SPATIO-TEMPORAL PARTITIONING AND DESCRIPTION OF FULL-LENGTH ROUTINE FETAL ANOMALY ULTRASOUND SCANS

H. Sharma[1], R. Droste[1], P. Chatelain[1], L. Drukker[2], A.T. Papageorghiou[2] and J.A.Noble[1]

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
[2]Nuffield Department of Women's and Reproductive Health, University of Oxford, UK

ABSTRACT

This paper considers automatic clinical workflow description of full-length routine fetal anomaly ultrasound scans using deep learning approaches for spatio-temporal video analysis. Multiple architectures consisting of $2D$ and $2D + t$ CNN, LSTM, and convolutional LSTM are investigated and compared. The contributions of short-term and long-term temporal changes are studied, and a multi-stream framework analysis is found to achieve the best top-1 accuracy=0.77 and top-3 accuracy=0.94. Automated partitioning and characterisation on unlabelled full-length video scans show high correlation (ρ=0.95, p=0.0004) with workflow statistics of manually labelled videos, suggesting practicality of proposed methods.

Index Terms— Fetal anomaly scan, spatio-temporal analysis, video classification, ultrasound, clinical workflow.

1. INTRODUCTION

Ultrasound imaging is widely used for monitoring pregnancy due to its non-invasiveness, absence of radiation, high accessibility, high reliability and low costs. In most countries, a routine ultrasound (US) scan is offered in the second trimester of pregnancy (18-22 weeks) to check for anomalies and assess fetal growth [1]. In the UK, scan guidelines are provided by the Fetal Anomaly Screening Programme (FASP) [2]. During a full scanning session, a sonologist or sonographer views required fetal anatomical structures including heart, abdomen, brain, head, spine and limbs, maternal structures such as the uterine arteries, and additional anatomy or activity such as fetal hands and feet, placenta, blood flows and umbilical cord insertion. These may be visualised in different viewing planes (transverse, coronal, sagittal) and imaging modes ($2D + t$, colour Doppler, $3D$ or $3D + t$). Hence, it is interesting to comprehensively analyse and quantify operator clinical workflow in a spatio-temporal context, *i.e.*, the type, duration and sequence of scanned anatomical structures and activities, in order to explore intra- and inter-operator correlation or variability, which may, for instance, offer insight into skill differences between experts and trainees. Such analysis requires temporal partitioning and anatomical labelling on full-length scans which, if performed manually, would be impractical due to the enormous amount of acquired video data. Hence, in the paper we address the problem of automating this task using spatio-temporal analysis methods.

In computer vision, video classification and activity recognition have proved to be challenging, and in ultrasound imaging, are further complicated by class imbalance, imaging artefacts and the relative complex interpretation of US video data. Previously, $2D$ standard plane detection and classification has been extensively studied for fetal US images [3], [4]. In contrast, this study involves natural partitioning and description of $2D+t$ US videos utilizing richer spatio-temporal information than $2D$ standard planes. Also, unlike previous US video-based studies [5], [6] that focussed on specific anatomical structures such as the heart, abdomen or skull in shorter-length US sweeps (clips), this work explores a natural partitioning for automated clinical workflow analysis in full-length scans, containing a comprehensive list of anatomical structures and activities.

The contributions of this paper are as follows. We perform extensive experiments to compare several deep learning architectures and propose methods for comprehensive $2D+t$ spatio-temporal description in fetal anomaly US video scans. Also, using a multi-stream framework, we investigate the effect of fine and coarse temporal changes in videos. The designed models are trained without pre-trained weights, as to our knowledge, no existing method represents this problem sufficiently. Finally, the trained models are applied to unlabelled full-length video scans to determine similarity with manually computed workflows of labelled video scans.

2. METHODS

2.1. Data Acquisition and Pre-processing

Clinical routine fetal anomaly US scans were acquired at 30 frames per second as part of a large-scale study *PULSE.*[1] Freeze frames (FF) were automatically detected using optical character recognition, and full-length videos were temporally partitioned to obtain clips of 5 seconds for ease of manual labelling, with two-thirds of the frames before FF and one-third afterwards. This partitioning was selected based on the observation that during scanning, the operator searches for a standard view in local proximity of an anatomical structure, and then freezes for inspection and biometric measurements.

This work is supported by ERC (ERC-ADG-2015 694581, Project PULSE) and EPSRC (EP/M013774/1, Project Seebibyte). ATP is supported by the NIHR Biomedical Research Centres funding scheme.

[1]Ethics approval was obtained for data acquisition.

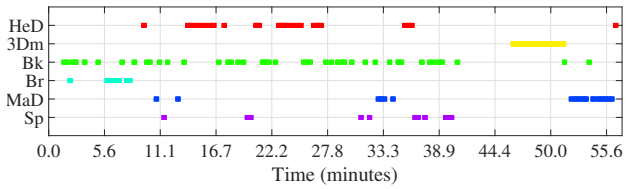

Fig. 1: Example of clinical workflow visualisation (timeline) of a fetal anomaly US scan showing the top-6 event classes

Extracted video clips from 25 full-length scans of different subjects were manually annotated. The average video duration was 45.7±11.6 minutes (82,219 ± 20,929 frames). 20 anatomical or activity categories (we call these "events") were identified by a clinical expert as [label] (% distribution): heart including Doppler[HeD] (20%), $3D$ or $3D + t$ mode[3Dm] (17.3%), background search or transition[Bk] (14.8%), brain with skull and neck[Br] (9.2%), maternal anatomy including Doppler[MaD] (9.2%), spine[Sp] (6%), abdomen[Ab] (3.6%), nose and lips[NL] (3.6%), kidneys[Ki] (2.6%), face side profile[Fa] (2.6%), femur[Fm] (1.8%) and other categories with low total (<10%) representation (umbilical cord insertion[Um], full body side profile[Fb], bladder including Doppler[BlD], feet[Fe], top head with eyes/nose[Th], girl or boy, hands[Ha], arms[Ar] and legs[Le]). Due to the high class imbalance, we select the 11 most dominant categories which include key FASP standard planes. An example of clinical workflow visualisation of a manually labelled scan with the top-6 classes (for illustration) is given in Figure 1.

To limit the GPU memory required for deep learning, each labelled video clip was approximated by 12 frames using uniform sampling, with a longer clip yielding ten unique smaller clips. $2D$ and $2D + t$ instances from full-length videos were subject-wise divided into training, validation and test, respectively (Table 1). Test data was held-out and not used for training, and validation data was used to monitor loss during training. Frames were pre-processed by cropping the relevant image area and resizing to 224×224 pixels. During training, a data augmentation method was randomly selected and applied to every frame of a clip, including rotation $[-30°, 30°]$, horizontal flip, vertical flip, Gaussian noise, and shear (≤ 0.2).

2.2. Learning Spatio-temporal Descriptions

Current deep learning methods for video classification and activity recognition can be broadly divided into three groups [7]: image (frame)-based methods, end-to-end convolutional neural networks (CNN), and modelling temporal dependency *via* recurrent neural networks (RNN). We explore these groups of methods as follows.

For spatial ($2D$) representation of individual frames, three deep CNN architectures, namely, VGG16, VGG19 [8] and SonoNet-64 [3] (a variant of VGG16), were considered due to their reported good classification performance on natural images and fetal anomaly US images, respectively. Deeper architectures were not considered to avoid intractable complexities of the spatio-temporal versions. Empirically, SonoNet-64 CNN consistently outperformed the other two on spatial subsets (column 3, Table 1), therefore, it was selected as the base $2D$ CNN on which to build the spatio-temporal models.

Table 1: Dataset distribution for deep learning experiments

Dataset	Full-length videos	Number (%) images	Number (%) clips
Train	19 (76%)	27,496 (77.2%)	11,004 (77.2%)
Validate	3(12%)	4,133 (11.6%)	1,657 (11.6%)
Test	3 (12%)	3,986 (11.2%)	1,593 (11.2%)
Total	**25 (100%)**	**35,615 (100%)**	**14,254 (100%)**

To analyse $2D + t$ data with end-to-end convolutional neural networks, $3D$ (or $2D + t$) CNNs were employed [9]. These utilize $3D$ convolutional kernels to learn motion (displacement) patterns between adjacent video frames. Another form of end-to-end model consists of two or more streams (for instance, an additional optical flow input), but this type of model was not considered due to the overhead of computing optical flow images for thousands of frames of an US scan.

To model temporal dependency, recurrent neural networks were investigated. Long-short term memory (LSTM) units have demonstrated their effectiveness in natural video classification *via* recurrent convolutional networks (RCN) [10]. Recently, the Convolutional LSTM (CLSTM) memory unit was introduced that extended the concept to the spatial domain [11]. We have implemented an LSTM-based RCN and a CLSTM-based RCN to study the comparative performance of both the memory units.

Several deep learning architectures were designed from the three groups (Figure 2). The architectures in Figure 2 (Block A) were trained in two ways. In the first case, spatial training was performed using $2D$ feature extraction and adaptation layers, and weights of these layers were fixed to extract $1D$ or $2D$ features of consecutive frames. The features were directly used to train spatio-temporal dependency models: LSTM-RNN with $1D$ features (*Feat1D-LSTM-RNN*), $2D + t$ CNN with $2D$ features (*Feat2D-2DtCNN*) and CLSTM-RNN with $2D$ features (*Feat2D-CLSTM-RNN*), followed by dense layers and softmax for final prediction. In the second case, three end-to-end models were built using spatio-temporal dependency units after the $2D$ base layers: hybrid $2D + t$ CNN (*Temp-Sono-2DtCNN*), RCN using LSTM (*Temp-Sono-LSTM-RCN*), and RCN using CLSTM (*Temp-Sono-CLSTM-RCN*). In Figure 2 (Block B), the pure $2D + t$ CNN (*Sono-2DtCNN*) was designed by temporally inflating subset of layers of SonoNet-64 $2D$ CNN, and only end-to-end trained.

Computational hardware used are NVIDIA GTX Titan X (12 GB) and NVIDIA GTX 1070 (8 GB) GPU. Deep learning models were implemented using Keras framework with TensorFlow backend. Stochastic gradient descent (SGD) with Nesterov momentum (μ=0.9) was used for optimisation during the training phase, with a learning rate of 0.01 and ≥ 100 epochs. Batch size was varied between 8, 16 or 32 depending on GPU memory availability. For model regularisation, dropout ($p_d \in \{0.2, 0.3, 0.5\}$), batch normalisation and data augmentation (Section 2.1) were employed.

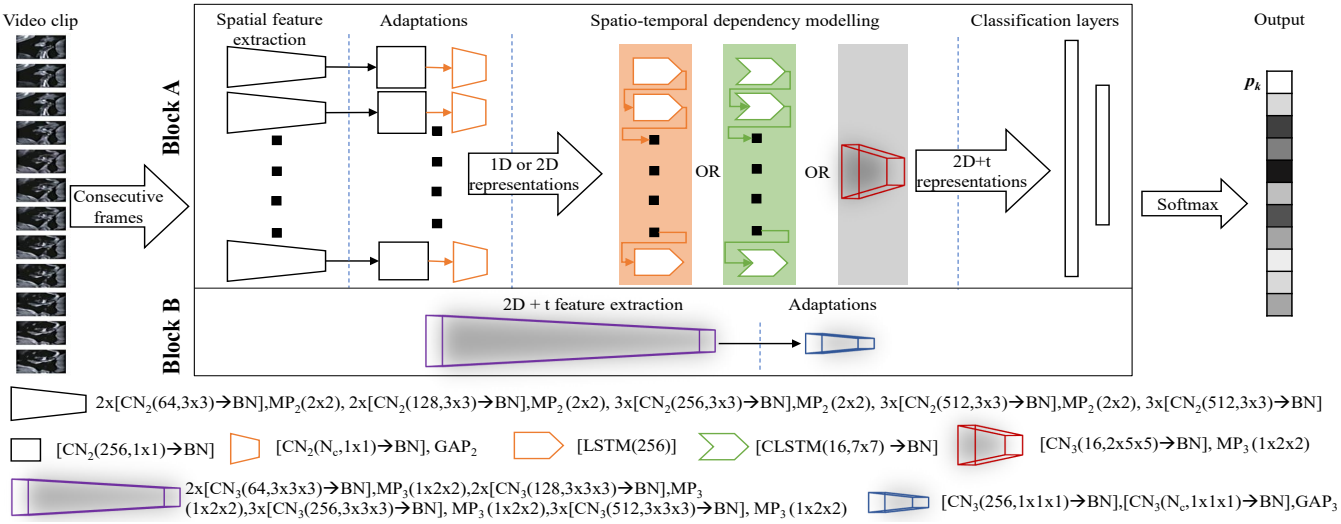

Fig. 2: Spatio-temporal deep learning architectures. In Block A, uncoloured convolutional and adaptation layer units are common but coloured adaptation layer units are only linked to the corresponding same coloured spatio-temporal layer units. CN (feature depth, kernel size): convolutions with ReLU activation, BN: batch normalisation, MP (kernel size): max-pooling, GAP: global average pooling, N_c: number of classes. Subscripts 2 and 3 represent operations in $2D$ and $2D+t$. Classification layers include 512, 256 or 128 and N_c elements, respectively.

2.3. Multi-stream Framework

The deep learning models, as described in Section 2.2, were built using video clips at one-fourth of the original frame rate ($FR/4$) representing short-term spatio-temporal motion. We selected the best performing model (*Sono-2DtCNN*) to learn and inspect spatio-temporal dynamics using near consecutive frames with fine temporal changes ($FR/2$) and long-term motion with coarse temporal changes ($FR/8$). A multi-stream framework with late fusion on learned model softmax probabilities was investigated, to observe whether individual streams with varying temporal dimensions can enhance learned knowledge due to the distinct characteristics of under-represented classes in the unbalanced datasets.

3. RESULTS AND DISCUSSION

3.1. Comparison of Trained Models

Precision (P), Recall (R), F1-score ($F1$), Top-1 accuracy (A_1) and Top-3 accuracy (A_3) were used as evaluation metrics. Image (frame)-based classification included baseline methods: 1) pre-trained weights of SonoNet-64 (PT) [3] with only overlapping categories; 2) SonoNet-64 trained using $2D$ image subsets of our dataset. VGGNets are reported for completeness, however, not considered for spatio-temporal analysis. Table 2 summarizes the number of parameters and resulting metrics for the investigated models.

Firstly, we observe that image-based training of SonoNet-64 CNN provides a good baseline and outperforms the pre-trained $2D$ model and the vanilla VGGNets. Among spatio-temporal models using LSTM and CLSTM memory units, feature-based RNN approaches perform better than corresponding end-to-end RCN methods. This can be explained as, in the first case, we fix the $2D$ feature extraction and adaptation layers, hence, these require fewer trainable parameters which scale to our datasets better than end-to-end RCNs with more parameters causing lower generalisation due to unbalanced classes. Furthermore, $2D$ CLSTM units show superior performance to $1D$ LSTM units consistently for feature-based models, and in accuracies for end-to-end models, demonstrating more powerful spatio-temporal representations for $2D+t$ data. However, due to the higher computational requirements, configurations involving CLSTM units are slower to converge than LSTM units. The best performing end-to-end model (for all evaluation metrics) is the $2D+t$ CNN *Sono-2DtCNN* that describes the spatio-temporal properties of video clips by directly using $2D+t$ convolutional and pooling operations. It outperforms the RCNs and hybrid *Temp-Sono-2DtCNN* which model temporal dynamics on $1D$ or $2D$ representations of consecutive frames. Also, the average evaluation metrics of $2D$ baseline model appear comparable to spatio-

Table 2: Comparative analysis of trained models

Model architecture	Params	P	R	$F1$	A_1	A_3
Baseline image(frame)-based methods						
VGGNet-16	134.3 M	0.45	0.32	0.33	0.47	0.76
VGGNet-19	139.6 M	0.36	0.29	0.30	0.45	0.77
SonoNet-64 (PT)	NA	0.52	0.48	0.46	0.55	0.55
SonoNet-64	14.8 M	**0.71**	**0.52**	**0.56**	**0.69**	**0.91**
Spatio-temporal methods						
Feat2D-2DtCNN	2.6M	0.64	**0.57**	0.59	0.68	0.86
Feat1D-LSTM-RNN	2.6M	0.67	0.52	0.54	0.68	**0.87**
Feat2D-CLSTM-RNN	5.7M	**0.68**	**0.57**	**0.60**	**0.71**	0.85
Temp-Sono-2DtCNN	15.5M	0.48	0.44	0.43	0.62	0.84
Temp-Sono-LSTM-RCN	15.7M	0.58	0.41	0.39	0.58	0.82
Temp-Sono-CLSTM-RCN	16.2M	0.55	0.39	0.41	0.58	0.82
Sono-2DtCNN	23.0M	**0.73**	**0.66**	**0.66**	**0.75**	**0.93**

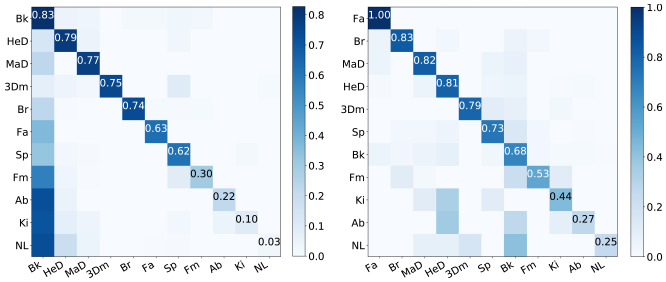

Fig. 3: Confusion matrix predicted vs. true label for image-based CNN (left) and spatio-temporal *Sono-2DtCNN*(right)

temporal models and lower than *Sono-2DtCNN*; the relative $2D$ and $2D+t$ performance is further investigated by careful inspection of the respective confusion matrices as follows.

Figure 3 shows confusion matrices for the image-based SonoNet-64 and for the spatio-temporal model *Sono-2DtCNN*. We observe that the spatial model shows greater bias towards the more commonly occurring classes 'Bk' and 'HeD'. All the remaining event classes (except 'Bk'), even when under-represented in the dataset, are more accurately described using $2D+t$ spatio-temporal information than only $2D$ spatial information, suggesting the contribution of temporal changes for classification in fetal US video scans. Specifically for the $2D+t$ method, 'Ki' and 'Ab' are misclassified as 'HeD' due to high visual similarity between these views, and the heart can sometimes appear in the other two views. Higher confusion is seen between 'Bk' and other classes, which is associated with non-compactness of 'Bk' as it may consist of multiple events. Also, 'NL' can be confused with '3Dm', which is explainable as a sonographer generally uses this mode to obtain a $3D$ reconstruction of the fetal head that includes the nose and lips, appearing similar to the nose and lips view in the $2D+t$ mode.

3.2. Multi-stream Classification

The results of the multi-stream framework are shown in Table 3. We observe that classification performance improves from near-consecutive frames and short-term dynamics to long-term changes with a higher temporal context. Furthermore, multi-stream classification combines effects of short-term and long-term temporal dynamics giving improvement in P and A_3. Detailed inspection of each confusion matrix confirmed that individual streams describe distinct categories more effectively, and the combined framework achieves a superior representation of the unbalanced dataset. For example, 'Sp' has higher discrimination at $FR/2$ with fine temporal changes, whereas 'Br' has increased detection accuracy at $FR/8$, indicating long-term temporal dependency.

Table 3: Multi-stream framework analysis

Model architecture	P	R	$F1$	A_1	A_3
$2D$ stream	0.71	0.52	0.56	0.69	0.91
$2D+t$ stream @ FR/2	0.63	0.59	0.58	0.70	0.89
$2D+t$ stream @ FR/4	0.73	**0.66**	0.66	0.75	0.93
$2D+t$ stream @ FR/8	**0.77**	**0.66**	**0.67**	**0.77**	0.92
Multi-stream late fusion	**0.77**	0.65	0.66	0.76	**0.94**

3.3. Comparison of Workflow Statistics

Automated classification using the best performing learnt prediction model was applied to ten unlabelled full-length videos of US fetal anomaly scans. Clinical workflow statistics were computed as mean $\pm$ standard deviation (duration in minutes) and mean percentage (95% confidence interval) per scan for the categories. Histograms of durations of categories were compared between manually and automatically labelled videos, and a high correlation with Pearson's correlation coefficient ρ=0.95 (p-value=0.0004) was observed. This suggests the suitability of proposed spatio-temporal methods to describe clinical workflow and composition of full-length US fetal anomaly scans.

4. CONCLUSION

Deep learning models were proposed, implemented and compared for automatic US video description. End-to-end spatio-temporal model *Sono-2DtCNN* was found to perform better than image-based and RNN-based analysis methods. A multi-stream framework was studied to explore the effects of temporal changes. Favourable correlation of workflow statistics was observed between manually labelled and automatically classified videos. We conclude that the proposed automatic spatio-temporal partitioning and description approach can be employed to generate video descriptions and workflow patterns in full-length fetal anomaly ultrasound scans.

5. REFERENCES

[1] Salomon et al., "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *Ultrasound in Obstetrics and Gynecology*, vol. 37(1), pp. 116–126, 2011.

[2] House et al., "About the NHS Screening Programmes," 2015.

[3] Baumgartner et al., "SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound," *IEEE TMI*, vol. 36(11), pp. 2204–2215, 2017.

[4] Cai et al., "SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking," in *IEEE ISBI*, 2018, pp. 1475–1478.

[5] Maraci et al., "A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat," *Medical image analysis*, vol. 37, pp. 22–36, 2017.

[6] Gao et al., "Describing ultrasound video content using deep convolutional neural networks," in *IEEE ISBI*, 2016, pp. 787–790.

[7] Wu et al., "Frontiers of multimedia research," chapter Deep Learning for Video Classification and Captioning, pp. 3–29. 2018.

[8] Simonyan et al., "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[9] Tran et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," in *IEEE ICCV*, 2015, pp. 4489–4497.

[10] Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE CVPR*, 2015, pp. 2625–2634.

[11] Shi et al., "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.