Thorax Disease Diagnosis Using Deep Convolutional Neural Network *

Jie Chen, *Member, IEEE*, Xianbiao Qi, Osmo Tervonen, Olli Silvén, Guoying Zhao, *senior Member, IEEE* and Matti Pietikäinen, *Fellow*, *IEEE*

Abstract **Computer aided diagnosis (CAD) is an important issue, which can significantly improve the efficiency of doctors. In this paper, we propose a deep convolutional neural network (CNN) based method for thorax disease diagnosis. We firstly align the images by matching the interest points between the images, and then enlarge the dataset by using Gaussian scale space theory. After that we use the enlarged dataset to train a deep CNN model and apply the obtained model for the diagnosis of new test data. Our experimental results show our method achieves very promising results.**

I. Introduction

In radiology, one of the widely used examinations is chest radiographs. They are important for the diagnosis of various diseases associated with high mortality. To improve the efficiency of radiologists, we propose a new algorithm to preprocess the thorax (chest) radiographic images and filter out the images taken from healthy persons. Only those images suspected abnormal will be further checked by doctors. It would save the time of doctors significantly since most of the chest x-ray images (e.g., 70-80%) are from healthy people. Some chest radiographic images are shown in Fig. 1.

The proposed framework is shown in Fig. 2, we first perform the image alignment and then we enlarge the dataset using Scale Space Theory (SST) [8]. After that, we augment the dataset by different cropping, flipping and rotating. These augmented images are used for a CNN model training and fine-tuning. Finally we employ the fine-tuned model to perform the diagnosis of new test data.

Specifically, the collected images show significant variations in, e.g., illuminations, poses and scales. These images might come from different hospitals, collected by different machines and different operators. In addition, some patients suffer from illness, which also makes them difficult to be in an ideal pose for image capturing. To remove these variations, we need to perform the image alignment. For this task, we detect the points of interest and then use scale-invariant feature transform (SIFT) as a feature descriptor [14] to perform image matching to align these images. We then augment the dataset to train a classifier for better performance. Specifically, we propose a Gaussian scale space theory to enlarge the dataset. We perform this step since the image labeling is a laborious work and it usually needs radiologists to perform this task. After augmenting the dataset, we use it to train a deep learning model, and use the obtained model to classify new testing thorax images for screening task. The experimental results show our method achieves very promising results.

(a) (b) (c)

Fig. 1. Chest x-ray (chest radiographic images). (a) and (b) are normal chest radiographs. (c) is Q fever pneumonia (Q fever is a disease caused by infection with *Coxiella burnetii*).

The novelty of our method is that we propose a new framework to augment the dataset dramatically. Using the augmented dataset to train a CNN model, we improve the model performance significantly.

II. Related work

In this section, we will shortly introduce some methods about deep learning and computer aided diagnosis (CAD) for medical images.

Deep Learning (DL) has performed very well, e.g., in hand-written digit recognition [3, 11], object recognition [10], edge detection [21], and face verification [19, 22]. The success of DL attributes to the development of parallel computing, cloud computing and big data.

The CAD systems have been widely used for medical disease diagnosis [1, 6, 9, 10, 11, 16, 17, 18, 20, and 23]. For example, Lin et al. [12] combined neural network and fuzzy logic for nodule detection. Korfiatis et al. [9] utilized a selective enhancement filter to extract some 2D candidates at first. Liu et al. [13] proposed to use deep learning for early diagnosis of Alzheimer's disease. Roth et al. [16] proposed to improve computer-aided detection using convolutional neural networks and random view aggregation. Bar et al. [1] examined the strength of deep learning approaches for pathology detection in chest radiograph data. In [1], they combined the CNN model trained with *ImageNet* and "Picture Codes" (PiCoDes) descriptor to test over 93 frontal chest x-ray images. Different from [1] we use medical images to train a CNN model. Furthermore, we augment the training set using the Gaussian scale space theory. The experimental results show that the augmented dataset improves the performance of CNN model significantly.

*Research supported by Academy of Finland and European Regional Development Fund.

Jie Chen is with the University of Oulu, Finland (phone: +358-29448-2898; e-mail: jiechen@ee.oulu.fi).

Xianbiao Qi, Olli Silvén, Guoying Zhao, and Matti Pietikäinen are with the University of Oulu, Finland (e-mail: qixianbiao@gmail.com; olli.silven@ee.oulu.fi; gyzhao@ee.oulu.fi; mkp@ee.oulu.fi).

Osmo Tervonen is with Oulu University Hospital, Finland. (e-mail: otervone@gmail.com).

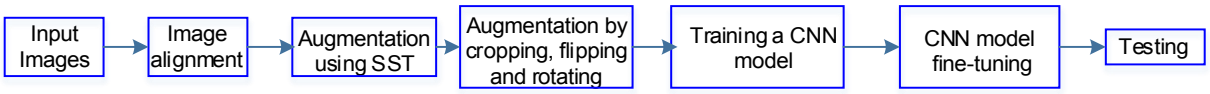


Fig 2. Framework of our method

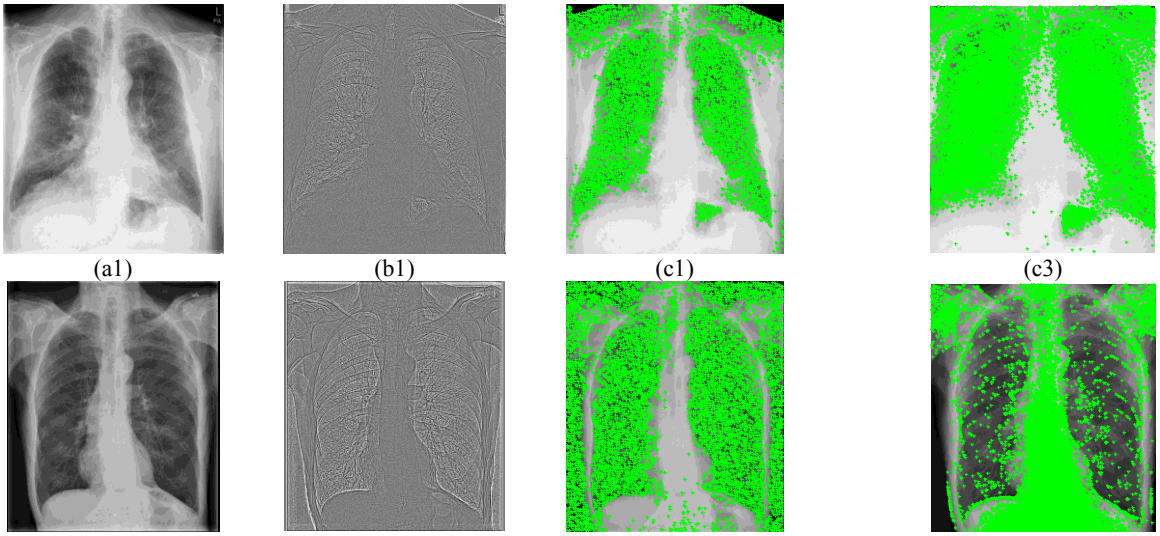


Fig. 3. Image alignment. (a1) and (a2) are two chest radiographic images. (b1) and (b2) are two images filtered by the WLD descriptor. (c1) and (c2) show the points of interest extracted by the Harris corner detector computed from (b1) and (b2). (c3) and (c4) are the results from Harris corner detectors computed from (a1) and (a2) directly.

Table 1: CNN architecture.

conv-1	conv-2	conv-3	conv-4	conv-5	full-6	full-7	full-8
96×7×7 st.2, pad 0 LRN, ×3 pool	256×5×5 st.1, pad 1 ×2 pool	512×3×3 st.1, pad 1 -	512×3×3 st.1, pad 1 -	512×3×3 st.1, pad 1 ×3 pool	4096 drop-out	4096 drop-out	2 soft-max

III. METHODS

In this section, we will introduce our framework for medical image analysis in details.

A. Image alignment

From Fig. 1, one can see that these chest radiographic images are collected in different scales and orientations. We need to align them to reduce the intra-class variance.

To align the thorax images effectively, we use a local descriptor, called the Weber Local Descriptor (WLD) to filter these images [3]. This descriptor consists of two components: differential excitation and orientation. The differential excitation component is a function of the ratio between two terms: one is the relative intensity differences of a current pixel against its neighbors; the other is the intensity of the current pixel. The orientation component is the gradient orientation of the current pixel. Here, we only use the differential excitation part. It can be computed quite fast and extracts the edges in the images very well although the contrast is very small (See Fig. 3, (b1) and (b2)). We then compute the Harris corners as shown in Fig. 3, (c1) and (c2). After that, we align the images using SIFT descriptors [14, 24].

In Fig. 3, (c3) and (c4) show the results from Harris corner detector computed directly from (a1) and (a2). From these two resulting images, we can find that the Harris corner detector is easily affected by illumination variations. However, after WLD filtering, we get much better results as shown in Fig. 3, (c1) and (c2).

B. Gaussian Scale Space Theory

After aligning the images, we enlarge the dataset using SST [8], which is a multi-scale signal representation framework. Given a 2-D image $I(x,y)$, the scale space of the images is defined as follows:

$$S(x,y,\sigma) = f(x,y,\sigma) * I(x,y), \tag{1}$$

where * is the convolution operation; $f(x,y,\sigma)$ is a filtering function, and σ is the scale factor.

In SST, Gaussian function is the most widely used filtering function [8]. Here, we propose to use Gaussian Scale Space (GSS) as a pre-processing technique for data augmentation. The 2-D Gaussian filter can be defined as:

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{2}$$

In our case, we use three scales (i.e., σ=1, 2 and 3). GSS has the following advantages. Firstly, we use it to remove noise in the thorax images. Secondly, we enlarge the dataset with different scales.

C. Data augmentation

Besides enlarging the dataset using GSS, we also

augment the dataset by the following strategies, i.e., combining cropping, flipping and rotating the images. We first get five crops from each original image. Specifically, these cropping windows are extracted from the four corners and the center of the image. 224×224 crops are sampled randomly from the training set. In addition, the crops are taken from the whole training image. If one image is smaller than 256, we use spatial padding as that of during CNN training process. These crops are then rotated clockwise, each time by 10° and 20°. Subsequently, these crops are flipped about the y-axis.

Although data augmentation (e.g., rotation) increase intra-class variance, all the images in the same set has the similar distributions, which speeds up the converging of CNN training confronting a large training dataset compared to without image alignment. On the other hand, it also improves the generalization of trained CNN model by increasing the diversity of the training set.

D. Deep learning model

During the training and classification, we use one of the deep learning models, i.e., CNN, for this task. For this model, we follow the strategy by [2] and use CNN- Slow because it achieved the best performance in [2].

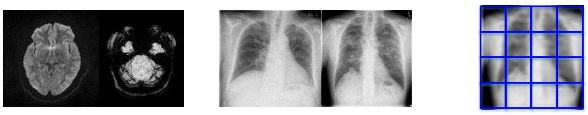
The CNN architecture representative is shown in Table 1. It is developed based on the Caffe framework [7]. It contains five convolutional layers (conv 1–5) and three fully-connected layers (full 6–8). The details of each of the convolutional layers are given in three sub-rows shown in Table 1: the first specifies the number of convolution filters and their respective field size as "num × size × size"; the second indicates the convolution stride ("st.") and spatial padding ("pad"); the third indicates if Local Response Normalisation (LRN) [10] is applied, and the max-pooling downsampling factor. For full 6–8, we specify their dimensionality in the last three columns of Table 1, which is the same for all three architectures. Full6 and full7 are regularized using dropout, while the last layer acts as a multi-way softmax classifier. The activation function for all weight layers (except for full8) is the Rectification Linear Unit (ReLU) [10]. The last fully-connected layer (full8) has output dimensionality equal to the number of classes.

IV. Experiments

A. Datasets

We collected two groups of datasets from local hospital. The first group is composed of magnetic resonance images (MRI) for brain (Fig. 4 (a)), which has 755,969 images and comes from 1,000 patients, with the dimensionality of 512×512. These images are provided with labels by radiologists, i.e., from a healthy or potential unhealthy person. Meanwhile, half of the images are from healthy persons and another half from unhealthy ones.

The second group consists of CT images for thorax (Fig. 4 (b)), which has 4000 images from 2000 patients, with the dimensionality about 2688*2688. These images have also been labeled, i.e., either from a healthy or potential unhealthy person. Likewise, half of the images are from healthy persons and half from unhealthy ones. All the labels are assigned by radiologists. For these CT images, we divide them into two parts. The first part consists of 2,000 images, half from healthy persons (negative) and half from unhealthy persons (positive), called CT_1. The second part also consists of 2,000 samples, half from positive samples and half from negative samples, called CT_2. Notes that the subjects/patients in CT_1 and CT_2 are not overlapped.



(a) (b) (c)

Fig. 4. (a) and (b) are example images collected in local hospital. Meanwhile, magnetic resonance images for brain are shown in (a) and computerized tomography images for thorax shown in (b). (c) is to divide a thorax image into patches.

Note that in our case, the first group of MRI dataset is augmented only by SST and then used for CNN model training. For the second group, CT_1 is used for fine-tuning and CT_2 is used for testing. Specifically, all the images in CT_1 are first aligned and then augmented by SST and by cropping, flipping and rotating as shown in Section III. After that we use the augmented CT_1 to fine tune the CNN model pre-trained using MRI images. The images in CT_2 are only aligned as shown in Section III-A and then used for testing without augmentation.

CNN training and fine-tuning. In our experiments, we train and fine-tune CNN using medical images. Specifically, we use the trained model in [2] as an initialization and then use augmented MRI images to re-train the CNN model and then use augmented CT_1 to fine tune the pre-trained CNN model.

B. Component analysis

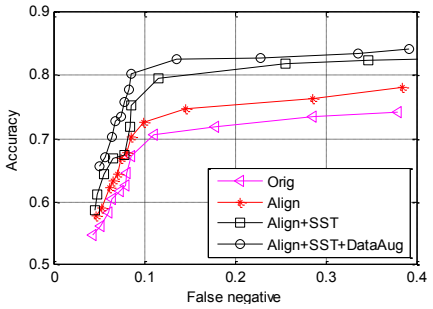
As shown in Fig. 5 (a), we show the performance of each component. Here, "*Orig*" means the original images without any pre-processing or augmentation. "*Align*" means we perform the alignment as shown in Section III-A. "*Align+SST*" means we align the images and then enlarge the dataset using SST. "*Align+SST+DataAug*" means we also perform the data augmentation as shown in Section III-C. All these datasets are used as training sets to train four CNN models respectively, and then the trained models are tested using CT_2.

In Fig. 5, the accuracy means that the percentage of both healthy and unhealthy samples are correctly classified. The false negative means that the percentage of unhealthy samples classified to be healthy ones. From Fig. 5 (a), we can find that aligning the images in the dataset improves the performance of the final CNN model by 2-3%. The SST improves another 3-4%. In addition, data augmentation achieves 2-3% more accuracy gain.

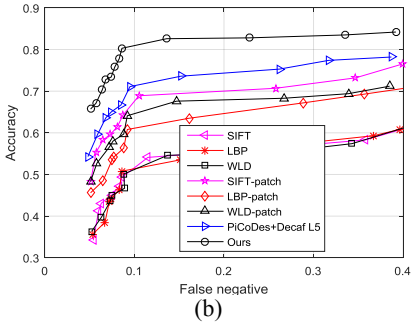
C. Comparison with other methods

In Fig. 5 (b), we compare our methods with widely used local descriptors, such as local binary pattern (LBP) [15], WLD and SIFT. For these local descriptors, we use the same datasets preprocessed and augmented from CT_1 as being used for CNN model training. After computing the descriptors from the images, we use support vector machine (SVM) with radial basis function (RBF) kernel for classification. *LBP_patch*, *WLD_patch* and *SIFT_patch* means that these

local descriptors are computed from CT image patches as shown in Fig. 4 (c). Dividing an image into patches and then computing the local descriptor for each patch for classification is widely used in pattern recognition for better performance. *PiCoDes+Decaf L5* is re-implemented by us following the idea of [1], which is trained using *ImageNet* and tested by the same set as our method, i.e., CT_2.



(a)



(b)

Fig. 5. (a) Component analysis. (b) Comparison with other methods.

Note that during the alignment, we only use differential excitation of WLD to filter the image for the next step of corner detection; we use sparse SIFT to represent local regions of points of interest for image matching. During classification, WLD (both differential excitation and orientation), LBP and SIFT are dense descriptors, which means to compute the features pixel by pixel and then compute these descriptor histograms for classification.

From Fig. 5 (b), we can see that the proposed method works much better than traditional local descriptors. Dividing the images into patches improves the performance of these local descriptors, e.g., LBP, WLD and SIFT. They still do not work as well as CNN model. One reason that local descriptors do not work well as CNN is that CNN model is trained using a large dataset. In addition, our method works better than *PiCoDes+Decaf L5*. One reason is that the data augmentation proposed in Section III improves the performance of the trained CNN model. The other reason is that the model in [1] (i.e., *PiCoDes+Decaf L5*) is trained with non-medical images which limited its performance in medical image analysis.

V. Conclusion

We propose a new framework to augment the dataset dramatically. Using the augmented dataset to train a CNN model for the thorax disease diagnosis, we improve the model performance significantly. One future work is to combine millions of images without labels collected from local hospital to improve the performance of the CNN models.

References

[1] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, Deep learning with non-medical training used for chest pathology identification, *Proc. of SPIE, Medical Imaging: Computer-Aided Diagnosis*, 2015

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, Return of the Devil in the Details: Delving Deep into Convolutional Nets, *BMVC*, 2014

[3] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, and W. Gao, WLD: A Robust Local Image Descriptor, *TPAMI*, 2010

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, *IEEE TPAMI*, 2010

[5] G. E. Hinton, S. Osindero and Y. Teh, A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

[6] A. Horvath and G. Horvath, Segmentation of chest X-ray radiographs, a new robust solution. *5th European Conference of the International Federation for Medical and Biological Engineering*, 2012.

[7] Y. Jia, E. Shelhamer, and J. Donahue, Caffe: Convolutional architecture for fast feature embedding, *ACM Inter. Conf. on Multimedia*, 2014.

[8] J. J. Koenderink and A. J. van Doorn, Surface shape and curvature scales, *Image and Vision Computing*, 1992

[9] P. Korfiatis, C. Kalogeropoulou, and L. Costaridou, Computer aided detection of lung nodules in multislice computed tomography, *International Conf. on Information Technology in Biomedicine*, 2006.

[10] A. Krizhevsky, Ilya Sutskever, and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *NIPS*, 2012.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[12] D. Lin, C. Van, and W. Chen, Autonomous detection of pulmonary nodules on CT images with a neural network-based fuzzy system, *Computerized Medical Imaging and Graphics*, 2005

[13] S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, Early Diagnosis Of Alzheimer's Disease With Deep Learning, *IEEE 11th International Symposium on Biomedical Imaging*, 2014

[14] D. Lowe, Distinctive Image Features from Scale Invariant Key Points, *International Journal of Computer Vision*, 2004.

[15] T. Ojala, M. Pietikäinen and T. Mäenpää, Multiresolution Gray Scale and Rotation Invariant Texture Analysis with Local Binary Patterns, *IEEE TPAMI*, 2002.

[16] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, Improving computer-aided detection using convolutional neural networks and random view aggregation. arXiv, 2016.

[17] M. Sofka, J. Zhang, S. Good, K. Zhou, and D. Comaniciu, Detection and Measurement of Structures in Fetal Head Ultrasound Volumes Using Sequential Estimation and Integrated Detection Network, *IEEE Trans. Medical Imaging*, 2014

[18] K. T. Spencer, B. J. Kimura, C. E. Korcarz, P. A. Pellikka, P. S. Rahko, and R. J. Siegel, Focused Cardiac Ultrasound: Recommendations from the American Society of Echocardiography, *Journal of the American Society of Echocardiography*, 2013

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification, *CVPR*, 2014

[20] P. Wang, O. Ecabert, T. Chen, M. Wels, M. Ostermeier, and D. Comaniciu, Image-based Co-Registration of Angiographjy and Intravascular Ultrasound Image, *IEEE Trans. Medical Imaging*, 2013

[21] S. Xie and Z. Tu, Holistically-Nested Edge Detection, *ICCV* 2015

[22] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T. Kim, Conditional Convolutional Neural Network for Modality-Aware Face Recognition, *ICCV* 2015

[23] Y. Zheng, D. Yang, M. John, and D. Comaniciu, Multi-Part Modeling and Segmentation of Left Atrium in C-Arm CT for Image-Guided Ablation of Atrial Fibrillation, *IEEE Trans. Medical Imaging*, 2014

[24] http://www.vlfeat.org/overview/sift.html