart tasks. A scan-wise holdout was implemented in each round of cross-validation. The reported metrics are the mean and standard deviation of the sensitivity (reference: NQ group), specificity, and accuracy for binary classification of operator experience group, computed over the cross-validation rounds. An ablation study was performed to test uni-modal CNN models. The experimental results for the Brain and Heart tasks are presented in Table 1 and Table 2, respectively.

Firstly, we observe that the overall cross-validation performance for the Heart task is slightly lower for most models

Table 1: Experimental results for the Brain task

Uni-modal CNNs	Parameters	Sensitivity	Specificity	Accuracy
Image-CNN	11.18 M	0.81 ± 0.05	0.52 ± 0.10	0.71 ± 0.02
Gaze-CNN	11.18 M	0.78 ± 0.05	0.54 ± 0.06	0.69 ± 0.02
Trajectory-CNN	11.18 M	0.62 ± 0.06	0.78 ± 0.03	0.68 ± 0.03
Pupillary-CNN	11.18 M	0.80 ± 0.06	0.64 ± 0.04	0.74 ± 0.04
Multi-modal CNNs	Parameters	Sensitivity	Specificity	Accuracy
LF-CNN	44.72 M	0.82 ± 0.04	0.79 ± 0.04	0.81 ± 0.02
IF-CNN	15.58 M	0.82 ± 0.03	0.83 ± 0.10	0.82 ± 0.02
HF-CNN	57.57 M	0.83 ± 0.06	0.72 ± 0.07	0.79 ± 0.03
TF-4M-CNN	44.79 M	0.66 ± 0.16	0.69 ± 0.07	0.67 ± 0.08
TF-3M-CNN	45.07 M	0.79 ± 0.04	0.44 ± 0.09	0.67 ± 0.08

Table 2: Experimental results for the Heart task

Uni-modal CNNs	Sensitivity	Specificity	Accuracy
Image-CNN	0.73 \pm 0.04	0.61 \pm 0.04	0.69 \pm 0.03
Gaze-CNN	0.85 \pm 0.08	0.46 \pm 0.11	0.71 \pm 0.02
Trajectory-CNN	0.72 \pm 0.07	0.71 \pm 0.04	0.71 \pm 0.04
Pupillary-CNN	0.77 \pm 0.07	0.61 \pm 0.03	0.71 \pm 0.05
Multi-modal CNNs	Sensitivity	Specificity	Accuracy
LF-CNN	0.74 \pm 0.06	0.74 \pm 0.07	0.73 \pm 0.06
IF-CNN	0.81 \pm 0.06	0.69 \pm 0.06	0.76 \pm 0.04
HF-CNN	0.72 \pm 0.03	0.78 \pm 0.07	0.74 \pm 0.03
TF-4M-CNN	0.62 \pm 0.12	0.64 \pm 0.20	0.63 \pm 0.06
TF-3M-CNN	0.84 \pm 0.05	0.31 \pm 0.06	0.65 \pm 0.05

compared to the Brain task. This shows a higher difficulty to discriminate operator skills from inspection of the heart, possibly due to heart being a smaller structure with higher number of standard planes to find, leading to more complex search and interpretation. Further, for both tasks, uni-modal CNNs achieve good results, suggesting value in each single modality to classify operator skills. There is no clear winner over all metrics for the uni-modal CNNs, as pupillary-CNN and trajectory-CNN are overall more accurate with higher specificities, but image-CNN and gaze-CNN show higher sensitivities in the two tasks respectively. A higher sensitivity can be considered clinically more valuable, as misclassifying a newly qualified operator as an expert can have a more serious consequence than vice-versa. We observe that most multi-modal fusion CNNs outperform the uni-modal CNNs. Intermediate fusion CNN, followed by late fusion CNN, show promising results for the Brain task, and hybrid fusion CNN shows a balanced performance for the Heart task. Lastly, for our classification problem, the two tensor fusion CNNs are not as accurate as the other fusions. Among tensor fusions, the 4-way fusion CNN outperforms the 3-way fusion CNN.

4. CONCLUSION

The paper describes a novel multi-modal learning framework to automatically predict operator expertise using heterogeneous spatio-temporal sensory cues, namely, acquired video, eye-tracking data, and pupillary response, in clinical obstetric ultrasound. Our preliminary findings, including an ablation study, confirm that the fusion models are more accurate compared to models built with single modalities, and we achieve reasonable performance using intermediate, late and hybrid fusion methods, modelling intra- and inter-modal dynamics.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and the ERC ethics committee.

6. ACKNOWLEDGMENTS

We acknowledge the ERC (ERC-ADG-2015 694581 project PULSE), the EPSRC (EP/MO13774/1), and the NIHR Oxford Biomedical Research Centre (BRC). We thank Pierre Chatelain and Richard Droste for their enabling contributions in developing the acquisition system and parameter extraction

tools for video and raw eye-tracker data, respectively. We are not aware of any financial conflicts of interest to be disclosed.

7. REFERENCES

- [1] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, et al., “Surgical data science for next-generation interventions,” *Nat. Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, 2017.
- [2] T. Drew, K. Evans, M. L.-H. Vö, F. L. Jacobson, et al., “Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?,” *Radiographics*, vol. 33, no. 1, pp. 263–274, 2013.
- [3] P. van der Wel and H. van Steenbergen, “Pupil dilation as an index of effort in cognitive control tasks: A review,” *Psychon Bull Rev*, vol. 25, no. 6, pp. 2005–2015, 2018.
- [4] A. Szulewski, D. Kelton, and D. Howes, “Pupillometry as a tool to study expertise in medicine,” *Frontline Learning Research*, vol. 5, no. 3, pp. 55–65, 2017.
- [5] H. Sharma, L. Drukker, R. Droste, P. Chatelain, et al., “OC10.02: Task-evoked pupillary response as an index of cognitive workload of sonologists undertaking fetal ultrasound,” *Ultrasound Obstet. Gynecol.*, vol. 56, no. S1, pp. 28–28, 2020.
- [6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.
- [7] W. Wang, D. Tran, and M. Feiszli, “What Makes Training Multi-Modal Classification Networks Hard?,” in *Proc. IEEE/CVF CVPR 2020*, 2020, pp. 12692–12702, IEEE.
- [8] A. Zadeh, M. Chen, S. Poria, E. Cambria, et al., “Tensor Fusion Network for Multimodal Sentiment Analysis,” in *Proc. EMNLP 2017*, 2017, pp. 1103–1114.
- [9] P. Chatelain, H. Sharma, L. Drukker, A. T. Papageorgiou, et al., “Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies,” *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 153–163, 2020.
- [10] Y. Cai, R. Droste, H. Sharma, P. Chatelain, et al., “Spatio-temporal visual attention modelling of standard biometry plane-finding navigation,” *Med. Image Anal.*, vol. 65, pp. 101762, 2020.
- [11] Y. Wang, R. Droste, J. Jiao, H. Sharma, et al., “Differentiating operator skill during routine fetal ultrasound scanning using probe motion tracking,” in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pp. 180–188. Springer, 2020.
- [12] H. Sharma, R. Droste, P. Chatelain, L. Drukker, et al., “Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans,” in *Proc. IEEE ISBI 2019*, 2019, pp. 987–990.
- [13] H. Sharma, L. Drukker, P. Chatelain, R. Droste, et al., “Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos,” *Med. Image Anal.*, p. 101973, 2021.
- [14] M. E. Kret and E. E. Sjak-Shie, “Preprocessing pupil size data: Guidelines and code,” *Behav. Res. Methods*, vol. 51, no. 3, pp. 1336–1342, 2019.
- [15] Z. Wang and T. Oates, “Imaging time-series to improve classification and imputation,” in *Proc. ICAI’15*, 2015, pp. 3939–45.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE/CVF CVPR 2016*, 2016, pp. 770–778.