

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/275355457>

Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks

Article in IEEE Journal of Biomedical and Health Informatics · April 2015

DOI: 10.1109/JBHI.2015.2425041 · Source: PubMed

CITATIONS

421

READS

1,684

7 authors, including:



Hao Chen

Hong Kong University of Science and Technology

376 PUBLICATIONS 33,007 CITATIONS

SEE PROFILE



Dong Ni

Shenzhen University

354 PUBLICATIONS 9,598 CITATIONS

SEE PROFILE



Jing Qin

The Hong Kong Polytechnic University

347 PUBLICATIONS 20,076 CITATIONS

SEE PROFILE



Xin Yang

Shenzhen University

216 PUBLICATIONS 10,227 CITATIONS

SEE PROFILE

Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks

Hao Chen, *Student Member, IEEE*, Dong Ni*, Jing Qin*, Shengli Li, Xin Yang, Tianfu Wang and Pheng Ann Heng, *Senior Member, IEEE*

Abstract—Automatic localization of the standard plane containing complicated anatomical structures in ultrasound (US) videos remains a challenging problem. In this paper, we present a learning based approach to locate the fetal abdominal standard plane (FASP) in US videos by constructing a domain transferred deep convolutional neural network (CNN). Compared with previous works based on low-level features, our approach is able to represent the complicated appearance of the FASP and hence achieve better classification performance. More importantly, in order to reduce the overfitting problem caused by the small amount of training samples, we propose a transfer learning strategy, which transfers the knowledge in the low layers of a base CNN trained from a large database of natural images to our task-specific CNN. Extensive experiments demonstrate that our approach outperforms the state-of-the-art method for the FASP localization as well as the CNN only trained on the limited US training samples. The proposed approach can be easily extended to other similar medical image computing problems, which often suffer from the insufficient training samples when exploiting the deep CNN to represent high-level features.

Index Terms—Ultrasound, standard plane, deep learning, domain transfer, knowledge transfer, convolutional neural network.

I. INTRODUCTION

ULTRASOUND (US) is a routine screening tool offered to all pregnant women because of its safety, relatively low cost and real-time manner. The main goal of a fetal US scan is to confirm fetal viability, establish gestational age accurately and look for malformation that could influence prenatal management. Recent study showed that the sensitivity for prenatal detection of malformations by US ranges from 27.5% to 96% in different medical institutes [1]. This wide variation indicates that US-based pregnant diagnosis is operator-dependent and requires a significant period of training before reaching competency. Among the pipeline of US diagnosis, acquisition of the standard plane is the prerequisite step and crucial for

H. Chen is with Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: jackie.haochen@gmail.com).

D. Ni, J. Qin, X. Yang and T. Wang are with National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Medicine, Shenzhen University, China (*corresponding author. e-mail: nidong@szu.edu.cn (D. Ni), jqin@szu.edu.cn (J. Qin)).

S. Li is with Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University, China (e-mail: lishengli63@126.com).

P. A. Heng is with Department of Computer Science and Engineering, The Chinese University of Hong Kong and Center for Human Computer Interaction, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China (e-mail: pheng@cse.cuhk.edu.hk).

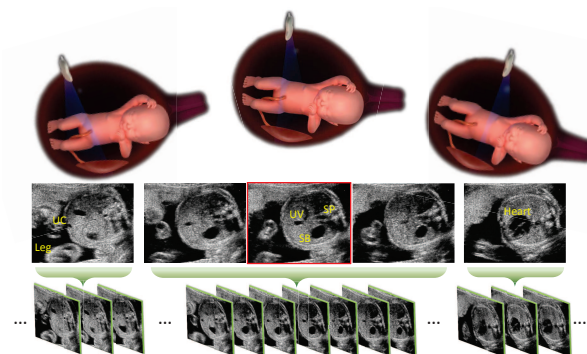


Fig. 1: Illustration of the FASP localization from 2-D US images (the frame with red rectangle is a FASP).

the subsequent biometric measurements and diagnosis [2], [3], [4]. In this regard, the reported wide sensitivity variation may be due, at least in part, to the quality of the standard plane obtained. In clinical practice, acquisition of the standard plane by clinicians often requires a thorough knowledge of human anatomy and substantial experience. Therefore, it is very challenging for novices and even difficult and time consuming for clinical experts. Thus, the development of automatic methods for locating standard planes from 2-D US images would assist the novices as well as improve the efficiency of experts.

In this study, we focus on automatically locating the fetal abdominal standard plane (FASP) from US videos (a preliminary version of this work has been reported in [5] and significant improvements have been made to the original paper). Clinically, to locate the FASP, a radiologist attempts to find the concurrent presence of three key anatomical structures (KASs): the stomach bubble (SB), the umbilical vein (UV) and the spine (SP) in one frame when moving the US probe across the patient body. The procedure is illustrated in Fig. 1. Based on the acquired FASP, the clinician can measure abdominal circumference (AC), which is the most important measurement for estimating fetal weight. The accuracy of AC measurement is heavily dependent on both the quality of the FASP and the manual measurement on the FASP by clinicians. Recently, commercial tools have been developed for the automatic AC measurement on several US scanners including Siemens Acuson S2000, GE LOGIQ S8, Mindray DC8, etc. However, little attention has been paid to the prerequisite step, that is, FASP acquisition.

Over the past few years, some methods have been proposed for locating standard planes from 2-D US images. Zhang
Caption Health Ex1016

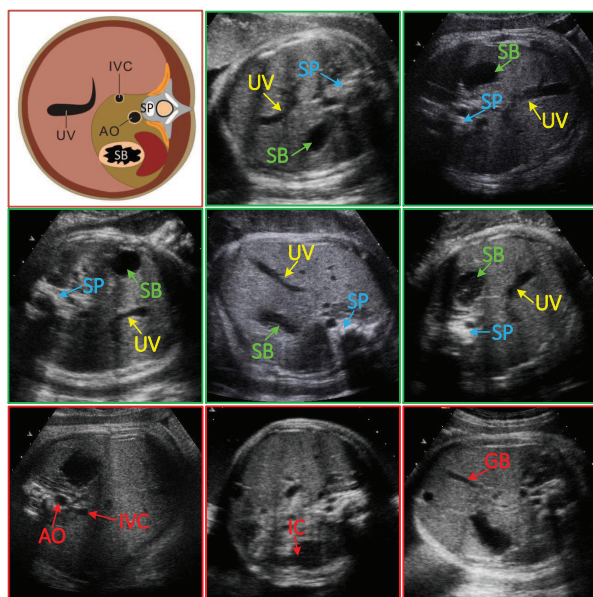


Fig. 2: Fetal abdominal anatomy (marked with brown rectangle), typical true FASPs (marked with green rectangles) and false FASPs (marked with red rectangles) with similar anatomical structures, e.g., GB and IC.

et al. [6] proposed to select the standard plane of early gestational sac (SPGS) from the US video by utilizing cascade AdaBoost classifiers trained on Haar features. However, the proposed system may not perform well in the detection of the FASP by only using Haar feature trained classifiers, since the FASP contains more complicated anatomical structures and has higher intra-class variations than the SPGS, which cannot be well captured by Haar features. Bahbibi *et al.* [7] proposed to detect KASs from the manually cropped US image of abdomen regions by combining local Haar features and a global multi-scale feature symmetry measure. This method is semi-automatic, and was primarily designed for detecting KASs rather than the FASP. Kwitt *et al.* [8] explored a kernel dynamic textures (KDTs) model to locate target structures by augmenting two configurations: raw intensity values and Bag of Words (BoW) representation of 3D-HOG. This method was only evaluated on the phantom data. As actual patient data are usually much more complex than phantom data, further investigation is needed to evaluate the efficacy of this method. Ni and Yang *et al.* [3], [4] presented a hierarchically supervised learning framework to locate FASP from US videos via the radial component model and vessel probability map (RVD). This method achieved acceptable detection accuracy by incorporating geometric constraints of KASs. However, more effective feature representation is still required for capturing the complicated appearance of FASPs and further boosting the performance.

The major issues need to be well addressed in the problem of locating the FASP from US videos can be summarized in threefold. First, as illustrated in the first and second rows of Fig. 2, the FASP often has high intra-class variations caused by the image artifacts (e.g., speckle noise and acoustic

shadows), deformations of soft tissues, as well as difference in gestational age, fetal posture and scanning orientation. Second, as shown in the third row of Fig. 2, the FASP and non-FASP often have low inter-class difference since large numbers of regions, e.g., acoustic shadows, abdominal aorta (AO), gall bladder(GB), intestinal canal(IC) and inferior vena cava (IVC), often share similar echogenicity appearance to the KASs. The low inter-class difference sometimes makes it very difficult to locate the FASP from US videos, even for experienced obstetric experts. Considering these challenges of high intra-class variation and low inter-class difference in our task, simple low-level features such as Haar, local phase and HOG used in previous studies [3], [4], [6], [7], [8] may not be able to represent the complicated appearance of the FASP and thus lower the classification performance. Third, as collecting large amount of data from patients and labeling them are very time-consuming, overfitting often occurs when employing the supervised learning based methods on small amount of training dataset. It is often one of the main challenges in medical computing community.

Recently, the advance of deep convolutional neural networks (CNN) have achieved a great success on a variety of applications including image classification [9], [10], [11], object detection [10], [12], [13] and image segmentation [14], [15], [16]. One of the main advantages of CNN is its powerful feature representation capability, which makes it a promising tool for automatic FASP localization. However, the performance of CNN classifiers heavily depends on the size of the training dataset, as small amount of training samples may probably result in overfitting in the fully-supervised deep architecture of CNN. Inspired by some recent studies on transfer learning [17], [18], [19], we propose a novel domain transferred CNN to address the three challenging issues in the problem of FASP localization from US videos. The main contributions of this work are threefold:

- 1) To the best of our knowledge, this is the first work to adopt deep learning methods for the automatic localization of standard planes from US videos. Deep learning with expressive power for feature representation can enlarge the inter-class distance and reduce the intra-class distance efficiently, hence suitable for the task.
- 2) A transfer learning strategy that constructs an efficient task-specific CNN with a limited specific dataset using the knowledge of a base CNN pre-trained on large cross-domain datasets is introduced to enhance the classification performance, and we believe this transfer strategy and its variants can bridge the gap between deep CNN and medical applications caused by the limitation of the training dataset, and expand its application in medical computing community.
- 3) A quantitative analysis is given to discuss the feasibility of sharing statistical strength between natural and medical domains. This analysis may provide empirical evidence for the knowledge transfer between different domains.

The remainder of this paper is organized as follows. Section II describes the proposed method and its background in detail. Experimental results are given in Section III. Section IV discusses the advantages and limitations of the proposed

method, as well as future research directions. Conclusions are drawn in Section V.

II. METHODS

Fig. 3 illustrates the pipeline of our proposed method. At the training stage, we first train a multi-layered base CNN from ImageNet detection data, which is an object-level annotated dataset containing a large number of natural images. Then we construct the domain transferred CNN for FASP localization in US videos by: (1) implanting the convolutional layers of the base CNN into the transferred CNN as the initial settings, and (2) jointly fine tuning the parameters of the convolutional layers and training the fully connected layers for the transferred CNN based on the US training samples. At the test stage, the trained classifier is utilized to generate a probability map for each frame in the US video and identify an US image as either a FASP or non-FASP.

In this section, we first experimentally demonstrate the feasibility of transferring knowledge from natural image domain to medical image domain through dictionary learning. Then we provide the implementation details of the domain transferred CNN for the FASP localization in US videos after a brief introduction of CNN.

A. Ultrasound Representation with Natural Image Bases

Domain transfer (knowledge transfer or transfer learning) often refers to the use of knowledge learnt from one source domain to efficiently develop a more accurate hypothesis for a new target task only consisting of small numbers of training examples in another domain [20]. Research in domain transfer began in the early 1980s [21] and has become an active topic in computer vision field with the development of deep learning methods [22]. Recently, Gupta *et al.* [23], for the first time, made use of cross-domain features to represent MRI data and achieved competitive performance for Alzheimer disease classification. However, with large differences between two domains, why and how the knowledge can be transferred from natural image domain to medical image domain to boost the classification performance for medical applications remains an open problem. In this study, we assume that although high level features of US and natural images are distinctively different, they do share similar statistical pattern in low level features, e.g., oriented edges, corners and junctions. In this regard, the low level representations learnt from natural images can be transferred to the medical image domain.

To validate this assumption, we calculate the residual errors of reconstructing target US images with the dictionaries learnt from the natural, US and corrupted images with random Gaussian noise, respectively. The corresponding residual errors are defined as E_n , E_u and E_g , respectively. If the E_n and E_u are small enough and approximately equal, while the E_g is much larger than E_n and E_u , we infer that dictionary bases learnt from natural image domain could reconstruct US data well, and hence the knowledge learnt from natural images is promising to be transferred to medical image domain.

In computer vision community, dictionary can learn a set of elementary bases or atoms from training data. These dictionary

bases can be linearly combined to well approximate a given signal [24]. In this study, we first randomly extracted three sets of patches (size 8×8) from natural images [25], US images and Gaussian noise images, respectively. Each set is composed of one million patches, where each patch is normalized with zero mean and unit variance. Then N_1 (750,000) patches of each set are used as training data to learn the dictionary bases $\mathbf{D}_n, \mathbf{D}_u, \mathbf{D}_g$ for natural, US and Gaussian noise images respectively. The left N_2 (250,000) patches are used for evaluating the reconstruction error. The dictionary bases are calculated according to [26] and defined as:

$$\arg \min_{\alpha \in \mathbb{R}^q, \mathbf{D} \in \mathcal{C}} \sum_{i=1}^{N_1} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\psi(\alpha_i) \quad (1)$$

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{p \times q}, \text{ s.t. } \forall j = 1, \dots, q, \mathbf{d}_j^T \mathbf{d}_j \leq 1\},$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is the i th sample patch of training set and can be represented by a sparse linear combination over a set of basis vectors from an over-complete dictionary \mathbf{D} with its corresponding sparse coefficient α_i ; $\psi(\alpha_i)$ is a sparsity-inducing regularizer, e.g., L_1 norm $\psi(\alpha_i) = \|\alpha_i\|_1$, and λ is the regularization parameter; \mathcal{C} is the convex set of matrixes, which is bounded on the column of dictionary $\mathbf{d}_j \in \mathbb{R}^p$ to prevent \mathbf{D} from having arbitrarily large values.

After the $\mathbf{D}_n, \mathbf{D}_u, \mathbf{D}_g$ are calculated, the reconstruction residual error E is computed by:

$$E = \frac{1}{2N_2} \sum_{i=1}^{N_2} \|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2^2 \quad (2)$$

where \mathbf{y}_i is the i th sample patch of testing set. It is worth noting that α_i in (2) is recalculated by minimizing (1) on testing data \mathbf{y}_i with the calculated \mathbf{D} from training data.

Fig. 4 shows the learnt dictionary bases from natural and US images. It is observed that the statistical patterns in low level features extracted from natural and US images are quite similar. The errors E_n , E_u and E_g of reconstructing US patches using dictionary bases learnt from natural, US and Gaussian noise images are 0.114, 0.111 and 0.263, respectively. The E_n and E_u are approximately equal and much smaller than E_g . In addition, we also performed the paired-sample t-test to illustrate the significance of differences (the null hypothesis set at the 5% significance level). The E_n and E_u shows no significant difference ($p = 0.249$), whereas the pairs of E_n and E_g , E_u and E_g both show significant difference ($p = 3.3 \times 10^{-4}$ and $p = 4.8 \times 10^{-3}$, respectively). These results demonstrate that although high level features of US and natural images are distinctively different, they do share similar statistical pattern in low level features. Hence, the knowledge transferred from natural image domain to medical image domain has the potential to enhance the learning performance with limited medical data.

B. Convolutional Neural Network

Convolutional neural networks (CNN) are biologically-inspired variants of multi-layered perceptrons, which exploit spatial correlation by extracting features generated from localized convolutional kernels [27]. Their ability to learn complex, high-dimensional and nonlinear mappings from large

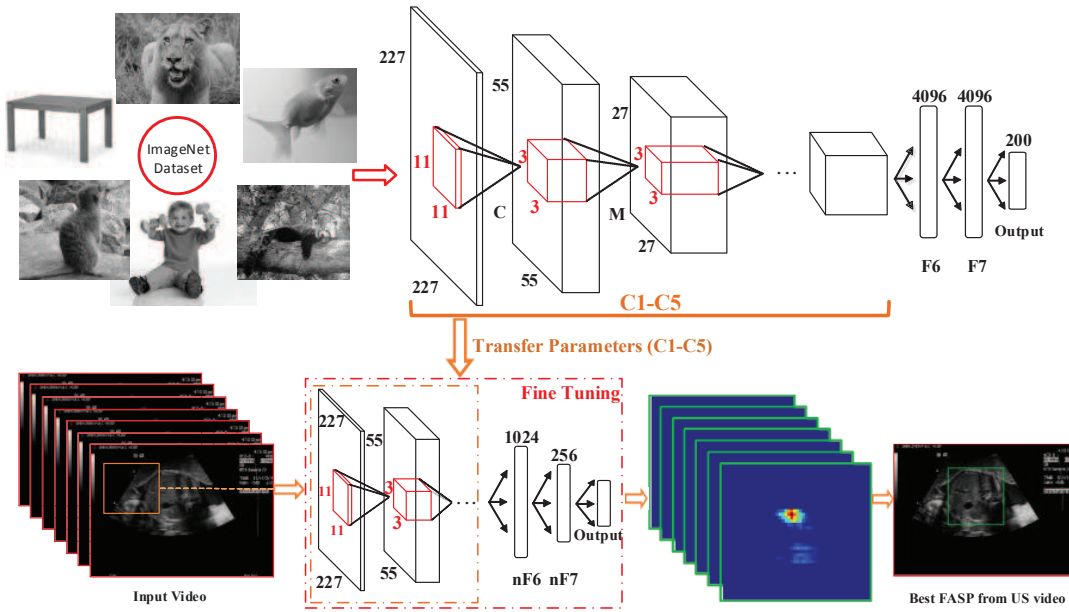


Fig. 3: The pipeline of our proposed method. The numbers in the upper and lower subfigures denote the network configuration; C, M and F denote the convolutional layer, the max-pooling layer and the fully connected layer, respectively.

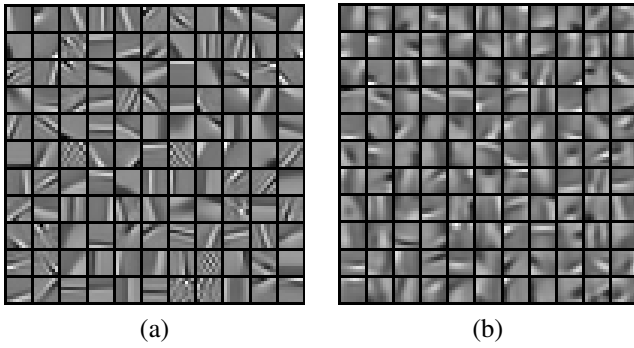


Fig. 4: Dictionary bases learnt from: (a) Natural images, (b) US images.

collections of training data makes them appropriate for a lot of complicated pattern recognition problems, as introduced in Section I.

The basic structure of CNN includes several pairs of convolutional and sub-sampling layers, followed by a shallow multi-layered perceptron for classification [28]. The convolutional layers (C layers) take local receptive fields in the previous layer as input and extract features while preserving spatial correlation. Suppose that \mathbf{h}_j^l is the j th feature map in the l th layer and \mathbf{h}_m^{l-1} ($m = 1, \dots, M$) is the m th feature map in the $(l - 1)$ th layer, the feature map \mathbf{h}_j^l can be computed by:

$$\mathbf{h}_j^l = \sigma\left(\sum_{m=1}^M \mathbf{W}_{jm}^l * \mathbf{h}_m^{l-1} + \mathbf{b}^l\right) \quad (3)$$

where \mathbf{W}_{jm}^l is the convolutional kernel connected to the m th feature map in the previous layer; \mathbf{b}^l is the bias in the l th layer; and $\sigma(\cdot)$ is the non-linear activation function.

We employ rectified linear unit (ReLU) $\sigma(x) = \max(x, 0)$

as the activation function in our implementation because it can create sparse representation with hard non-linearity and achieve better performance compared with the traditional sigmoid function [29].

Sub-sampling layers are used to reduce the resolution of feature maps by pooling over local neighborhood. In our study, max-pooling layers (M layers) are used to perform non-max suppression and down-sample the resolution of the feature maps generated by the C layers. The multi-layered perceptron with fully connected layers (F layers) follows after several altering convolutional and max-pooling layers. The final layer of the multi-layered perceptron outputs the posterior probability for each class with softmax function.

In order to train the CNN classifier, the parameters of CNN, including the convolution kernel \mathbf{W} , the bias \mathbf{b} and the weights in the multi-layered perceptron, should be automatically learnt from the training samples \mathbf{I}_i ($i = 1, \dots, N$). In practice, the parameters are calculated by minimizing a loss function $\mathcal{L}(\mathbf{I}_i, y_i; \theta)$ via mini-batch gradient descent (SGD) with momentum [30], where θ represents the set of parameters that should be trained. The loss function is defined as:

$$\mathcal{L}(\mathbf{I}_i, t_i; \theta) = -\sum_{i=1}^N \sum_{k=1}^K (\mathbf{1}\{t_i = k\} \log p(c_k = 1 | \mathbf{I}_i, \theta)) \quad (4)$$

$$p(c_k = 1 | \mathbf{I}_i, \theta) = \frac{e^{o_k}}{\sum_{j=1}^K e^{o_j}}, k = 1, \dots, K \quad (5)$$

where $p(c_k = 1 | \mathbf{I}_i, \theta)$ is the posterior probability of predicting the training sample \mathbf{I}_i as the k th class c_k among total K classes; $\mathbf{1}\{\cdot\}$ is the indicator function and t_i is the class label of training sample \mathbf{I}_i , $\mathbf{1}\{t_i = k\}$ evaluates to 1 when \mathbf{I}_i belongs to the k th class, otherwise 0; o_k is the output of neural network before softmax given the training sample \mathbf{I}_i . Readers can refer to [9], [27] for more details of the implementation of CNN.

C. Domain Transferred CNN for FASP Localization

Although the CNN has the advantage of learning powerful feature representations, with limited training data in many medical applications, the fully-supervised deep architectures may overfit the training data and hence degrade the learning performance. To the end, the small amount of training samples in many medical applications prohibits its use in medical domain. Recently, a number of studies [17], [18], [19] have demonstrated that transfer learning is a powerful tool to reduce overfitting by first training a base network on a base dataset and task, and then transferring the learnt features of the base network to a new target network to be trained on a target dataset and task. Yosinski *et al.* [18] further pointed out that the fundamental principle for transfer learning is that features learnt from lower layers (such as C layers) of CNN are general so that they can be applicable to datasets across different image domains, while the features computed from higher layers (such as F layers) are more specific to the particular dataset. In this regard, the knowledge acquired from the lower layers has potential to be used in other tasks.

Inspired by these studies, we attempt to investigate if the knowledge acquired from CNN trained by a dataset of natural images can be transferred to a medical application where the training dataset is limited and directly employing CNN on the dataset will probably result in overfitting to some extent. In this study, we aim to locate the standard planes in fetal US videos by employing such kind of transferred CNN. In previous subsection A, we have experimentally demonstrated that low level cues learnt from natural images can be transferred and reused in the domain of fetal US. In this section, a transfer learning strategy is proposed to construct a transferred CNN that takes the advantage of the base CNN, which was trained from a large dataset of natural images. The transferred CNN can effectively identify the FASPs from US videos.

TABLE I: The architecture of N-CNN

Layer	Feature maps	Kernel size	Stride
Input	227x227x1	-	-
C1	55x55x96	11	4
M1	27x27x96	3	2
C2	27x27x256	5	-
M2	13x13x256	3	2
C3	13x13x384	3	-
C4	13x13x384	3	-
C5	13x13x256	3	-
M5	6x6x256	3	2
F6	4096	-	-
F7	4096	-	-
F8	200	-	-

Since we are interested in the object-level features, we first train the base CNN on the 2014 ImageNet detection dataset [25], which contains 478,807 and 20,121 objects in the training and validation dataset, respectively. The base CNN is denoted as N-CNN. Our N-CNN's architecture shares the same spirit with the widely used BVLC CaffeNet Model [31]. The N-CNN includes five C layers and several M layers between C layers, followed by three F layers, as shown in Table I. Noting that we converted the natural images into gray images since US images are single-channel images. Then we construct

the domain transferred CNN for FASP detection, namely T-CNN. In order to transfer the knowledge from the base CNN, we implant the trained C and M layers of the base CNN to the same positions of T-CNN. Then we jointly fine tune these layers and train following three new F layers (called nF layers to distinguish them from F layers in the base CNN) on US training samples. These three nF layers consist of 1024, 256, 2 neurons, respectively. It is worth noting that during the training of T-CNN, parameters of transferred layers in T-CNN are initialized by the pre-trained N-CNN, while parameters of three nF layers in T-CNN are randomly initialized with Gaussian distribution. Meanwhile, the strategies of data augmentation and dropout [32] are implemented in the training process for regularization in order to improve the generalization ability.

Once the T-CNN is trained, the probability map p for each image in the US video is calculated with a sliding window method. This approach involves scanning the image with a rectangular window after excluding non-US regions, and applying a T-CNN classifier to the sub-image defined by the window. In order to improve the robustness, each sub-image is further augmented by cropping its center and corners as well as its mirrored versions, resulting in 10 inputs (size 227×227) for the T-CNN. The score of the sliding window at its center is obtained by averaging the scores of these 10 inputs. After the probability map p is obtained, we further smooth it to eliminate noise with a bilateral filter. Then the final score of the image is the highest value of the smoothed probability map. Finally, the US image in the video with the highest detection score is identified as the FASP when the score is above a threshold T .

III. EXPERIMENTS AND RESULTS

A. Dataset and System Implementation

The study protocol was reviewed and approved by the ethics committee of our institution, and informed consent was obtained from all subjects.

Training Dataset. To train the T-CNN classifier, we used 11942 expert annotated fetal abdominal US images from 300 videos for generating the training samples. First, 1991 positive samples were generated by cropping the image regions that contained the anatomical objects from FASPs, while 3160 negative samples were extracted randomly from the non-FASPs and the background of FASPs. Some of these had an overlap of 20% to 40% with a positive sample. Note that all training samples were further rotated, translated and mirrored to augment the training database.

Testing Dataset. A conventional US sweep was performed by three graduate students with a half-year of training in an obstetric US department to obtain 219 videos with a total of 8718 US images on 219 pregnant women (the fetal gestational age ranged from 18 to 40 weeks) in the supine position for evaluation of the T-CNN classifier. Such study protocol considers variations of operators in clinical practice and makes the FASP detection more challenging than the protocol reported in [4]. Each sweep lasted approximately 2–5s and each video contained 17–48 frames. During image acquisition, the US imaging parameters were adjusted by the students to obtain a desired imaging quality and were not required to conform to

a pre-defined criterion. Then a radiologist with more than 5 years of experience in obstetric US carefully checked to see if one or more FASPs was in each video. Of the 219 videos, 199 videos contained at least one FASP, the remaining 20 videos contained no FASP. All the images and videos used in our experiments were acquired using a Siemens Acuson Sequoia 512 US scanner from Shenzhen Maternal and Child Healthcare Hospital.

System Implementation. The Caffe package [31] was used to implement the CNN architecture. Our system was implemented with the mixed programming technology of MATLAB and C++. The running time for locating the FASP from one video depended primarily on the number of frames in the video. It generally took 1 min for detecting the FASP from a video containing 40 frames using a workstation with a 2.50 GHz Intel(R) Xeon(R) E5-2609 CPU and a NVIDIA GeForce GTX TITAN GPU.

In order to evaluate our system, we first qualitatively evaluated the performance of our T-CNN classifier by visualizing the high-level features learnt from T-CNN and showing typical FASPs detected from different US videos. Then, we evaluated our system quantitatively on US images and videos separately. The threshold T used to identify an US image as FASP or non-FASP was set as 0.68 by testing the algorithm on a set of samples from the training set.

B. Qualitative Performance Evaluation

Data visualization is an effective technique to directly show the discriminant capability of feature representations, hence can indicate the performance of the classifier. We employed the Barnes-Hut-SNE method [33] to reduce dimensions of raw training datasets, raw testing datasets, and features of intermediate layers of T-CNN extracted from the training and testing datasets. Fig. 5 shows these four data sets, where red and blue points represent FASP and non-FASP, respectively. The mixed distribution of training and testing data in the original domain illustrates the high intra-class variation of FASP and the low inter-class variation between FASP and non-FASP, which makes the FASP detection very challenging, as discussed in section I. In comparison, the features of $nF7$ layer extracted from training and test data can be used to classify FASP and non-FASP very well.

Fig. 6 shows three typical examples of true FASPs detected from three different videos and the corresponding probability maps generated by our T-CNN. The scores of these US images represented by the highest scores in the probability maps are all larger than the threshold T . Three KASs including UV, SB and SP are all contained in each image. These results demonstrate the efficacy of our proposed T-CNN qualitatively.

C. Quantitative Performance Evaluation and Comparison

We quantitatively evaluated the performance of our method in two different experiments for different purposes. In the first experiment, to determine the efficacy of the proposed T-CNN classifier, we compared the performance of three methods for detecting the FASP from a set of 2-D US images, including the

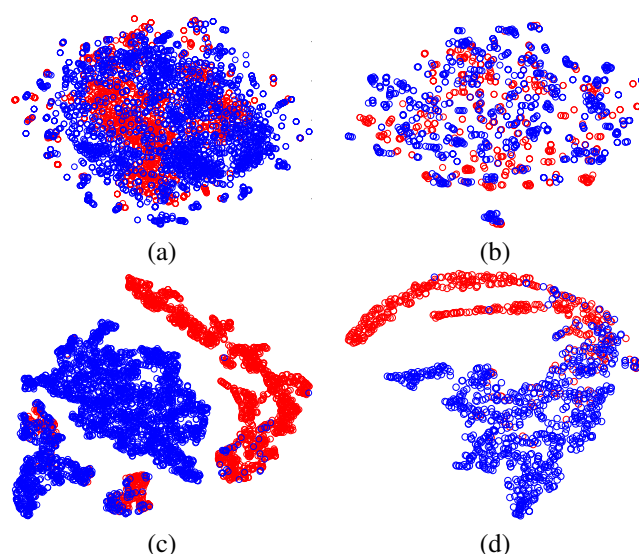


Fig. 5: Barnes-Hut-SNE visualizations of four data sets. Red and blue points represent FASP and non-FASP, respectively. (a) Raw training data, (b) Raw testing data, (c) Extracted features of $nF7$ layer from training dataset, (d) Extracted features of $nF7$ layer from testing dataset.

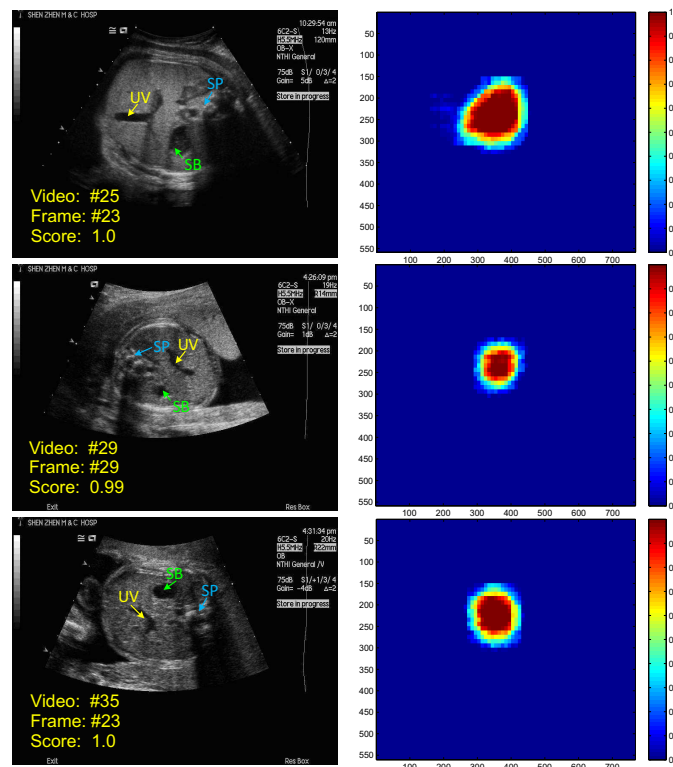


Fig. 6: Typical FASPs (left column) classified by our method and the corresponding probability maps (right column).

proposed T-CNN method, the state-of-the-art RVD method [4] and the R-CNN method which directly trained the CNN on the training US dataset without exploiting the proposed transfer learning strategy (here R denotes randomly initializing the parameters of CNN at the beginning of training). In the second

TABLE II: Results of FASP localization on US images

Method	Accuracy	Precision	Recall	F1
T-CNN	0.896	0.714	0.710	0.712
R-CNN	0.857	0.594	0.681	0.635
RVD [4]	0.833	0.532	0.693	0.602

experiment, we investigated the performance of these three methods for detecting the FASP from US videos, which was consistent with the clinical practice.

We computed the following evaluation measurements [34]: recall (R), precision (P), F_β score ($\beta = 1$, i.e., F1 score in our experiment) and accuracy (A), as shown in (6)–(9).

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}} \quad (6)$$

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}} \quad (7)$$

$$F_\beta = (1 + \beta^2) \cdot \frac{RP}{R + \beta^2 P} \quad (8)$$

$$A = \frac{N_{tp} + N_{tn}}{N_{tp} + N_{tn} + N_{fp} + N_{fn}} \quad (9)$$

where N_{tp} , N_{tn} , N_{fp} and N_{fn} are the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), respectively.

1) *Evaluation on US Images:* In this experiment, the US image is classified as a FASP when its score is larger than the threshold T . As shown in Table II, the accuracy, precision, recall and F1 score of the proposed T-CNN on testing images were 0.896, 0.714, 0.710 and 0.712, respectively, which significantly outperformed the state-of-the-art method [4]. In addition, the results shown in Table II further demonstrates that the T-CNN with knowledge transfer outperformed the R-CNN, which indicates that the knowledge transferred from natural image domain can obviously improve the results of our specific medical task with limited dataset for training. The Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves evaluated on US images were also shown in Fig. 7. The areas under the ROC curve (AUC) obtained by the T-CNN, R-CNN, and RVD methods were 0.93, 0.9 and 0.8, respectively. Our proposed T-CNN method achieved the best performance and the result of the R-CNN method was better than the RVD method, which further illustrates the efficacy of both the deep CNN algorithm and the proposed knowledge transfer strategy.

2) *Evaluation on US Videos:* For the evaluation of the FASP localization in US videos, we used the same rules of [4] to define the true positive and true negative. Each video was treated as one testing sample. A true positive was obtained when the correct FASP was detected from a video containing at least one FASP. And a true negative was obtained when no FASP was detected from a video containing no FASP. As shown in Table III, the accuracy, precision, recall and F1 score of the T-CNN on testing videos were 0.904, 0.908, 0.995 and 0.950, respectively. Our proposed T-CNN method outperformed the other two methods significantly. The R-CNN performed better than the state-of-the-art RVD method. These results demonstrate the efficacy of the proposed domain

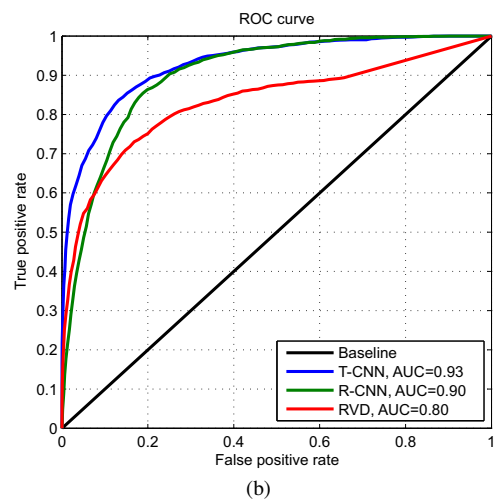
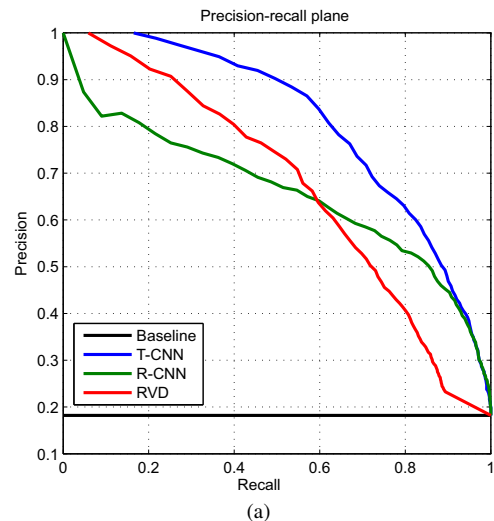


Fig. 7: (a) Precision-recall plane, (b) ROC Curve.

TABLE III: Results of FASP localization on US videos

Method	Accuracy	Precision	Recall	F1
T-CNN	0.904	0.908	0.995	0.950
R-CNN	0.822	0.826	0.994	0.902
RVD [4]	0.762	0.823	0.913	0.865

transferred deep learning method and the system equipped with this method is promising for clinical practice.

IV. DISCUSSIONS

In this paper, a novel domain transferred CNN model for the automatic localization of the FASP from US videos is presented. Specifically, we first train a base CNN on the 2014 ImageNet detection dataset [25]. Then the knowledge embedded in the convolutional layers of the base CNN is transferred to our task-specific CNN by implanting these trained convolutional layers into the proposed T-CNN model. The T-CNN model is generated by jointly fine tuning the transferred layers and training the fully connected layers. The proposed knowledge transferred deep CNN can reduce the overfitting of the classifier with limited size of training

samples while taking the advantage of CNN for powerful feature representation. The experimental results presented in section III show that the proposed T-CNN method achieved better classification results than the state-of-the-art method [4]. The high accuracy of locating the FASP from US videos demonstrates that our system has great potential for clinical applications.

Generally speaking, the data size of medical domain for learning-based approaches is much smaller than that of natural domain, since collecting data from patients need the approval of the ethics committee and is very time-consuming, especially for some rare diseases. Therefore, it is a main challenge faced in the medical computing community that overfitting induced by limited training dataset may affect the performance of the learning system and make it difficult for clinical applications. Nowadays, data-driven medicine enabling the discovery of new treatment options with analysis on massive amounts of data has become the next frontier for modern medicine [35], while deep-learning models are breakthroughs over traditional methods in addressing longstanding computing problems and has been applied in a variety of applications. Our proposed domain transferred method will benefit the medical computing community in addressing the challenge of limited size of dataset and boost the application of the CNN in medical domain.

In this study, all US images and videos were acquired by three graduate students with a half-year of training in an obstetric US department. Thus the operator variation is taken into account for the system evaluation. It possibly lowered the performance of the RVD method compared with the results reported in [4]. Our proposed T-CNN method has the superior performance on this challenging dataset. In addition, the RVD method is only designed for the specific task, i.e., the FASP localization. Although we applied the proposed method for the same task of [4] in this work, our method is a general learning framework and can be easily adapted to the automatic localization of other standard planes in US videos. It can also be easily extended to other classification problems suffering from the small size of the training dataset in medical domain.

There are several limitations for this study. First, it takes approximately 1 min for our system to locate the FASP from one US video. It is much faster than the system developed in [4]. However, our system is still not in real time and cannot be directly used in a clinical examination. In our system, the most time consuming step is the sliding window operation on the US image with the original size. We plan to further increase the speed of our system by first generating the sub-sampled feature maps of the US image from the T-CNN model and then performing the sliding window classifier on the feature maps. Second, although the high detection accuracy of our system makes it applicable for clinical practice, additional work is needed to further improve the system performance. Fig. 8 illustrates two examples of false FASPs detected from two videos by our method. This incorrect detection is possibly caused by the acoustic shadow and the GB being misidentified as the SB and UV, respectively, due to very similar appearance. Previous studies suggested that one possible way to improve the performance of CNN is to stochastically

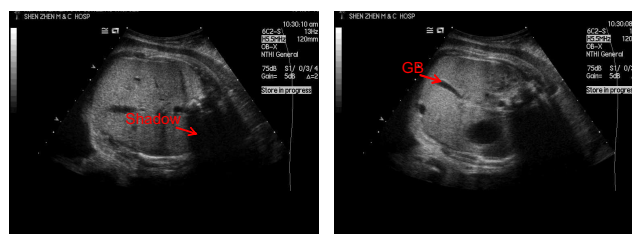


Fig. 8: Two examples of false FASPs detected from videos.

corrupt the training samples in the learning process [9], [36], [37]. Similarly, we may randomly corrupt the training data with acoustic shadows in the training process to improve the robustness of CNN. On the other hand, we observe that clinicians often select the FASPs with contextual clues in the US videos. Accordingly, the false positive findings identified by the computerized method can be possibly eliminated with the incorporation of the temporal context from consecutive frames into the mathematical detection model.

V. CONCLUSIONS

We propose an automatic approach to locate the fetal abdominal standard plane (FASP) from US videos. In contrast with existing approaches based on low-level features, our approach exploits the deep convolutional neural network (CNN) to represent the complex anatomical structures appearing in the FASP. We further propose a transfer learning strategy to reduce the overfitting problem resulted from inadequate training samples by leveraging the knowledge of a base CNN trained on a large database of natural images. We implant the knowledge into the transferred CNN model as initial settings, which can improve the performance of the transferred CNN on localization of FASP. Both qualitative and quantitative experiments demonstrate the efficacy of the proposed approach. We believe this approach is also appropriate for other medical image computing problems, where the sophisticated anatomical structures cannot be well captured by low-level features and the insufficiency of training samples makes it difficult to train a robust deep CNN model to represent the high-level features. The transfer learning strategy is promising to expand the applications of deep CNN in medical domain.

ACKNOWLEDGMENT

The work described in this paper was supported in part by the National Natural Science Foundation of China (Nos. 81270707 and 61233012), in part by the Shenzhen Key Basic Research Project (Nos. JCYJ20130329105033277 and JCYJ20140509172609164), in part by the Shenzhen-Hong Kong Innovation Circle Funding Program (Nos. JSE201109150013A and SGLH20131010151755080), in part by the Hong Kong Innovation and Technology Fund (Nos. GHP/003/11SZ and GHP/002/13SZ), and in part by a grant from the Research Grants Council of Hong Kong (No. CUHK412510).

REFERENCES

- [1] L. Salomon, N. Winer, J. Bernard, and Y. Ville, "A score-based method for quality control of fetal images at routine second-trimester ultrasound examination," *Prenatal diagnosis*, vol. 28, no. 9, pp. 822–827, 2008.
- [2] 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*. Springer, 2013, pp. 203–211.
- [3] 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*. IEEE, 2014, pp. 1180–1183.
- [4] 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.
- [5] 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*. Springer, 2014, pp. 125–132.
- [6] 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*, vol. 39, no. 8, pp. 5015–5027, 2012.
- [7] 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*. Springer, 2012, pp. 402–409.
- [8] R. Kwitt, N. Vasconcelos, S. Razzaque, and S. Aylward, "Localizing target structures in ultrasound video—a phantom study," *Medical image analysis*, vol. 17, no. 7, pp. 712–722, 2013.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.
- [11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [12] D. C. Cireřan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. Springer, 2013, pp. 411–418.
- [13] C. Lu, H. Chen, Q. Chen, H. Law, Y. Xiao, and C.-K. Tang, "1-hkust: Object detection in ilsvrc 2014," *arXiv preprint arXiv:1409.6155*, 2014.
- [14] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.
- [15] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.
- [16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.
- [17] 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.
- [18] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.
- [19] 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.
- [20] D. Silver, I. Guyon, G. Taylor, V. Lemaire, and G. Dror, "Icml2011 unsupervised and transfer learning workshop," *Journal of Machine Learning Research*, vol. 27, pp. 1–16, 2012.
- [21] T. M. Mitchell, *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ., 1980.
- [22] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *ICML Unsupervised and Transfer Learning*, 2012, pp. 17–36.
- [23] 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)*, 2013, pp. 987–994.
- [24] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.
- [25] 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.
- [26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.
- [27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [28] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Ciresan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 342–347.
- [29] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, vol. 15, 2011, pp. 315–323.
- [30] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1139–1147.
- [31] 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.
- [32] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.
- [33] L. Maaten, "Barnes-hut-sne," in *Proceedings of the International Conference on Learning Representations*, 2013.
- [34] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval*. Springer, 2005, pp. 345–359.
- [35] N. H. Shah and J. D. Tenenbaum, "The coming age of data-driven medicine: translational bioinformatics' next frontier," *Journal of the American Medical Informatics Association*, 2012.
- [36] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," *arXiv preprint arXiv:1312.5402*, 2013.
- [37] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.



Hao Chen (S'14) received the B.S. degree in Information Engineering from Beihang University (BUAA) in 2009. He is currently a PhD student in the Department of Computer Science and Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. His research interests include medical image analysis, deep learning, object detection and segmentation, etc.



Dong Ni received his Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong in 2009. He is currently an Associate Professor in Department of Biomedical Engineering, School of Medicine, Shenzhen University. His research interests mainly include ultrasound image analysis, image guided surgery and pattern recognition.



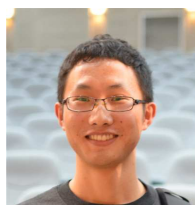
Pheng Ann Heng (M'92-SM'06) received the Ph.D. degree in computer science from Indiana University, Indianapolis, IN. He is currently a Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, where he is also the Director of the Virtual Reality, Visualization, and Imaging Research Centre. He is also the Director of the Research Center for Human-Computer Interaction, Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include virtual reality applications in medicine, visualization, medical imaging, human-computer interfaces, rendering and modeling, interactive graphics, and animation.



Jing Qin received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Shatin, Hong Kong. He is currently an associate professor in the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Medicine, Shenzhen University. His research interests include physically based modeling, medical image processing and computer-assisted surgery.



Shengli Li received his master degree in radiology from Xiang Ya School of Medicine in 1994. He is currently a Chief Physician and a Professor in Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University. His research interests focus on ultrasound diagnosis.



Xin Yang received the B.S. degree in biomedical engineering from South Central University for Nationalities (SCUN) in 2012. He is currently a master student in Department of Biomedical Engineering, School of Medicine, Shenzhen University. His research mainly focuses on intelligent ultrasound screening and diagnosis.



Tianfu Wang received his Ph.D. degree in biomedical engineering from Sichuan University in 1997. He is currently a Professor in Department of Biomedical Engineering, School of Medicine, Shenzhen University and the Associate Chair of School of Medicine. His research interests include ultrasound image analysis, medical image processing, pattern recognition and medical imaging.