Ultrasound Standard Plane Localization via Spatio-Temporal Feature Learning with Knowledge Transfer

Hao Chen [1], **Dong Ni** [2,1,*], **Lingyun Wu** [2], **Jing Qin** [2,4], **Shengli Li** [3], **Pheng Ann Heng** [1,4]
[1] Dept. of Computer Science and Engineering, The Chinese University of Hong Kong
[2] National Key Laboratory for Medical Ultrasound, Shenzhen University
[3] Dept. of Ultrasound, Shenzhen Maternal and Child Healthcare Hospital
[4] Human Computer Interaction Research Center,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Abstract

Acquisition of ultrasound standard planes containing key anatomical structures is crucial for subsequent clinical diagnosis and biometric measurements. However, it requires a thorough knowledge of human anatomy and expensive manual labor. Unlike previous works that designed different methods for various anatomical standard planes respectively, we propose a general framework to automatically locate standard planes from consecutive 2D ultrasound images. Furthermore, we analyze the shared statistical strength between medical and natural image modalities, which provides the evidence for knowledge transfer strategy. Instead of utilizing low-level features, we propose a *M*ulti-stage deep neural network to explore the *S*patio-*T*emporal feature representation with knowledge transfer; we name it as *MST-net*. Extensive experiments on a large number of ultrasound images bear out that our method can achieve high localization accuracy, which outperforms state-of-the-art methods significantly. The performance on different ultrasound standard planes demonstrates the generalization capability of the proposed framework for various clinical applications.

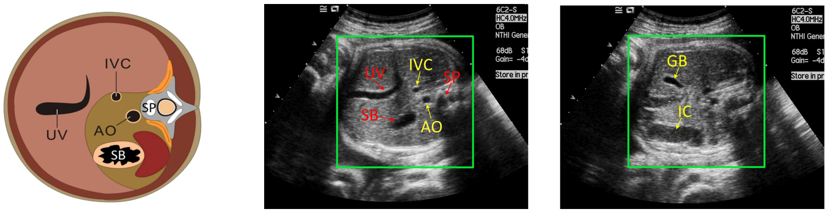
1 Introduction

Ultrasound (US) imaging is widely used in obstetric diagnosis with advantages of low-cost, conveniently portable and real time imaging capability. Generally, US-based pregnancy diagnosis includes the following procedures: image scanning, US standard plane selection, biometric measurement and diagnosis [1]. The accurate acquisition of the ultrasound standard plane e.g. fetal abdominal standard plane (FASP) and fetal face axial standard plane (FFASP) is crucial for the biometric measurements and ultrasound diagnosis. In clinical practice, the ultrasound standard plane is manually acquired with the presence of key anatomical structures (KASs). For example, the FASP as shown in Fig. 1, is determined by three KASs: stomach bubble (SB), umbilical vein (UV), and spine (SP) located in the region of interest (ROI) by experienced clinicians [13]. However, this process requires a thorough knowledge of human anatomy and substantial clinical experience, which makes it challenging for novices and time-consuming for experts. Hence, the development of automatic methods for locating the ultrasound standard plane would enhance the ability of non-experts to operate US devices and improve the examination efficiency for experts [10]. However, this task is very challenging for several reasons. First, the ultrasound standard plane often has high intra-class variations due to acoustic shadows, deformations and scanning orientations [18]. Second, some anatomical structures often carry similar appearance to the KASs. In the case of FASP as shown in

---

[*] Corresponding author: nidong@szu.edu.cn

Fig. 1, abdominal aorta (AO), gall bladder (GB), intestinal canal (IC) and inferior vena cava (IVC) are difficult to be distinguished from KASs, even for some experienced obstetric experts.



(a) (b) (c)

Figure 1: (a) Fetal abdominal anatomy, (b) True FASP, (c) False FASP with similar anatomical structure GB and IC (ROI marked with green rectangle).

Over the past few years, a number of studies have contributed to the automatic localization of standard plane from 2D US images. Zhang et al. [22] proposed to locate the standard plane of early gestational sac (SPGS) from US videos by utilizing cascade AdaBoost classifiers. Kwitt et al. [10] explored the use of a kernel dynamic textures (KDTs) model to locate target structures by augmenting two configurations: raw intensity values and mid-level Bag of Words (BoW) representation. Ni and Yang et al. [21, 14] presented a hierarchically supervised learning framework to locate FASP via a radial component-based detection (RCD) model with geometric constrains and selective search.

Although previous methods presented their efficacy in locating standard planes from 2D US images, the main limitation was that they only considered hand-crafted features by observation and experiences. Recently, deep neural networks have made breakthroughs in object recognition with expressive feature representation [2, 8, 9, 11]. Instead of manually designing a feature detector according to different tasks, the feature representation learned from deep convolutional neural networks (DCNN) have been shown to work well when re-applied to generic tasks across different domains [20, 4]. Another limitation of previous works was that relatively insufficient training data in the medical domain may lead to the overfitting problem and degrade the learning performance. Knowledge transfer has been proved to be able to improve the performance by making use of the knowledge obtained in cross domains [5, 17]. Chen et al. [3] achieved a high accuracy of FASP localization by utilizing the deep learning based spatial feature representation with knowledge transfer. However, only considering spatial feature representation may not be the optimal solution since temporal information in time-series video could provide more context for discrimination. Previous work [8] gained significant improvement on video action recognition by taking advantage of CNN with spatio-temporal information. Since regions outside of ROI in US images are useless for FASP decision, input entire consecutive frames into neural network can be computationally expensive and less accurate. Therefore, we propose a multi-stage framework by exploiting the spatio-temporal feature representation (MST-net), which divides this localization problem into separate stages and conquer them respectively. Different with [3], we first analyze the shared statistical strength between medical and natural images in low-level structural patterns, which provides the evidence for the strategy of knowledge transfer. In addition, temporal information is carefully considered in our multi-stage framework. Experimental results on different US standard planes, including fetal abdomen, face and four-chamber view of heart prove the effectiveness and generalization capability of our proposed framework.

The outline of this paper is as follows. Section 2 analyzes the shared statistical strength between medical and natural images. Section 3 takes FASP localization as an example and describes our multi-stage localization framework. Experimental results are evaluated qualitatively and quantitatively in Section 4. Section 5 validates the generalization ability of our approach on other US standard planes and Section 6 concludes the paper.

2 Ultrasound image representation with natural bases

Although medical and natural images are two different modalities and high level abstraction information is distinct (e.g. fetal abdomen and wild cat in Fig. 2), they do share statistical strength in low-level structural patterns. A previous study [6] utilized cross-domain features to represent Magnetic Resonance Imaging (MRI) data and achieved competitive performance in classifying

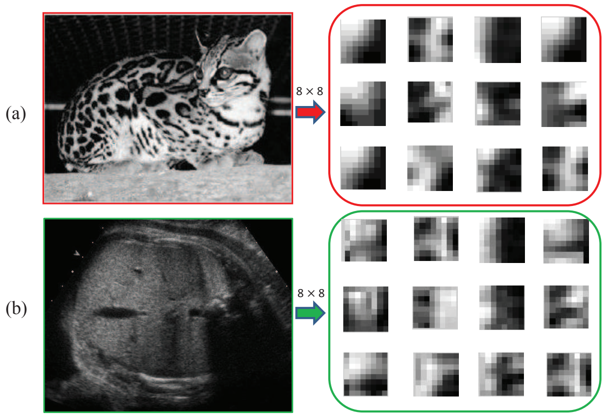
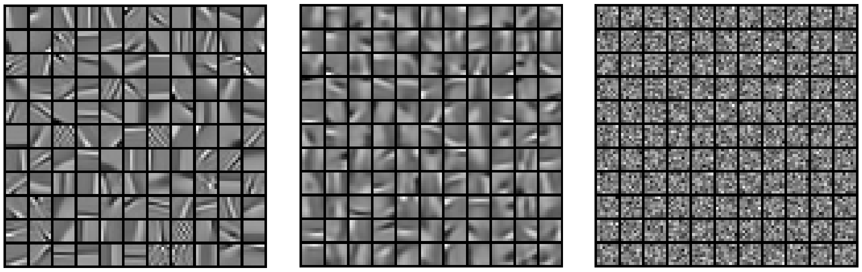


Figure 2: (a) Natural image of wild cat, (b) Ultrasound image of fetal abdomen.



(a) (b) (c)

Figure 3: Dictionary bases: (a) Natural images, (b) Ultrasound images, (c) Gaussian noise.

Alzheimer diseases. Different from this study, which represented MRI data empirically, we gave a quantitative analysis of the shared statistical strength between two modalities with dictionary learning. We reconstructed the medical test data $test_m$ by the learned natural dictionary bases D_n and self-modality learned dictionary bases D_m, respectively. If the reconstruction errors of these two approaches are approximately equal and small, we conclude that dictionary bases learned from natural image domain could reconstruct medical data well, which means knowledge learned from low level patches of natural images could be transferred to medical modality. The dictionary bases D_m and D_n are first learned by minimizing Eq. 1 on the low level patches of medical and natural images. Then patches of one modality are reconstructed by the dictionary bases learned from patches of self-domain modality or cross-domain modality. And patches of a normal distribution noise t_{noise} are also generated for comparison,.

$$min_{\alpha_i, D} \sum_{i=1}^{n} ||x_i - D\alpha_i||_2^2 + \lambda\psi(\alpha_i) \;\; s.t. \; d_k^T d_k \leq C, \forall \; k = 1, ..., q \tag{1}$$

where $x_i \in \mathbb{R}^p$ is the i^{th} sample patch of training set $\Gamma_1 = \{x_1, ..., x_n\}$. Each x_i is a sparse linear combination over a set of basis vectors from an over-complete dictionary $D \in \mathbb{R}^{p \times q}$ with its corresponding sparse coefficient $\alpha_i \in \mathbb{R}^q$. The d_k denotes k^{th} column of dictionary D, $\psi(\alpha_i)$ is a sparsity-inducing regularizer (e.g. L_1 norm $\psi(\alpha_i) = ||\alpha_i||_1$) and λ, C are constants. The reconstruction residual error is defined by Eq. 2, and α_j is solved by minimizing Eq. 1 with fixed D on testing set $\Gamma_2 = \{x_1, ..., x_m\}$.

$$e = \frac{1}{2m} \sum_{j=1}^{m} ||x_j - D\alpha_j||_2^2 \tag{2}$$

We randomly extracted 1,000,000 patches (size 8×8) from natural images (a subset of ILSVRC12 [19]), medical ultrasound images and noise data, respectively. The learned dictionaries are visualized in Fig. 3, both D_n and D_m resemble the receptive fields of neurons in mammalian primary

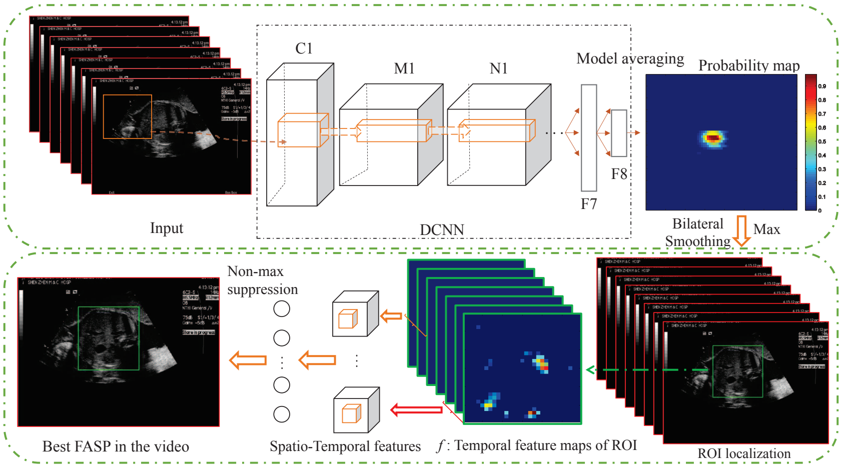


Figure 4: Overview of the proposed framework.

visual cortex [15, 16]. Reconstruction residual errors in Table 1 show the dictionary bases trained on natural images D_n can reconstruct medical data $test_m$ well, but perform poorly on the reconstruction of noise data. This certifies our hypothesis that the learning bases from natural images could help to discover the underlying distribution of medical data, thus benefit related medical tasks. An interesting discovery is that the learned dictionary bases D_m could not reconstruct the natural data $test_n$ well, we conjecture that D_n is more overcomplete than D_m since the natural image is usually more informative than medical image.

Table 1: Reconstruction errors of different modalities

Data	D_m	D_n	D_{noise}
$test_m$	**0.111**	**0.114**	0.263
$test_n$	0.186	0.158	0.263
t_{noise}	0.405	0.338	0.263

3 Multi-stage neural network for US standard plane localization

The pipeline of our proposed method consists of two sequential stages as shown in Fig. 4. In the first stage, the DCNN ROI classifier is trained with knowledge transfer on the extracted ROIs of fetal abdomen from the expert-annotated US images. Given an input image, the probability map M_p is produced by model averaging. Subsequently it's further smoothed by bilateral filter for outlier elimination and ROI of fetal abdomen is located at position C_m. For the second stage, spatio-temporal features are extracted in the ROI and input into the convolution neural network. Finally, the test image is classified as the FASP when the output score is larger than one threshold. The best FASP in the video is located by the non-max suppression.

3.1 Deep learning based feature representation with knowledge transfer

Previous studies [4, 20] have indicated that the pre-trained model in other domains could well initialize the training of the DCNN in the domain of interest. But how to adapt it into medical applications with knowledge transfer hasn't been well studied. Based on the analysis in Section 2, we adapted the model trained on ImageNet [7] into our DCNN architecture by initializing the parameters of previous layers with the pre-trained model. This process can be seen as a supervised pre-training prior, which disentangles variation factors for our medical task and reduces overfitting due to relatively limited medical data. Then the model was subsequently refined on the medical training data with softmax regression:

$$p(c = k|f) = \frac{e^{o(f_k)}}{\sum_{l=1}^{K} e^{o(f_c)}} \tag{3}$$

where $o(f)$ is the output of neural network, and $p(c = k|f)$ is the kth-class posterior probability given the feature vector f.

3.2 Compute probability map of ROI with DCNN

Given the input image I, the probability map M_p was computed by the trained DCNN ROI classifier in a sliding window way. Specifically, each sliding window was augmented by cropping the center and corners of the sliding window as well as its mirrored versions, resulting in 10 inputs to the DCNN. The final score of the sliding window was obtained by averaging the scores of these 10 inputs. Ideally, we want M_p to be zero everywhere except at the center of the ROI of FASP, but it could be noisy in practice. In this regard, M_p was further smoothed with a bilateral filter. Then the maximum probability value at the position C_m was calculated by performing the non-max suppression on the smoothed probability map M_s, defined in Eq. 4. If $M_s(C_m)$ was larger than the threshold value t obtained by the cross validation ($t = 0.68$ in our experiments), the test image was regarded as containing ROI of FASP, or not if smaller.

$$\begin{aligned} M_s(x, y) &= \frac{\sum_{x_i, y_i \in \Omega} M_p(x_i, y_i) w_i}{\sum_{x_i, y_i \in \Omega} w_i} \\ w_i &= f_r(||M_p(x_i, y_i) - M_p(x, y)||_2^2) g_s(||x_i - x||_2^2 + ||y_i - y||_2^2) \\ C_m &= \operatorname*{argmax}_{x,y} M_s(x, y) \end{aligned} \tag{4}$$

Here, (x, y) is the pixel coordinate, Ω is the window centered at (x, y), f_r is the range smoothing kernel and g_s is the spatial smoothing kernel.

3.3 Standard plane localization with spatio-temporal features

Spatio-temporal features in time-series video could provide more context information for the localization of US standard planes. However, the ROI containing anatomical structures is most discriminatively informative for FASP assessment while regions outside of ROI are redundant and uselss. Therefore, spatio-temporal features of DCNN in ROI are extracted and input into the convolutional neural network for classification. Specifically, the ROIs are localized at C_m on the previous output of DCNN. The deep learning based feature maps are cropped in the ROIs and resized as 3D temporal volume ($27 \times 27 \times n_t$, where $n_t = 38 \sim 45$ is the number of time-series frames). And the neural network (C1: 5×5, M1: 3×3, N1: 3×3, C2: 3×3, M2: 2×2, C for convolution, M for max pooling, N for normalization) is trained on the extracted spatio-temporal features f_n (one sample is a 3D volume: $27 \times 27 \times m$, m is the number of consecutive frames and set as 3 in our experiment) by minimizing the loss function:

$$\mathcal{L} = \sum_{n=1}^{N} ||r(f_n \otimes w) - t_n||_2^2 + \xi \cdot \gamma(w) \tag{5}$$

where first term is sum-of-squares error between output of MST-net $r(f_n \otimes w)$ and ground truth t_n, and $\gamma(w)$ is a weight decay term controlled by ξ for regularization.

4 Experiments and Results

4.1 Materials

Ultrasound images were acquired by performing the conventional US sweep on the pregnant women (fetal gestational age from 18 to 40 weeks) using Siemens Acuson Sequoia 512 US scanner. For DCNN ROI classifier, we first generated the training samples (1911 positive and 3160 negative ones) by manually extracting the ROI of fetal abdomen from the expert-annotated US images. Note that the training samples were further rotated and mirrored to augment the training database. In addition, 219 videos with total 8718 US images were manually labeled for the performance evaluation by a clinical radiologist with more than five years experience in US obstetrics.

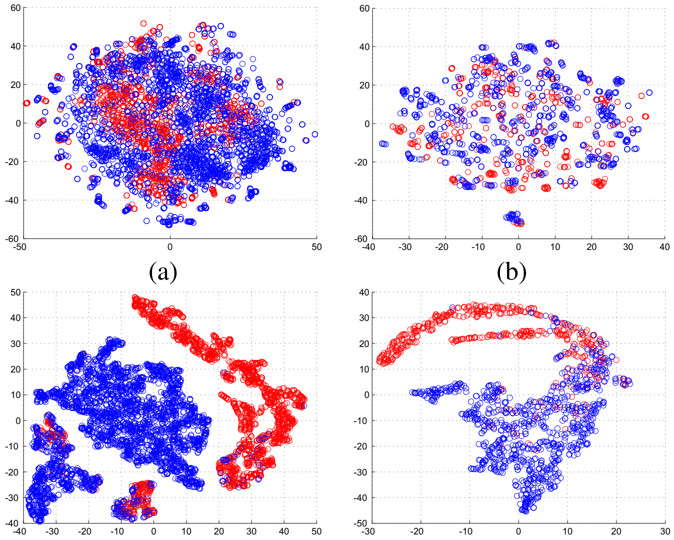


Figure 5: Feature embedding and visualization (red and blue points represent FASP and non-FASP respectively). (a) Raw training data, (b) Raw testing data, (c) Intermediate layer of training data, (d) Intermediate layer of testing data.

4.2 Qualitative performance evaluation

In order to show the powerful feature representation of the DCNN, we first illustrate the automatically learned features of the intermediate layers before output by reducing the dimensions of the features utilizing the Barnes-Hut Stochastic Neighbor Embedding (BH-SNE) method [12]. As shown in Fig. 5 (a) and (b), the low inter-class difference of raw data between FASP and nonFASP makes the classification very challenging. However, Fig. 5 (c) and (d) show that the automatically learned features make it easier to classify the FASP and nonFASP. This result visually certifies that extracted features encoding high level information could disentangle the variation factors and benefit more to our task than using the original data. Fig. 6 (a) and (b) show two typical FASPs correctly classified by our method, where the KASs including UV, SB and SP are contained. Fig. 6 (e) and (f) are the corresponding probability maps of Fig. 6 (a) and (b), respectively. Fig. 6 (c) shows one false FASP classified by our method, due to its similar appearance with the true FASP.

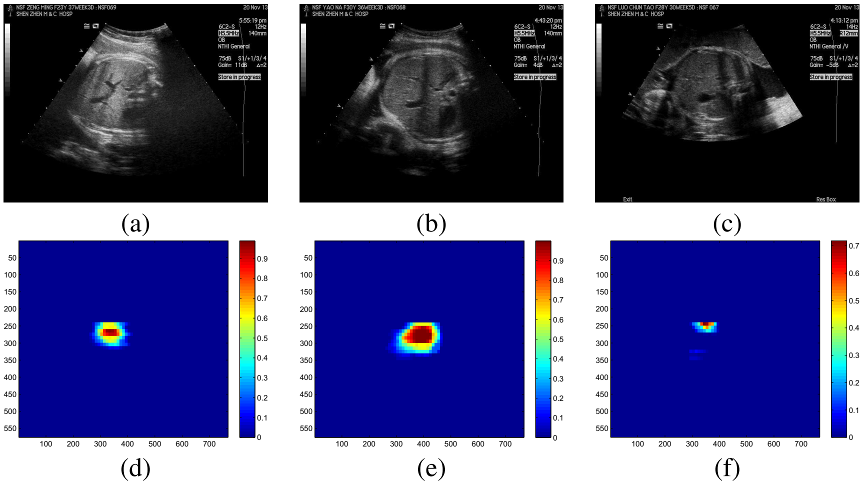
4.3 Comparison of Quantitative Performance

Table 2: Results of FASP Localization in US Images and Videos

Method	A_i	A_v
MST-net	**0.918**	**0.927**
DCNN with pre-train [3]	0.910	0.904
DCNN without pre-train	0.857	0.822
RCD [14, 21]	0.775	0.762

A_i : accuracy of US images, A_v: accuracy of US videos

We compared the performance of our method on the US images and videos with other methods [3, 14, 21]. For the localization of FASP from one video, the US image with the highest score was selected as the FASP. In [14, 21], a novel radial component-based detection (RCD) framework was developed to improve the localization performance. In order to show the efficacy of the knowledge transfer strategy employed in our method, we further compared the performance of our method with the DCNN method without using the pre-trained parameters. The Precision-Recall plane and Receiver Operating Characteristic (ROC) curve are shown in Fig. 7. The result of the DCNN method with pre-training is better than the RCD method and DCNN without pre-training, which proves the



(a) (b) (c)

(d) (e) (f)

Figure 6: Examples of ROI probability maps: the first row is original US images, the second row is the corresponding ROI probability maps (the color bar indicates the probability value). (a)-(b) are two typical true FASPs, (c) is one false FASP.

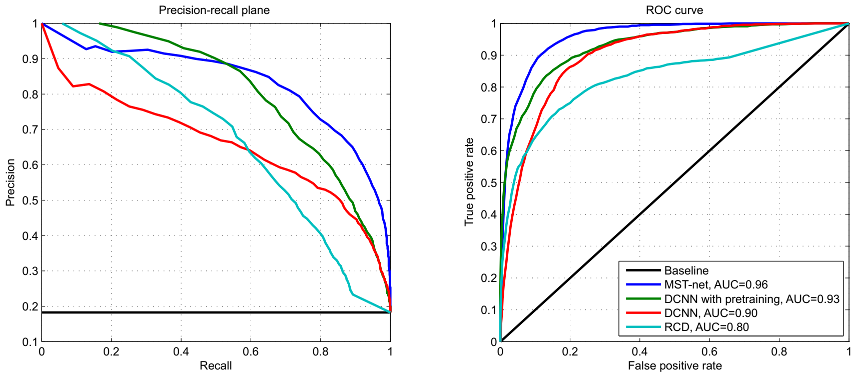
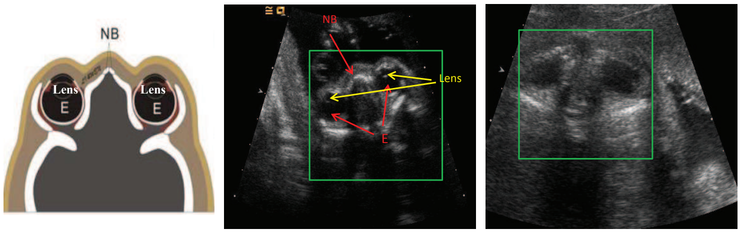


Figure 7: Precision-recall plane (left) and ROC curve (right) of different methods.

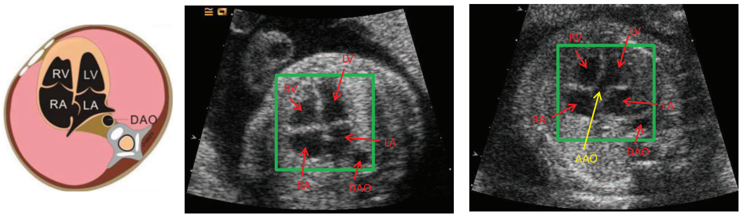
efficacy of both the deep learning based feature representation and the knowledge transfer strategy. As shown in Table 2, MST-net achieves the best performance on US images and videos with classification accuracies 0.918 and 0.927, respectively. In addition, compared with pre-trained DCNN, MST-net gains obvious improvement on video accuracy, which further indicates spatio-temporal features benefit the task of US standard plane localization with more contextual information.

5 Generalization to other US standard plane localization



(a) (b) (c)

Figure 8: (a) Fetal face anatomy, (b) True FFASP, (c) False FFASP due to lack of NB and Lens.



(a) (b) (c)

Figure 9: (a) Fetal four-chamber view anatomy, (b) True FFVSP, (c) False FFVSP.

For validating the generalization capability of our approach, we extend it to other US standard planes i.e. Fetal Face Axial Standard Plane (FFASP) and Fetal Four-chamber View Standard Plane (FFVSP) of heart. The anatomical structure of FFASP and FFVSP can be seen in Fig. 8 and Fig. 9. The FFASP is determined by three KASs: Nose Bone (NB), Eyes (E) and Lens. The FFVSP is determined by five KASs: left atrium (LA), right atrium (RA), left ventricle (LV), right ventricle (RV) and descending aorta (DAO). A false FFVSP is shown in Fig. 9 (c) because an ascending aortic (AAO) connects to LV, which increases the confusion with true FFVSP. In the training process, the FFASP training samples were generated from 300 videos with a total of 13,091 images and FFVSP training samples were generated from 300 videos with a total of 12,343 images. Performance of our method on FFASP and FFVSP localization were evaluated on 52 videos with 2278 images and 60 videos with 2252 images, respectively. Results in Table 3 show that MST-net outperforms other methods by a large margin on different US standard planes from US images and videos, especially for FFVSP with much closer temporal information in consecutive hear-beating frames. It further demonstrates that our approach can be extended to localization of other US standard planes.

Table 3: Results of FFASP and FFVSP Localization in US Images and Videos

Method	FFASP		FFVSP	
	A_i	A_v	A_i	A_v
MST-net	**0.925**	**0.883**	**0.857**	**0.852**
DCNN with pre-train [3]	0.876	0.817	0.784	0.75
DCNN without pre-train	0.844	0.767	0.756	0.731

A_i : accuracy of US images, A_v: accuracy of US videos

6 Conclusions

In this paper, we presented a multi-stage framework to automatically locate the standard plane from US images and videos by exploring the spatio-temporal features with deep neural network and knowledge transfer strategy. Experimental results demonstrate the efficacy of our approach on this challenging clinical problem. In the future, we will apply this framework clinically and compare its performance with obstetric experts.

Acknowledgements

The work described in this paper was supported in part by the National Science Foundation of China (No. 61101026 and 81270707), in part by the Shenzhen Key Basic Research Project (No. JCYJ20130329105033277 and 201101013), in part by the Shenzhen-Hong Kong Innovation Circle Funding Program (No. JSE201109150013A) and the Hong Kong Innovation and Technology Fund (Project No. GHP/002/13SZ).

References

[1] J. Bamber and M. Tristam. Diagnostic ultrasound. *Webb S., The Physics of Medical Imaging*, 351, 2012.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[3] H. Chen, D. Ni, X. Yang, S. Li, and P. A. Heng. Fetal abdominal standard plane localization through representation learning with knowledge transfer. In *Machine Learning in Medical Imaging*, pages 125–132. Springer, 2014.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[5] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

[6] A. Gupta, M. Ayhan, and A. Maida. Natural image bases to represent neuroimaging data. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 987–994, 2013.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] R. Kwitt, N. Vasconcelos, S. Razzaque, and S. Aylward. Localizing target structures in ultrasound video–a phantom study. *Medical image analysis*, 17(7):712–722, 2013.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] L. Maaten. Barnes-hut-sne. In *Proceedings of the International Conference on Learning Representations*, 2013.

[13] D. Ni, T. Li, X. Yang, J. Qin, S. Li, C.-T. Chin, S. Ouyang, T. Wang, and S. Chen. Selective search and sequential detection for standard plane localization in ultrasound. In *Abdominal Imaging. Computation and Clinical Applications*, pages 203–211. Springer, 2013.

[14] D. Ni, X. Yang, X. Chen, C.-T. Chin, S. Chen, P. A. Heng, S. Li, J. Qin, and T. Wang. Standard plane localization in ultrasound by radial component model and selective search. *Ultrasound in medicine & biology*, 2014.

[15] B. A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[16] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[17] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[18] B. Rahmatullah, A. T. Papageorghiou, and J. A. Noble. Integration of local and global features for anatomical object detection in ultrasound. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 402–409. Springer, 2012.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.

[20] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[21] X. Yang, D. Ni, J. Qin, S. Li, T. Wang, S. Chen, and P. A. Heng. Standard plane localization in ultrasound by radial component. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 1180–1183. IEEE, 2014.

[22] L. Zhang, S. Chen, C. T. Chin, T. Wang, and S. Li. Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination. *Medical physics*, 39(8):5015–5027, 2012.