ISSN 1674-1056
http://cpb.iphy.ac.cn
http://iopscience.iop.org/cpb
中国物理 B
Chinese Physics
中国物理
Chinese Physics B

Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data

Jian Tang(汤键), Zhibing Jin(金志斌), Xue Zhou(周雪), Weijing Zhang(张玮婧), Min Wu(吴敏), Qinghong Shen(沈庆宏), Qian Cheng(程茜), Xueding Wang(王学鼎), Jie Yuan(袁杰)

Citation: Chin. Phys. B . 2019, 28(3): 038701. **doi:** 10.1088/1674-1056/28/3/038701
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Synthesization of high-capacity auto-associative memories using complex-valued neural networks

Yu-Jiao Huang(黄玉娇), Xiao-Yan Wang(汪晓妍), Hai-Xia Long(龙海霞), Xu-Hua Yang(杨旭华)
Chin. Phys. B . 2016, 25(12): 120701. **doi:** 10.1088/1674-1056/25/12/120701

Multistability of delayed complex-valued recurrent neural networks with discontinuous real-imaginary-type activation functions

Huang Yu-Jiao, Hu Hai-Gen
Chin. Phys. B . 2015, 24(12): 120701. **doi:** 10.1088/1674-1056/24/12/120701

Stability analysis of Markovian jumping stochastic Cohen–Grossberg neural networks with discrete and distributed time varying delays

M. Syed Ali
Chin. Phys. B . 2014, 23(6): 060702. **doi:** 10.1088/1674-1056/23/6/060702

Further studies on stability analysis of nonlinear Roesser-type two-dimensional systems

Dai Xiao-Lin
Chin. Phys. B . 2014, 23(4): 040701. **doi:** 10.1088/1674-1056/23/4/040701

Novel delay-distribution-dependent stability analysis for continuous-time recurrent neural networks with stochastic delay

Wang Shen-Quan, Feng Jian, Zhao Qing
Chin. Phys. B . 2012, 21(12): 120701. **doi:** 10.1088/1674-1056/21/12/120701

Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data*

Jian Tang(汤键)[1,†], Zhibin Jin(金志斌)[2,†], Xue Zhou(周雪)[1], Weijing Zhang(张玮婧)[2], Min Wu(吴敏)[2], Qinghong Shen(沈庆宏)[1], Qian Cheng(程茜)[3], Xueding Wang(王学鼎)[3], and Jie Yuan(袁杰)[1,2,3,‡]

[1] *School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China*
[2] *Nanjing Drum Tower Hospital, Nanjing 210023, China*
[3] *Institution of Acoustics, Tongji University, Shanghai 200092, China*

(Received 18 October 2018; revised manuscript received 19 December 2018; published online 23 January 2019)

The gray-scale ultrasound (US) imaging method is usually used to assess synovitis in rheumatoid arthritis (RA) in clinical practice. This four-grade scoring system depends highly on the sonographer's experience and has relatively lower validity compared with quantitative indexes. However, the training of a qualified sonographer is expensive and time-consuming while few studies focused on automatic RA grading methods. The purpose of this study is to propose an automatic RA grading method using deep convolutional neural networks (DCNN) to assist clinical assessment. Gray-scale ultrasound images of finger joints are taken as inputs while the output is the corresponding RA grading results. Firstly, we performed the auto-localization of synovium in the RA image and obtained a high precision in localization. In order to make up for the lack of a large annotated training dataset, we performed data augmentation to increase the number of training samples. Motivated by the approach of transfer learning, we pre-trained the GoogLeNet on ImageNet as a feature extractor and then fine-tuned it on our own dataset. The detection results showed an average precision exceeding 90%. In the experiment of grading RA severity, the four-grade classification accuracy exceeded 90% while the binary classification accuracies exceeded 95%. The results demonstrate that our proposed method achieves performances comparable to RA experts in multi-class classification. The promising results of our proposed DCNN-based RA grading method can have the ability to provide an objective and accurate reference to assist RA diagnosis and the training of sonographers.

Keywords: rheumatoid arthritis, convolutional neural network, medical ultrasound images

PACS: 87.57.–s, 07.05.Mh, 43.35.Wa **DOI:** 10.1088/1674-1056/28/3/038701

1. Introduction

Rheumatoid arthritis (RA) is one of the most common autoimmune diseases, which is characterized by pain, swelling, and stiffness of the joints, affecting almost 1% of the population around the world.[1–4] Detecting the diseased finger joint region of RA patients and quantifying its severity in progressive stages are important for diagnosis, treatment decisions, and prediction of disease progression. Clinically, RA severity is assessed using the gray-scale-evaluated synovial hyperplasia score, which is a four-grade semi-quantitative criterion.

Most of the related studies about RA severity grading concentrate on the detection of finger joints or the identification of bone erosions on the dataset of x-ray images instead of ultrasound images. Sungmin Lee *et al.* proposed a novel approach for the detection of all finger joints from hand radiograph images.[5] Seiichi Murakami *et al.* proposed a method to automatically identify bone erosions on hand radiographs with segmentation using the multiscale gradient vector flow snakes algorithm and an identification classifier.[6] Since ultrasound imaging is a more popular RA detection approach than x-ray computed tomography (CT) in RA grading, the research of detecting and quantifying the severity of RA with ultrasound images is recognized as helpful to clinical diagnosis.

The assessment of RA can be considered as an image classification[7,8] problem. Previous automatic medical image analysis methods are based on low-level pixel processing and mathematical modeling, such as region growing and edge detection.[9] Later, statistical classifiers and analysis methods have started to use data training with handcrafted features, such as the histogram of oriented gradient (HOG)[10] and the scale-invariant feature transform (SIFT).[11] The critical step in these methods is the extraction of discriminant features, which is difficult to be inducted from the processed images. Recently, deep learning algorithms represented by the convolutional neural network (CNN)[12,13] have made significant breakthroughs in medical image analysis. Studies have demonstrated that deep learning algorithms outperform traditional statistical learning methods, achieving high accuracy and stability.[14]

However, training a CNN from scratch requires an enormous amount of training data while the process of annotating medical images is time consuming, requiring medical

*Project supported by the National Key Research and Development Program of China (Grant No. 2017YFC0111402) and the Natural Science Funds of Jiangsu Province of China (Grant No. BK20181256).

†These authors contributed equally to this work.

‡Corresponding author. E-mail: yuanjie@nju.edu.cn

© 2019 Chinese Physical Society and IOP Publishing Ltd

specialists.[15] The sizes of medical datasets are usually at a level of hundreds or thousands, which is much smaller than those million-level datasets in computer vision. The lack of annotated medical data is a crucial challenge for the application of deep learning in medical image analysis, thus inspiring the employment of transfer learning. Essentially, transfer learning aims to extract knowledge from one or more source tasks and apply it to a new target task. The main transfer learning strategy for deep neural network training is using a pre-trained network (typically on natural image datasets such as ImageNet[16]) as a feature extractor and fine-tuning the pre-trained network on the target medical dataset. This strategy has been widely applied in the deep learning for medical image analysis.

In this study, we proposed an enhanced CNN-based grading scheme to solve the automatic grading of RA ultrasound images. The rest of the paper is organized as follows. In Section 2, the methods of data acquisition, data augmentation and the employed methods of localization and classification are all presented. Section 3 describes the experimental details and discusses the results. Finally, conclusions are given in Section 4.

2. Methods

2.1. Data acquisition

The RA ultrasound images were acquired from the Philips IU22 and the Philips EPIQ 7 ultrasound scanners in Nanjing Drum Tower Hospital by two professional clinicians, employing the high-frequency transducer (12–5 MHz) seen in Fig. 1. The scanning depth was set at 2.5 cm and the two-dimensional gain was adjusted to 50%–60% to compensate for variations in the acoustic properties of the patients. The metacarpophalangeal (MCP) and proximal-interphalangeal (PIP) joints were scanned on the dorsal aspect in the longitudinal plan, with the joint in a neutral position (Fig. 2). During the acquisition process, the ultrasound probe clung to the skin without tissue deformation.

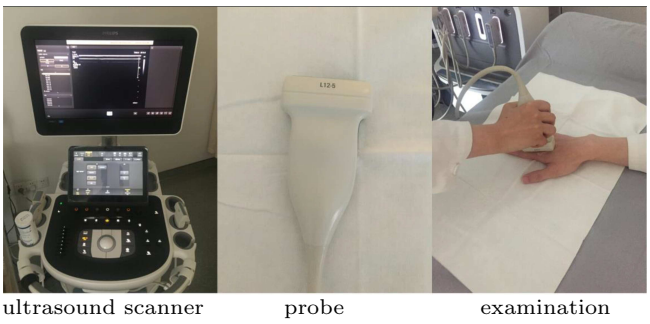


Fig. 1. The acquisition process of RA ultrasound images.

The gray-scale-assessed synovial hyperplasia scoring system defines four grades to classify RA severity from the ultrasound images.[3] Grade 0 corresponds to no synovial thickening. Grade 1 corresponds to the minimal synovial thickening that fills the angle between the periarticular bones without bulging over the line linking tops of the bones. Grade 2 corresponds to the moderate synovial thickening that bulges over the line linking tops of the periarticular bones but without extension along the bone diaphysis. Grade 3 corresponds to the severe synovial thickening that bulges over the line linking tops of the periarticular bones with extension to at least one of the bone diaphysis, as shown in Fig. 3.

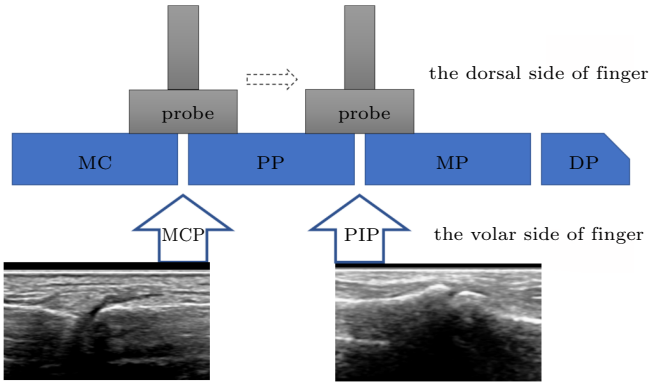


Fig. 2. The acquisition process of MCP images and PIP images. The MCP joint is the joint between metacarpal phalangeal bone and proximal phalange bone. The PIP joint is the joint between proximal and intermediate phalanx.

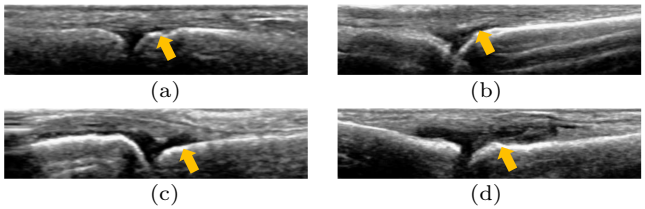


Fig. 3. Synovial changes on ultrasonography: (a) Grade 0, (b) Grade 1, (c) Grade 2, and (d) Grade 3.

2.2. Data augmentation

We employed two RA ultrasound image datasets for the network training. One contained 662 RA images of MCP joints while the other one contained 315 images of PIP joints. The samples were split into 80% training and 20% test sets.

Hundreds of ultrasound images are not sufficient for the training of CNN and may lead to a serious over-fitting problem.[17] Therefore, we preformed data augmentation on training samples to improve the generalization ability of the network. Data augmentation is essentially expanding the training set through different kinds of transformations which should preserve the data label. The data augmentation we took mainly included the following three methods.

2.2.1. Image crops

The size of original ultrasound images offered by clinicians is 1024×768. Through careful observation, we found that the region of interest (ROI) around the finger joint could always be included in a 520×120 rectangle. We cropped four

520×120 ROI areas from the original images as shown in Fig. 4. By this means, we augmented the size of the training set by four times.

2.2.2. Adding Gaussian noises

We added Gaussian noise to the cropped images as a common data augmentation method. Related studies, such as Refs. [18] and [19], employed this method to generate sufficient training examples and avoid the over-fitting problem.

2.2.3. Histogram equalization

We performed histogram equalization on the cropped images as another choice of augmentation which enhances the contrast of the processed images. Histogram equalization is also an effective approach for image augmentation which was adopted in the related studies such as Refs. [20] and [21].

By data augmentation, the size of training sets was increased by twelve times (see Fig. 4). The dataset statistics before and after data augmentation are shown in Table 1.

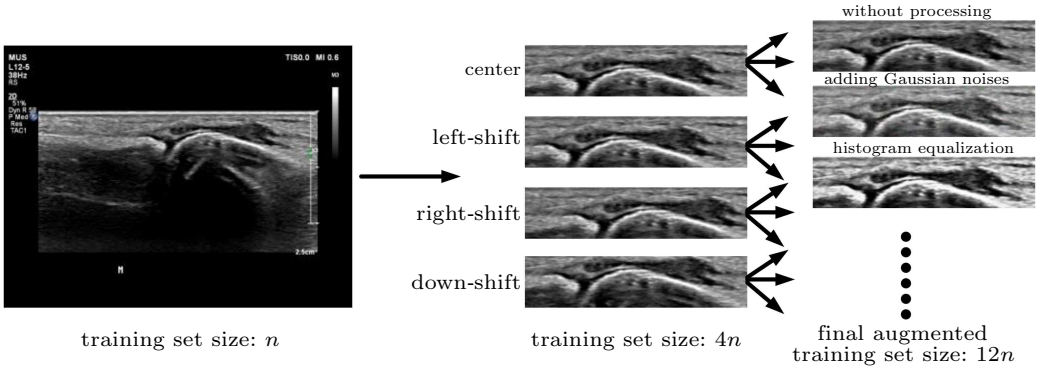


Fig. 4. The process of data augmentation.

Table 1. Statistics on MCP & PIP datasets (the slash character refers to the split of training set and test set).

MCP	Grade 0 to Grade 3		Grade 0 vs. Grade 1/2/3		Grade 1 vs. Grade 3	
	primary	augmented	primary	augmented	primary	augmented
Grade 0	244 (196/48)	2400 (2352/48)	244 (196/48)	2400 (2352/48)	–	–
Grade 1	179 (144/35)	1763 (1728/35)			179 (144/35)	1763 (1728/35)
Grade 2	158 (127/31)	1555 (1524/31)	418 (336/82)	4114 (4032/82)	–	–
Grade 3	81 (65/16)	796 (780/16)			81 (65/16)	796 (780/16)
Train	532	6384	532	6384	209	2508
Test	130	130	130	130	51	51
Total	662	6514	662	6514	260	2559
PIP	Grade 0 to Grade 3		Grade 0 vs. Grade 1/2/3		Grade 1 vs. Grade 3	
	primary	augmented	primary	augmented	primary	augmented
Grade 0	204 (164/40)	2008 (1968/40)	204 (164/40)	2008 (1968/40)	–	–
Grade 1	58 (47/11)	575 (564/11)			58 (47/11)	575 (564/11)
Grade 2	37 (30/7)	367 (360/7)	111 (90/21)	1101 (1080/21)	–	–
Grade 3	16 (13/3)	159 (156/3)			16 (13/3)	159 (156/3)
Train	254	3048	254	3048	60	720
Test	61	61	61	61	14	14
Total	315	3109	315	3109	74	734

2.3. Localization of synovium

To implement the auto localization of the synovium, we employed faster R-CNN,[22] which is a state-of-the-art object detection method. It consists of two modules, which are the region proposal network (RPN) and the fast R-CNN network[23] (see Fig. 5). The RPN is a fully convolutional network which predicts the object bounding box and its score. The fast R-CNN network uses the region proposals generated by RPN for detection. The RPN shares convolutional feature maps with the fast R-CNN, which can reduce its computing cost for generating region proposals. The first 3×3 convolutional layer takes a 3×3 sliding window of the convolutional feature map as input and each window is mapped to a lower-dimensional feature vector (such as 256-dimension for the Zeiler and Fergus network (ZFNet)[24]). Then this feature vector is sent into the next two parallel 1×1 convolutional layers for the classification and bounding box regression of the region proposals.

Anchors are applied in the RPN to detect objects of different scales and aspect ratios. An anchor is a bounding box centered at the sliding 3×3 window. Each anchor is denoted as positive or negative depending on whether it contains an object. Each sliding position in the convolutional feature map corresponds to k anchors (by default $k = 9$). For a convolu-

tional feature map of a size $W \times H$, there will be $W \times H \times k$ region proposals (see Fig. 6).

The faster R-CNN employs a multi-task loss on each anchor to jointly train for classification and bounding box regression,

$$L = \frac{1}{N_{\text{cls}}} \sum L_{\text{cls}} + \lambda [u \geqslant 1] \frac{1}{N_{\text{reg}}} \sum L_{\text{loc}}, \tag{1}$$

$$L_{\text{cls}} = -\log p_u, \tag{2}$$

$$L_{\text{loc}} = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L1}\left(t_i^u - v_i\right), \tag{3}$$

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2, & |x| > 1, \\ |x| - 0.5, & \text{otherwise}, \end{cases} \tag{4}$$

where p_u is the predicted probability of an anchor being an object and L_{cls} is the log loss for true class u. The localization class L_{loc} is defined over both the bounding box ground truth $v = (v_x, v_y, v_w, v_h)$ and the predicted bounding box coordinates $t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$ for class u. It is noted that the background class is labeled as $u = 0$ and hence L_{loc} is ignored for background anchors. L_{cls} is normalized by the mini-batch size N_{cls} and L_{loc} is normalized by the number of anchor locations N_{reg}. The balancing parameter λ is set to 10 by default. We emphasize that the bounding box regression loss is calculated with the smooth_{L1} function which is more robust to outliers than the commonly used $L2$ loss.

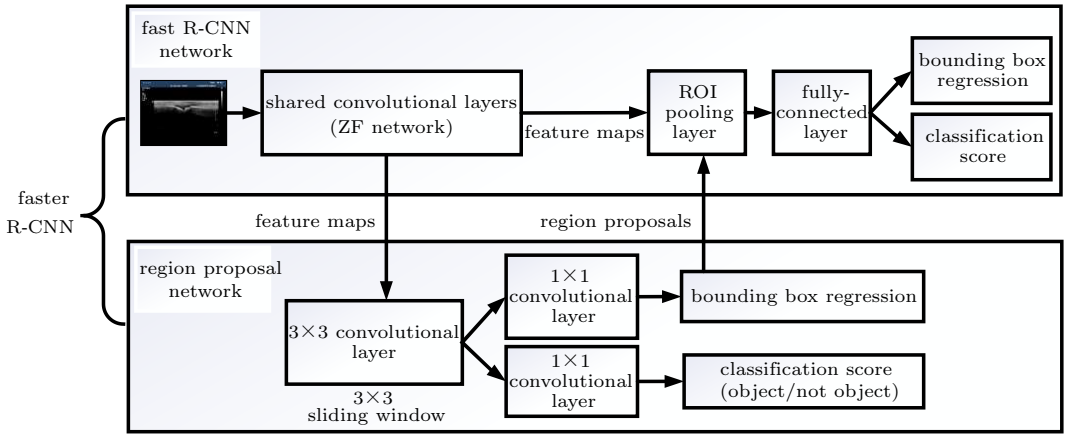


Fig. 5. The network structure of faster R-CNN.

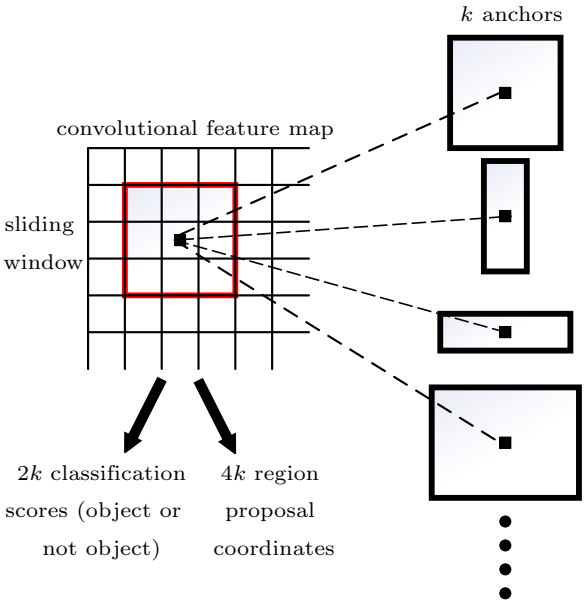


Fig. 6. The illustration of sliding window.

2.4. Grading of rheumatoid arthritis

We adopted a 22-layer GoogLeNet inception V1 network[25] to implement the RA severity grading. GoogLeNet is characterized by its inception module which is designed to use the available dense components to approximate an optimal local sparse structure in CNN. As shown in Fig. 7, the inception module is a parallel structure of 1×1 convolutional layer, 3×3 convolutional layer, 5×5 convolutional layer, and 3×3 max pooling layer. The convolution kernels with different sizes correspond to the different receptive fields, and the final combination means the fusion of features with different scales. An additional 1×1 convolutional layer is added before the convolution layers and after the max pooling layer to reduce the dimension of the input data. This implementation can considerably reduce the calculation parameters compared to a direct convolution. A rectified linear activation function is followed by increasing the representation ability of the network.

To alleviate the vanishing gradient problem,[26,27] two extra losses are added in the intermediate layers of the network with a discount weight during the training process. This implementation is based on the fact that features from the intermediate layers are also discriminative. The total loss of the network is calculated as

$$\text{loss}_{\text{total}} = 0.3 \times \text{loss}_1 + 0.3 \times \text{loss}_2 + \text{loss}_3, \tag{5}$$

where loss_1 and loss_2 are the losses of auxiliary classifiers in

the intermediate layers weighted by 0.3 and loss_3 is the loss of the final layer. Each loss function is calculated in the form of cross-entropy

$$P(y_i|x_i, w_t) = \text{softmax}(x_i, y_i, w_t) = e^{f_{y_i}} / \sum_j e^{f_j}, \quad (6)$$

$$\text{loss} = -\log P(y_i|x_i, w_t) = -\log\left(e^{f_{y_i}} / \sum_j e^{f_j}\right), \quad (7)$$

where f_j is the j-th element in the output vector of the final fully-connected layer, y_i is the label of the input image x_i, and w_t is the weight matrix of the network. The softmax function squashes the input vector to a vector of values between zero and one that sum to one.

Motivated by the transfer learning strategy, we employed the GoogLeNet models pre-trained on ImageNet, which is a large-scale image database with millions of annotated natural images. Then we fine-tuned it on our own MCP&PIP datasets. As Fig. 8 shows, we copied the weight of each layer in the pre-trained model to initialize our own fine-tuned model except the fully-connected layers. The fully-connected layers were re-named to initialize with random weights when the fine-tuned network started training.

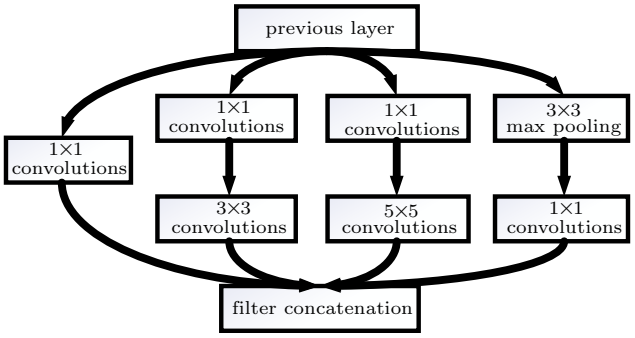


Fig. 7. Structure of the inception module.

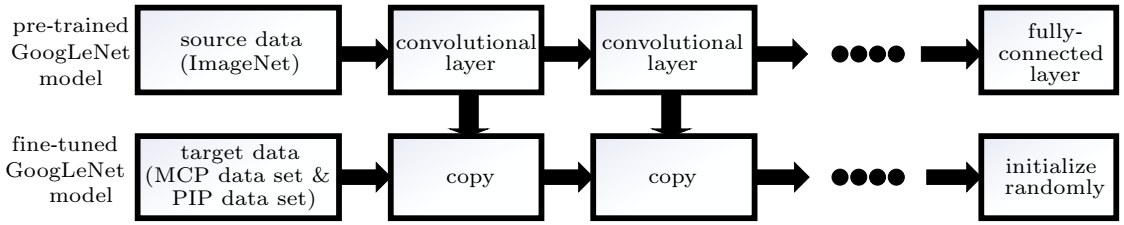


Fig. 8. The illustration of fine-tuning.

3. Experiments and discussion

3.1. Assessment method

To ensure the stability and reliability of the testing results, we preformed five-fold cross validation in the synovium localization and RA severity grading for all datasets (MCP&PIP: Grade 0 to Grade 3, Grade 0 vs. Grade 1/2/3, Grade 1 vs. Grade 3). For each dataset, we split the whole dataset D into five mutually exclusive subsets with equal size, which means

$$D = D_1 \cup D_2 \cup D_3 \cup D_4 \cup D_5, \quad D_i \cap D_j = \emptyset, \ (i \neq j). \quad (8)$$

The five subsets were obtained by stratified sampling to maintain the consistency of data characteristics. Each time we used the union of four subsets as the training set and performed data augmentation on it and took the remaining one as the test set. Through such methods, we obtained five groups of training/test sets. The final result was averaged over all five test results (see Fig. 9).

To quantitatively evaluate the performance of our grading model, we calculated four common machine-learning assessment metrics: accuracy, precision, recall, and $F1_{\text{measure}}$. These metrics can be calculated by the prediction result of all tested images. Taking the binary classification problem of normal people and the patients with synovial thickening (Grade 0 vs. Grade 1/2/3) for an example,

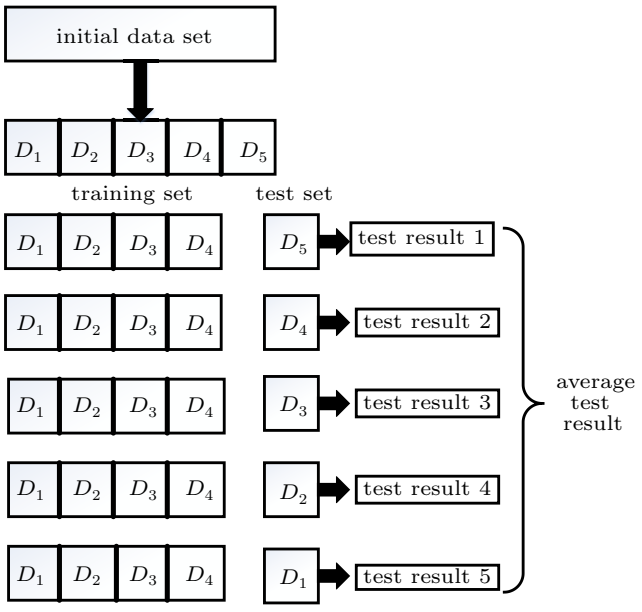


Fig. 9. The diagram of five-fold cross validation.

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \times 100\%, \quad (9)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\%, \quad (10)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\%, \tag{11}$$

$$F1_{\text{measure}} = \frac{2 \times \text{percision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{12}$$

where the TP, FP, TN, FN are explained in Table 2.

Table 2. The explanation of TP, FP, TN, and FN.

	Actual synovial thickening	Actual normal
Predicted synovial thickening	true positive (TP)	false positive (FP)
Predicted normal	false negative (FN)	true negative (TN)

3.2. Experiments

3.2.1. Localization of synovium

To perform synovium localization, we took the original 1024×768 images of MCP&PIP as the input of the faster R-CNN. The choice for such datasets is to provide guidance for clinicians on recognizing the synovium area from the original images. The size of the MCP dataset was 662 while the size of the PIP dataset was 315. We chose ZFNet as the basic network to extract the features shared by the RPN and the fast R-CNN. The employed ZFNet model was pre-trained on ImageNet. We chose the four-step alternating training strategy to jointly train the RPN and fast R-CNN network on Caffe framework.[28]

Step 1 Train the RPN initialized with the ZFNet model pre-trained on ImageNet (iteration number = 8000).

Step 2 Train the fast R-CNN network which is also initialized with the ZFNet model pre-trained on ImageNet. The inputs are the region proposals generated by the RPN in step 1 (iteration number = 4000).

Step 3 Train the RPN initialized with the weights of the fast R-CNN in step 2. Then the shared convolutional layers are fixed (learning rates are set to zero) while the unique layers of RPN are fine-tuned during the training process (iteration number = 8000).

Step 4 Train the fast R-CNN network. We fixed the shared convolutional layers and meanwhile fine-tuned the unique layers of the fast R-CNN (iteration number = 4000).

All experiments were performed on the HP xw8600 workstation equipped with NVIDIA Quadro K5200. The operation system is ubuntu 16.04 while the CUDA version is 7.5. Each training of the faster R-CNN took about one hour.

We took the average precision (AP) as a quantitative measure for the detection of synovium, which is the sum of the area under the precision-recall curve. Detections are considered true or false positives based on the intersection over union (IOU)

$$\text{IOU} = \frac{B_{\text{gt}} \cap B_{\text{p}}}{B_{\text{gt}} \cup B_{\text{p}}}, \tag{13}$$

where B_{p} is the predicted bounding box while B_{gt} is the ground truth bounding box. Correct detections must have an IOU over 0.5.

3.2.2. Grading of rheumatoid arthritis

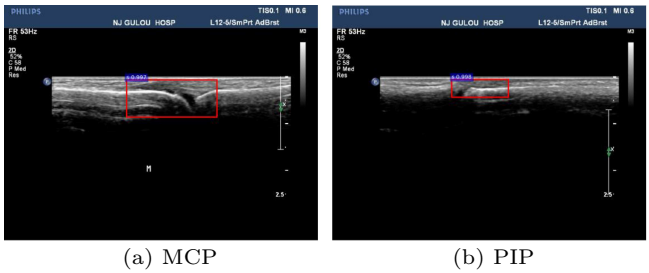
We employed the Caffe deep learning framework to implement and train the GoogLeNet. The pre-trained model was downloaded from the Caffe Model Zoo. We renamed the three fully-connected layers while the output number was changed from 1000 to our own grading number. Specifically, the output number for Grade 0 to Grade 3 classification is 4, while the output numbers for Grade 0 vs. Grade 1/2/3 classification and Grade 1 vs. Grade 3 classification are 2. We modified the basic learning rate of the whole network from 0.01 to 0.001 to match our training dataset. However, for the fully-connected layers, we kept the learning rates unchanged to fit our own RA severity grading rapidly. Then we started training with the pre-trained Caffe model files. We chose stochastic gradient descent (SGD) as the optimization method to minimize training loss with back-propagation.[29]

To evaluate the effect of data augmentation and fine-tuning with the pre-trained model, we performed contrast experiments for the Grade 0 to Grade 3 classification, Grade 0 vs. Grade 1/2/3 classification classification (normal vs. abnormal), and Grade 1 vs. Grade 3 classification (minimal vs. severe) on MCP&PIP datasets, which are without augmentation, without fine-tuning, performing both techniques, and preforming neither technique. The running environment was the same as the training of the faster R-CNN and each training of GoogLeNet took about 2.5 hours.

3.3. Experimental results and discussion

3.3.1. Localization of synovium

We first discuss the final detection results averaged by five-fold cross validation. The average precision (AP) of MCP is 94.37% while the average precision of PIP is 91.57%. Two detection results are displayed in Fig. 10. The red bounding box is associated with an "s" label and a softmax score in [0,1]. The s label refers to the synovium area and the score represents the confidence probability of being an object. As we can see, the synovium areas along the finger joints are precisely surrounded by the bounding box.



(a) MCP (b) PIP

Fig. 10. The detected synovium in (a) MCP and (b) PIP ultrasound images.

3.3.2. Grading of rheumatoid arthritis

The test results averaged by five-fold cross validation are shown in Tables 3 and 4. The standard deviations are also given in the tables.

Table 3. The classification accuracy of MIP/PIP datasets under the circumstance of non-augmentation & non-fine-tuning, non-augmentation & fine-tuning, augmentation & non-fine-tuning, and augmentation & fine-tuning. The slash character is used to distinguish between MCP dataset and PIP dataset and the standard deviations are given in parentheses.

Classification accuracy	Grade 0 to Grade 3	Grade 0 vs. Grade 1/2/3	Grade 1 vs. Grade 3
Non-augmentation & non-fine-tuning	86.00%/84.92%	90.30%/91.48%	90.20%/70.00%
	$(8.427\times10^{-3}/7.331\times10^{-3})$	$(4.213\times10^{-3}/7.331\times10^{-3})$	$(1.386\times10^{-2}/9.313\times10^{-2})$
Non-augmentation & fine-tuning	87.38%/88.20%	92.92%/93.11%	94.12%/77.00%
	$(6.880\times10^{-3}/1.372\times10^{-2})$	$(8.427\times10^{-3}/1.372\times10^{-2})$	$(1.386\times10^{-2}/7.825\times10^{-2})$
Augmentation & non-fine-tuning	90.00%/91.15%	96.00%/94.43%	97.25%/94.29%
	$(5.439\times10^{-3}/8.979\times10^{-3})$	$(8.427\times10^{-3}/8.979\times10^{-3})$	$(1.074\times10^{-2}/3.194\times10^{-2})$
Augmentation & fine-tuning	90.30%/95.00%	98.60%/97.00%	98.82%/98.57%
	$(4.213\times10^{-3}/1.159\times10^{-2})$	$(6.436\times10^{-3}/2.137\times10^{-2})$	$(2.148\times10^{-2}/3.194\times10^{-2})$

Table 4. The precision, recall, and $F1$-measure of MCP/PIP datasets under the circumstance of augmentation & fine-tuning. The slash character is used to distinguish between MCP dataset and PIP dataset and the standard deviations are given in the parentheses.

	Grade 0 to Grade 3	Grade 0 vs. Grade 1/2/3	Grade 1 vs. Grade 3
Precision	91.78%/93.60%	98.55%/96.63%	97.62%/96.88%
	$(5.974\times10^{-3}/3.644\times10^{-2})$	$(7.854\times10^{-3}/2.298\times10^{-2})$	$(3.043\times10^{-2}/5.590\times10^{-2})$
Recall	92.23%/92.15%	98.47%/96.85%	98.86%/99.00%
	$(9.248\times10^{-3}/3.222\times10^{-2})$	$(6.966\times10^{-3}/2.498\times10^{-2})$	$(1.565\times10^{-2}/2.033\times10^{-2})$
$F1_{\text{measure}}$	92.00%/92.87%	98.50%/96.74%	98.23%/97.93%
	$(6.670\times10^{-3}/2.888\times10^{-2})$	$(6.872\times10^{-3}/2.362\times10^{-2})$	$(2.314\times10^{-2}/3.889\times10^{-2})$

It is shown that CNN does not perform well without any preprocessing. In this case, the grading accuracy is basically not higher than 90%. After we employed the data preprocessing and transfer learning techniques, the four-grade classification accuracy exceeded 90% while the binary classification accuracies of Grade 0 vs. Grade 1/2/3 and Grade 1 vs. Grade 3 both exceeded 95%. Such comparison results prove that data augmentation and loading of pre-trained models for fine-tuning are capable of boosting the generalization ability of CNN to increase classification accuracy.

We also assessed the contributions of the network performance promotion. Data augmentation plays a crucial role by boosting the initial accuracy by 5% in most cases while the pre-training & fine-tuning only boosts the initial accuracy by less than 5%. The Grade 1 vs. Grade 3 classification shows the highest accuracy as the contrast between minimal and severe synovium thickening is quite obvious. The two binary classification accuracies are also higher than the four-grade classification accuracy. The precision, recall, and $F1$-measure of each augmented dataset after fine-tuning also exceed 90%, which indicate the feasibility and effectiveness of our proposed method from other aspects.

The classification results of five tests tend to be stable overall. As the dataset size of Grade 1 vs. Grade 3 is smaller than the other two, so its standard deviations look bigger.

4. Conclusions

Quantifying RA severity has always been a difficult diagnostic problem due to its semi-quantitative score criterion. Based on the CNN algorithm, we proposed an automatic classification method of RA ultrasound images. Firstly, we performed the auto-localization of the synovium lesion area along the finger joints. The detection results showed fine visual effects with an average precision exceeding 90%, albeit without making classifications in the finest steps. Then we implemented the RA severity grading. The four-grade classification accuracy exceeded 90% while the binary classification accuracies exceeded 95%. The stability and high accuracy of CNN provide a possible solution to such a classification problem of semi-quantitative criterion and also make up for the inconsistency of diagnosis made by clinicians. Most critically, the combination of data augmentation and transfer learning techniques employed in our work alleviated the problem of lacking large, annotated datasets, a frequently-encountered problem in medical image analysis. This reveals the promising application value of our proposed CNN-based RA grading method in assisting professional clinicians to diagnose.

Future work will focus on the detection of joint effusion and the blood flow grading of synovial joints. We will keep on increasing the training samples with the cooperation of sonographers and prepare to try other data augmentation methods such as generating new samples with generative adversarial networks (GAN). In addition, the research can be extended to other joints such as those of the wrist, ankle, metatarsophalanges, and elbow.

References

[1] Tang M W, Garcia S, Gerlag D M, Tak P P and Reedquist K A 2017 *Front. Immunol.* **8** 720

[2] Wakefield R J, Balint P V, Szkudlarek M *et al.* 2005 *J. Rheumatol.* **32** 2485

[3] Szkudlarek M, Court-Payen M, Jacobsen S, Klarlund M, Thomsen H S and Østergaard M 2003 *Arthritis Rheum.* **48** 955

[4] Colebatch A N, Edwards C J, Ostergaard M *et al.* 2013 *Ann. Rheum. Dis.* **72** 804

[5] Lee S, Choi M, Choi H S, Park M S and Yoon S 2015 *IEEE Biomedical Circuits and Systems Conference*, October 22–24 2015 Atlanta, USA, p. 1

[6] Murakami S, Hatano K, Tan J, Kim H and Aoki T 2018 *Multimed. Tools Appl.* **77** 10921

[7] Krizhevsky A, Sutskever I and Hinton G E 2012 *Advances in Neural Information Processing Systems 25*, December 3–6, 2012, Lake Tahoe, USA, p. 1097

[8] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R and Lecun Y 2013 arXiv:1312.6229

[9] Pavlidis T and Liow Y T 1988 *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 208

[10] Dalal N and Triggs B 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 20–25, 2005, San Diego, USA, p. 886

[11] Lowe D G 2004 *Int. J. Comput. Vis.* **60** 91

[12] Schmidhuber J 2015 *Neural Netw.* **61** 85

[13] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436

[14] Litjens G, Kooi T, Bejnordi B E, Setio A A A, Ciompi F, Ghafoorian M, Laak J A W M V D, Ginneken B V and Sánchez C I 2017 *Med. Image Anal.* **42** 60

[15] Pan S J and Yang Q 2010 *IEEE Trans. Knowl. Data Eng.* **22** 1345

[16] Deng J, DongW, Socher R, Li L J, Li K and Fei-Fei L 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, June 20–25, 2009, Miami, USA, p. 248

[17] Zhang G P 2000 *IEEE Trans. Syst. Man Cybern. Pt. C Appl. Rev.* **30** 451

[18] Hussein S, Gillies R, Cao K, Song Q and Bagci U 2017 *IEEE 14th International Symposium on Biomedical Imaging*, April 18–21, 2017, Melbourne, Australia, p. 1007

[19] Jung H, Lee S, Yim J, Park S and Yim J 2015 *15th IEEE International Conference on Computer Vision*, December 11–18, 2015, Santiago, Chile, p. 2983

[20] Bayramoglu N, Kannala J and Heikkilä J 2015 *IEEE 15th International Conference on Bioinformatics and Bioengineering*, November 2–4, 2015, Belgrade, Serbia, p. 1

[21] Rodrigues L F, Naldi M C and Mari J F 2017 *30th Conference on Graphics, Patterns and Images*, October 17–20, 2017, Niteroi, Brazil, p. 170

[22] Ren S, He K, Girshick R and Sun J 2017 *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 1137

[23] Girshick R 2015 *IEEE International Conference on Computer Vision*, December 7–13, 2015, Santiago, Chile, p. 1440

[24] Zeiler M D and Fergus R 2014 *European Conference on Computer Vision*, September 6–12, 2014, Zurich, Switzerland, p. 818

[25] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, June 7–12, 2015, Boston, USA, p. 1

[26] Bengio Y, Simard P and Frasconi P 1994 *IEEE Trans. Neural Netw.* **5** 157

[27] Glorot X and Bengio Y 2010 *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, May 13–15, 2010, Sardinia, Italy, p. 249

[28] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T 2014 *ACM Conference on Multimedia*, November 03–07, 2014, Orlando, USA, p. 675

[29] LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W and Jackel L D 1989 *Neural Comput.* **1** 541