Medical Image Analysis 69 (2021) 101973

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Challenge Report

Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos

Harshita Sharma[a,*], Lior Drukker[b], Pierre Chatelain[a], Richard Droste[a], Aris T. Papageorghiou[b], J. Alison Noble[a]

[a] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[b] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 12 January 2020
Revised 18 November 2020
Accepted 11 January 2021
Available online 23 January 2021

Keywords:
Clinical workflow
Fetal ultrasonography
Ultrasound image analysis
Video understanding
Knowledge representation
Skill assessment
Spatio-temporal analysis
Deep learning
Convolutional neural networks

ABSTRACT

Ultrasound is a widely used imaging modality, yet it is well-known that scanning can be highly operator-dependent and difficult to perform, which limits its wider use in clinical practice. The literature on understanding what makes clinical sonography hard to learn and how sonography varies in the field is sparse, restricted to small-scale studies on the effectiveness of ultrasound training schemes, the role of ultrasound simulation in training, and the effect of introducing scanning guidelines and standards on diagnostic image quality. The Big Data era, and the recent and rapid emergence of machine learning as a more mainstream large-scale data analysis technique, presents a fresh opportunity to study sonography in the field at scale for the first time. Large-scale analysis of video recordings of full-length routine fetal ultrasound scans offers the potential to characterise differences between the scanning proficiency of experts and trainees that would be tedious and time-consuming to do manually due to the vast amounts of data. Such research would be informative to better understand operator clinical workflow when conducting ultrasound scans to support skills training, optimise scan times, and inform building better user-machine interfaces.

This paper is to our knowledge the first to address sonography data science, which we consider in the context of second-trimester fetal sonography screening. Specifically, we present a fully-automatic framework to analyse operator clinical workflow solely from full-length routine second-trimester fetal ultrasound scan videos. An ultrasound video dataset containing more than 200 hours of scan recordings was generated for this study. We developed an original deep learning method to temporally segment the ultrasound video into semantically meaningful segments (the video description). The resulting semantic annotation was then used to depict operator clinical workflow (the knowledge representation). Machine learning was applied to the knowledge representation to characterise operator skills and assess operator variability.

For video description, our best-performing deep spatio-temporal network shows favourable results in cross-validation (accuracy: 91.7%), statistical analysis (correlation: 0.98, $p < 0.05$) and retrospective manual validation (accuracy: 76.4%). For knowledge representation of operator clinical workflow, a three-level abstraction scheme consisting of a Subject-specific Timeline Model (STM), Summary of Timeline Features (STF), and an Operator Graph Model (OGM), was introduced that led to a significant decrease in dimensionality and computational complexity compared to raw video data. The workflow representations were learnt to discriminate between operator skills, where a proposed convolutional neural network-based model showed most promising performance (cross-validation accuracy: 98.5%, accuracy on unseen operators: 76.9%). These were further used to derive operator-specific scanning signatures and operator variability in terms of type, order and time distribution of constituent tasks.

* Corresponding author.
E-mail address: harshita.sharma@eng.ox.ac.uk (H. Sharma).

https://doi.org/10.1016/j.media.2021.101973

1. Introduction

Ultrasonography is one of the most widely used medical imaging technologies worldwide and the preferred choice for monitoring pregnancy due to its non-invasiveness, absence of ionising radiation, high accessibility, high reliability, and low costs. The United States Bureau of Labor Statistics predicted that demand for sonographers will increase by 14% between 2018 and 2028 (much faster than average) (Bureau of Labor Statistics, 2019), showing that the patient throughput is outnumbering the rate of sonographer employment and training. Efforts to train new sonographers have seen some success but have not met demand. Ultrasound simulators have emerged as powerful sonography training tools (Gibbs, 2015) but still have limited usage due to high costs and the intrinsic difference to real-life scanning. An alternative solution is to re-design ultrasound (US) imaging technology to be easier to use by trainees and non-specialists, but this is non-trivial to do. In this paper, we go back to first principles and measure what is done in a US clinic (by recording full-length fetal US scan videos), and then automatically analyse operator clinical workflow for skill characterisation and variability assessment. To our knowledge, this paper is the first to consider *Sonography Data Science* which aims to inform US imaging technology design, offer insights into how to use hospital resources efficiently, assist operator training, increase operator efficiency, improve human-computer interfaces with US machines, and determine when and how automated analysis may assist manual scanning. Specifically, this study considers three questions:

1. Can large-scale US video datasets containing hundreds of recorded scan hours be automatically analysed in an accurate and efficient manner for knowledge representation of operator clinical workflow?
2. Can operator skills be characterised by learning from workflow representations of full-length routine second-trimester US scan video recordings?
3. Do US operators have specific scanning signatures and show variability in terms of type, ordering and time distribution of tasks?

We consider these questions in the context of second-trimester fetal US screening. In many countries, a second-trimester (gestational age of 18–22 weeks) US scan is offered to pregnant women for a detailed assessment of the fetal anatomy and growth. For instance, in the UK, the second-trimester scan guidelines are regulated by the National Health Service (NHS) under the Fetal Anomaly Screening Programme (FASP) (Kirwan, 2010). During a full-length routine second-trimester US scanning session, an operator (a sonographer or fetal medicine doctor) views defined fetal anatomical structures including the head and brain, the heart, the abdomen, the limbs, the spine, and additional anatomy such as fetal hands and feet, umbilical cord insertion, and maternal structures such as the uterine arteries. These may be visualised in different viewing planes (e.g. axial, coronal, or sagittal) and US imaging modes (e.g. Two dimensional 2*D* B-mode, colour Doppler, three-dimensional 3*D*, or four-dimensional 4*D*, which is real-time 3*D* mode). Hence, the FASP clinical protocol defines a fixed number of tasks of varying complexity that need to be conducted, but their *order* and their *duration* are not fixed, along with presence of *additional tasks* as preferred by the operator, though there is usually a practical constraint on the full scan time. The operator may also repeat tasks if they choose, or record tasks that were not satisfactorily recorded earlier. These properties may vary due to changing fetal position and fetal movement, poor acoustic windows, operator preferences, and opportunity-grabbing abilities based on operator's skill and experience. Hence, it is interesting to consider if it is feasible to comprehensively analyse and quantify *operator clinical workflow* representing the type, duration, and order of the sequential tasks by retrospectively analysing recorded full-length US scan videos. Clinical workflow analysis solely from full-length US scan videos would require semantic (anatomical) labelling of the scanned events, which, if performed manually would be impractical due to the enormous amount of acquired raw video data. It is known that data annotation is resource intensive and may require medical expertise; hence, we propose automated annotation as the solution. Further, working directly with large-scale video datasets containing hundreds of recorded hours is challenging due to high storage and computational requirements. We address this challenge by formulating a simplified knowledge representation of the operator clinical workflow at three levels, for characterising skill and assessing variability.

Contribution In this study, we analyse operator clinical workflow during full-length routine second-trimester fetal US scans solely based on video recordings. To the best of our knowledge, no previous work has been reported for automatic clinical workflow analysis in obstetric ultrasound. The specific contributions of the paper are as follows:

1. *Semi-automatic semantic annotation.* We propose a semi-automatic semantic video annotation method including video clip extraction and annotation into 23 label categories. This provides the representative labelled dataset for training deep spatio-temporal networks for video description.
2. *Video description.* We automate the temporal semantic segmentation of full-length US scan videos using spatio-temporal deep learning. To reduce the requirement for large-scale labelled video datasets for supervised learning, the proposed deep spatio-temporal network combines spatial features extracted from pre-trained or fine-tuned layers of existing networks previously learnt on large-scale labelled image datasets, and transfer-learns spatio-temporal characteristics from a representative labelled video dataset. We compare several architectures and deploy the best-performing learnt model to classify sequential events in unlabelled full-length US scan videos, and temporally regularise the predicted result.
3. *Knowledge representation of clinical workflow.* We describe an original method for knowledge representation of operator clinical workflow focussing on task type, order, and distribution, which includes three levels of abstraction in decreasing order of dimensionality and complexity. We demonstrate how a complex raw video dataset requiring several Gigabytes for storage and high computational power (e.g. requiring GPU) can be represented by operator clinical workflow, thereby, reducing the dimensionality by a 10^4 order of magnitude to hundreds of Kilobytes, and less computations, for instance, training deep networks with *ca.* 150 times fewer parameters (e.g. feasible on CPU).
4. *Learning for skill characterisation and variability assessment.* We demonstrate the use of the proposed abstractions for learning distinguishing features for operator skill classification and achieve favourable results in differentiating between expert and newly-qualified operators. We obtain the most probable scan-path for each operator yielding operator-specific scanning signatures, for instance, identifying activities when longest and shortest time was spent for each operator. We also reveal intra- and inter-operator variability using the derived knowledge representations.

2. Related technical work

This section describes the related work on automated clinical workflow analysis, and image and video analysis.

2.1. Automated clinical workflow analysis

Automated clinical workflow analysis using data science and machine learning has enabled applications such as decision support, context-aware assistance, and skill assessment predominantly in the field of surgery, sometimes referred to as the emerging discipline of *Surgical Data Science* (Maier-Hein et al., 2017). For instance, earlier works in surgical workflow analysis include automatic generation and visualisation of surgical workflows using Hidden Markov Models (HMM) in laparoscopic cholecystectomy (Blum et al., 2008); development of context-aware operating rooms by modelling and monitoring the workflow of surgical interventions using dynamic time warping and HMM, also in laparoscopic cholecystectomy (Padoy et al., 2012); and automatic workflow segmentation using Markov models and SVM for an unknown sequence of tasks, for tracked needle interventions collected from ultrasound-guided epidural injections and lumbar punctures (Holden et al., 2014). Prediction of the remaining intervention time using patient-individual and generalised surgical process models based on a layered model structure of low-level surgical tasks in discectomies and brain tumor resections is explored (Franke et al., 2013). (März et al., 2015) modelled heterogeneous data comprising of patient-individual, factual, and practical knowledge for surgical decision support in liver surgery. Recent literature explicitly acknowledges the role of artificial intelligence and deep learning for surgical workflow analysis in computer-assisted interventions, using prior knowledge and sensory inputs from the clinical environment (Vercauteren et al., 2020). For instance, deep learning is explored for surgical workflow recognition on laparoscopic videos (Twinanda et al., 2017). Image- and video-based surgical workflow analysis using active learning with Deep Bayesian Networks is addressed in laparoscopy (Bodenstedt et al., 2019).

Lately, objective and computer-aided methods for automated surgical skill assessment and evaluation have been introduced, for example, under the framework of OCASE (Vedula et al., 2017). Related works include the analysis of 3*D* movement trajectories of trainees and experts during birth using forceps delivery training system (Sielhorst et al., 2005); automatic skill assessment in robotic surgery using HMM modelling for surgical gestures (Varadarajan et al., 2009); surgical skill assessment to differentiate between novices and experts in laparoscopic training based on statistical features derived from videos capturing instrument motion (Uemura et al., 2016); description of surgical tool motion trajectories for the classification of gestures and skills in robotic surgery (Ahmidi et al., 2017); and holistic features for automated skill assessment using only robot kinematic data (Zia and Essa, 2018). Ultrasound operator skill assessment and characterisation have not been extensively studied in the clinic using objective computer-aided methods. For example, in fetal ultrasound, the probe motion of operators has been investigated for automatic skill assessment (Wang et al., 2020). However, most of the above work is based on motion tracking (kinematics) or camera-based action analysis in surgical data science. In contrast, in this paper, we analyse operator clinical workflow in fetal ultrasound solely from routine scan video recordings, without the requirement of any motion tracking or camera-based action data.

2.2. Automated image and video analysis

Ultrasound image analysis deals with the automatic extraction of information from ultrasound images and videos. Early work focussed on segmentation (Noble and Boukerroui, 2006), tracking (Sanchez-Ortiz et al., 2000), detection and classification (Yaqub et al., 2015; Chen et al., 2015). Recent work considers application-specific tasks within the context of obstetric ultrasound for standard plane detection (Chen et al., 2017; Baumgartner et al., 2017; Cai et al., 2018; Droste et al., 2019), image quality assessment (Wu et al., 2017a), and fetal biometry measurement and safety assessment (Carneiro et al., 2008; Noble, 2010; Khan et al., 2016; Sinclair et al., 2018). However, such analysis focuses on only the image interpretation task (e.g. detection, classification, and measurement), and does not say anything about operator clinical workflow during real-time ultrasound scanning.

Video classification and activity recognition have been extensively studied in computer vision on public benchmarks (e.g. YouTube videos, sports datasets) (Wu et al., 2017b). Under surgical workflow analysis, a novel convolutional neural network (CNN) architecture called EndoNet is introduced to perform phase recognition and tool presence detection tasks from laparoscopic cholecystectomy videos (Twinanda et al., 2017). Cataract surgical videos are analysed using multilevel statistical modelling for surgical phase or step recognition (Charriére et al., 2017). Another work proposes active learning *via* Deep Bayesian Networks to reduce the large-scale annotated data requirement of machine learning for laparoscopic videos (Bodenstedt et al., 2019). A small number of video analysis studies are explored under ultrasound video analysis. For instance, (Maraci et al., 2017; Gao et al., 2016) focusses on automated annotation of specific anatomical structures such as the heart, the abdomen, and the skull in shorter-length US sweeps (clips). Cai et al. (2020) uses eye-tracking with biometric video sequences to navigate and find standard planes. However, these prior studies consider a limited number of anatomical structures and portions of the scan where the anatomy has already been found by the operator, ignoring information in rest of the full scan where the operator is searching or fine-tuning around a potential anatomy of interest. Recently, we performed a preliminary comparative analysis for classifying constituent video clips in full-length routine second-trimester US scan videos using spatial and spatio-temporal deep neural networks trained from scratch (2*D* CNN, 3*D* CNN, LSTM and convolutional LSTM) (Sharma et al., 2019), and concluded that the spatio-temporal models, specifically 3*D* CNN, outperformed the spatial models. In the current paper, we first perform video description by: 1) exploring more advanced model fusion configurations combining spatial CNNs, spatio-temporal CNNs, and temporal dependency models *i.e.* recurrent neural networks (RNN); 2) using transfer learning approaches such as pre-training and fine-tuning; and 3) temporally regularising the predictions. We achieve a significant improvement over the baseline to reliably automate the temporal semantic segmentation of full-length US scan videos for large-scale clinical workflow analysis. Then, we propose a knowledge representation scheme for operator clinical workflow analysis, and perform skill characterisation and variability analysis in routine fetal ultrasound.

3. Method overview

Fig. 1 presents an outline of the whole clinical workflow analysis pipeline. Video data used in this paper was acquired as part of the Perception Ultrasound by Learning Sonographic Experience (PULSE) study[1]. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051). For the purposes of this study, pregnant women with a singleton pregnancy undergoing pregnancy care at the Oxford University Hospitals NHS Foundation Trust were prospectively enrolled. Written informed consent was given by all participating pregnant women, as well as operators who participated in the study. Data were stored according to approved data governance rules.

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

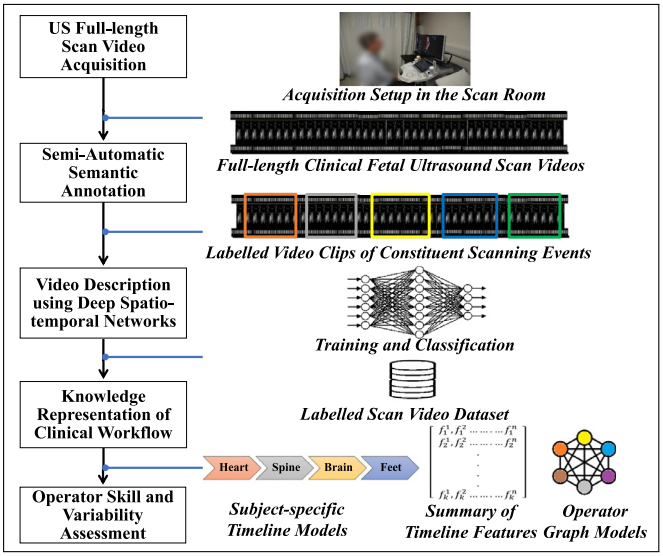

Fig. 1. Fetal ultrasound clinical workflow analysis pipeline.

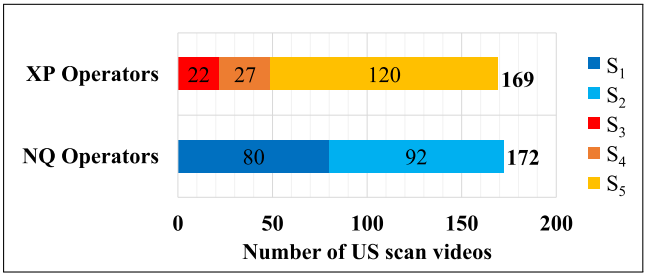

Fig. 2. Data distribution for the individual operators and the two skill groups NQ and XP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Women who agreed were consented to have their full-length routine second-trimester US scan videos recorded. A large-scale dataset of**205 hours** of US scan video recording from 341 entire fetal ultrasound examinations from the same number of women, undertaken by five sonographers or fetal medicine doctors, was analysed. All US scans included in this study were performed using a commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machine. The LCD monitor has a resolution of 1920 $\times$ 1080 pixels and refreshes at a frequency of 60 Hz. The video signal was recorded from the scanner using a lossless compression and sampled at the rate of 30 frames per second (Chatelain et al., 2018). A full-length second-trimester routine examination was on average 36.2 $\pm$ 11.6 minutes in length, with an average of 65,089 frames per scan video.

The US scan videos were acquired by five operators S_1, S_2, S_3, S_4 and S_5 and separated into two groups based on operator experience, namely, data from **newly-qualified (NQ)** operators and **experienced (XP)** operators. The NQ group consists of operators with less than two years of scanning experience (S_1 and S_2), and XP group has operators with more than two years of scanning experience (S_3, S_4 and S_5). The data distribution for the operators is summarised in Fig. 2.

A subset of the full-length US scan videos (62/341 subjects) was manually annotated using automatic clip extraction and visual inspection. The semantically labelled dataset of video clips was used to train spatio-temporal deep neural networks, to automatically derive a **semantic annotation** of sequential events in the remaining unlabelled US scan videos.

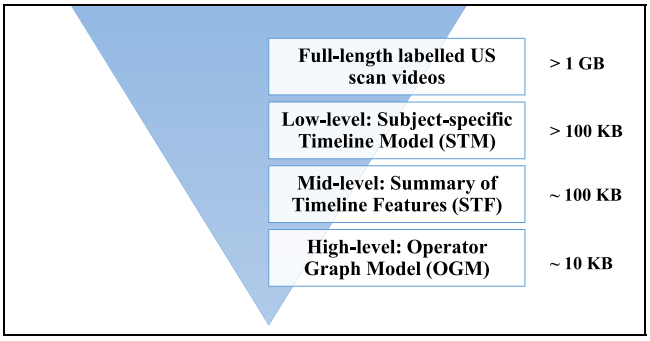

Fig. 3. Proposed knowledge representation scheme for operator clinical workflow analysis.

Having obtained a large number of manually and automatically labelled full-length US scan videos, the next step was to derive a knowledge representation for clinical workflow analysis. The combined large-scale dataset of manually and automatically labelled full-length US scan videos was used in a **three-level knowledge representation** scheme, namely, **low-level subject-specific timeline models, mid-level summary of timeline features**, and **high-level operator graph models**. The scheme, summarised in Fig. 3, illustrates the reduction in data dimensionality at each successive level. The low- and mid-level representations are used to learn skill differences between newly-qualified and experienced operators. The high-level representations are used to analyse operator variability and operator-specific scanning signatures.

Based on the above pipeline, we have divided the remaining paper into four modules (sections) for readability, namely, Video Description (Section 4), Subject-specific Timeline Model (Section 5), Summary of Timeline Features (Section 6), and Operator Graph Model (Section 7). Each section consists of the data, methods, experiments, results and discussion for the corresponding module.

4. Video description

4.1. Semi-automatic semantic annotation

A method of semi-automatic semantic annotation was developed to obtain non-overlapping labelled video clips from full-length US scan videos depicting the individual scanning events based on viewed anatomy. These clips were used for training deep spatio-temporal networks (see Section 4.2). The semi-automatic semantic annotation method is divided into two steps: **video clip extraction** and **manual annotation**.

In the first step, scanning parameters were automatically extracted for each video frame in the full-length US scan video using optical character recognition (Kay, 2007) on the US machine screen. Under these, the 'freeze' state for each frame was automatically detected and recorded as a technical annotation. A video clip is defined with respect to a freeze frame. Frames are typically frozen when a standard plane according to the UK FASP protocol (Kirwan, 2010) is found. Specifically, a video clip corresponds to approximately 5 seconds in time or 151 frames total. 100 frames are selected before a freeze frame and 50 after this frame. The video clip definition is based on the observation that in each *scanning event*, the operator searches an anatomy of interest for a standard view, freezes to perform tasks, and then moves to find the next anatomy. The operator performs, for example, the following activities after freezing:

- Diagnostic inspections (e.g. Heart, Face, Feet).
- Biometric measurements (e.g. Head, Femur, Abdomen).

Table 1
Manual labels for semantic annotation method.

Label name	Abbreviation	Description
3*D* and 4*D* Mode	*3Dm*	Views taken in static or real-time 3*D* mode, showing surface rendering of the fetal head and face.
Abdomen	*Ab*	Fetal abdomen (with biometric measurements).
Arms	*Ar*	Fetal arms.
Background Search	*Bk*	The operator quickly froze-unfroze as they did not finalise the frozen frame as standard view during their search.
Bladder including Doppler	*BlD*	Fetal bladder (including Doppler mode).
Brain with Skull, Head and Neck	*Br*	Fetal brain (with biometric measurements).
Face-side Profile	*Fa*	Side (sagittal) view of the fetal face.
Feet	*Ft*	Fetal feet.
Femur	*Fm*	Fetal femur (with biometric measurements).
Full Body Side Profile	*Fb*	Full-body sagittal views of the fetus. May include face, hands, heart, ribs, spine, diaphragm.
Girl or Boy	*GoB*	Views to determine fetal sex.
Hands	*Ha*	Fetal hands.
Heart including Doppler	*HeD*	Fetal heart (including Doppler mode).
Kidneys	*Ki*	Fetal kidney (including Doppler mode).
Legs	*Le*	Fetal lower legs.
Maternal Anatomy including Doppler	*MaD*	Maternal uterine artery (including Doppler mode).
Mixed	*Mx*	Clip containing views (frames) of more than one label, representing abrupt scene changes.
Nose and Lips	*NL*	Fetal front (coronal) view of the face showing nose or lips or both.
Placenta	*Pl*	Placenta (with biometric measurements).
Situs	*Si*	Situs
Spine	*Sp*	Fetal spine (may be full spine or part of spine).
Top Head with Eyes and Nose	*Th*	Top (axial) view of the fetal head showing eye sockets and/or nose.
Umbilical Cord Insertion	*Um*	Insertion of the umbilical cord.

Fig. 4. Representative example images for each manual label and corresponding abbreviations. *Bk* and *Mx* cannot be illustrated by a single image due to their spatio-temporal characteristics.

- Measurements related to Doppler or Pulse Doppler (e.g. Heart, Maternal uterine artery, Bladder).
- Getting the optimal surface rendering (e.g. 3*D* mode).

A higher number of frames are used before freezing as operators are often refining the view selection (fine-tuning) over this time period. Having frozen the frame, they typically move out of the anatomy quite quickly. Using the video clip definition and extracted 'freeze' states, a full-length US scan video was automatically segmented in time to extract video clips. Additional technical annotations were automatically detected from the screen indicating screensaver, anonymisation, missing probe signal, 2*D*, 3*D* and 4*D* modes for each frame.

After video clip extraction, the extracted video clips were visually inspected and manually annotated. Twenty-three labels were used, as identified by a fetal medicine specialist, as provided in Table 1. Representative example images are shown in Fig. 4. *Bk* and *Mx* labels cannot be illustrated by a single image due to their spatio-temporal characteristics.

Validation of Manual Annotations The semantic manual annotations have been predominantly done by three annotators; two engineering researchers and one fetal medicine specialist qualified in fetal sonography. Before starting the annotation process, the engineering annotators attended routine second-trimester scanning sessions with the medical specialist in the hospital to understand the scan video contents. The majority of the full-length US scan videos annotated by the three annotators are mutually exclusive. In order to evaluate inter-annotator agreement of manual annotations, two to four full-length US scan videos were annotated by multiple (two or three) annotators, and the overlap between annotations of each pair of annotators was calculated. A high average inter-annotator agreement (78.7%) was found between the engineering annotators and the medical specialist. Confusion matrices were computed between each pair of annotators, confirming a high agreement for most labels. For some labels, confusion was higher, for example, 1) hands and arms, as these are anatomically close, 2) hands and face-side profile as sometimes the hand of the fetus is placed close to the face such that both are visible, 3) face-side profile and full-body side profile as these are visually similar. Overall, the validation confirms good agreement between manual annotations. For the small number of overlapping scan videos annotated by multiple annotators, we selected the intersection of the multiple annotations, while giving the highest precedence to the fetal medicine specialist (*i.e.* selecting any additional and uncommon annotations from the specialist). As a mismatch was observed between some anatomical categories due to different interpretations of class definitions (e.g. similar views, multiple visible structures), a possible future direction to address this challenge could be to consider other sensory cues (e.g. eye-tracking data) to determine which labels more accurately represent the video clip, for instance, where the operator is actually looking when multiple anatomical structures are visible.

Clip-level Label Distribution Sixty-two full-length US scan videos were manually labelled for training the deep networks for video description. Only two operators (S_1, S_5) initially acquired the US scan videos on which the manual clip-level annotations were made. Incidentally, S_1 is a newly-qualified operator and S_5 is an ex-

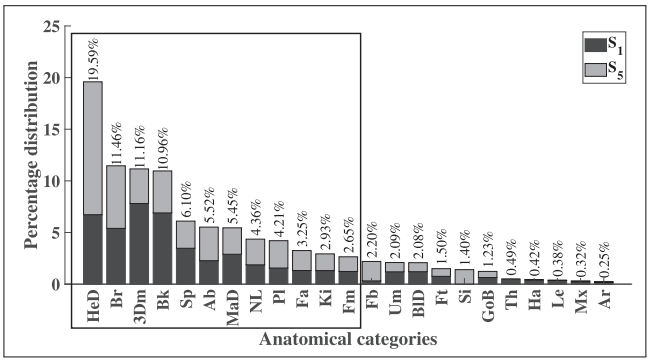

Fig. 5. Percentage distribution of manually labelled video clip dataset among 23 anatomical categories.

perienced operator. Note that the manually labelled dataset for automatic video description is generated purely by visual inspection and used as a training dataset for machine learning of the deep spatio-temporal networks. The real-world dataset was designed to include representative video clip examples from each anatomical category; thus, it will not affect the appearance-based automatic classification in unlabelled videos acquired by other operators.

The percentage distribution of the labelled video clips in the 23 anatomical categories for the manually labelled video clip dataset is depicted in Fig. 5. A high class imbalance can be observed, with the box depicting the top-12 dominant categories, and other categories representing $< 15\%$ of the total annotated frames. These include key FASP-based anatomies and three additional classes with frequency more than other FASP-based anatomies in the dataset, namely, *3Dm, Bk* and *MaD*. Therefore, due to higher representation and relative clinical importance, these 12 classes were selected for training the deep networks. The full-length US scans consist of anatomical structures which can either be one of the primary categories (mandatory according to the scanning protocol) or secondary categories (optional). From the class distribution, it can be observed that some secondary structures were less frequent, or scanned by one of the two operators (e.g. *Si, Ha, Th*). Their imbalanced representation suggests how scanning styles can vary among individual operators in terms of the type and the duration of the scanned anatomical structures during the full-length US scan, which emphasizes the importance to analyse operator clinical workflow in routine fetal ultrasound.

For training the deep learning models, the 62 labelled full-length US scan videos were used to obtain a total of 6,387 event video clips for the 12 dominant categories. These were subject-wise divided into training, validation and test datasets constituting 47 (+ 1), 8 (- 1) and 8 (- 1) full-length US scan videos, respectively, in the initial exploratory stage and the four-fold cross validation stage (randomly selected in both stages). To reduce model complexity during deep learning, the 151-frame video clips were subsampled into 12-frame clips by retaining every eighth frame, where skip length of 8 frames was decided after an empirical evaluation of multiple skip lengths (Sharma et al., 2019). Each longer clip was sampled to obtain five unique shorter clips by initialising a random seed frame followed by uniform sampling. This gave on average 28,425 clips for evaluating the deep learning networks per training experiment.

4.2. Network architectures

We have developed a deep learning network architecture for automatic temporal semantic annotation of the fetal full-length routine second-trimester US scan videos. A deep spatio-temporal network architecture for general video classification (Diba et al., 2018) using a 3*D* DenseNet with a pre-trained 2*D* convolutional neural network (CNN) branch, is used to supervise transfer learning between spatial (2*D*) and spatio-temporal ($2D+t$) branches. Similarly, to overcome the requirement of large-scale labelled video datasets, and to leverage the capabilities of existing image analysis networks in supervised learning, we utilise transfer learning (pre-training and fine-tuning) and model fusion approaches.

Spatial Networks Convolutional neural networks are selected as the spatial (2*D*) representation of individual video frames. Empirically, we found that a SonoNet-64 CNN (Baumgartner et al., 2017) consistently outperformed other existing CNN architectures on US images (spatial data). Therefore, it was selected as the base 2*D* CNN on which to build the spatio-temporal models (Sharma et al., 2019). We use three configurations of the spatial CNN: SonoNet-64 trained on our spatial data from randomly initialised weights *SonoNet-64 (RI)*, SonoNet-64 with pre-trained weights (Baumgartner et al., 2017) *SonoNet-64 (PT)*, and SonoNet-64 fine-tuned on our spatial data *SonoNet-64 (FT)*.

Spatio-temporal Networks To model long-term temporal dependency, recurrent neural networks were investigated. Long-short term memory (LSTM) units have demonstrated effectiveness in video classification *via* recurrent convolutional networks (RCN) (Donahue et al., 2015). LSTM units (Hochreiter and Schmidhuber, 1997) are preferred over vanilla (ungated) RNN units due to their ability to learn long-term dependencies by preventing vanishing or exploding gradients with the help of gating mechanisms. Such mechanisms control how much information from a previous hidden state and input should be used to predict next states, and is achieved by the input gate, forget gate, output gate and memory cell in LSTM units. In this work, the LSTM-based spatio-temporal architecture is called *Sono-2Dt-LRCN (RI)* following the long-term RCN (LRCN) method (Donahue et al., 2015), with consecutive video clip frames jointly learnt by spatial feature extractors and adaptation layers (based on SonoNet-64 CNN), and temporal dependency modelled *via* a recurrent LSTM layer. A single LSTM layer is considered based on previous findings (Soh, 2016) that 1–2 RNN layers are favourable, and more than two layers were found to overfit on large-scale general image data. The weights are trained from random initialisation, as the layer architectures are different from any pre-trained spatial layers, and hence, cannot be used in transfer learning from other models. Recently, convolutional LSTM units were introduced, extending the LSTM to the spatial domain (Xingjian et al., 2015), with potential to address image and video analysis problems. Convolutional LSTM units were earlier studied (Sharma et al., 2019), but are not selected here due to their high computational complexity scaling poorly to limited labelled data, observed slower convergence during training, and negligible performance improvement. To analyse spatio-temporal ($2D+t$) data with end-to-end convolutional neural networks, 3*D* CNNs were employed (Tran et al., 2015). These utilise 3*D* convolutional kernels to learn motion (displacement) patterns between adjacent video frames. The method to convert available 2*D* CNN architectures to 3*D* called *temporal inflation* was previously used on general images (Carreira and Zisserman, 2017). Intuitively, temporal inflation of the spatial layers of the base 2*D* CNN can provide a spatio-temporal representation of the video clips. The resulting $2D+t$ spatio-temporal CNN architecture is called *Sono-2Dt-CNN (RI)*. The weights are trained from random initialisation, with the same reason as the LRCN counterpart.

Model Fusion The main challenge for training spatio-temporal models is the requirement of a large-scale labelled video dataset. However, the knowledge learnt in spatial architectures can be utilised to transfer the learnt knowledge from pre-trained or fine-tuned 2*D* CNNs to a spatio-temporal network, which can be more effective to train with a limited number of labelled samples. Hence, fusion configurations of fixed spatial feature extractors and train-

Table 2
Fused spatio-temporal model configurations.

Model Configuration	Fusion Components
Sono-2D(PT)-2Dt-LRCN(RI)	Pre-trained *SonoNet-64(PT)* and randomly initialised *Sono-2Dt-LRCN(RI)*.
Sono-2D(FT)-2Dt-LRCN(RI)	Fine-tuned *SonoNet-64(FT)* and randomly initialised *Sono-2Dt-LRCN(RI)*.
Sono-2D(PT)-2Dt-CNN(RI)	Pre-trained *SonoNet-64(PT)* and randomly initialised *Sono-2Dt-CNN(RI)*.
Sono-2D(FT)-2Dt-CNN(RI)	Fine-tuned *SonoNet-64(FT)* and randomly initialised *Sono-2Dt-CNN(RI)*.
Sono-2D(PT)-2Dt-LRCN-CNN(RI)	Pre-trained *SonoNet-64(PT)*, randomly initialised *Sono-2Dt-LRCN(RI)* and *Sono-2Dt-CNN(RI)*.
Sono-2D(FT)-2Dt-LRCN-CNN(RI)	Fine-tuned *SonoNet-64(FT)*, randomly initialised *Sono-2Dt-LRCN(RI)* and *Sono-2Dt-CNN(RI)*.

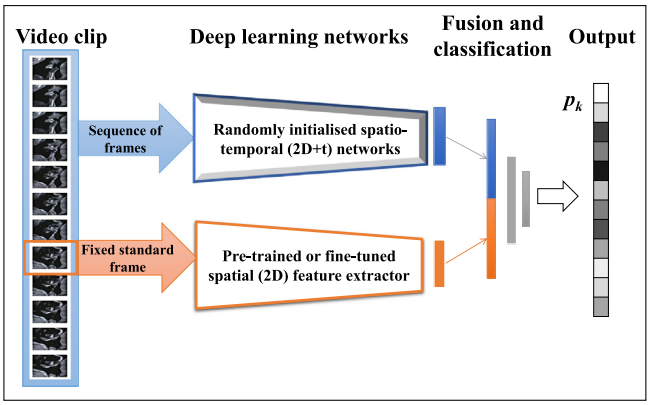

Fig. 6. Model fusion of spatial feature extractor and spatio-temporal networks. Here, one spatio-temporal network is shown but may exceed to more networks, as tested in our experiments.

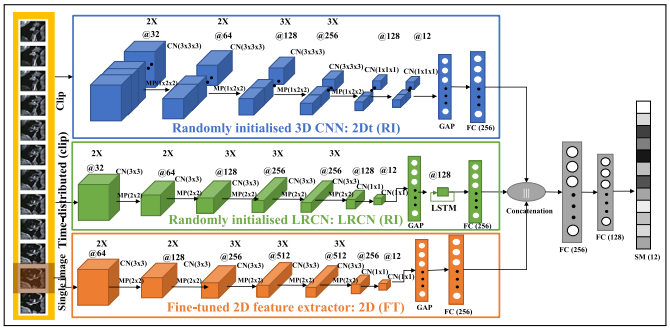

Fig. 7. Selected spatio-temporal deep network. (CN: 2*D* or 3*D* convolution layer, MP: 2*D* or 3*D* max-pooling layer, GAP: 2*D* or 3*D* global average pooling layer, FC: fully connected layer, LSTM: LSTM unit, SM: softmax layer).

able spatio-temporal networks are evaluated in the model fusion architectures.

In general, model fusion is performed as follows. A video clip is denoted as vc_i and a fixed standard plane image (frame) from the clip as I_i. Let f_{im} and f_{vc} be the feature extractors for images and video clips respectively, we consider networks of the form,

$$f_{fus}(vc_i) = fusion(f_{im}(I_i), f_{vc}(vc_i)) \quad (1)$$

such that $f_{vc} \in \{f_{LRCN}, f_{2DtCNN}\}$. f_{fus} is the network combining the spatial and spatio-temporal networks. Concatenation is found to be the most successful fusion method compared to other methods such as addition (Huang et al., 2017). Hence, in the fusion layer, fully connected layer outputs of the two or more branches are concatenated, followed by more fully connected layers.

The fusion method is illustrated in Fig. 6. Specifically, the base SonoNet-64 CNN pre-trained or fine-tuned models *SonoNet-64(PT)* or *SonoNet-64(FT)* are used in a merged configuration with the spatio-temporal architectures *Sono-2Dt-LRCN(RI)* and *Sono-2Dt-CNN(RI)* to give six fused model configurations as given in Table 2.

In all the fusion configurations, weights of the spatial layers are fixed (pre-trained or fine-tuned) to obtain 2*D* features for a fixed standard frame in each video clip whereas the spatio-temporal branch is randomly initialised for training. The spatial branch could also be made trainable. However, this would greatly increase the model complexity, which is not desirable with a fixed computational budget ($< 25M$ trainable parameters based on computational hardware). In the last two fusion architectures involving both the spatio-temporal LRCN and CNN networks, the number of channels has been reduced by half in each layer of the spatio-temporal branches compared to highly computationally expensive models if full constituent layers are used. In this way, model fusion architectures help in combining representations from spatial layers trained on large-scale datasets, and the spatio-temporal layers trained on the acquired dataset.

A preliminary analysis of the individual spatial (2*D*) and spatio-temporal ($2D+t$) features was previously reported for US video clip classification (Sharma et al., 2019). However, the individual network architectures were randomly initialised without any pre-training or fine-tuning. We extended the analysis here, with an ablation study of the individual features (spatial: 2*D* CNN, spatio-temporal: $2D+t$ LRCN, and spatio-temporal: 3*D* CNN), followed by the design and comparison of several model fusion configurations utilising transfer learning (pre-training and fine-tuning). This exploratory analysis comparing the spatial and spatio-temporal deep networks for video description is discussed in detail in the **Supplementary Material**, where the network architectures are explained in Section 1. The comparison results in Section 2 show that the fusion CNN architectures outperform the individual networks, and the best-performing CNN architecture *Sono-2D (FT)-2Dt-LRCN-CNN (RI)* presents a significant improvement over the other model fusion configurations.

We selected the deep network *Sono-2D (FT)-2Dt-LRCN-CNN (RI)* for automatic annotation due to most promising classification results. The network architecture of *Sono-2D (FT)-2Dt-LRCN-CNN (RI)* is shown in Fig. 7. In our evaluation, this network consistently outperformed the other spatial and spatio-temporal architectures that were considered. Two reasons for the success of this architecture are the combination of a fine-tuned spatial CNN with randomly initialised spatio-temporal layers, and a comparatively less complex model that scales better to the training data under a fixed computational budget.

4.3. Training and classification

Focal Loss is used as the loss function in deep network training, to address class imbalance of our datasets. The focal loss L_f is given by (Lin et al., 2017),

$$L_f(p_k) = -(1-p_k)^{\gamma} \log(p_k) \quad (2)$$

A modulating factor $(1-p_k)^{\gamma}$ is introduced to the standard cross-entropy loss $-log(p_k)$, where p_k is the softmax probability of class k. For a $\gamma > 0$, the relative loss for well-classified or easy examples is reduced, and harder (misclassified) examples are more focussed during training. It has been shown that focal loss works well for training object detectors in the presence of several background

classes (Lin et al., 2017). After empirical evaluation, the value of γ was set to 2.

All frames were pre-processed by cropping the relevant image area to 224 × 288 pixels and resizing to 224 × 224 on-the-fly during training. In addition, all frames were normalised to zero-mean and unit-variance. Image augmentation was consistently applied to all frames of a clip, including rotation with angle randomly sampled from $[-30°, 30°]$, flipping, random Gaussian noise (σ=0.01), and shear (≤ 0.2). Regularisation was achieved using batch normalisation and dropout (p_d=0.5). Adaptive Moment Estimation (Adam) was used for optimisation, with initial learning rate of 10^{-4} and decay 10^{-6}. Batch size was varied between 8, 16 or 32 depending on GPU memory availability for the particular model. All models were trained for 100 epochs (200 for spatial models) and a checkpoint was created for the lowest validation error. During the classification stage, the checkpoint weight profile was used to classify unlabelled video clips in unseen full-length US scan videos.

The networks were trained for the top-12 occurring classes in the manually labelled datasets, due to wide class imbalance (see Section 4.1). These classes include *HeD, Br, 3Dm, Bk, Sp, Ab, MaD, NL, Pl, Fa, Ki*, and *Fm*. These also represent the most anatomically relevant classes according to the FASP protocol. As a result, the networks predict these 12 classes in the unseen full-length US scan videos. The other 11 classes are represented by a cluster called 'Other Anatomical Classes' with abbreviation *Oth*. Intuitively, the *Oth* cluster is not compact as it contains multiple constituent subclasses with variable appearances that will lead to high confusion during training if used as a single class. Hence, the deep networks were not trained for classifying a clip as *Oth* cluster in the first instance, but this was addressed by post-processing the classification results of multiple networks. For each video clip, the majority vote of N networks ($N = 4$ in our case, as we use four cross-validation models) is calculated from the softmax probability p_k^i for a class k and network i as

$$C_i = \arg\max_k (p_k^i) \tag{3}$$

$$C = \begin{cases} \text{mode}\{C_i\}_{i=1}^N, & \text{if } freq(\text{mode}\{C_i\}) \geq 2 \\ Oth, & \text{otherwise} \end{cases} \tag{4}$$

where $freq()$ represents the frequency, C_i is the classifier result of the i^{th} network, and C is the final result after majority voting for a video clip. Majority voting was applied to capture the uncertainty of the individual networks in the ensemble of trained networks, where, if majority of the trained networks are uncertain about the label of the video clip, the label is considered as the *Oth* cluster.

4.4. *Post-processing* via *temporal regularisation*

After automatic classification of the video clips constituting the full-length US scan videos, temporal information and posterior (softmax) classification probability scores of the video clips were used to regularise the classification results and smoothen temporal over-segmentation for each US scan video in a post-processing stage (Maraci et al., 2017). This was performed by constructing a conditional random fields (CRF) graphical model (Lafferty et al., 2001), with each video clip in the US scan video as one node of the graph.

The joint probability of assignment to the node vc_j in the graph with J number of nodes and K number of edges is defined as the normalised product of two non-negative potentials as

$$P(vc_1, vc_2.vc_J) = \frac{1}{M} \prod_{j=1}^{J} \psi_j(vc_j) \prod_{k=1}^{K} \psi_k(vc_{k_a}, vc_{k_b}) \tag{5}$$

where unary (node) potential $\psi_j()$ is the posterior classification probability score, and binary (edge) potential $\psi_k()$ is the probability of a node transitioning from one state to another, where (vc_{k_a}, vc_{k_b}) represents an edge between nodes a and b. The binary potentials were empirically set by computing a transition matrix from the training (manually labelled) US scans. The normalisation constant M ensured that distribution sums to one over all possible joint configurations of variables. The most probable label path was obtained in each test (automatically labelled) scan using Viterbi decoding (Forney, 1973). Hence, this setting smoothed out the classification result by considering neighbouring video clips in the full-length US scan video. The combined dataset of manually labelled and post-processed automatically labelled full-length US scan videos was used for large-scale clinical workflow analysis, as described Section 5 onwards.

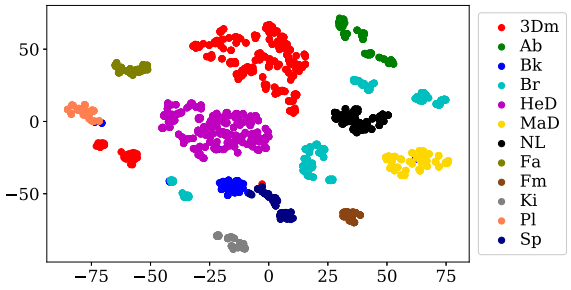

Fig. 8. t-SNE Feature visualisation of the penultimate layer of *Sono-2D (FT)-2Dt-LRCN-CNN (RI)* model for test dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5. *Results and discussion*

The comparative evaluation and discussion of the different tested networks are reported in Section 2 of the **Supplementary Material**. Here, we discuss the results of the best-performing video description network *Sono-2D (FT)-2Dt-LRCN-CNN (RI)*. The spatio-temporal network was tested using four-fold cross validation. Due to limited availability of the manually labelled video datasets, we evaluated the trained network on unseen data, where the deep network trained on 62 labelled videos was used to infer clip-level labels for the remaining 279 unlabelled full-length US scan videos using the methods in Section 4.3 and Section 4.4. The quality of the automatic labels was assessed using statistical analysis and retrospective manual validation.

Cross Validation Standard multi-class classification evaluation metrics are used to evaluate performance the trained deep video description network, namely, Precision (P), Recall (R), F1-score ($F1$), Top-1 accuracy (A_1) and Top-3 accuracy (A_3). Mean and standard deviations of evaluation metrics were obtained as $P = 0.88 \pm 0.10$, $R = 0.88 \pm 0.09$, $F1 = 0.88 \pm 0.10$, $A_1 = 0.92 \pm 0.08$, $A_3 = 0.98 \pm 0.02$. This result establishes the suitability of the selected model to solve the video clip classification problem for clinical fetal US scan videos. The misclassification is highest for the classes *Bk, Pl* and *Ki*. *Bk* includes searching or plane-finding that may contain multiple structures; hence, the class has higher confusion probability with the other anatomical categories. *Pl* and *Ki* are misclassified due to a lower number of labelled instances in the training dataset. In particular, there is a higher binary confusion between *Pl* and *MaD*, and *Ki* and *HeD*, due to their similar appearance.

Feature visualisations from the penultimate layer of this model using t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) are depicted in Fig. 8. The clusters mostly show a clear distinction between categories and the extracted CNN features are representative of the anatomical classes. We can observe a small number of outliers which are the misclassifications discussed earlier.

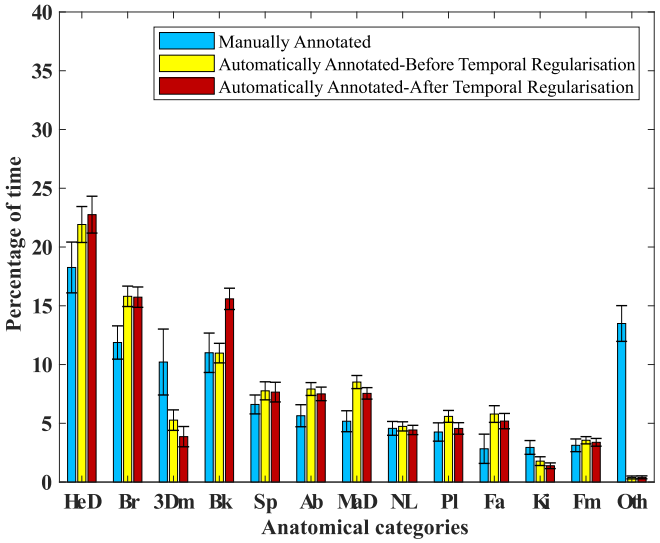

Fig. 9. Percentage mean statistical distribution of manually labelled US scans, automatically labelled US scans before post-processing for temporal regularisation, and automatically labelled US scans after post-processing for temporal regularisation. Error bars represent 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

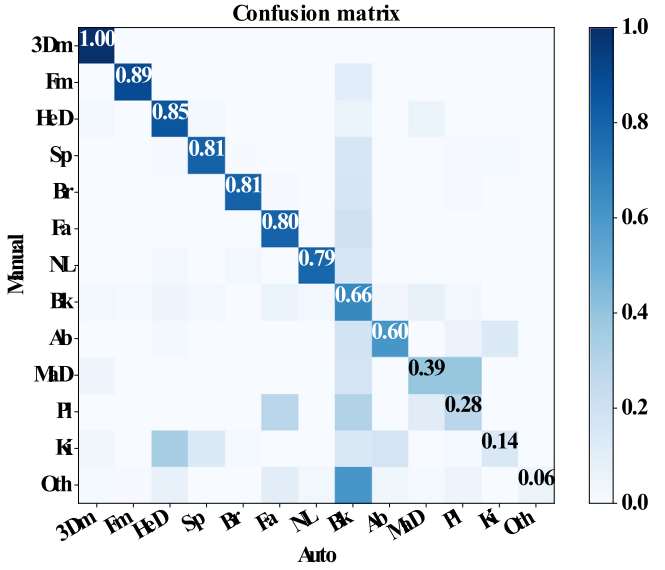

Fig. 10. Confusion matrix between manual and automatic labels for 10% randomly selected US scans for retrospective manual validation.

Statistical Analysis Statistics were computed as mean ± standard deviation (duration in minutes) and mean percentage (95% confidence interval) per scan for each of the $n = 13$ given anatomical categories. The automatically labelled US scans were validated against manually labelled data by computing the correlation between the anatomical class histograms of manually and automatically labelled US scans. A high correlation with Pearson's correlation coefficient $\rho = 0.98$, ($p = 2.83 \times 10^{-10}$) was observed. Fig. 9 depicts the percentage statistical distribution of manually labelled US scans, automatically labelled US scans before post-processing for temporal regularisation, and automatically labelled US scans after post-processing for temporal regularisation. This was calculated as a percentage of the whole scan.

We observe that there is a one-to-one correspondence between occurrence of most event classes for manual and automatic labels, and for the automatic labels before and after temporal regularisation. However, the highest mismatch is in *Oth* (other cluster), as the video clips in *Oth* cluster have a lower proportion in the automatically labelled US scans compared to manually labelled ones. The main reason for this behaviour is that, the *Oth* cluster consists of multiple constituent sub-classes with variable appearances, and the deep networks were not trained to recognise their appearance as a single class, thus, video clips belonging to *Oth* cluster were most likely classified as one of the remaining trained classes. Moreover, there was an overestimation of background search class *Bk* after temporal regularisation due to a higher number of transitions in the full-length scans from or to *Bk* clips. Furthermore, it is interesting to see that for both manually and automatically labelled datasets, the highest percentage durations observed are for anatomies *HeD* (Heart including Doppler) and *Br* (Brain), which is expected, as Heart and Brain are important anatomical structures for inspection and measurements during the fetal US scan, so the operators have spent the longest part of their scan time on these anatomical tasks.

Retrospective Manual Validation From the 279 unseen and automatically labelled US scans, a random 10% (28 US scans) were manually annotated. After comparative evaluation of the manual and automatic labels, the total accuracy for all US scans was computed as 0.76, with an average accuracy per scan as 0.75 ± 0.08. The confusion matrix between manual and automatic labels is shown in Fig. 10. The highest confusion is found in *Oth* cluster being classified as Background Search (*Bk*). This is because *Bk* can include multiple anatomies, so the sub-classes in the *Oth* cluster, which were not recognised by the deep networks based on their appearance, were probably classified as *Bk*. We believe the solution is to use all the sub-classes in *Oth* cluster separately during training. However, this would require acquisition of a significant amount of data. Currently, the data distribution in sub-classes of the *Oth* cluster is inadequate to train the deep networks with all the 23 anatomical classes (Fig. 5). Also, high confusion exists between *Ki* and *HeD*, as well as between *Pl* and *MaD* due to their similar appearance.

While we observe a lower accuracy of *MaD, Pl, Ki* and *Oth* in the automatically labelled US scans, these labels are included in clinical workflow analysis for two reasons. Firstly, most of these classes have lower number of instances in the manually labelled dataset, which was also the reason for a lower accuracy in the automatic labels. Secondly, the lower accuracy is because the clips were detected as other types of classes (e.g. *Ki* mostly detected as *HeD, Oth* as *Bk*). But there are negligible false positives for these classes themselves, which means, the clips classified into these classes have mostly correct labels, which makes the analysis of these clips useful. Hence, any additionally labelled samples are valuable and would increment the knowledge of the labelled data pool.

5. Subject-specific timeline model

In this section we begin to form representations of operator clinical workflow. We start by describing the subject-specific timeline model.

5.1. Method

At the lowest level of abstraction, we generate a **Subject-specific Timeline Model (STM)** representing operator clinical workflow for the full-length US scan video of each subject. In this context, clinical workflow can be defined as the fine-grain representation of a subject's US scan video consisting of distinguished successive scanning events as a time sequence. Thus, the STM of a scan is the ordered set of K scanning events $\{E_1, E_2, ., E_K\}$, where each scanning event E_i ($i \in \mathbb{N}$) is a video clip represented

by its numerically coded anatomical label $L_i \in [1, 2, ., n]$, where n is the number of unique anatomical tasks (label classes). The STM is computed by ignoring all non-anatomical frames (labelled as anonymised, screensaver and missing probe signal after automatic detection), and considering the top-12 anatomical classes and the *Oth* cluster ($n = 13$). The non-anatomical frames are removed as these do not constitute the operator's visual scanning experience, and usually reflects their time spent on, for example, asking subject details (labelled as anonymised), doing other administrative jobs or conversing with the subject (labelled as missing probe signal and screensaver), which do not constitute the anatomically relevant events of the US scan. The process of creating an STM from a US scan video involves two additional steps.

Firstly, all the frames of the US video were not labelled manually or automatically, as video clip extraction follows a strict protocol of detecting 'freeze' frames and partitioning fixed-length clips around these frames. Thus, there are gaps of unlabelled frames in the video. In general, we know that after freezing, the scan is focussed on a particular anatomy, and then the operator quickly moves out of the anatomy to find the next anatomy. Therefore, to handle unlabelled frames, we post-process the STM by two sequential steps, namely, 1) annotating subsequent unlabelled 'freeze' frames directly after the clip with the clip label, since these would be the same anatomical task; and 2) labelling unlabelled frames between two scanning events with the label of the next sequential scanning event- this is a *backward filling extrapolation* operation.

Secondly, there are a high number of background search (*Bk*) clips in the labelled videos (we observe and discuss reasons for this in Section 4.5.0.2). To reduce the overestimation of *Bk*, we assume the following: if a scanning event (or series of scanning events) with the label *Bk* is encountered between two (series of) scanning events of the same anatomical task, *i.e.*, with the same numerical code, and has a shorter runlength (total number of successive events of the same task) than the neighbouring scanning events, it is replaced by the label of the neighbouring scanning events. This assumption is intuitive, as the minor *Bk* task located in between the same major task on both sides is most likely a part of the major task.

An example of a typical STM with corresponding video frames, and the color-coded STMs for operator S_4 are shown in Figs. 11(a) and 11(b) respectively. An STM is represented similar to a multi-colored bar code, where a color corresponds to an anatomical category. For visualisation purposes, in Fig. 11(b), the total duration of US scans is normalised to one, and all STMs are arranged in the increasing order of scan duration.

5.2. Observation and discussion

In a standalone way, the STMs provide a shorthand representation of each full-length US scan video for each operator. This representation can be utilised for analysing operator clinical workflow to classify scanning skills and analyse variabilities. We can make visual observations by only looking at the color-coded STMs of each operator, such as patterns of the most prominent tasks, task distributions, and task ordering. For example, from Fig. 11(b), we can conclude that the operator prefers to look at the brain (*Br*) and nose-lips (*NL*) in the beginning of the scan, and mother's anatomy (*MaD*) and face-side profile (*Fa*) during the latter part of the scan.

6. Summary of timeline features

Using the foundations of Section 5, we now consider how to derive a quantitative mid-level representation of assessing clinical workflow.

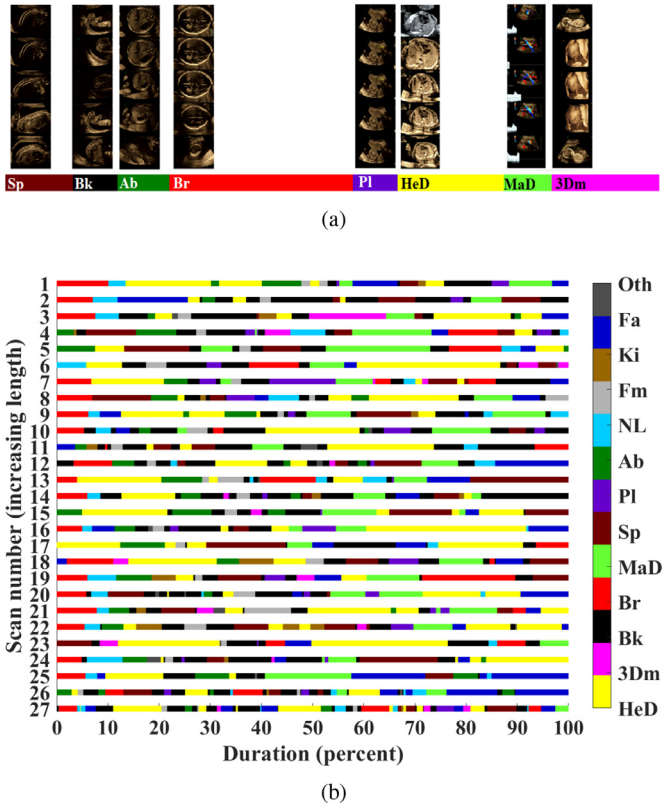

Fig. 11. Examples of (a) a typical STM with corresponding video frames for each task (b) STMs for the operator S_4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.1. Method

From the literature, we can summarise the three dominant categories of methods for knowledge representation of recorded sequential clinical data as *time-series models* (e.g. Markov models and HMM) (Oropesa et al., 2011), *summary of features models* (e.g. feature analysis with SVM, LDA, neural networks) (Horeman et al., 2014), and *formalised models* based on dictionaries (e.g. UML, flowcharts, petrinets) (Basu and Blanning, 2000). The performance of these three categories is compared in (Vedula et al., 2017) for several surgical skill assessment studies, and superior performance is reported for summary of features models. Intuitively, and based on this evidence, in this paper we explore hand-engineered features with classical machine learning for clinical workflow analysis of full-length fetal US scan videos. Advantages of deriving features include a dimensionality reduction to a fixed number of features compared to time-series of varying length, and a simpler approach compared to the formalised models which may involve the use of specialised data management tools. We carefully engineer the features for clinical workflow analysis based on operator skills in routine fetal US scans. State-of-the-art surgical workflow features (e.g. described in (Vedula et al., 2017)) were not included due to the inherent difference between a fetal US scanning workflow and a surgical workflow. Specifically, a fetal US scan consists of arbitrary task types, distributions, and ordering with repetitions, whereas, a surgical workflow usually has a defined number and order of tasks without repetitions.

Considering the STMs (introduced in Section 5) as time-series, we also evaluate HMMs and deep learning methods directly on the STMs. Furthermore, OGMs (introduced in Section 7) can be considered analogous to a formalised model of the sequential data. Therefore, we capture the three categories of description methods

through our knowledge representation scheme consisting of STMs, STFs and OGMs, respectively.

6.1.1. Feature engineering

The mid-level representation of the STMs is achieved by designing and extracting discriminating**Summary of Timeline Features** from each STM that relate to operator skill. Discussions with specialists led to hypothesise possible differences between the skill groups XP and NQ to be the following.

1. The XP operators are more familiar with knowing where to look next for an interesting anatomy compared to the NQ operators. Hence, we expect that the XP group will have shorter *search duration* between anatomies than the NQ group.
2. The XP operators should be quicker to recognise an opportunity than the NQ operators. For instance, if they see a certain view within an anatomy, then they should get a good standard plane. This would lead to smaller *task durations* and less *task repetitions* of the individual anatomical tasks for the XP group compared to the NQ group.
3. The XP operators are expected to conduct exams in a more structured way compared to the NQ operators. In other words, the XP group will follow a specific *task ordering* more frequently (e.g. head first, face second, spine third, and so on), with less *task transitions*.
4. The XP operators are expected to have a faster acceptance of their momentary inability to acquire a certain standard plane and proceed to the next anatomy more quickly. For instance, an XP operator would try to find the nose-lips plane for a few seconds but is more likely to recognise to abandon the task if the fetal position is unsuitable. In contrast, an NQ operator might not appreciate this and continue to try to find a good imaging plane for longer. In this case, the XP group will have a shorter *fine-tuning duration* within the anatomy.

We capture the above properties quantitatively from the STMs by the following Summary of Timeline Features (STFs). Firstly, the total duration (**point 1.**) of the US scan is an important discriminating feature, which is computed as the length of the STM, normalised to the range [0,1] for a specific operator. The normalisation is considered important, as otherwise this value can be highly variable depending on different operator preferences and styles. Intuitively, the **normalised scan duration** depends on the operator skills and position of the fetus. We have ignored fetal position variations assuming a normal fetal position, hence, the normalised scan duration can be a good indicator of operator skills.

The **relative anatomical task durations (point 2.)**, including the total, mean, maximum and minimum duration of each of the 13 anatomical tasks are computed for each STM, giving a total of 13×4=52 features.

Task ordering **(point 3.)**, as the extent of how organised the STMs are, is calculated computing the **Shannon entropy** of each STM. Since, an STM is a stream of numerically-coded labels, using information theory, Shannon entropy (Shannon, 1948) can be utilised to determine the diversity and redundancy of each STM. Entropies are normalised to the range [0,1] before feature selection and machine learning.

A feature describing the type of search where the operator performs quick freeze-unfreeze but doesn't finalise the view is computed using the relative duration of *Bk* task, was already considered earlier. However, to more effectively capture the search durations between anatomical tasks, fine-tuning durations within anatomical tasks, and durations of activities (e.g. diagnostic inspection, biometric measurements) performed after finding the standard plane **(point 1., point 4.)**, we consider the **relative non-freeze and freeze durations**, respectively, for each anatomical task in the STM. For this purpose, we create a vector of the same length as the number of frames in the scan video, containing the corresponding freeze states of frames in each of the scanning events from the extracted technical annotations. The relative durations of activities after finding the standard plane (freeze state=1), search and fine-tuning (freeze state=0) for each anatomical task is calculated, giving rise to 2×13=26 additional features.

We use a directed graph-representation of the STM, where each unique anatomical task L_i is represented by a node, and transitions from task L_i to task L_j by a directed edge $i \rightarrow j$. We count the loops for each node, representing the number of repetitions of each anatomical task **(point 2.)**, leading to 13 more features. We also count the in-degree and out-degree of each node to determine the number of transitions to and from each anatomical task, respectively **(point 3.)**, giving a total of 13×2=26 more features.

As a result, we have defined a total of 119 timeline features of the STM that describe operator skill.

6.1.2. Feature selection

To identify the most discriminative features, we perform feature ranking and visualisation by using two filter methods, namely, **ReliefF** (Robnik-Šikonja and Kononenko, 2003) and **Neighbourhood Component Analysis** (NCA) (Yang et al., 2012). We select filter methods over wrapper or embedded methods (Liu and Motoda, 2012) for feature selection for the following reasons. Firstly, filter methods depend on the general characteristics of data (for example, statistical tests for correlations with the output variable), without depending on any chosen machine learning algorithm unlike the other two types of methods. Hence, a filter method can be used as a pre-processing step with any machine learning algorithm. Secondly, filter methods are computationally less expensive as they do not involve model training. Thirdly, filter methods are not prone to over-fitting compared to the other methods. A recognised weakness of filter methods is that there is a lower possibility to find the optimal subset of features (Kohavi and John, 1997). To address this, we use the common results of two filter methods to ensure the selected features are most discriminative for the given dataset.

6.1.3. Machine learning

The hand-engineered features (STFs) are used within traditional feature-based machine learning methods, namely, support vector machines (SVMs) and random forests using (i) all the STFs and (ii) the selected STFs, to build a model for skill classification, distinguishing between the two skill groups. The SVM models are trained using radial basis function (RBF) kernels. Random forest classifiers are trained after empirically setting the number of trees to 300 with a minimum leaf size to 6. For completeness of the analysis, we also report the results from two basic prediction models, a Hidden Markov Model (HMM) trained directly on the STMs, and a logistic regression model trained on all the STFs.

To investigate the usefulness of the low-level STM representation for operator skill characterisation, we perform deep learning directly on the STMs using convolutional neural networks (CNN) based architectures. For this purpose, we adapt ideas that have been used in deep learning for genomics, specifically DNA sequence classification (Yue and Wang, 2018), due to similar ordinal characteristics of DNA sequences to our STMs. Particularly, motivated by a CNN architecture called Viraminer CNN (Tampuu et al., 2019) that uses a two-branch 1*D* CNN, namely, a pattern branch with global max pooling and frequency branch with global average pooling, we propose a three-branch 1*D* CNN architecture that we call *SonoSkillClassifier CNN*, designed to learn the STM characteristics such as task length, scanning pattern, transitions and repetitions. The CNN architecture is defined using the vocabulary for

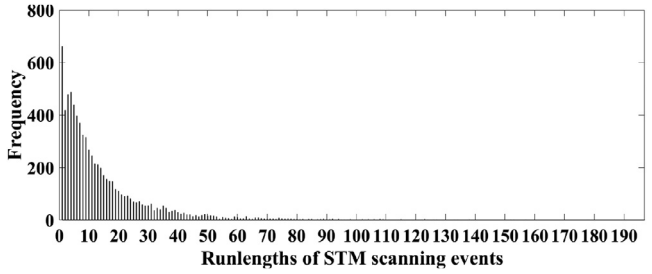

Fig. 12. Histogram plot of STM sequence lengths.

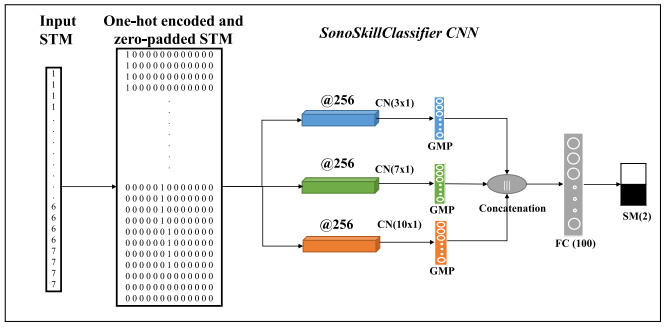

Fig. 13. The proposed *SonoSkillClassifier CNN* architecture for skill classification. The multi-scale temporal CNN has three branches of different filter lengths, 3 for fine-scale, 7 for medium-scale and 10 for coarse-scale representation. (CN: 1D convolutional layer, GMP: global max pooling layer, FC: fully connected layer, SM: softmax layer).

DNA and text classification (Nguyen et al., 2016), as follows. The CNN consists of three branches, where the inputs to each branch are one-hot encoded STMs, each of size $k \times n$ such that the sequence length (k) is the maximum possible STM length, and the dictionary size n is the number of tasks ($n = 13$ anatomical classes in our case). Each branch consists of a 1*D* convolutional layer followed by a global max pooling operation. There are 256 convolutional feature maps in each layer, and convolutional filter lengths of the three layers are selected as 3, 7 and 10 where the filter lengths represent the number of words convolved during each filter operation (field of view); in our case, these are the sequential scanning events in the STM. To justify the size selection, in Fig. 12 we plot the histogram of sequence lengths (or runlengths) of scanning events in the STM.

We observe that the sequence lengths are frequently found to be in the range [1,10]. Hence, the convolutional filters of lengths 3 (fine-scale) and 7 (medium-scale) will capture anatomical task lengths and task transitions, and length 10 (coarse-scale) captures task repetitions in the larger seen window. The features from each branch are concatenated, followed by a 100-neuron fully connected layer and a softmax layer. The proposed **multi-scale temporal CNN** architecture is shown in Fig. 13 and has only 144,430 trainable parameters. In comparison, other existing CNN architectures are much larger, for instance, the Viraminer CNN has 2.42 million parameters. Such a 1*D* CNN architecture has the advantage of significantly reduced learnable parameters compared to the spatio-temporal analysis networks for raw video description, as the number of parameters of the selected network, even after fixing the computational budget, was *ca.* 23 million (Section 4.2).

To train the deep CNN models, the STMs were augmented to prevent over-fitting. The augmentation strategy involved a random dynamic window selection and warping method (Le Guennec et al., 2016) with a random sampling factor ($\lambda \in [0.5, 1.5]$), followed by one-hot encoding and zero padding. For the one-hot encoding, each sequential scanning event was considered as an individual word. Hence, a sequence of words was obtained as a sequence of one-hot encoded vectors, each of length n. Zero-padding was performed on shorter sequences to obtain a fixed sequence length k. The data augmentation strategy is illustrated with an example in Fig. 14. Also, dropout (p_d=0.5) was used between the fully connected layers for model regularisation, and binary crossentropy loss was used during backpropagation. A batch size of 16 was used during training.

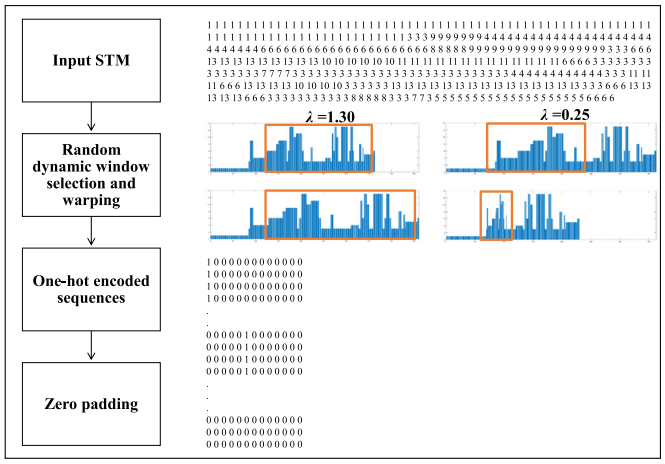

Fig. 14. Data augmentation of STMs for deep learning involving a random dynamic window selection and warping method (λ: random sampling factor). Orange box represents the dynamic window selected in the STM for re-sampling.

6.2. Results and discussion

6.2.1. Feature selection

Feature selection was performed using ReliefF and NCA algorithms, where a subset of 20% random STM samples (68/341) was used as the holdout set.

For the ReliefF algorithm, the number of nearest neighbours for this algorithm were empirically fixed as 5. The feature ranks were computed in decreasing order of their importance weights. A fixed number of features (one-third of the total features) with the highest weights were selected. These were the top-40 features, shown in Fig. 15(a). The set of selected features is denoted by $F_{reliefF}$. The feature names are abbreviated from Section 6.1.1, with the prefixes 'Fr' and 'Ft' denoting freezing and fine-tuning respectively, and suffix representing the label L_i.

For the NCA algorithm, the regularisation parameter was empirically fixed as 0.01. Just like ReliefF, the ranks of features were computed in decreasing order of importance weight, and a fixed number of features (one-third of the total features) with the highest weights were selected. These were the top-40 features, shown in Fig. 15(b). The set of selected features is denoted by F_{NCA}.

We found that 21 features are common in the sets $F_{reliefF}$ and F_{NCA}. We consider the combined set $F_{comb} = F_{reliefF} \cap F_{NCA}$ for further analysis. It can be observed that **freeze and fine-tuning (or searching) durations** of different anatomical tasks are the most important discriminating factors for skill classification, along with **total duration**. It is intuitive that total scan duration and **relative durations (total, mean, maximum, and minimum)** of certain anatomical tasks are significant factors to distinguish between newly-qualified and experienced skill groups. **Shannon entropy**, a measure to determine disorder in the STMs, is considered discriminative by both feature selection algorithms. The **number of repetitions** of some tasks is also discriminatory, along with **transitions to and from the anatomical tasks**. Anatomical tasks such as background search, brain, nose and lips, kidneys, and other cluster

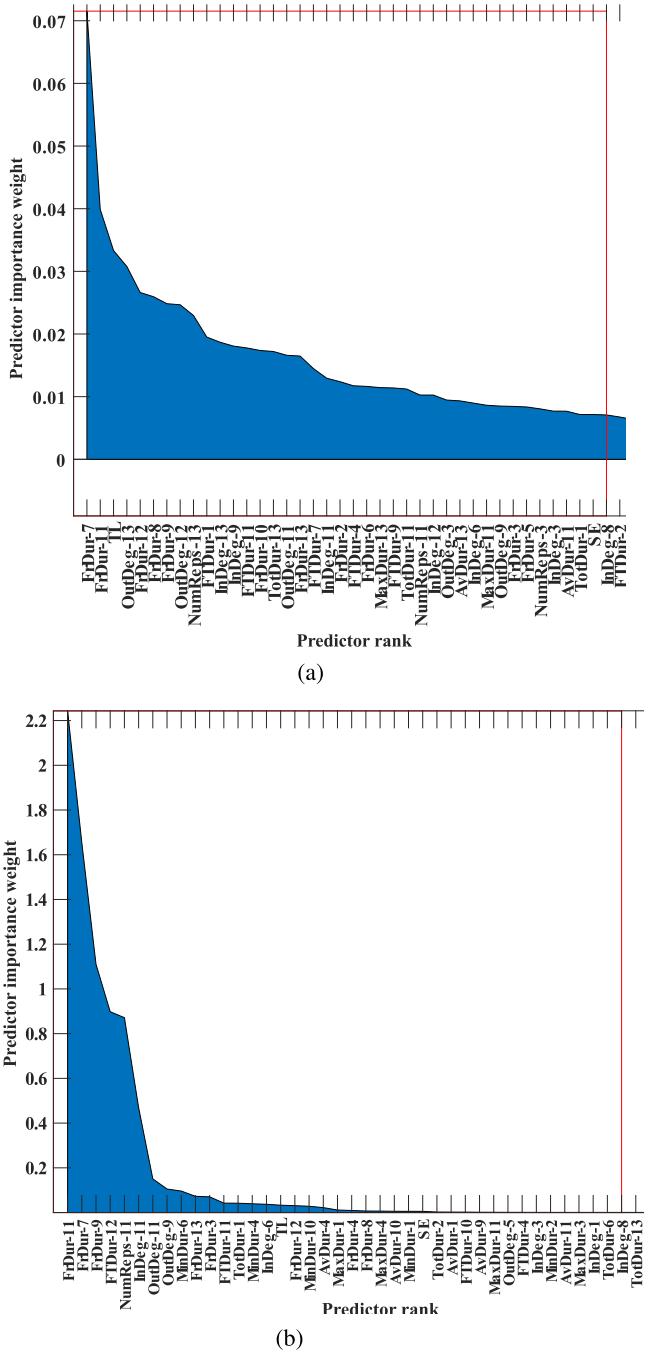

Fig. 15. Feature selection using (a) ReliefF algorithm (b) NCA algorithm. The red box highlights the top-40 features with the highest importance weights. Remaining features are not shown in the figure. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

of tasks are given more importance by the feature selection algorithms.

Two features worthy of further discussion are the total scan duration and average task durations for the two groups, NQ and XP. Fig. 16 shows the plot of the average durations of all the anatomical tasks, for individual operators and the average operator for each group, namely S_{avnq} and S_{avxp}, respectively.

Firstly, we observe that the NQ group has longer scans (average length=29.84 $\pm$ 8.73 min) compared to the XP group (average length=22.69 $\pm$ 7.54 min). This observation suggests that total scan duration is a suitable feature to discriminate between the two

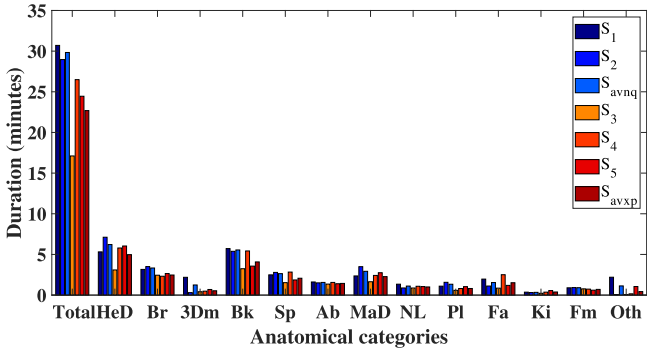

Fig. 16. Average total and individual anatomical task durations for the two groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.), the reader is referred to the web

groups. This feature was also selected by both feature selection methods. This is perhaps an intuitive finding, as one might expect that a newly-qualified operator takes longer to search and localise the different anatomies due to lower expertise compared to experienced operators. Secondly, we observe higher average durations of all anatomical tasks for the NQ group compared to the XP group, again for the same reason, which may have led to more repetitions and transitions, and more time spent in each of the anatomical tasks. Notably, the duration of background search (*Bk*) is much higher for NQ than XP, which suggests that newly-qualified operators spend a longer time to search and reach an optimal plane, or are not easily satisfied with their acquired views, leading to quick freeze-unfreeze actions. In contrast, experienced operators are quick to recognise an opportunity to get a good standard plane, the ability that leads to a much shorter duration of background search.

6.2.2. Machine learning

The effectiveness of the STFs with classical machine learning methods, and STMs with deep learning methods was evaluated through five-fold cross validation for skill classification. For comparative evaluation with existing methods, we compared the *SonoSkillClassifer CNN* with the Viraminer CNN. To demonstrate the enhancement using the multi-scale temporal architecture over individual scales, we performed an ablation study for each of the three branches of the multi-scale temporal *SonoSkillClassifer CNN*. The results are summarised in Table 3, and best results are highlighted in boldface, both for classical machine learning and deep learning. Reported standard metrics are sensitivity, specificity, accuracy and F1-score.

From the results in Table 3 for the two skill groups, we observe the following. As expected, the simplest classical models such as HMM and logistic regression are not observed to be as accurate as the SVMs and random forests. For the classical SVM and random forests machine learning with the hand-engineered STFs, a small overall improvement is seen with the selected features compared to all features, showing good feature selection. However, for SVM, the NQ scans are classified more accurately using all the features, suggesting that some discriminative features may not be captured in the reduced set. Random forests are not superior compared to SVMs but have a balanced classification performance in both operator groups. This result suggests that the hand-engineered timeline features can be applied to the skill classification problem.

Overall, deep learning outperforms the hand-engineered features and classical machine learning (as is often found to be the case). Among the deep learning methods, the proposed *SonoSkillClassifier CNN* is superior to the Viraminer CNN. This is because, its architecture was specifically designed for learning operator skill

Table 3
Five-fold cross validation results for skill classification.

Method	Features (or Params)	Sensitivity	Specificity	Accuracy	F1 Score
STFs and Classical Machine Learning					
HMM (STMs)	NA	0.59 ± 0.24	0.56 ± 0.20	0.58 ± 0.05	0.57 ± 0.12
Logistic regression (all features)	119	0.64 ± 0.34	0.44 ± 0.45	0.54 ± 0.09	0.55 ± 0.18
SVM (all features)	119	**0.87 ± 0.06**	0.82 ± 0.10	0.85 ± 0.03	0.85 ± 0.02
SVM (selected features)	21	0.83 ± 0.06	**0.90 ± 0.05**	**0.86 ± 0.04**	**0.86 ± 0.04**
Random forests (all features)	119	0.80 ± 0.06	0.81 ± 0.07	0.81 ± 0.03	0.81 ± 0.03
Random forests (selected features)	21	0.83 ± 0.05	0.81 ± 0.09	0.82 ± 0.05	0.82 ± 0.05
STMs and Deep Learning					
Viraminer CNN (Tampuu et al., 2019)	2,422,789	0.904 ± 0.114	0.850 ± 0.118	0.876 ± 0.114	0.880 ± 0.109
SonoSkill-Classifier CNN (scale=3)	36,142	0.987 ± 0.022	**0.978 ± 0.024**	0.984 ± 0.015	0.983 ± 0.015
SonoSkill-Classifier CNN (scale=7)	49,454	0.991 ± 0.013	0.972 ± 0.018	0.981 ± 0.011	0.981 ± 0.011
SonoSkill-Classifier CNN (scale=10)	59,438	0.988 ± 0.012	0.972 ± 0.021	0.980 ± 0.010	0.980 ± 0.009
SonoSkill-Classifier CNN (multi-scale)	144,430	**0.998 ± 0.005**	0.972 ± 0.019	**0.985 ± 0.010**	**0.985 ± 0.010**

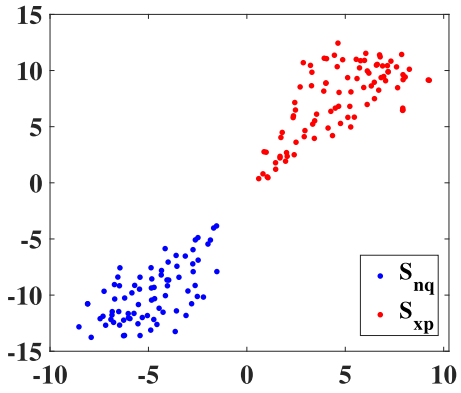

Fig. 17. t-SNE visualisation of the automatically generated test features from a trained *SonoSkillClassifier* CNN. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

assessment from the STMs with lower number of trainable parameters. It has an overall higher performance compared to each single-scale branch tested in the ablation study, showing an improvement in the NQ group from all the individual branches, and in the XP group from scales 7 and 10 and comparable accuracy for scale 3. The ablation study confirms that the multi-scale temporal CNN model is well-suited to combine different STM characteristics such as task patterns, transitions and repetitions. The t-SNE visualisation of the automatically generated features of test STMs fed to one trained *SonoSkillClassifier* CNN model, extracted from the fully-connected layer, is depicted in Fig. 17. From the feature visualisation of the auto-encoded CNN features, it can be clearly observed that, the extracted CNN features show distinct clustering between the XP and NQ groups.

To test the robustness of the proposed CNN method with respect to unseen operators, it is important to perform experiments in leave-operator-out settings. However, an issue in doing this is that the highly unbalanced dataset (Fig. 2) can lead to insufficient data for training both skill groups. For instance, leaving out S_1, S_2 or S_5 was observed to cause over-fitting in the learnt models because of their high contribution to the scan data, hence, removing these scans caused the highest data imbalance (e.g. leaving out S_5 leads to only 49 scans in XP group). For the remaining two operators S_3 and S_4 with lowest data contribution, the leave-operator-out experiment gave the accuracies 70.0% and 76.4% respectively. From this result, it may be argued that with a limited dataset of operators and scans per operator, for dominant operators, the current models may learn operator-specific scanning signatures to classify their unseen STMs more accurately. Nevertheless, an average accuracy of 73.2% for the less dominant operators indicates that the proposed CNN can perform operator skill classification. Furthermore, we test the trained model on an unseen dataset consisting of 13 scan videos undertaken by three expert (XP) operators; these operators were not considered in the previous experiments. We achieve an accuracy of 76.9% on the unseen scans of the new operators, which suggests a good generalisability of the trained models for the skill classification problem.

7. Operator graph model

In this section we describe how to derive an operator-specific model of workflow from the STMs.

7.1. Method

A directed relational graph which we call an **Operator Graph Model (OGM)** is constructed for each operator to model the clinical workflow pattern in a US scan for that operator. A similar idea has been used to model event relations and represent related events as connected graphs for extracting workflow for an individual during interaction with an information management system (Abeta and Kakizaki, 1999).

To derive an OGM representation for an operator, we first cluster the STMs for each operator. However, we first remove the *Bk* class, as background search is not defined as a unique anatomical task but occurs in between other pairs of anatomical tasks. This means an OGM representation is constructed from $n = 12$ anatomical task classes. Then, we calculate the **anatomical task-start probabilities, anatomical task-occurrence probabilities**, and **task transition probability matrix** for the STMs of each operator using the following heuristic approach. Assuming there are K operators, we consider the k^{th} operator, where $k \in \{1, 2, ., K\}$.

1. *Anatomical task-start probability* is calculated for each of the anatomical tasks by counting the relative occurrence of each task at the beginning of the STM. The task-start probability for the i^{th} anatomical task is given as $P_k(i)$ such that $\sum_{i=1}^{n} P_k(i) = 1$.
2. *Anatomical task-occurrence probability* is calculated for each of the anatomical tasks by finding the total relative duration of each task from the corresponding STMs. The task-occurrence probability for the i^{th} anatomical task is given as $O_k(i)$ such that $\sum_{i=1}^{n} O_k(i) = 1$.
3. *Task transition probability matrix* is calculated as the probability to transition from anatomical task i to anatomical task j for non-identical tasks $i \neq j$, and 0 for $i = j$. The transition probability matrix $T_k(i, j)$ is stochastic, *i.e.* $\sum_{j=1}^{n} T_k(i, j) = 1$.

Using the above definitions, an attributed relational directed graph is a 4-tuple $G_k =< N, E, A, B >$ defined for each operator,

with N nodes, E edges, node attributes $A = \{P_k, O_k\}$ and edge attributes $B = \{T_k\}$. All nodes in the OGM are reachable. However, there are no self-loops ($T_k(i,i) = 0$), as in the high-level representation, transitions to the same node are not meaningful.

7.1.1. Most probable non-repeating path

In an ideal scan with no repetitions, each node would be traversed exactly once, leading to an acyclic flow of tasks representing the most-probable non-repeating path. To obtain the most probable non-repeating path for each operator k, we consider all the paths in G_k as first order Markov chains. Hence, the probability $Pr(x)$ of a given path x from edge 1 to edge L is given by,

$$Pr(x) = Pr(x_1) \prod_{i=2}^{L} Pr(x_i | x_{i-1}) \tag{6}$$

From the above definition of the first order Markov chain probability, and our start and transition probabilities, for a given start node i we can define the probability of any existing path in the OGM G_k as,

$$Pr(path) = P_k(i) \prod_{i,j \in E} T_k(i,j) \tag{7}$$

To calculate the most probable path for a given operator, the most probable start nodes (nodes with maximum task-start probability) are given highest preference, because task initialisation can be an important prior to determine the remaining task ordering. Also, finding all the possible paths of the OGM is a hard problem. To solve this, we compute and analyse all the possible paths emanating from the most probable start node for each of the remaining end nodes. The algorithm efficiently finds complete paths for each start node-end node pair using recursive programming. The overall most probable paths are selected as the ones having the highest path probability $Pr(path)$ calculated from Eq. 7. It is found that multiple paths may have the same value of the highest path probability due to repeating transition matrix entries and a (commutative) product operation. Thus, to select one of these paths as the most probable path, we first find subset of paths with most probable end nodes for each operator and compare these paths with the approximate non-repeating paths from each operator's STMs. The approximate non-repeating paths are generated from each STM by assuming that the last instance of an anatomical task is the actual instance, *i.e.*, when the task was successfully completed. Hence, the most probable path for each operator is selected as the one with the maximum overlap with the approximate STM-derived non-repeating paths.

The Viterbi algorithm (Forney, 1973) is not applicable for finding the most probable non-repeating paths because it may lead to task repetitions from the OGM, which was not desired in our case. Other algorithms for shortest path problem were not applied as we want to traverse all the nodes of the graph exactly once, whereas these algorithms may not consider all the graph nodes. Furthermore, if cyclic workflow of events is allowed (with repetition of events), the computation of most-probable paths will incur a higher computational complexity.

By way of illustration, an OGM for operator S_1 is shown in Fig. 18. The relative durations (task-occurrence probability) of each anatomical task for the operator are proportional to the node size of the OGM. The most probable non-repeating path is highlighted in red.

7.1.2. Intra-operator variability

Intra-operator variability could not be computed from the OGM as each graph G_k represents all the STMs for a single operator k. Hence, to find the workflow variability within the scans for a given operator, we go back to their STMs and measure the standard deviations in terms of task type, order, and time-distributions. For task type, the standard deviation in the number of unique tasks are reported. For task order, standard deviation of the Shannon entropy is computed. For task time distributions, the deviation from the mean relative durations of anatomical tasks are found. The three quantities are reported in the range [0,1] as a normalised deviation from the respective means.

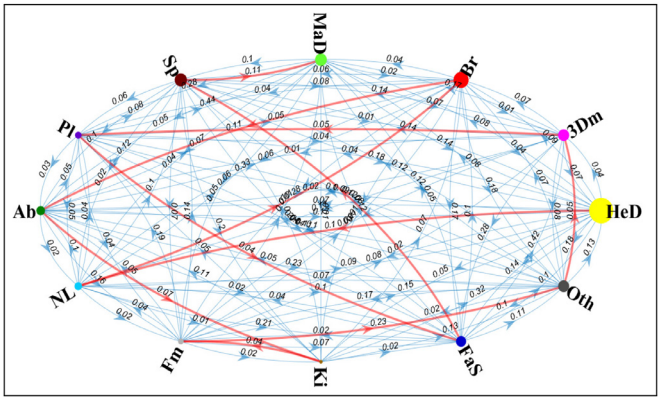

Fig. 18. Example of an OGM with most probable path (red) for operator S_1. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

7.1.3. Inter-operator variability

Given the OGM representation, we can evaluate the inter-operator variability *via* **graph matching** of the operator graphs as follows. A graph-mismatch distance $d_{tot} \in \mathbb{R}$ is obtained for operator x and y as the sum of the three distances d_1, d_2 and d_3 given by,

$$d_1(x,y) = \frac{1}{N} \sum_{i=1}^{N} |P_x(i) - P_y(i)| \tag{8}$$

$$d_2(x,y) = \frac{1}{N} \sum_{i=1}^{N} |O_x(i) - O_y(i)| \tag{9}$$

$$d_3(x,y) = \frac{1}{3} \left\{ \frac{1}{N_{xy}} \sum_{i,j \in E_{xy}} |T_x(i,j) - T_y(i,j)| + \frac{1}{N_{ox}} \sum_{i,j \in E_{ox}} T_x(i,j) + \frac{1}{N_{oy}} \sum_{i,j \in E_{oy}} T_y(i,j) \right\} \tag{10}$$

where N_{xy} is the number of common edges E_{xy} of the OGMs of x with edges E_x and y with edges E_y, N_{ox} and N_{oy} are the number of edges E_{ox} and E_{oy} only found in the OGM of x and y respectively. The last two terms in d_3 help in penalising the distance with extra edges in either of the OGMs. The distance d_{tot} is a measure of the dissimilarity between two given OGMs. The components d_1 and d_2 are related to task type and time-distribution, whereas d_3 is an edge-matching distance representing task transitions for task order. The distance can be calculated for each operator pair, normalised and assembled in a symmetric variability matrix V where each entry $v(i,j)$ represents the normalised distance in the range [0,1] for each pair of operators i, j. Average inter-group and intra-group variabilities are computed from V.

7.2. Results and discussion

7.2.1. Most probable non-repeating paths

The most probable non-repeating paths of all operators are depicted in Fig. 19, where the size of each node is directly proportional to the task-occurrence probability of each anatomical task.

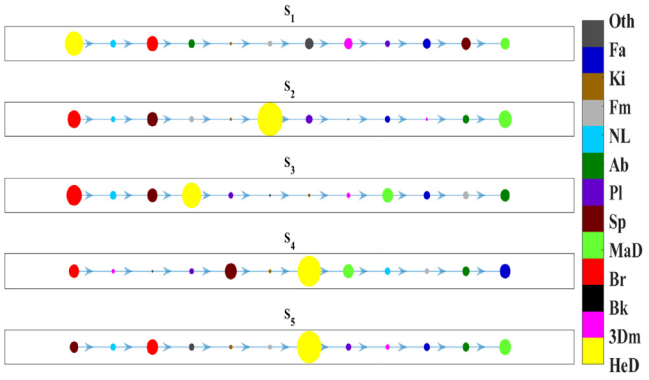

Fig. 19. Examples of most probable non-repeating paths of each operator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We find that majority of the operators prefer to view the brain, the heart or the spine as the first anatomical task in the second-trimester fetal US scan. As expected, the operators are initially interested to examine these three organs first, as these are most important to diagnose any abnormalities in fetal growth. These three anatomical tasks also have the highest task occurrence probabilities for the same reason. Maternal anatomy, face-side, and 3*D* mode images are examined towards the latter part of the scan by most operators, as the operators wish to see these anatomical structures at a lower priority (*i.e.* at a later time during the scan). Perhaps, as expected, the anatomical structures closer to each other are scanned in succession. For example, nose-lips and head (S_1, S_2, S_3 and S_5), kidney and heart (S_2 and S_4), nose-lips and spine (S_2, S_3 and S_5), 3*D* mode and face side (S_5), abdomen and kidney (S_1), and kidney and spine (S_4). This is an efficient strategy, and specifically observed by operators in the XP group, which suggests that the experts are more experienced and skilled to follow a systematic and opportunistic scanning approach forming a mental checklist, which may not be the case of NQ operators.

It should be noted that the number of available scans for each operator is an important consideration in finding the accurate OGM and most probable non-repeating paths, as a higher number of scans per operator will give a path more closely resembling their actual clinical workflows. Moreover, the resulting task orders may depend on fetal position and movement, which reinstates that a high number of scans for each operator will lead to more accurate depictions of the operator clinical workflows.

Furthermore, this representation suggests that the operator clinical workflow depends on not only the skill and experience of operators, but also their personal preferences and priorities of anatomical tasks during the scan. For instance, operator S_1 has higher durations of *3Dm* compared to other operators, and S_3 has a higher interest in viewing *Ab*, hence spends relatively more time on that anatomy. Hence, most probable non-repeating paths derived from the high-level OGMs can be considered as operator-specific scanning signatures of their clinical workflow.

7.2.2. Intra-operator variability

Table 4 shows the results of intra-operator variability for each operator. It can be observed that, in general, intra-operator variabilities are low, with an average of 11%. Task types lead to the highest variability among the scans of the same operator, followed by order and time-distribution. Among the operators, S_2 shows the highest intra-operator variability which may be linked to their less refined skills compared to the XP operators. Interestingly, even though S_1 shows a comparatively lower variability in task type and order, the average intra-operator variability of the NQ group is higher than the XP group.

Table 4
Intra-operator variability for task type, order and distribution.

Operator	Type	Order	Time distribution
S_1	0.11	0.08	0.05
S_2	0.21	0.17	0.06
S_3	0.17	0.12	0.05
S_4	0.14	0.09	0.05
S_5	0.16	0.14	0.05

Table 5
Variability matrix V showing inter-operator variability.

Operator	S_1	S_2	S_3	S_4	S_5
S_1	0.00	0.75	0.87	0.96	0.68
S_2	0.75	0.00	0.87	0.82	0.65
S_3	0.87	0.87	0.00	1.00	0.62
S_4	0.96	0.82	1.00	0.00	0.71
S_5	0.68	0.65	0.62	0.71	0.00

7.2.3. Inter-operator variability

The results of graph matching of the OGM for each operator are shown as the variability matrix V in Table 5.

Inter-operator variabilities are lower between the pair of NQ operators, which suggests that the clinical workflow characteristics are more similar among the newly-qualified operators, with average intra-group variability as 0.75. The next higher value is the average intra-group variability of the XP group, *i.e.* 0.78, which suggests that the experienced operators have more distinct workflow characteristics (type, order and time distributions) and scanning signatures among themselves than the NQ operators. We observe that, the average inter-group variability between the two groups is found to be 0.81, as expected, relatively higher than the intra-group variabilities. The highest variability is found between operators S_3 and S_4 (XP operators), and lowest between S_3 and S_5 (also XP operators).

8. Discussion and conclusion

This study describes automatic clinical workflow analysis in full-length routine second-trimester fetal ultrasound scan videos, including semi-automatic generation of labelled datasets, automatic temporal semantic annotation by training deep spatio-temporal networks in a video description pipeline, and subsequent knowledge representation for operator clinical workflow using simpler models, *i.e.* low-level subject-specific timeline models (STM), mid-level summary of timeline features (STF), and high-level operator graph models (OGM). At each step, the proposed scheme reduces the required dimensionality and computational load for the large-scale raw US scan video dataset. The video description stage involving temporal semantic segmentation of US scan videos using spatio-temporal deep network architectures, achieved a cross-validation accuracy of 91.7%, correlation of 0.98 ($p < 0.05$) with the manually labelled data, and a retrospective validation accuracy of 76.4% on unseen data. The low-level STM provides a shorthand representation of the operator clinical workflow for each US scan, and makes observations on large-scale datasets easier and insightful. The mid-level STFs, consisting of hand-crafted features, give a good baseline with classical machine learning with a best cross-validation accuracy of 86.1%, which is improved by deep learning with an accuracy of 98.5% and generalizability of 76.9% on unseen operators. The high-level OGMs further simplify the clinical workflow for each operator, and analysis of the most probable paths and operator variability shows interesting findings.

This study helps us to look at the scan recordings using different lenses, not only at the video level, but three levels of knowledge representation through which we have demonstrated applications such as operator skill characterisation and inter-operator variability assessment. The proposed clinical workflow analysis methods provide a proof-of-concept for larger sonography studies, that may in turn lead to new methodology and tools to assess newly-qualified operators, compare different scanning protocols in more formal ways, improve human-machine interfaces, identify scan room limitations and develop efficient context-aware workflow management systems in scan rooms. Despite the complexity to train skilled operators for routine scanning, and the growing popularity of ultrasonography in obstetrics, ultrasound operator skill assessment is currently subjective and inconsistent, with limited computational research addressing automated and objective methods. A long-term goal of our operator skill characterisation study is to develop objective computer-aided technical skill evaluation and assessment methods in ultrasonography that can provide actionable feedback to the operators, such as suggesting ways to improve scanning, helping with training resources, and reducing their mental and physical workload in real-time in the scan room.

Limitations of our study and possible future directions of research include the following. Firstly, the results show that the proposed methods enable operator workflow analysis in routine ultrasound imaging, but these need to be tested on a larger dataset containing more operators in each skill group and more scans per operator to be able to draw stronger conclusions. Moreover, in the current paper, we have designed our methods to model the entire clinical workflow for all the anatomical tasks in a full-length routine fetal US scan, to determine the effect of their arbitrary types, order, and duration, for analysing operator skills and variability. In another work, we have studied how motion data recorded *via* probe tracking can be useful for skill assessment (Wang et al., 2020), where we consider specific ultrasonographic tasks, namely, the Heart and the Brain. The results show that operator skill levels can be differentiated for these individual tasks. A natural multimodal extension to our analysis would be to combine video and probe data, both for the specific ultrasonographic tasks and the full-length US scans. Moreover, the current definitions of the NQ and XP operator groups with a threshold at 2 years could be further refined. The choice of 2 years as the threshold follows the recommendation from fetal ultrasound specialists. It was also chosen to not amplify the imbalance in our datasets which might result if other thresholds are selected. An interesting future direction would be to vary this threshold, if more data was available; this would involve a higher number of skill groups to define operator experience. Finally, it would be interesting to study the factors affecting the operators' perceived scanning difficulty based on our findings here, for instance, mental workload of different tasks, fetal position, etc.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Harshita Sharma: Conceptualization, Methodology, Validation, Investigation, Data curation, Writing - original draft. **Lior Drukker:** Investigation, Writing - review & editing. **Pierre Chatelain:** Data curation, Writing - review & editing. **Richard Droste:** Data curation, Writing - review & editing. **Aris T. Papageorghiou:** Supervision, Conceptualization, Writing - review & editing. **J. Alison Noble:** Supervision, Conceptualization, Project administration, Funding acquisition, Writing - review & editing.

Acknowledgments

This work is supported by ERC (ERC-ADG-2015 694581, Project PULSE) and EPSRC (EP/M013774/1, Project Seebibyte). Aris T. Papageorghiou is supported by the Oxford Partnership Comprehensive Biomedical Research Centre funded by the NIHR Biomedical Research Centre (BRC) funding scheme.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.media.2021.101973

References

Abeta, A., Kakizaki, K., 1999. Implementation and evaluation of an automatic personal workflow extraction method. In: Proceedings. Twenty-Third Annual International Computer Software and Applications Conference (Cat. No.99CB37032), pp. 206–212. doi:10.1109/CMPSAC.1999.812702.

Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D., 2017. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans. Biomed. Eng. 64 (9), 2025–2041. doi:10.1109/TBME.2016.2647680.

Basu, A., Blanning, R.W., 2000. A Formal Approach to Workflow Analysis. Information Systems Research 11 (1), 17–36. doi:10.1287/isre.11.1.17.11787. https://pubsonline.informs.org/doi/abs/10.1287/isre.11.1.17.11787

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging 36 (11), 2204–2215. doi:10.1109/TMI.2017.2712367.

Blum, T., Padoy, N., Feußner, H., Navab, N., 2008. Workflow mining for visualization and analysis of surgeries. International Journal of Computer Assisted Radiology and Surgery 3 (5), 379–386. doi:10.1007/s11548-008-0239-0. http://link.springer.com/10.1007/s11548-008-0239-0

Bodenstedt, S., Rivoir, D., Jenke, A., Wagner, M., Breucha, M., Müller-Stich, B., Mees, S.T., Weitz, J., Speidel, S., 2019. Active learning using deep bayesian networks for surgical workflow analysis. Int. J. Comput. Assist. Radiol. Surg. 14 (6), 1079–1087. doi:10.1007/s11548-019-01963-9.

Cai, Y., Droste, R., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2020. Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. Med. Image Anal. 65, 101762.

Cai, Y., Sharma, H., Chatelain, P., Noble, J.A., 2018. SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1475–1478. doi:10.1109/ISBI.2018.8363851.

Carneiro, G., Georgescu, B., Good, S., Comaniciu, D., 2008. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree. IEEE Trans. Med. Imaging 27 (9), 1342–1355. doi:10.1109/TMI.2008.928917.

Carreira, J., Zisserman, A., 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp. 4724–4733. doi:10.1109/CVPR.2017.502.

Charriére, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., Cochener, B., 2017. Real-time analysis of cataract surgery videos using statistical models. Multimed. Tools Appl. 76 (21), 22473–22491. doi:10.1007/s11042-017-4793-8.

Chatelain, P., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2018. Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies. IEEE Trans Cybern 1–11. doi:10.1109/TCYB.2018.2866274.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A., 2015. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J Biomed Health Inform 19 (5), 1627–1636. doi:10.1109/JBHI.2015.2425041.

Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J., Ni, D., Heng, P., 2017. Ultrasound standard plane detection using a composite neural network framework. IEEE Trans. Cybern. 47 (6), 1576–1586. doi:10.1109/TCYB.2017.2685080.

Diba, A., Fayyaz, M., Sharma, V., Hossein Karami, A., Mahdi Arzani, M., Yousefzadeh, R., Van Gool, L., 2018. Temporal 3D ConvNets using Temporal Transition Layer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1117–1121.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

Droste, R., Cai, Y., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2019. Ultrasound image representation learning by modeling sonographer visual attention. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 592–604.

Forney, G.D., 1973. The viterbi algorithm. Proc. IEEE 61 (3), 268–278.
Franke, S., Meixensberger, J., Neumuth, T., 2013. Intervention time prediction from surgical low-level tasks. J. Biomed. Inform. 46 (1), 152–159. doi:10.1016/j.jbi.2012.10.002.
Gao, Y., Maraci, M.A., Noble, J.A., 2016. Describing ultrasound video content using deep convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 787–790. doi:10.1109/ISBI.2016.7493384.
Gibbs, V., 2015. The role of ultrasound simulators in education: an investigation into sonography student experiences and clinical mentor perceptions. Ultrasound 23 (4), 204–211.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Holden, M.S., Ungi, T., Sargent, D., McGraw, R.C., Chen, E.C.S., Ganapathy, S., Peters, T.M., Fichtinger, G., 2014. Feasibility of real-time workflow segmentation for tracked needle interventions. IEEE Trans. Biomed. Eng. 61 (6), 1720–1728. doi:10.1109/TBME.2014.2301635.
Horeman, T., Dankelman, J., Jansen, F.W., Dobbelsteen, J.J.v.d., 2014. Assessment of laparoscopic skills based on force and motion parameters. IEEE Trans. Biomed. Eng. 61 (3), 805–813. doi:10.1109/TBME.2013.2290052.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.
Kay, A., 2007. Tesseract: an open-source optical character recognition engine. Linux Journal 2007 (159), 2.
Khan, N.H., Tegnander, E., Dreier, J.M., Eik-Nes, S., Torp, H., Kiss, G., 2016. Automatic measurement of the fetal abdominal section on a portable ultrasound machine for use in low and middle income countries. In: 2016 IEEE International Ultrasonics Symposium (IUS), pp. 1–4. doi:10.1109/ULTSYM.2016.7728557.
Kirwan, D., 2010. NHS Fetal anomaly screening programme. 18+ 0 to 20+ 6 Weeks Fetal Anomaly Scan National Standards and Guidance for England.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97 (1–2), 273–324.
Bureau of Labor Statistics, U., 2019. Diagnostic Medical Sonographers and Cardiovascular Technologists and Technicians, Including Vascular Technologists. https://www.bls.gov/ooh/healthcare/diagnostic-medical-sonographers.htm?__hstc=182781753.07430159d50a3c91e72c280a7921bf0d.1542067200111.1542067200112.1542067200113.1&__hssc=182781753.1.1542067200114&__hsfp=998628806
Lafferty, J.D., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.
Le Guennec, A., Malinowski, S., Tavenard, R., 2016. Data augmentation for time series classification using convolutional neural networks. ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data.
Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.
Liu, H., Motoda, H., 2012. Feature selection for knowledge discovery and data mining, 454. Springer Science & Business Media.
Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9 (Nov), 2579–2605.
Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G.D., Jannin, P., 2017. Surgical data science for next-generation interventions. Nature Biomedical Engineering 1 (9), 691–696. doi:10.1038/s41551-017-0132-7. https://www.nature.com/articles/s41551-017-0132-7
Maraci, M.A., Bridge, C.P., Napolitano, R., Papageorghiou, A., Noble, J.A., 2017. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. Med. Image Anal. 37, 22–36.
März, K., Hafezi, M., Weller, T., Saffari, A., Nolden, M., Fard, N., Majlesara, A., Zelzer, S., Maleshkova, M., Volovyk, M., Gharabaghi, N., Wagner, M., Emami, G., Engelhardt, S., Fetzer, A., Kenngott, H., Rezai, N., Rettinger, A., Studer, R., Mehrabi, A., Maier-Hein, L., 2015. Toward knowledge-based liver surgery: holistic information processing for surgical decision support. Int. J. Comput. Assist. Radiol. Surg. 10 (6), 749–759. doi:10.1007/s11548-015-1187-0.
Nguyen, N.G., Tran, V.A., Ngo, D.L., Phan, D., Lumbanraja, F.R., Faisal, M.R., Abapihi, B., Kubo, M., Satou, K., 2016. Dna sequence classification by convolutional neural network. J. Biomed. Sci. Eng. 9 (05), 280.
Noble, A., Boukerroui, D., 2006. Ultrasound image segmentation: a survey. IEEE Trans. Med. Imaging 25 (8), 987–1010.
Noble, J.A., 2010. Ultrasound image segmentation and tissue characterization. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 224 (2), 307–316. doi:10.1243/09544119JEIM604.
Oropesa, I., Sánchez-González, P., Lamata, P., Chmarra, M.K., Pagador, J.B., Sánchez–Margallo, J.A., Sánchez-Margallo, F.M., Gómez, E.J., 2011. Methods and tools for objective assessment of psychomotor skills in laparoscopic surgery. Journal of Surgical Research 171 (1), e81–e95.
Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O., Navab, N., 2012. Statistical modeling and recognition of surgical workflow. Medical Image Analysis 16 (3), 632–641. doi:10.1016/j.media.2010.10.001. http://www.sciencedirect.com/science/article/pii/S1361841510001131
Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. 53 (1–2), 23–69.
Sanchez-Ortiz, G.I., Declerck, J., Mulet-Parada, M., Noble, J.A., 2000. Automating 3d echocardiographic image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 687–696.
Shannon, C.E., 1948. A mathematical theory of communication. Bell system technical journal 27 (3), 379–423.
Sharma, H., Droste, R., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2019. Spatio-Temporal Partitioning And Description Of Full-Length Routine Fetal Anomaly Ultrasound Scans. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 987–990. doi:10.1109/ISBI.2019.8759149.
Sielhorst, T., Blum, T., Navab, N., 2005. Synchronizing 3D movements for quantitative comparison and simultaneous visualization of actions. In: Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05), pp. 38–47. doi:10.1109/ISMAR.2005.57.
Sinclair, M.D., Martinez, J.C., Skelton, E., Li, Y., Baumgartner, C.F., Bai, W., Matthew, J., Knight, C.L., Smith, S., Hajnal, J., King, A.P., Kainz, B., Rueckert, D., 2018. Cascaded Transforming Multi-task Networks For Abdominal Biometric Estimation from Ultrasound. https://openreview.net/forum?id=r1ZGQW2if
Soh, M., 2016. Learning CNN-LSTM architectures for image caption generation. Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep.
Tampuu, A., Bzhalava, Z., Dillner, J., Vicente, R., 2019. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS ONE 14 (9). doi:10.1371/journal.pone.0222271. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6738585/
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning Spatiotemporal Features with 3d Convolutional Networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497. doi:10.1109/ICCV.2015.510.
Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M.d., Padoy, N., 2017. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Transactions on Medical Imaging 36 (1), 86–97. doi:10.1109/TMI.2016.2593957. Conference Name: IEEE Transactions on Medical Imaging
Uemura, M., Jannin, P., Yamashita, M., Tomikawa, M., Akahoshi, T., Obata, S., Souzaki, R., Ieiri, S., Hashizume, M., 2016. Procedural surgical skill assessment in laparoscopic training environments. Int. J. Comput. Assist. Radiol. Surg. 11 (4), 543–552. doi:10.1007/s11548-015-1274-2.
Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G., 2009. Data-Derived Models for Segmentation with Application to Surgical Assessment and Training. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009. Springer, Berlin, Heidelberg, pp. 426–434. doi:10.1007/978-3-642-04268-3_53.
Vedula, S.S., Ishii, M., Hager, G.D., 2017. Objective assessment of surgical technical skill and competency in the operating room. Annu. Rev. Biomed. Eng. 19, 301–325. doi:10.1146/annurev-bioeng-071516-044435.
Vercauteren, T., Unberath, M., Padoy, N., Navab, N., 2020. CAI4CAI: The Rise of Contextual Artificial Intelligence in Computer-Assisted Interventions. Proceedings of the IEEE 108 (1), 198–214. doi:10.1109/JPROC.2019.2946993. Conference Name: Proceedings of the IEEE
Wang, Y., Droste, R., Jiao, J., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2020. Differentiating Operator Skill During Routine Fetal Ultrasound Scanning Using Probe Motion Tracking. In: Hu, Y., Licandro, R., Noble, J.A., Hutter, J., Aylward, S., Melbourne, A., Abaci Turk, E., Torrents Barrena, J. (Eds.), Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer International Publishing, Cham, pp. 180–188.
Wu, L., Cheng, J., Li, S., Lei, B., Wang, T., Ni, D., 2017. FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybern. 47 (5), 1336–1349. doi:10.1109/TCYB.2017.2671898.
Wu, Z., Yao, T., Fu, Y., Jiang, Y.-G., 2017. Deep learning for video classification and captioning. In: Frontiers of Multimedia Research. Association for Computing Machinery and Morgan & Claypool, pp. 3–29.
Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp. 802–810.
Yang, W., Wang, K., Zuo, W., 2012. Neighborhood component feature selection for high-dimensional data.. JCP 7 (1), 161–168.
Yaqub, M., Kelly, B., Papageorghiou, A.T., Noble, J.A., 2015. Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 687–694.
Yue, T., Wang, H., 2018. Deep learning for genomics: a concise overview. arXiv preprint arXiv:1802.00810.
Zia, A., Essa, I., 2018. Automated surgical skill assessment in RMIS training. Int. J. Comput. Assist. Radiol. Surg. 13 (5), 731–739. doi:10.1007/s11548-018-1735-5.