Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent Neural Networks

Hao Chen[1], Qi Dou[1], Dong Ni[2,⋆], Jie-Zhi Cheng[2], Jing Qin[2], Shengli Li[3], and Pheng-Ann Heng[1]

[1] Dept. of Computer Science and Engineering, The Chinese University of Hong Kong
[2] School of Medicine, Shenzhen University, China
[3] Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University

Abstract. Accurate acquisition of fetal ultrasound (US) standard planes is one of the most crucial steps in obstetric diagnosis. The conventional way of standard plane acquisition requires a thorough knowledge of fetal anatomy and intensive manual labors. Hence, automatic approaches are highly demanded in clinical practice. However, automatic detection of standard planes containing key anatomical structures from US videos remains a challenging problem due to the high intra-class variations of standard planes. Unlike previous studies that developed specific methods for different anatomical standard planes respectively, we present a general framework to detect standard planes from US videos automatically. Instead of utilizing hand-crafted visual features, our framework explores spatio-temporal feature learning with a novel knowledge transferred recurrent neural network (T-RNN), which incorporates a deep hierarchical visual feature extractor and a temporal sequence learning model. In order to extract visual features effectively, we propose a joint learning framework with knowledge transfer across multi-tasks to address the insufficiency issue of limited training data. Extensive experiments on different US standard planes with hundreds of videos corroborate that our method can achieve promising results, which outperform state-of-the-art methods.

1 Introduction

Obstetric ultrasound (US) examination generally involves the procedures of image scanning, standard plane selection, biometric measurement and diagnosis. Accurate acquisition of US standard planes, e.g., fetal abdominal standard plane (FASP), fetal face axial standard plane (FFASP) and fetal four-chamber view standard plane (FFVSP) of heart, is one crucial step for the subsequent biometric measurement and obstetric diagnosis. Clinically, US standard plane is manually acquired by searching the view with concurrent presence of key anatomical structures (KASs) in the regions of interest (ROI) [1]. Fig. 1 illustrates the KASs for

[⋆] Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 507–514, 2015.
DOI: 10.1007/978-3-319-24553-9_62

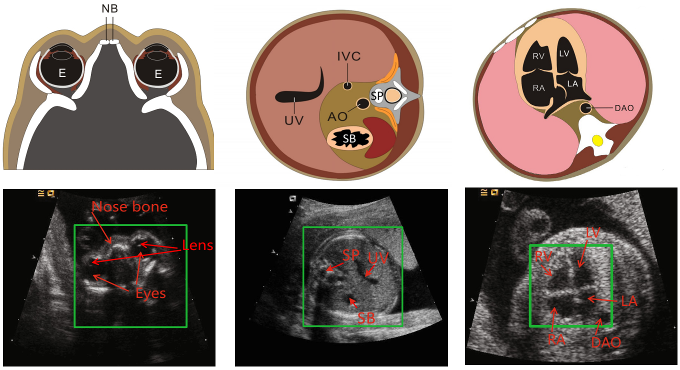

Fig. 1. Left: FFASP containing nose bone, eyes and lens; middle: FASP containing stomach bubble (SB), umbilical vein (UV) and spine (SP); right: FFVSP containing left atrium (LA), right atrium (RA), left ventricle (LV), right ventricle (RV) and descending aorta (DAO) (green rectangles denote the ROIs).

FFASP, FASP and FFVSP, respectively. The manual acquisition of standard planes heavily relies on clinical experience and is also very laborious. Hence, automatic detection methods are highly demanded to boost the examination efficiency [2]. However, this computerized detection task is quite challenging due to the high intra-class variations of US standard planes resulting from acoustic shadows, deformations of soft tissues and various transducer orientations [3].

Over the past few years, several methods have been proposed to address this challenging problem. Most of them either utilized hand-crafted features by observation [2,3,4] or incorporated component-based geometric constraints for a specific standard plane detection task, e.g., the radial component model and vessel probability map detection (RVD) method in [5]. However, these low level features may not accurately represent the complicated characteristics of standard planes. In addition, the insufficiency of training data in the medical domain usually leads to the overfitting problem in supervised learning based methods, and hence degrades the generalization performance. Chen *et al.* [6] compared the performance of randomly initialized convolutional neural network (R-CNN) and that of transferred convolutional neural network (T-CNN) on FASP detection. The method of T-CNN achieved a high accuracy by using deep learning based spatial feature representations with knowledge transfer from natural images. However, the cross-domain knowledge transfer may boost the detection performance with limited improvement due to the larger domain gap. Besides, only considering spatial features may not be the optimal solution, since temporal information of consecutive sequences in US videos could provide extra contextual clues for better discrimination.

Recently, the recurrent neural network (RNN), especially the long short-term memory (LSTM) model, has achieved success in sequence learning tasks, such as speech recognition [7] and video recognition [8]. In order to meet above challenges, we propose a knowledge transferred recurrent neural network (T-RNN) by exploring spatio-temporal feature learning. The major contributions of this paper are three-fold. First, to our best knowledge, this is the first work that considers spatio-temporal feature representations under the framework of deep learning for the detection of standard planes from US videos. Second, a joint learning model for effective spatial feature learning across multi-tasks is presented, which reduces the overfitting problem caused by the inadequacy of training data. Third, the proposed T-RNN is a general framework and can be easily extended to other US standard plane or anatomical structure detection problems. Extensive experiments on different US standard plane detection tasks with large scale datasets demonstrated the efficacy of our method.

2 Method

Fig. 2 (left) shows the architecture of the proposed T-RNN, which is a hybrid model integrating deep convolutional neural networks (CNN) and recurrent neural networks (LSTM model). A ROI classifier is first trained based on the joint learning of convolutional neural networks (J-CNN) across multi-tasks to locate the most discriminative regions for US standard plane detection. Then, the temporal information is explored via the LSTM model based on the features of ROIs in consecutive frames extracted from the J-CNN model. Finally, the score of each frame is obtained by averaging all predictions from the LSTM model and the frame is classified as the standard plane when the output score is larger than a threshold T_0.

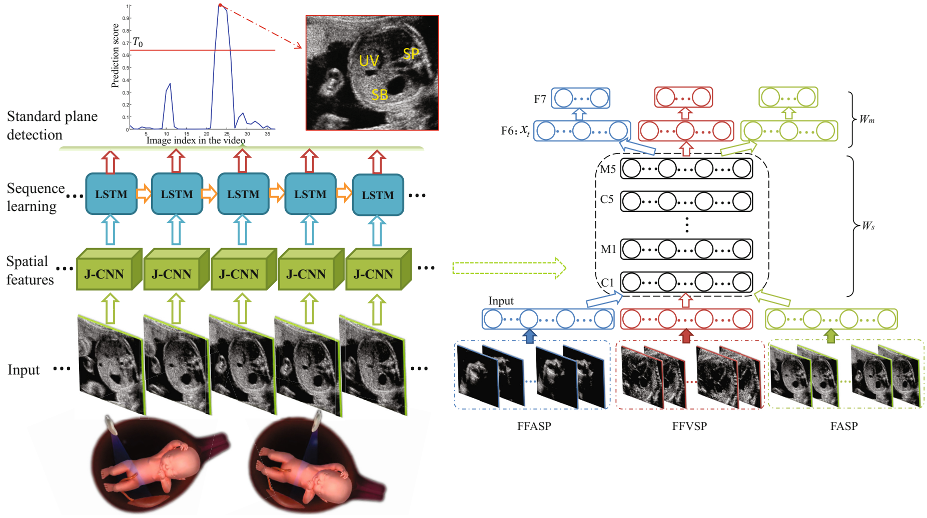

Fig. 2. Left: architecture of the proposed T-RNN; right: the proposed J-CNN.

2.1 Joint Learning with Knowledge Transfer across Multi-tasks

The basic structure of CNN includes several pairs of alternating convolutional (C) and max-pooling (M) layers, followed by fully-connected (F) layers. Previous studies have indicated that the knowledge learned from one domain or task via CNN could benefit the training for another domain or task with limited annotated data [6]. Inspired by these studies, it is reasonable to speculate that leveraging the transferred knowledge across similar US detection tasks can mitigate the challenge of insufficient training data for a specific task as well as improve the generalization performance of the learning. To the end, we propose a joint learning model with CNN across multiple detection tasks of US standard planes, as illustrated in Fig. 2 (right).

In the figure, the matrix W_s denoting the parameters of layers from C1 to M5 is trained from all training samples of the three detection tasks and shared among these tasks. The W_m ($m = 1, 2, 3$ represents the task of FFASP, FFVSP and FASP, respectively) denotes the parameters of F6 and F7 layers and is trained individually on each task for the discrimination of different standard planes. These parameters can be optimized by minimizing the following joint max-margin loss function $\mathcal{L}_1$:

$$\mathcal{L}_1 = \frac{\lambda}{2}(\sum_m ||W_m||_2^2 + ||W_s||_2^2) + \sum_m \sum_k max(0, 1 - y_{mk} F_m(f_{mk}^s; W_m))^2 \quad (1)$$

$$f_{mk}^s = F_s(I_{mk}; W_s) \quad (2)$$

where the first part of $\mathcal{L}_1$ is the regularization penalty term and the second part is the data loss term. The tradeoff between these two terms is controlled by the hyperparameter λ, which is determined by cross-validation in our experiments. The F_s denotes the shared feature extraction function while the F_m denotes the discriminant function for different US standard planes individually. The I_{mk} is the kth input frame of mth task and f_{mk}^s is the output of shared section (i.e., activations of M5 layer). The $y_{mk} \in \{-1, 1\}$ is the corresponding ground truth. The architecture of J-CNN model can be seen in Table 1 (padding and non-linear activation layers are not shown).

Table 1. Architecture of J-CNN

Layer	Feature maps	Kernel size	Stride
input	227x227x1	-	-
C1	55x55x24	11	4
M1	27x27x24	3	2
C2	14x14x24	5	2
M2	7x7x24	3	2
C3	7x7x24	3	1
C4	7x7x24	3	1
C5	7x7x24	3	1
M5	3x3x24	3	2
F6	100	-	-
F7	2	-	-

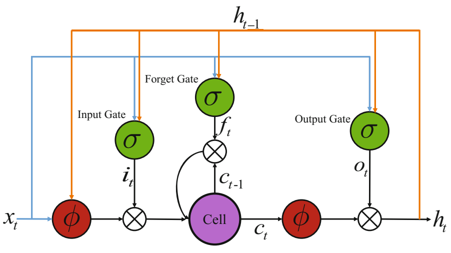

Fig. 3. LSTM model

2.2 US Standard Plane Detection via T-RNN

Temporal information in time-series videos could provide additional contextual clues for the improvement of detection performance. In our T-RNN model, spatio-temporal features in ROIs, which have been detected by the J-CNN model, are further explored by the LSTM. Given the input frame I_{mk}, the probability map of the ROI is computed by the J-CNN model in a sliding window way and the center of the ROI is located at the position with maximal value in the probability map. Features in the penultimate layer (i.e., activations of F6 layer) of the J-CNN model are then extracted from the ROI of each frame. Before inputting features into the LSTM, we manually clip each video into separated clips with the same number of T frames. Thus, each input video can be transformed into sequenced samples, where each sample is represented by a vector sequence $\mathbf{x} = \{x_1, ..., x_t, ..., x_T\}$ and $x_t \in \mathbb{R}^q$ ($q = 100$ in our experiments). The corresponding labelling vector is $\mathbf{y} = \{y_1, ..., y_t, ..., y_T\}$, where $y_t \in \{0, 1\}$.

In the traditional RNN, the back-propagation algorithm may result in vanishing or exploding gradients. The LSTM model tackles this problem by incorporating memory cells that allow the network to learn when to forget previous hidden states and when to update hidden states given the new input [7]. A simplified version of LSTM model is shown in Fig. 3. The element-wise nonlinear functions $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ squash their inputs into [0,1] and [-1,1], respectively. The gates serve to modulate the interactions between the memory cell c_t and its environment. The input gate i_t can allow incoming input x_t to alter the state of the memory cell or block it. The output gate o_t can allow the state of the memory cell to have an effect on hidden neurons or prevent it. The forget gate f_t can modulate the self-recurrent connection of the memory cell, allowing the cell to remember or forget its previous state c_{t-1}. All the gates and memory cells have the same vector size with hidden state $h_t \in \mathbb{R}^H$ (H is the number of hidden units). Specifically, they are updated with following equations:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ h_t &= o_t \odot \phi(c_t) \end{aligned} \tag{3}$$

where $h_0 = 0$, W denotes the weight matrix (e.g., W_{xi} is the input-input gate matrix and W_{hi} is the hidden-input gate matrix), b is corresponding bias term, and $\odot$ denotes the element-wise multiplication. The predictions can be obtained by feeding h_t into a softmax classification layer. Thus, the parameters θ (including all W and b) of the model can be trained by minimizing the negative logarithm loss function with stochastic gradient descent method [9]:

$$\mathcal{L}_2 = -\sum_{n=1}^{N}\sum_{t=1}^{T} \log p_n(y_t|x_t, h_{t-1}; \theta) \tag{4}$$

where N is the total number of sequenced training samples after clipping.

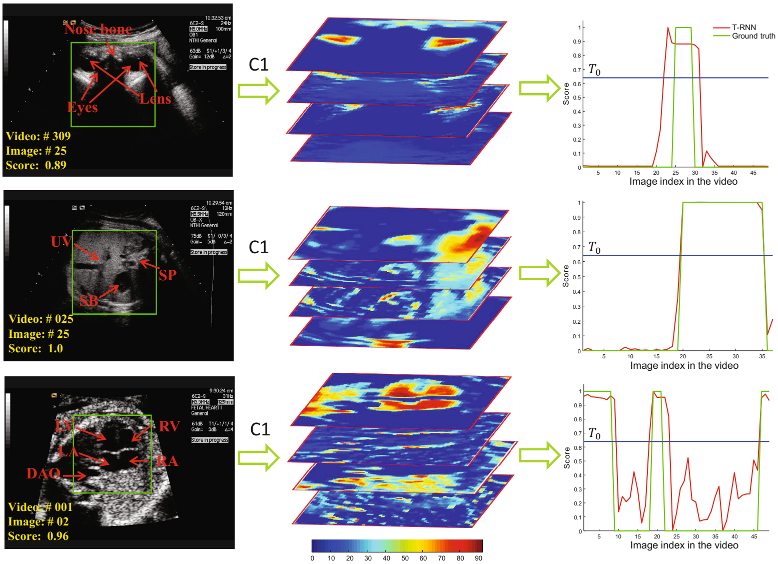

Fig. 4. Left: typical US standard plane detection results; middle: several feature maps of ROIs in C1 layer; right: sequenced predictions in the video.

3 Experiments and Results

Materials. Ultrasound videos were acquired by performing a conventional US sweep on the pregnant women (fetal gestational age from 18 to 40 weeks) in the supine position using a Siemens Acuson Sequoia 512 US scanner. Each video was acquired from one patient and contained 17-48 frames. They were manually annotated by an experienced obstetrician. For training the ROI classifier under the framework of J-CNN, training samples of FASP, FFASP and FFVSP were generated from 300 videos with a total of 11,942, 13,091 and 12,343 US images, respectively. In addition, 219 videos with 8718 US images of FASP, 52 videos with 2278 images of FFASP and 60 videos with 2252 images of FFVSP were used for the performance evaluation, respectively.

Qualitative Performance Evaluation. Fig. 4 (left) shows the typical detection results of three US standard planes. All the detected standard planes contained the KASs and the predicted scores were above the threshold T_0 (determined with cross validation). In addition, we input the ROIs of the detected standard planes into the J-CNN and visualized their feature maps in C1 layer in Fig. 4 (middle). It is observed that large responses of feature maps were excited in the regions of KASs, revealing the model captured the discriminative structures. Furthermore, as shown in Fig. 4 (right), the whole sequenced predictions of three videos by T-RNN demonstrated a good consistency with ground truth.

Table 2. Results of Standard Plane Detection

Method	FASP				FFASP				FFVSP			
	A	P	R	$F1$	A	P	R	$F1$	A	P	R	$F1$
T-RNN	**0.908**	**0.748**	**0.747**	**0.747**	**0.867**	**0.634**	**0.598**	**0.615**	**0.867**	**0.770**	**0.612**	**0.682**
J-CNN	0.902	0.729	0.739	0.734	0.854	0.605	0.513	0.555	0.835	0.718	0.611	0.660
T-CNN[6]	0.896	0.714	0.710	0.712	0.847	0.582	0.503	0.535	0.831	0.708	0.606	0.653
R-CNN[6]	0.857	0.594	0.681	0.635	0.831	0.530	0.443	0.482	0.826	0.688	0.608	0.651
RVD[5]	0.833	0.532	0.693	0.602	-	-	-	-	-	-	-	-

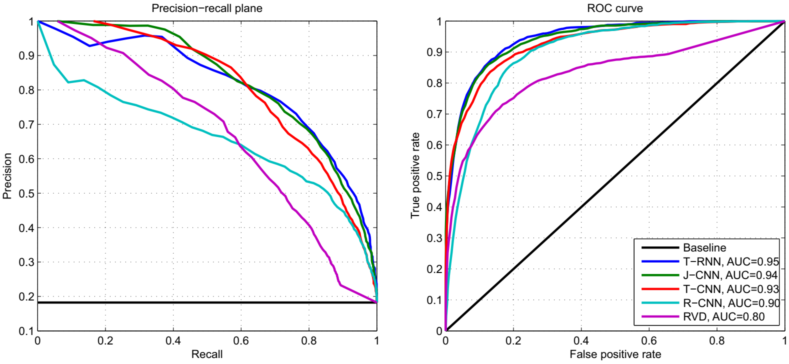

Fig. 5. The PR plane and ROC curves of different methods on FASP detection.

Comparison of Quantitative Performance. We compared our method with state-of-the-art methods [5,6] and the J-CNN that relies only on the spatial feature representations from three detection tasks. The evaluation measurements include *accuracy* (A), *precision* (P), *recall* (R), and $F1$ score. The results of different methods are shown in Table 2. The four deep learning based methods achieved better results than the method of RVD [5] on the FASP detection, which evidenced the efficacy of exploiting deep learning based feature representations. The detection results of J-CNN and T-CNN [6] outperformed those of R-CNN [6] on most measurements, demonstrating the advantages of the knowledge transfer strategy on reducing overfitting caused by the inadequacy of training data. In addition, the results of J-CNN were better than those of T-CNN, indicating that the knowledge transferred from images of the same domain reduced the gap between cross-domains (e.g., natural images used in the T-CNN). Compared with other methods, our T-RNN achieved the best performance on different measurements, which further highlighted the superiority of exploring spatio-temporal feature learning with knowledge transfer in standard plane detection from US videos. The precision-recall (PR) plane and receiver operating characteristic (ROC) curves of different methods on FASP detection are shown in Fig. 5, further demonstrating the advantages of the proposed T-RNN. The T-RNN method generally took less than 1 minute to detect the standard planes from a video containing 40 frames using a workstation equipped with a 2.50 GHz Intel(R) Xeon(R) E5-2609 CPU and a NVIDIA Titan GPU.

4 Conclusion

In this paper, we presented a knowledge transferred RNN to automatically detect fetal standard planes from US videos by exploring spatio-temporal feature learning. Experimental results on three US standard planes demonstrate the efficacy of our approach quantitatively on this challenging problem. Furthermore, our approach is a general framework and can be extended to the detection of other US standard planes or anatomical structures. In the future, we will accelerate the detection process and apply it in clinical practice.

Acknowledgements. The work described in this paper was supported in part by the Shenzhen Key Basic Research Project (No. JCYJ20140509172609164) and in part by the Shenzhen-Hong Kong Innovation Circle Funding Program (Nos. GHP/002/13SZ and SGLH20131010151755080). We also gratefully acknowledge the NVIDIA Corporation for the donated GPU used in this research.

References

1. Chen, H., Ni, D., Yang, X., Li, S., Heng, P.A.: Fetal abdominal standard plane localization through representation learning with knowledge transfer. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 125–132. Springer, Heidelberg (2014)
2. Kwitt, R., Vasconcelos, N., Razzaque, S., Aylward, S.: Localizing target structures in ultrasound video–a phantom study. Medical Image Analysis 17(7), 712–722 (2013)
3. Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A.: Searching for structures of interest in an ultrasound video sequence. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 133–140. Springer, Heidelberg (2014)
4. Rahmatullah, B., Papageorghiou, A.T., Noble, J.A.: Integration of local and global features for anatomical object detection in ultrasound. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 402–409. Springer, Heidelberg (2012)
5. Ni, D., Yang, X., Chen, X., Chin, C.-T., Chen, S., Heng, P.A., Li, S., Qin, J., Wang, T.: Standard plane localization in ultrasound by radial component model and selective search. Ultrasound in Medicine & Biology 40(11), 2728–2742 (2014)
6. Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.: Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE Journal of Biomedical and Health Informatics (2015)
7. Graves, A.: Supervised Sequence Labell. with Recur. Neur. Networks. SCI, vol. 385, pp. 5–13. Springer, Heidelberg (2012)
8. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint:1411.4389 (2014)
9. Williams, R.J., Zipser, D.: Gradient-based learning algorithms for recurrent networks and their computational complexity. In: Back-Propagation: Theory, Architectures and Applications, pp. 433–486 (1995)