IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 25, 2019, accepted October 16, 2019, date of publication October 30, 2019, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2950387

Deep Learning-Based Methodology for Recognition of Fetal Brain Standard Scan Planes in 2D Ultrasound Images

RUOWEI QU[1], GUIZHI XU[1], (Member, IEEE), CHUNXIA DING[2], WENYAN JIA[3], AND MINGUI SUN[3], (Fellow, IEEE)

[1]State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin 300130, China
[2]Ultrasound Department, Zhangjiakou Maternal and Child Health Hospital, Hebei 075000, China
[3]Laboratory for Computational Neuroscience, University of Pittsburgh, Pittsburgh, PA 15213, USA

Corresponding author: Guizhi Xu (gzxu@hebut.edu.cn)

This work was supported in part by the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, and in part by the National Natural Science Foundation of China under Grant 51677053 and Grant 51737003.

ABSTRACT Two-dimensional ultrasound scanning (US) has become a highly recommended examination in prenatal diagnosis in many countries. Accurate detection of abnormalities and correct fetal brain standard planes is the most necessary precondition for successful diagnosis and measurement. In the past few years, support vector machine (SVM) and other machine learning methods have been devoted to automatic recognition of 2D ultrasonic images, but the performance of recognition is not satisfactory due to the wide diversity of fetal postures, shortage of data, similarities between standard planes and other reasons. Especially in the recognition of fetal brain images, the features of fetal brain images such as shape, texture, color and others are very similar, which presents great challenges to the recognition work. In this study, we proposed two main methods based on deep convolutional neural networks to automatically recognize six standard planes of fetal brains. One is a deep convolutional neural network (CNN), and the other one is CNN-based domain transfer learning. To examine the performance of these algorithms, we constructed two datasets. Dataset 1 consists of 30,000 2D ultrasound images from 155 subjects between 16 and 34 weeks. Dataset 2, containing 1,200 images, was acquired from a research participant throughout 40 weeks, which is the entire pregnancy. Experimental results show that the proposed solutions achieve promising results and that the frameworks based on deep convolutional neural networks generally outperform the ones using other classical deep learning methods, thus demonstrating the great potential of convolutional neural networks in this area.

INDEX TERMS Medical image processing, CNN, transfer learning.

I. INTRODUCTION

Ultrasound Scans (US) are now widely used in many countries as highly recommended examinations in prenatal diagnosis because they are painless, low-cost, and possible without harmful radiation, and they can be carried out at any stage of pregnancy [1]–[3]. In most countries, guidance for how to select and examine these standard planes is defined in the fetal standard plane (FSP) handbook. Those standard planes contain detailed information such as biometric measurements and possible abnormalities. Biometric measurement results such as head circumference on the trans-ventricular head view may indicate fetal development and detect dysplasia; possible abnormalities such as lesions in the posterior skin edge on the standard sagittal spine view may provide early warning for physicians and pregnant mothers to make future plans [4].

During the development of fetal formation, doctors may observe ventricular dilatation, intracranial hematoma, enhanced echo of brain tissue, intracerebral calcification, hydrocephalus, congenital brain atrophy, sub ependymal cyst and other notable features by ultrasound examination [5], [6]. These abnormalities require a high level of attention because they may represent the manifestations of intracranial hemorrhage, intracranial infection and ischemic-hypoxic encephalopathy. If these indicators are observed

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

during pregnancy, it will be of great and far-reaching significance to discover the causes of death and disability of immature infants in order to assist clinicians in choosing appropriate diagnostic and therapeutic schemes in time and to evaluate the health and disease status of fetuses. Therefore, the objective accurate recognition of the fetal brain standard plane plays a key role in the diagnosis of fetal brain diseases [7]. Currently, physicians make diagnoses and determine therapies through visual means. The 2D ultrasound images of each patient are reviewed by independent reviewers, with disagreements resolved by the senior author. In this way, human error caused by fatigue or other reasons will be effectively prevented. However, because of different display parameters, poor signal to noise ratio and image artifacts such as shadowing, digital medical images present different display states in different section offices of different hospitals. Second, guiding the ultrasound probe to a correct standard plane through the complex anatomical structure is a highly sophisticated task which requires years of learning and training for an experienced physician. Furthermore, if the fetal position is not perfect, it can be difficult for physicians to obtain a clear image of a desired view. Hence, it is a challenging task to select the fetal standard plane, especially for automatic recognition algorithms.

To accurately detect fetal brain standard planes and obtain correct measurements, correct fetal brain standard planes must first be acquired, which requires high-level expertise in fetal anatomy and intensive manual labor by sonographers. Therefore, in the past few years, considerable efforts have been devoted to automate recognition of B-ultrasonic images [8], [9]. Typical methods include feature extraction, feature selection, feature encoding and classification. Traditional methods mainly used hand-crafted visual features [10]–[12], such as morphological features and textural features. The morphological features, also called shape features, are among the most important empirical classification criteria to detect tissue in ultrasound images. Almost twenty significant shape features [13], [14], such as scale-invariant feature transform (SIFT), Dense-SIFT (DSFT), Haar-like features, and histogram-of-gradient (HOG), and other combinations of these feature representations with intensity and edges as image descriptors are utilized to represent the images. Textural features are features that are based on image texture. More than ten kinds of textural features, including gray-level histogram (HIS) [15], gray-level co-occurrence matrix (GLCM) [16], [17] and Gabor wavelets-based spectral texture features are higher-level image representations in some sense. Not all of these features can be used in the classification system. Some features are selected by researchers based on experience, while some are selected by the laws texture energy (LTE) [18], which represents the laws that empirically determine that several masks of appropriate sizes were useful for discriminating between different kinds of texture. The method is based on applying such masks to the image and then estimating the energy within the pass region of filters. These selected features are then encoded by algorithms including bag of visual words (BoVW), vector of locally aggregated descriptor (VLAD) [19], Fisher vector (FV) [20], and multi-layer Fisher vector (MFV) to enhance the effectiveness of classification. Finally, support vector machine (SVM) [21], [22] and other machine learning methods are applied to classify these features [5], [16], [23]. Compared with the traditional manual method, automatic recognition of fetal standard planes can not only reduce the visual fatigue for physicians but also enhance the precision of diagnosis.

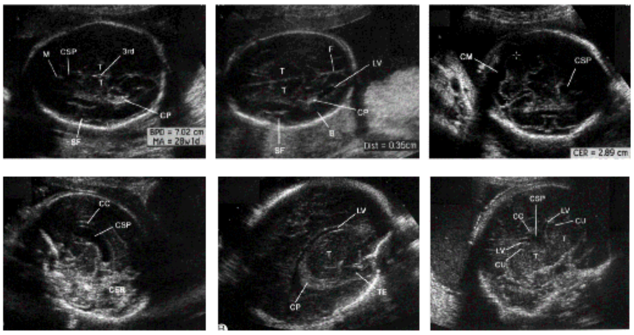


FIGURE 1. Samples showing the six fetal brain standard planes.

However, these types of methods exhibit some shortcomings: first, the recognition performance is still unsatisfactory due to the wide diversity of fetal postures and the high degree of visual similarity between standard fetal planes and others; second, they are ad hoc solutions, i.e., a specific set of hand-crafted features is needed for each standard plane; third, the final result and accuracy hinge on whether the hand-crafted feature is suited for this specific SVM or other machine learning algorithm, which is an uncertain factor for the whole algorithm.

In the past few years, neural networks have been proven to be very successful in solving image classification tasks because of the available large-scale labeled datasets, powerful representation ability of deep neural networks, especially convolutional neural networks (CNN) [24], and distributed workstations with powerful computing power. Deep neural networks for AI have been successfully applied to the fields of fetal standard planes [25]–[27], such as image classification, contour detection, nidus localization, object measurement, and target segmentation. However, automatic recognition of fetal brain standard planes from the data acquired by the color Doppler ultrasonic diagnosis apparatus is still a challenge because of practical factors such as low image resolution, motion-caused blur and different fetal positions.

Consequently, to overcome these problems, we proposed four more generic methods to automatically recognize standard planes of the six fetal brain standard planes shown in Fig. 1, which are the horizontal transverse section of thalamus, horizontal transverse section of lateral ventricle, transverse section of cerebellum, midsagittal plane, paracentral sagittal section, and coronal section of the anterior horn of the lateral ventricle. The conventional machine learning solutions using clustering and support vector machine are first proposed; then, to extract more representative latent features and to increase the recognition accuracy, we further proposed two frameworks based on deep convolutional neural networks.

To evaluate the effect of these machine learning algorithms, large sets of fetal brain standard plane images were acquired from free-living pregnant woman using a HITACHI ALOKA ARIETTA 70 colored Doppler ultrasonic diagnosis apparatus. 30000 fetal brain standard plane images and other fetal images containing six standard planes and other planes, which formed Dataset 1, were constructed from 155 subjects. Dataset 2 contained 300 images and was acquired from one research participant throughout 40 weeks, which represents the entire pregnancy. It is worth mentioning that the images in Dataset 2 do not intersect with Dataset 1.

The rest of the paper is organized as follows. Section 2 introduces the methodology of the proposed deep learning algorithm. Experimental results and analysis are provided in Section 3. The positive and negative aspects of our proposed method are articulately listed and compared in Section 4. Finally, in Section 5, we conclude the paper and discuss the future work.

II. METHODOLOGY

In image analysis problems, the descriptiveness and discriminative power of extracted features are critical to achieve good analysis performance. The remarkable advantage of deep learning is that this type of algorithm can be extended to difficult problems with relatively complex features, because the features for recognition can be automatically extracted via training.

A. CONVOLUTIONAL NEURAL NETWORK

Deep learning technology for AI has attracted great attention as it constantly breaks records in a variety of common benchmark tests [28]–[31]. As one type of method that satisfies the requirements of deep learning, CNN [28] is now a state-of-the-art deep learning structure applied to a wide range of fields, such as automatic machine translation, computer vision, and speech recognition. Great performance in object detection or classification from images [32], [33] has been shown because CNN networks provide a faster, more robust and more convenient algorithm than traditional neural networks. In CNN, image pixels could be directly used as input to the standard feed-forward neural networks. Although thousands of pixels from even small image patches result in a very large number of connection weight parameters to be trained, CNN models combine weights into much smaller kernel filters that dramatically simplify the learning model.

The **convolutional layer** is the key component of CNN. The process to compute a single output matrix is defined as:

$$A_j = f\left(\sum_{i=1}^{N} I_i * K_{i,j} + B_j\right) \tag{1}$$

where I_i is the input matrix, which is convoluted with a corresponding n × n kernel $K_{i,j}$ ($n <$ input size). The sum of all convoluted matrices is then computed, and a bias value B_j is added to each element of the resulting matrix. f is a non-linear activation function that is applied to each element of the previous matrix to produce one output matrix A_j.

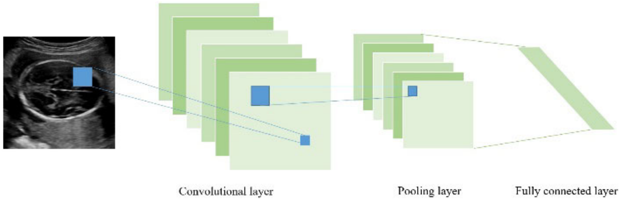


FIGURE 2. The framework of CNN architecture.

The **pooling layer** performs down-sampling by dividing the input into rectangular pooling regions in order to reduce the number of output neurons in the convolutional layer. Commonly used pooling algorithms include max-pooling and average-pooling.

Activation function. To solve more complex problems and to converge quicker, an activation function is used to add non-linear factors to the neural network. Commonly used activation functions such as the sigmoid and hyperbolic tangent functions are saturating non-linear functions for which the output gradient drops close to zero as the input increases. Some recent studies suggested that non-saturating non-linear functions, such as the rectified linear function $f(x) = max(0, x)$ (ReLU), improve both learning speed and classification performance in CNN applications [34]. The overall framework design is illustrated in Fig. 2.

B. TRANSFER LEARNING

Although the CNN offers the advantage of learning powerful feature representations, there are limits on the actual implementation of these networks in the field of application. The application of deep learning in medical image processing is developing slowly. The small amount of training data is the main limitation. With limited training data in many medical applications, the fully supervised deep architectures may overfit the training data and degrade the learning performance, and hence limit the development of deep learning in the area of medical image processing. [35] shows that transfer learning is a powerful tool to reduce over-fitting by first training a base network on a similar dataset and task, and then transferring the learned architecture and features of the base network to a new target network to be trained on a target dataset and task.

In this paper, we use fine-tuning as the strategy of transfer learning. We used the proposed DCNN network which was already pretrained on Dataset 1, and then trained all layers on the target dataset, Dataset 2. The reason why we chose to fine-tune the whole network is that although the images in the two datasets are similar, the details are slightly different. If we only train the last few layers and freeze the first few layers, the detailed features will not be learned adequately. Moreover, the structure and parameters of our proposed

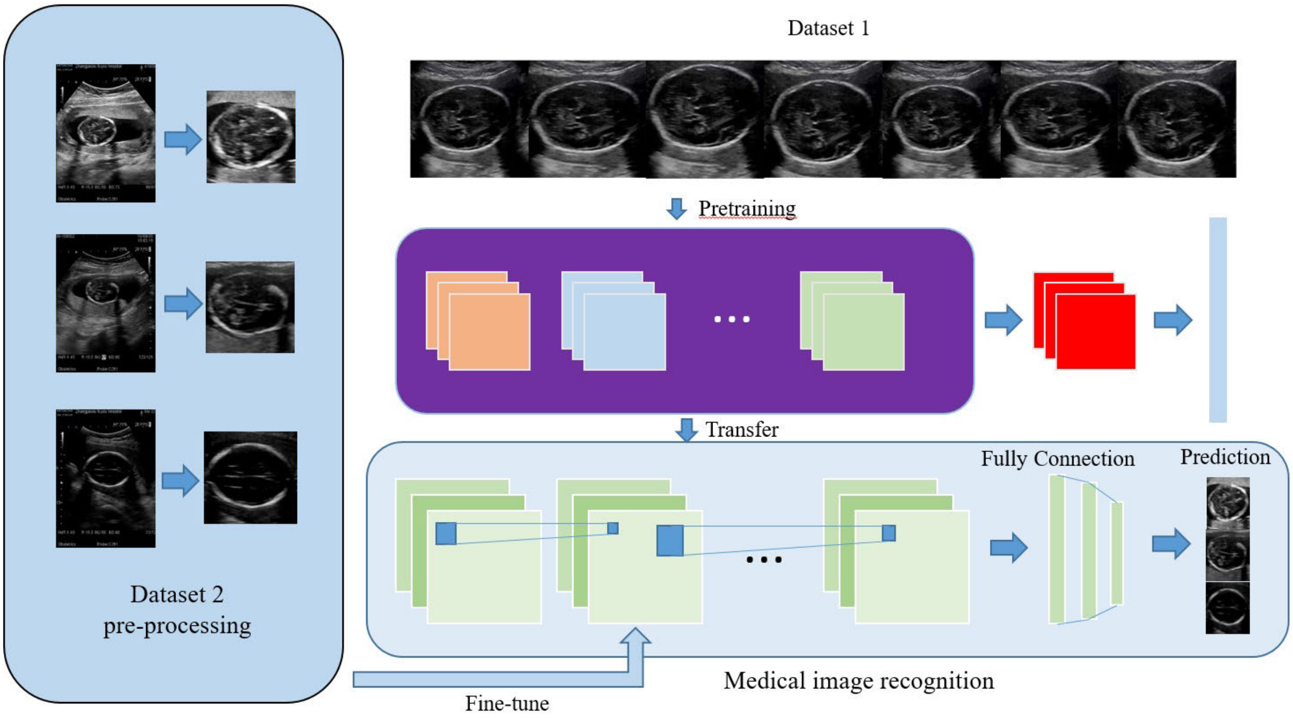


FIGURE 3. Flowchart of transfer learning.

DCNN in this paper are not complicated. The experimental results show that this learning strategy has not resulted in overfitting.

In this study, it is a challenge to transfer the knowledge acquired from the CNN trained by Dataset 1, which consists of 30,000 2D ultrasound images acquired from 155 subjects between 16 and 34 weeks, to Dataset 2, which contains 300 images acquired from one research participant over 40 weeks. Although the source dataset and target dataset both consist of fetal brain US images, we were curious regarding whether the feature similarities will confuse the deep learning system and bring about errors. The flowchart of transfer learning is shown in Fig. 3.

III. EXPERIMENTAL SETUP

A. PROPOSED DCNN MODEL

Appropriate network architecture design can improve network performance significantly. Our DCNN model contains five convolutional layers and five average-pooling layers between convolutional layers, followed by three fully connected layers, as shown in Table 1. *C* represents the convolutional neural network, *P* represents the pooling layer (average pooling layer in this structure), and *F* represents the fully connected layer. Initial values of the feature maps in convolutional neural networks are randomly generated by the system.

Once the proposed CNN is trained, the probability weight, also referred to as the feature map, can be calculated by the gradient descent method during forward and back propagation. The feature map rolls with a sliding window method. During the window sliding, the stride of sliding is according to the difference between two adjacent layers. To make the training processing fast and efficient, we choose the average pooling layers to down-sample the image. The stride of pooling, also called the down-sampling rate, is chosen to fit the size of two adjacent layers. At last, a final one-dimensional probability vector is obtained. After obtaining the final probability vector, we further smooth it and use bilateral filtering to eliminate noise. The final score of the image is the highest value of the smoothed probability vector. Finally, when the detection score is higher than the threshold, the highest detection US image with the highest detection score is identified as the brain standard plane.

TABLE 1. Architecture of proposed DCNN model.

Layer	Numbers of feature map	Kernel size	stride	layer
Input		11		1020*1020
C1	12	2	2	500*500*12
P1	1	5		250*250*12
C2	5	2	1	250*250*60
P2	1	6		125*125*60
C3	5	3	1	120*120*300
P3	1	3		40*40*300
C4	2	2	1	40*40*600
P4	1	3		20*20*600
C5	1	3	1	18*18*600
P5	1			6*6*600
F1				21600
F2				2520
F3				200

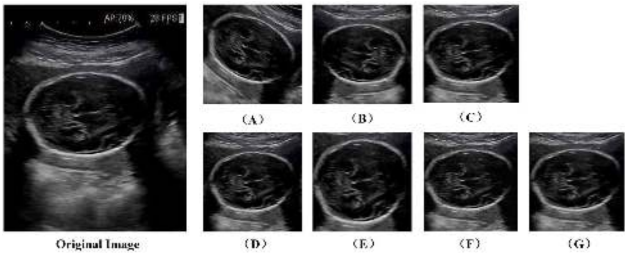


FIGURE 4. Samples showing results of four methods transformed from the original image. (A-G) respectively represent rotation, reflection, Gaussian white noise, flip, zoom, cubic spline interpolation and bilinear interpolation.

B. DATASET 1 AND DATA AUGMENTATION

A total of 19,142 fetal brain standard planes and other fetal plane images, which were constructed from 155 subjects containing 6 standard planes and other planes, were acquired by a Hitachi ARIETTA 70 B-mode ultrasonic apparatus with probe frequency of 4-6 MHz to form Dataset 1. Since the ultrasonic apparatus recorded images as an image sequence in one pregnancy examination, adjacent images in the sequences were usually similar. Therefore, we down-sampled the image sequences by a factor of 12. Even after this down-sampling, some images were still quite similar. Doctors at Zhangjiakou Maternal and Child Health Care Hospital helped us to select fetal brain standard planes and further deleted similar ones manually to keep the number of images recorded from the same event to less than 15. The images which were too blurry were also removed. All private information was removed from the images. All the images were annotated to provide the ground truth (such as horizontal transverse section of thalamus, horizontal transverse section of lateral ventricle, and transverse section of cerebellum) and some detailed information, such as thalamus (T), lateral ventricle (LV), and Sylvian fissure (SF).

Although CNN presents the advantage of learning powerful feature representations, the fully supervised deep architectures may overfit the training data in the case of limited training data in many medical applications. When the dataset is small, excessive parameters fit all characteristics of the dataset, rather than the commonalities between data. This makes the generation of the network perform terribly on the test dataset. Therefore, in order to prevent overfitting, data augmentation is increased. In this paper, we use the following methods to extend Dataset 1, as shown in Fig. 3. (1) Rotation: random rotation of the image at a certain angle; (2) Reflection: changing the orientation of the image content; (3) Gaussian white noise: adding random noise with normal distribution and the same energy density at all frequencies to the image; (4) Flip: flipping the images horizontally or vertically; (5) Zoom: zooming in or out of an image in a certain proportion; (6) Cubic spline interpolation: constructing a new image by the cubic spline interpolation method; (7) Bilinear interpolation: constructing a new image by bilinear interpolation. Figure 4 shows these four image transformations for augmentation.

After all the processing of the original dataset, a total of 30,000 1020 × 1020 fetal brain standard planes (including horizontal transverse section of thalamus, horizontal transverse section of lateral ventricle, transverse section of cerebellum, midsagittal plane, paracentral sagittal section, and coronal section of the anterior horn of the lateral ventricles) and other fetal plane images formed Dataset 1. In Dataset 1, there are 4,000 images for each standard plane and 6,000 images for other planes. We set the ratio of training set, validation set and testing set as 6:2:2, resulting in 18,000 images, 6,000 images and 6,000 images, respectively.

C. DATASET 2 AND SYSTEM IMPLEMENTATION

After Institutional Review Board (IRB) approval, we recorded a set of images corresponding to a complete pregnancy. With the permission of the pregnant woman, her US video during the whole pregnancy was recorded and kept. Since the adjacent images in the video change little, we down-sampled the image sequences by a factor of 10, so that the resulting images were separated by 12-15 seconds. Even after this down-sampling, some images were still quite similar. Due to this persistent similarity in the images, doctors helped us to further manually delete similar images, blurry images and excessively dark images to keep the number of images from the same event to less than 10. Finally, we collected 1,200 images to form Dataset 2. In Dataset 2, there are 150 images for each of the standard planes and 300 images for other planes. We set the training set, validation set and testing set ratio as 6:2:2, resulting in 720 images, 240 images and 240 images, respectively. The images in Dataset 2 do not intersect with those of Dataset 1.

Overfitting of the learning performance due to the small amount of image data was considered. In recent years, many studies have demonstrated that transfer learning is a powerful tool which can reduce overfitting by first training the basic network on the basic datasets and tasks, then transferring the learning features of the basic network to the new target network, and then training the target datasets and tasks. Inspired by these studies, we attempt to investigate whether the knowledge acquired from Dataset 1 formed by 155 subjects can be transferred to a dataset formed from one subject during the entire pregnancy, where the training dataset is limited and directly using CNN on Dataset 2 may lead to some extent of overfitting.

Noting that all ultrasonic scan images are converted into gray images, the input of CNN is thus single channel. To transfer the knowledge from Dataset 1, we construct a CNN-based on the proposed DCNN as mentioned earlier. Then, the pretrained convolutional layers are implanted to the same positions. Meanwhile, the parameters in fully connected layers are randomly initialized with Gaussian distribution. The strategy of dropout is applied for regularization to improve the generalization ability. This means that only the structures of CNN and pretrained convolutional layers are transferred from the proposed DCNN trained by Dataset 1, while other information including pooling layers and fully connected layers is trained as usual.

IV. RESULTS

We carried out the experiment based on the HP Z640 Workstation, Intel Xeon E5-2620 v4 2.1 2133 8C CPU and the GPU, NVIDIA Quadro M4000 8GB GeForce. The framework we used to establish the CNN architecture is TensorFlow. Our system was also implemented with the mixed programming technology of MATLAB. The running time for detecting one fetal brain standard scan plane in a testing set from the image flow is approximately 1.2 s.

A. QUANTITATIVE PERFORMANCE EVALUATION AND COMPARISON

We quantitatively evaluated the performance of our method in two experiments. In the first experiment, in order to determine the effectiveness of the proposed DCNN classifier, we compared the performance of four methods for detecting standard planes from Dataset 1, including three classical machine learning methods which perform outstanding with respect to automatic recognition, such as K-means clustering, support vector machines and radial component-based model (RCM) methods. **K-means clustering** is a method of vector quantization. In machine learning, it aims to divide n observations into k clusters, each of which belongs to the clustering with the nearest mean. In addition to k-means, **support-vector machine (SVM)** is a supervised learning model. It maps the points in space in order to divide the examples of individual categories into an obvious gap which is as wide as possible. **RCM [14]** is a novel method that was incorporated in the detection procedure to improve the performance.

To compare the differences between different algorithms, we computed the following statistical indices: Accuracy (A), Precision (P), Recall (R) and F1-measure (F_1).

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 = \frac{2RP}{R + P} \tag{5}$$

Here, TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives, respectively.

The precision-recall (PR) curves and receiver operating characteristic (ROC) curves are used to evaluate the classification performance of machine learning algorithms for a given dataset. Each dataset contains a fixed number of positive and negative samples. There is a deep relationship between ROC curve and PR curve:

For a given dataset containing positive and negative samples, there is a one-to-one correspondence between ROC plane and PR plane; i.e., if recall is not equal to 0, the planes contain completely identical obfuscation matrices. We can convert an ROC curve into a PR curve. For a given number of positive and negative sample datasets, one curve has an advantage over another curve in the ROC plane if and only if the first curve has an advantage over the second curve in the PR plane. (Here, ''a curve has advantages over other curves'' means that all parts of other curves coincide with or are below this curve.)

TABLE 2. Results of fetal brain standard plane recognition for dataset 1.

Method	Accuracy	Precision	Recall	F1–measure
Cluster	0.607	0.711	0.741	0.704
SVM	0.780	0.812	0.846	0.897
RCM	0.733	0.845	0.913	0.863
DCNN	**0.910**	**0.855**	**0.901**	**0.900**

In the precision-recall plane, the more convex the ROC curve, the better the effect it exhibits. In comparison with the left upper convex ROC curve, the right upper convex PR curve is superior.

Area under curve (AUC) refers to the proportion of the area under the ROC curve to the total square value. Sometimes, the ROC curves of different classification algorithms intersect, so the AUC value is often used as the criterion to judge the performance of the algorithm. The larger the area, the better the classification performance.

1) EVALUATION OF DATASET 1

In this experiment, the image is classified as a certain category when its final vector score is the highest. As shown in Table 2, the accuracy, precision, recall, and F1-measure values of the proposed DCNN on the testing data were 0.910, 0.855, 0.901 and 0.900, respectively, which significantly outperformed the other methods. The precision-recall and receiver operating characteristic (ROC) curves were also shown in Fig. 6. The areas under the ROC curve (AUC) obtained by the proposed DCNN, RCM, SVM and cluster methods are 0.90, 0.87, 0.75 and 0.71, respectively. The DCNN method we proposed in this paper achieved the best performance.

Also, we compared classification results between the proposed DCNN that includes data augmentation and the DCNN that without using data augmentation.

Table 3 shows that simple image data augmentation methods such as rotation and stretching have little impact on the output effect of SVM method, with the accuracy increased by only 1.2% from 77.1% to 78.0%. This is because SVM method is very sufficient for the feature extraction of small sample image data, and the image deformation of training set and testing set is unified. However, the accuracy of K-means clustering method decreased by 13% from 70.1% to 60.7%. This is because K-means clustering algorithm use the Euclidean distance of vector to measure the similarity between pixels. When the Euclidean distance is greater than a certain threshold, it is an unsupervised learning method to allocate the pixels to similar image areas. This method

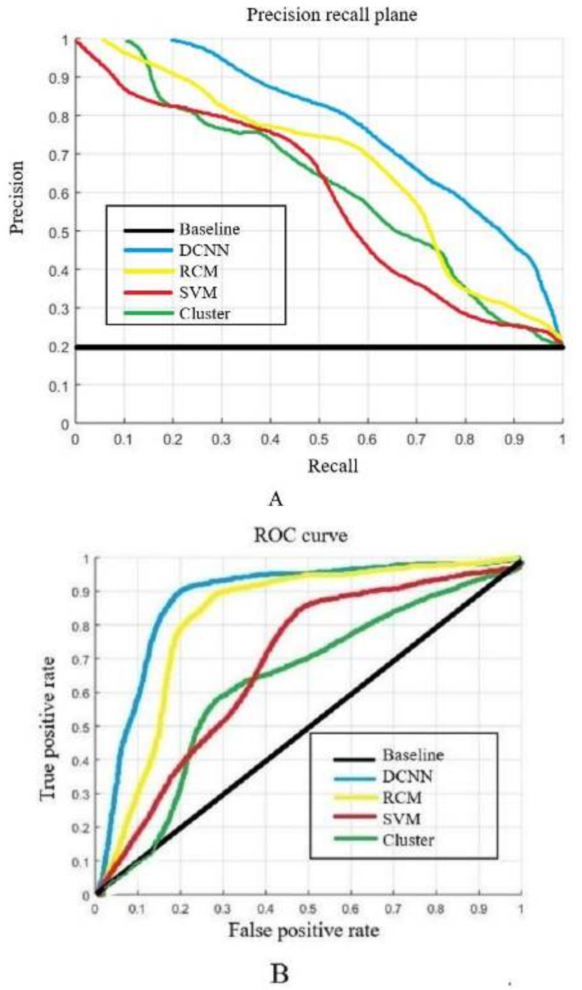


FIGURE 5. A: Precision recall plane. B: ROC curve.

TABLE 3. Comparison of K-means, SVM and CNN on dataset 1 with data augmentation and without data augmentation.

Algorithm	Data augmentation?	Accuracy	Precision	Recall	F1-measure
K-means	Yes	0.607	0.711	0.741	0.704
	No	0.701	0.720	0.761	0.758
SVM	Yes	0.780	0.812	0.846	0.897
	No	0.771	0.801	0.845	0.876
DCNN	**Yes**	**0.910**	**0.855**	**0.901**	**0.900**
	No	0.651	0.601	0.685	0.661

is easy to be interfered by image deformation. When the clustering center changes, the clustering effect will be quite different. This change will not improve the accuracy of image classification, but will mistakenly classify the input image due to the interference of image deformation, thus affecting the classification effect.

Because the local receptive field of the convolutional neural network can obtain some basic features of the image, such as the edge and angle in the image, the maximum pooling has the consistency of expression, which makes the convolutional neural network have a certain degree of relative invariance to geometric transformations such as image displacement, stretching and rotation. The data augmentation method used in this paper raised the output accuracy of the convolutional neural network by 40.0%.

TABLE 4. Results of fetal brain standard plane recognition for dataset 2.

Method	Accuracy	Precision	Recall	F1-measure
Cluster	0.532	0.715	0.691	0.704
SVM	0.882	0.844	0.956	0.897
RCM	0.734	0.785	0.813	0.851
DCNN	0.833	0.842	0.887	0.831
DCNN (data augmentation)	0.854	0.838	0.791	0.825
Transfer learning	**0.891**	**0.853**	**0.864**	**0.901**

2) EVALUATION OF DATASET 2

In the second part of the experiment, we used the same rules to evaluate the performance of the transfer learning-based method on Dataset 2.

As shown in Table 4, the accuracy, precision, recall, and F1-measure values of the proposed transfer learning-based method for the testing data were 0.891, 0.853, 0.864 and 0.901, respectively, which significantly outperformed the other methods. The precision-recall and receiver operating characteristic (ROC) curves were also shown in Fig. 7. The areas under the ROC curve (AUC) obtained by the proposed transfer learning method, DCNN, RCM, SVM and cluster methods are 0.90, 0.83, 0.79, 0.87 and 0.69, respectively. The DCNN method we proposed in this paper achieved the best performance. Additionally, we found that the SVM method can extract image features more comprehensively on a small dataset, and so its performance is relatively good. Meanwhile, an inappropriate use of DCNN on the small dataset causes network out-fitting on the testing set, resulting in relatively poor performance. Even after data augmentation, the CNN system is still difficult to extract enough features via the small data sample.

V. DISCUSSION

In this paper, we proposed two CNN-based deep learning methods for automatic recognition of fetal brain standard planes from US image sequences. In fact, inaccurate recognition results are mainly caused by the following reasons: (1) the differences between features in different planes are very small, and the sizes of feature areas are small; (2) there are fewer data to be trained. Based on these reasons, we selected small feature maps and manipulated the data by extending the depth of the CNN to make the network more sensitive to small features. Meanwhile, we proposed image transformation and domain transfer learning to solve the problem of overfitting caused by the lack of training data.

In fact, the main challenge of using CNN-based deep learning in medical applications is the lack of training data. Generally, the availability of data from medical applications is much less than that in other areas, and the underlying reason for this is that collecting data from patients requires the approval of ethics committees, which requires long periods

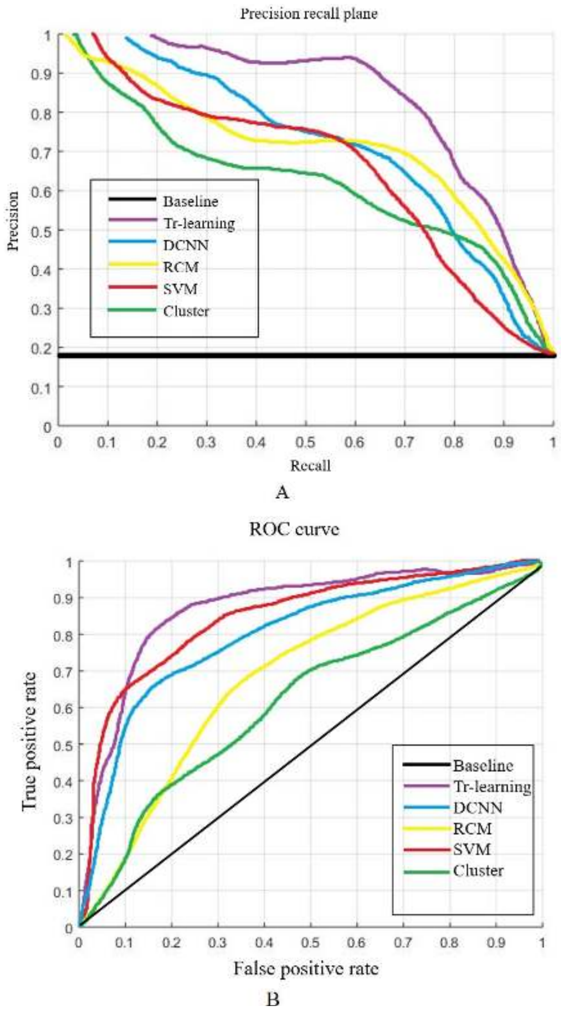


FIGURE 6. A: Precision recall plane. B: ROC curve.

of time. In fact, many pregnant women refuse to share their private US data for scientific research. Therefore, the overfitting problem caused by limited training datasets will affect the performance of the learning system and bring difficulties to clinical application, which is a major challenge faced in the medical computing community. Currently, investigating new therapies through the analysis of massive amounts of data has become the next frontier of modern medicine. Deep learning models represent a breakthrough over traditional methods in solving long-term computing problems, and they have been applied for various applications. Our proposed data augmentation and domain transferred learning methods will help the medical computing community to solve the challenges of limited size of datasets and promote the application of CNN in the medical field.

VI. CONCLUSION

To assist doctors in improving automatic diagnosis efficiency and accuracy, we proposed a novel deep CNN-based method and a domain transferred CNN model for automatic recognition of the fetal brain standard planes from US image sequences. Data augmentation and knowledge transfer were also adopted to reduce the overfitting for improvement of recognition performance. To elaborate the proposed methods and investigate their effectiveness, we collected and established two datasets. Numerous experiments have been carried out on the fetal brain plane dataset collected by our group, which proves that our method is superior to the traditional classification model. In addition, our experiments showed the effectiveness of data augmentation, especially in the case of insufficient training data. The proposed method demonstrated great prospects for deep learning in clinical application.

REFERENCES

[1] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Trans. Med. Imag.*, vol. 25, no. 8, pp. 987–1010, Aug. 2006.

[2] B. Rahmatullah, A. Papageorghiou, and J. A. Noble, "Automated selection of standardized planes from ultrasound volume," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2011, pp. 35–42.

[3] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 507–514.

[4] B. Rahmatullah and J. A. Noble, "Anatomical object detection in fetal ultrasound: Computer-expert agreements," in *Proc. Int. Conf. Biomed. Inform. Technol.*, 2014, pp. 207–218.

[5] S. Bodzioch and M. R. Ogiela, "New approach to gallbladder ultrasonic images analysis and lesions recognition," *Comput. Med. Imag. Graph.*, vol. 33, pp. 154–170, Mar. 2009.

[6] Q. Huang, E. Yang, L. Liu, and X. Li, "Automatic segmentation of breast lesions for interaction in ultrasonic computer-aided diagnosis," *Inf. Sci.*, vol. 314, pp. 293–310, Sep. 2015.

[7] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, pp. 299–317, Jan. 2010.

[8] B. Rahmatullah, A. T. Papageorghiou, and J. A. Noble, "Integration of local and global features for anatomical object detection in ultrasound," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2012, pp. 402–409.

[9] R. Kwitt, N. Vasconcelos, S. Razzaque, and S. Aylward, "Localizing target structures in ultrasound video—A phantom study," *Med. Image Anal.*, vol. 17, no. 7, pp. 712–722, 2013.

[10] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.

[11] B. Karimi and A. Krzy ak, "A novel approach for automatic detection and classification of suspicious lesions in breast ultrasound images," *J. Artif. Intell. Soft Comput. Res.*, vol. 3, no. 4, pp. 265–276, 2008.

[12] Y.-L. Huang, D.-R. Chen, Y.-R. Jiang, S.-J. Kuo, H.-K. Wu, and W. K. Moon, "Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound," *Ultrasound Obstetrics Gynecol.*, vol. 32, no. 4, pp. 565–572, Sep. 2008.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[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 Med. Biol.*, vol. 40, no. 11, pp. 2728–2742, 2014.

[15] J. K. Udupa, V. R. LaBlanc, H. Schmidt, C. Imielinska, P. K. Saha, G. J. Grevera, Y. Zhuge, L. M. Currie, P. Molholt, and Y. Jin, "Methodology for evaluating image-segmentation algorithms," *Proc. SPIE*, vol. 4684, pp. 266–278, May 2002.

[16] M. W. Attia, F. E. Z. Abou-Chadi, H. El-Din Moustafa, and N. Mekky, "Classification of ultrasound kidney images using PCA and neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 4, pp. 53–57, 2015.

[17] S. Manikandan and V. Rajamani, "A mathematical approach for feature selection and image retrieval of ultra sound kidney image databases," *Eur. J. Sci. Res.*, vol. 24, no. 2, pp. 163–171, 2008.

[18] K. I. Laws, "Rapid texture identification," *Proc. SPIE*, vol. 0238, pp. 376–382, Dec. 1980.

[19] E. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 143–156.

[20] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3304–3311.

[21] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis," *Ultrasound Med. Biol*, vol. 29, no. 5, pp. 679–686, 2003.

[22] W.-J. Wu, S.-W. Lin, and W. K. Moon, "Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images," *Comput. Med. Imag. Graph.*, vol. 36, no. 8, pp. 627–633 2012.

[23] C. P. Bridge, C. Ioannou, and J. A. Noble, "Automated annotation and quantitative description of ultrasound videos of the fetal heart," *Med. Image Anal.*, vol. 36, pp. 147–161, Feb. 2017.

[24] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang, and B. Hu, "Liver fibrosis classification based on transfer learning and FCNet for ultrasound images," *IEEE Access*, vol. 5, pp. 5804–5810, 2017.

[25] B. Lei, L. Zhuo, S. Chen, S. Li, D. Ni, and T. Wang, "Automatic recognition of fetal standard plane in ultrasound image," in *Proc. IEEE 11th Int. Symp. Biomed. Imag.*, Apr./May 2014, pp. 85–88.

[26] 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," *Med. Phys.*, vol. 39, no. 8, pp. 5015–5027, 2012.

[27] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[31] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[32] A. Karpath and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[33] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.

[34] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2012, pp. 25–30.

[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

RUOWEI QU received the B.S. degree from the School of Science, University of Science and Technology Beijing, Beijing, China, in 2012. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin, China. Her research interests include medical image processing, pattern recognition, and machine learning.

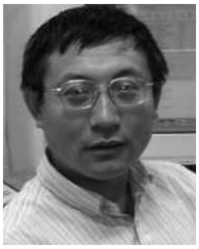
GUIZHI XU (Member, IEEE) was born in 1962. She received the Ph.D. degree from the School of Electrical Engineering, Hebei University of Technology, Tianjin, China, in 2002.

She is currently a Professor, a Doctoral Advisor, and the Dean of the School of Electrical Engineering, Hebei University of Technology. She is also the Head of key subjects at the provincial level of biomedical engineering, a Professor of metaoptics with the Hebei University of Technology, and the Head of the National top-quality course of engineering electromagnetic field. She published more than 90 academic articles retrieved by SCI and EI, and published three monographs. She presided over one key project of the National Natural Science Foundation, three projects of the National Natural Science Foundation, and one preresearch project of the Ministry of General Equipment, and completed two key projects of the National Natural Science Foundation in cooperation with Tsinghua University and the Fourth Military Medical University.

She received Hebei Science and Technology Outstanding Contribution Award, Hebei Natural Science Second Prize and Third Prize, Hebei Science and Technology Progress Second Prize and Third Prize, and Hebei Excellent Teaching Achievement Second Prize. She received the honorary titles of the first famous teaching teacher in Hebei Province, the outstanding young and middle-aged experts in Hebei Province, the outstanding young and middle-aged backbone teachers in Hebei Province, the advanced individuals in Hebei Province, and the outstanding Communist Party members in Hebei Province's education system.

CHUNXIA DING received the Bachelor of Medicine degree from the Zhangjiakou Medical College, in 2003. She has been engaged in ultrasound diagnosis for 16 years and worked on prenatal screening for more than ten years. She is skilled in gynecology and obstetrics, heart, blood vessels, abdomen, superficial small organs, and neonatal brain, and other conventional diagnosis, prenatal screening, and fetal malformation diagnosis.

WENYAN JIA received the B.S. and M.S. degrees in biomedical engineering from Capital Medical University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree in biomedical engineering from Tsinghua University, Beijing, in 2005. She is currently a Research Assistant Professor with the Department of Neurosurgery, University of Pittsburgh, Pittsburgh, PA. Her current research interests include biomedical signal processing and brain–computer interface.



MINGUI SUN (Fellow, IEEE) received the B.S. degree in instrumental and industrial automation from the Shenyang Chemical Engineering Institute, Shenyang, China, in 1982, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 1986 and 1989, respectively. He is currently a Professor in neurosurgery, electrical and computer engineering, and bioengineering with the University of Pittsburgh. His current research interests include advanced biomedical electronic devices, biomedical signal and image processing, sensors and transducers, and artificial neural networks.