Ultrasound Volume Projection Image Quality Selection by Ranking from Convolutional RankNet

Juan Lyu[a], Sai Ho Ling[b,*], S. Banerjee[b], J.Y. Zheng[c], K.L. Lai[d], D. Yang[d], Y.P. Zheng[d], Xiaojun Bi[e,a], Steven Su[b], Uphar Chamoli[b]

[a]*College of Information and Communication Engineering, Harbin Engineering University, Harbin, China*
[b]*School of Biomedical Engineering, University of Technology Sydney, Ultimo, NSW2007, Australia*
[c]*Department of Computer Science, Imperial College London, UK*
[d]*Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hung Hum, Hong Kong*
[e]*College of Information Engineering, Minzu University of China, Beijing, China*

Abstract

Periodic inspection and assessment are important for scoliosis patients. 3D ultrasound imaging has become an important means of scoliosis assessment as it is a real-time, cost-effective and radiation-free imaging technique. With the generation of a 3D ultrasound volume projection spine image using our Scolioscan system, a series of 2D coronal ultrasound images are produced at different depths with different qualities. Selecting a high quality image from these 2D images is the crucial task for further scoliosis measurement. However, adjacent images are similar and difficult to distinguish. To learn the nuances between these images, we propose selecting the best image automatically, based on their quality rankings. Here, the ranking algorithm we use is a pairwise learning-to-ranking network, RankNet. Then, to extract more efficient features of input images and to improve the discriminative ability of the model, we adopt the convolutional neural network as the backbone due to its high power of image exploration. Finally, by inputting the images in pairs into the proposed convolutional RankNet, we can select the best images from each case based on the output ranking orders. The experimental result shows that convolutional RankNet achieves better than 95.5% top-3 accuracy, and we prove that this performance is beyond the experience of a human expert.

Keywords:
Scoliosis, 3D ultrasound imaging, image selection, convolutional RankNet

1. Introduction

Scoliosis is a lateral curvature of the spine greater than 10 degrees in the coronal plane. It is a common spinal deformity occurring in adolescents aged between 10 and 18 years (Dunn et al., 2018), especially in females Dunn et al. (2018); Konieczny et al. (2012). Adolescent idiopathic scoliosis (AIS) accounts for over 80% cases if scoliosis Popko et al. (2018). During growth, spine curve progression occurs in around 67% of patients and increases the risk of pulmonary function disorder, severe scoliosis can be disabling Popko et al. (2018). Therefore, before skeletal maturity, early screening and treatment of scoliosis is recommended Law et al. (2016).

For the diagnosis of scoliosis, X-ray is a traditional and typical way to detect spine deformity. However, most scoliosis patients are children and adolescents who require frequent routine care for diagnosis, monitoring, and treatment Law et al. (2018, 2016); Larson et al. (2019). For these repetitive measurements, whole spine radiography has been the gold standard Hwang et al. (2018). Young patients are more sensitive to radiation than adultsHwang et al. (2018); Larson et al. (2019), and the cumulative radiation exposure and dose will increase their cancer risk. Studies showed that repeated radiography will raise the lifetime risk of breast cancer and heritable defects, by approximately 2% and 3%, respectively Hwang et al. (2018); Hui et al. (2016). Although several methods were proposed to reduce radiation exposure Geijer et al. (2003); Luo et al. (2015); Ben-Shlomo et al. (2016), such as the EOS system Hui et al. (2016); Wybier and Bossard (2013); Faria et al.

*Corresponding author.
Email address: Steve.Ling@uts.edu.au (Sai Ho Ling)

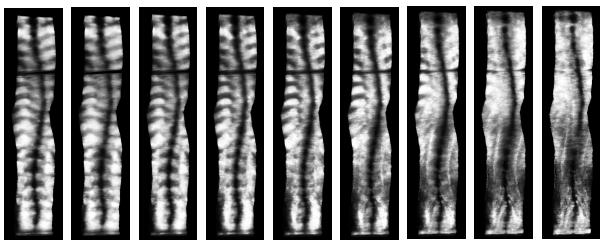

Figure 1: The samples of nine 2D ultrasound coronal images in the order of the depth increasing in a 3D spine volume.

(2013) (EOS Imaging, Paris, France), the radiation dose and cancer induction risks still exists. It has also been reported that the EOS system is highly costly, time consuming, and not easy to operate Lee et al. (2019); Larson et al. (2019). Moreover, even though the Cobb's method Cobb (1948); Knott et al. (2014) has been the gold standard for scoliosis assessment, the Cobb angle measurement is affected by the intra-observer and the inter-observer variations which can reach 3-5° and 6-9°, respectively Pruijs et al. (1994).

Compared to X-ray measurement, radiation-free technologies are more beneficial to scoliosis patients. Magnetic resonance imaging (MRI) is a technique with no radiation. MRI produces high-resolution spine images for diagnosis. However, it has been reported that MRI is controversial due to its low detection rate of abnormalities for adolescent scoliosis Popko et al. (2018). MRI is also an expensive and time-consuming technique Lee et al. (2019).

Alternatively, ultrasound imaging is relatively cost effective, is conducted in real-time, and is radition-free. In our previous studies Zhou and Zheng (2015); Zhou et al. (2017); Jiang et al. (2019), we proposed a radiation-free, freehand 3D ultrasound system, Scolioscan, to detect scoliosis using volume projection imaging (VPI). From 3D volume rendering, nine 2D coronal images are extracted with the depth increasing of the cut plane; samples are shown in Fig. 1. As shown, those nine 2D images have different qualities but are very similar, especially the adjacent images. For spinal curvature measurements and further assessment we need to select good images, but as far as we know, selection has been manually done by humans until now. Therefore, we propose selecting the best images using the most popular technique available, artificial intelligence (AI), and we are the first group to do spine ultrasound image selection automatically.

However, there is no relative research on medical imaging area. The most relative task is the image quality assessment (IQA) on computer vision. RankIQA Liu et al. (2017) was proposed for the no-reference image quality assessment (NR-IQA) utilizing the RankNet and a backbone of VGG-16. The ranking method in that paper was mainly used to augment dataset by artificially adding Gaussian blur distortion to the images; the ranking law between these artificial images becomes simpler than with the natural images. Then, they fine-tuned the learned network using the squared Euclidean distance to regress the image quality scores. Po et al. Po et al. (2019) proposed to apply a variance-based weighting for the original regression image quality scores to avoid homogenous image patches for the network training and quality score estimation. Ahmad et. al Ahmed and Asif (2019) proposed to use a learning rate scheduler to produce a set of suboptimal models. The final ensemble CNNs were selected from them with weighted averaging. Yan et. al Yan et al. (2018) designed a two-stream CNN for the NR-IQA, where the inputs of two streams are original image and gradient image, to extract more effective features of inputs in different levels compared with one stream structure. However, compared with our task, their databases have exact ground-truth for each image, therefore, they used the regression network to get the image quality scores for testing images. While for our task, the human expert only annotated the best one, it is difficult to do regression. There is another kind of IQA task, called aesthetic or attractiveness image ranking. Deep RankNet was applied to the aesthetic ranking task Tian et al. (2018). They proposed to use the visual similar-based method to generate training pairs and use a dual CNNs to learn the rankings of the images. Ma et. al Ma et al. (2019) proposed a Bayesian ranking cost function for the deep ranking network, DARN. DARN directly learned an attractiveness score mean and variance for each image, and all the images are human labelled. The images they ranked are aesthetic similar images, while our ultrasound images are more similar in structures and have the same contents and the same clarity. Moreover, we do not have the exact human labelled score for each image. Hence, our task is more challenging.

Nevertheless, inspired by the ranking idea, we propose to select the good images of each case by learning their probability ranking scores of good images. Even though we do not have the labelled score for each image, we know which one is the best image in each case. Then, we propose to label each image based on the structure similarity index (SSIM) Wang et al. (2004) score with the best image, and the score of the best image is set to 1.0. The reason is that the SSIM score is judging the structure similarity between two images and is closer to the human's subjective perception. How-

ever, we do not exactly regress these scores, we only utilize them for obtaining the rankings of the images. Accordingly, we propose to use a learning-to-rank algorithm to solve it. We utilize the combination of RankNet Burges et al. (2005); Cao et al. (2007); Burges (2010) and the convolutional neural network (CNN) to rank images. RankNet is a conventional pairwise learning-to-rank algorithm, which trains the inputs in a paired manner using a Siamese neural network to learn their final rankings. As we know, traditional neural network is a fully-connected network that each of the input nodes is interacted with each of the output nodes, and it is called shallow neural network since there are generally one or two hidden layers. Especially, the input images are generally large with thousands of pixels in most of the tasks, therefore, there are huge number of weights using traditional neural network and, which causes the costly hardware memory and training time but low performance LeCun et al. (1998). Compared to it, CNNs typically adopt the sparse connections between the inputs and outputs by filters with much fewer parameters, meanwhile, the parameter sharing mechanism further reduce the numbers of parameters between them LeCun et al. (2015). Even, CNNs usually have deeper structures than the traditional neural networks, when they have similar size of layers, CNNs are still easier to train Krizhevsky et al. (2017). Furthermore, if CNNs are initialized by learned features instead of random initialization, which named as transfer learning, the time required for model convergence is further reduced Hussain et al. (2018). Therefore, we replace the neural network of the conventional RankNet with CNN, which has shown the more excellent performance and effectiveness on computer vision and medical imaging.

We designed a convolutional RankNet for the spine ultrasound image selection based on the following ideas. Firstly, as there are several images with a large number of similarities for each patient, it is difficult to distinguish them using classification methods. Instead, we consider this as a ranking problem. By ranking them in sequence from 1st to 9th (in descending picture quality), it is easier to select the best image according to the ranking result. Secondly, RankNet traditionally requires a large amount of computational power, as its network structure uses a conventional artificial neural network (ANN). Convolutional neural networks (CNNs) can reduce the computational requirements through their weight sharing strategies and extract features more effectively than ANN LeCun et al. (1998, 2015); Krizhevsky et al. (2017). We design a specific CNN architecture for the feature extraction. Thirdly, for further discrimination of similar images, the loss function we adopt is the hinge loss function rather than the cross-entropy loss function in the conventional RankNet. The hinge loss has more discriminating ability when the estimated output scores of two images are very close. Above all, we solve this task by combining the CNN and the principle of RankNet.

The subsequent sections are arranged as follows. In Section 2, we introduce conventional RankNet and the proposed convolutional RankNet in detail. Then, in Section 3, we describe our data and the implementation details of the experiment, presenting the experimental results with several comparisons. Finally, we discuss the testing result in Section 4.

2. MATERIALS AND METHODS

2.1. Data

The ultrasound spine images we used are collected by the Scolioscan system (Model SCN801, Telefield Medical Imaging Ltd, Hong Kong) which is developed using a 3D ultrasound imaging method to obtain spine VPI for the assessment of scoliosis Zheng et al. (2016); Cheung et al. (2013, 2015). The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board. The subjects gave informed consent to their inclusion in the study as required, the work adheres to the Declaration of Helsinki. In this work, we have studied 400 cases of 400 patients. For each case, we select good images from the nine 2D vertebral anatomical images at different imaging depths. The generation method of the 2D coronal projection images from Scolioscan is the narrow-band, non-planar volume rendering algorithm which was introduced in our previous work and named as volume projection imaging (VPI) Cheung et al. (2015). An illustration of the VPI method is shown in Fig. 2. All the 2D images are in bitmap format. The original image sizes of the 2D coronal images vary, but are around 640×2466. However, it is time-consuming to load such high resolution images. Therefore, we resize them uniformly to the same size of 200×60. Although there is a 100 times resolution reduction, it has been shown experimentally that the reduction does not decrease selection accuracy. In addition, our labels of the best images are marked manually by an experienced operator who based on the criteria of a clear, dark line in the middle representing the spine profile, and also other spinal features as clearly as possible in the image including transverse processes and ribs.

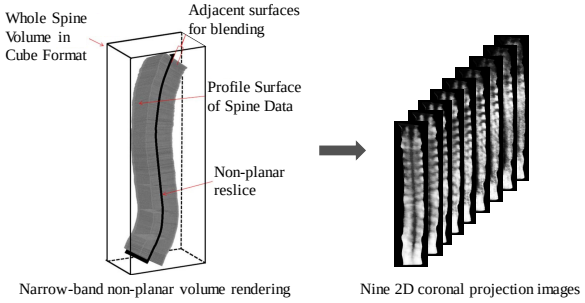

Figure 2: Illustration of the narrow-band, non-planar volume rendering algorithm for the generation of 2D projection images.

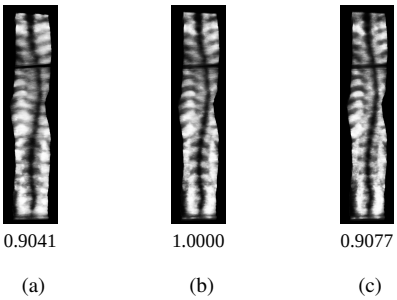

Figure 3: Illustration of the similarity between the best image and its adjacent images. (b) is the best image, we set its label as 1, (a) and (c) are its adjacent similar images, the values at their bottoms are their similarities with the best image, respectively, by SSIM.

2.2. Why RankNet

For our further study, our task is to select the images that have sharp, clear, continuous mid dark spine lines, as well as clear transverse processes and ribs. Before this study, we selected the best images using a human expert. However, selection by a human expert is affected by his/her experience and it is inevitably subjective. Moreover, the whole process of this type of selection is time-consuming. Therefore, we propose selecting them automatically using an artificial intelligence (AI) technique.

In our ultrasound imaging system, a 3D ultrasound volume is produced for each patient that can then be further processed into nine 2D coronal images from different imaging depths. The images at different depths have different imaging qualities. As shown in Fig. 3, several images have similarities with the best image when they are adjacent, especially for the mid dark line. As a result, it is difficult to discriminate them with a deep classification model. Hence, we consider selecting them based on their probability rankings to learn the nuanced distinctions. Conventional RankNet is the classic pairwise learning-to-rank algorithm which has been widely applied in search engines Burges (2010). In this work, we combine RankNet with CNN to build a convolutional RankNet for solving our problem; the former is good at ranking, and the latter has great performance in image processing.

2.3. Overview of Conventional RankNet

The traditional RankNet Burges et al. (2005) is a pairwise learning-to-rank algorithm, its structure is based on the traditional artificial neural network (ANN). The main body of RankNet is a Siamese network that consists of two streams of neural networks with shared weights. They also use the same initialization and the same gradient during the back-propagation process.

For each pair of inputs x_i, x_j, their corresponding output scores are $s_i = f(x_i)$ and $s_j = f(x_j)$, respectively. If s_i is larger than s_j, x_i is ranked higher than x_j. The prediction probability that x_i is ranked higher than x_j is defined as

$$P_{ij} = \frac{e^{s_{ij}}}{1 + e^{s_{ij}}}, \tag{1}$$

where $s_{ij} \equiv f(x_i) - f(x_j)$, the difference of the outputs of two branches. We can see that the prediction probability is a sigmoid function of s_{ij}.

The original loss function is cross entropy which is defined as

$$L(s_{ij}) = -\bar{P}_{ij} \log P_{ij} - (1 - \bar{P}_{ij}) \log(1 - P_{ij}), \tag{2}$$

where $\bar{P}_{ij}$ is the target probability. Then, the network learns the ranking using a stochastic gradient descent (SGD) algorithm.

The conventional RankNet has two main drawbacks when solving our problem. On the one hand, as we all know, compared with ANN, CNN is more powerful when extracting efficient features of the images and has exponential reduction of trained parameters. On the other hand, the prediction probability of RankNet is the sigmoid function of the difference between a pair of inputs. However, when the difference is small, meaning the data have great similarities, such as less than 0.1 in our task, the prediction probabilities will be very close, between around 0.475 and 0.525. Thus, it is hard to distinguish between similar images. Meanwhile, in our task, most of the images are quite similar to their neighbors.

2.4. Convolutional RankNet

To tackle the above problems involving the conventional RankNet, we extend the RankNet to a convolutional RankNet by replacing the Siamese ANN with

a Siamese CNN as the backbone. The convolutional RankNet is a combination of the CNN and the pairwise learning-to-rank algorithm, which is much more able to find small distances between images.

The backbone of the proposed convolutional RankNet is shown in Fig. 4, which is a dual branched CNN with shared weights. Each branch has six layers, and the first 5 layers have the same components, each layer consisting of a convolutional layer, a batch normalisation layer and a max pooling layer. The activation function we used are ReLU. All the pooling layers have the same filter size of 2, as well as a stride of 2. In terms of the convolutional layers, they have the same filter size of 3 and the same stride of 1, but the numbers of the filters are different, as they are increased exponentially from 2^5 to 2^9. The last layer consists of a convolutional layer, a batch normalisation layer, ReLU, and a global average pooling layer. The parameters of the convolutional layer are the same as those in the fifth layer. Then, the global average pooling layer is followed by a fully connected layer, the final score is activated by Sigmoid function.

The whole process for image ranking is shown in Fig. 5. Firstly, the nine images of each case are paired with each other, and each of the two images are only paired once. Then, inputting them into the Siamese CNN in pairs, each branch outputs a prediction score for each input. The two streams of the Siamese CNN are sharing weights. The loss function layer of the convolutional RankNet uses a hinge loss function, which connects two outputs together to train the network. After training, we get the final training score for each image. For testing, we input any new case of nine images to any one of the two streams, and their corresponding output scores are obtained. The images are ranked based on their scores from highest to lowest, thereby selecting the best images.

Before introducing the loss function, we firstly show how we get the target probabilities of good images. In fact, we only created the label of the best image in each group in the beginning. However, if we label the best image as 1, and the others as 0, the most similar image and the least similar image will have the same score. The network may find it difficult to identify them. Therefore, we set the target probability of the best image as 1.0, and others are the similarity values with it. To measure the similarity between two images, structural similarity index (SSIM) is closer to the human preferences Wang et al. (2004). Subsequently, to overcome the sensitivity of the SSIM to image distortions, such as scaling, translation, and rotation of images, the complex wavelet structural similarity index (CW-SSIM) is

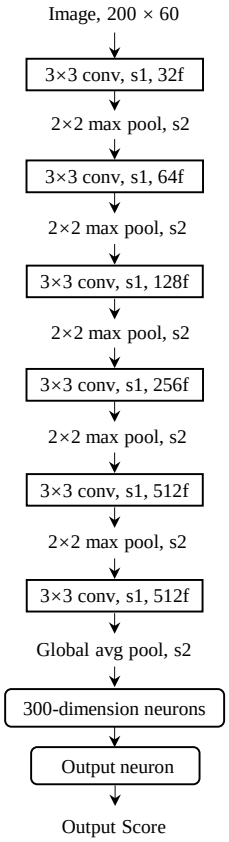

Figure 4: The backbone architecture of the proposed convolutional RankNet. The rectangles present the convolution layers. 3×3 is the filter size, and conv is the logogram of convolution + Batch normalization + ReLU. s1 means the convolution stride is one, and the *f means *features are generated. All the first 5 pooling layers use a 2×2 max pooling with a stride of 2. The final convolution layer is followed by a global average pooling layer to summarize the features. The rounded rectangle is the fully-connected layer with 300 neurons. Finally, the corresponding output score of the input image is obtained.

proposed and performs better for the distorted images Sampat et al. (2009); Wang and Simoncelli (2005). Despite the images we used are scaled, the purpose of measuring their similarities is to rank them instead of obtaining more accurate similarity scores. Moreover, experimentally, we can obtain the same ranking by using any of two indexes. Hence, we adopt the most widely used SSIM to calculate the similarity values of images, which is defined as

$$SSIM(z_i, z_{best}) = \frac{(2\mu_{z_i}\mu_{z_{best}} + c_1)(2\sigma_{z_i z_{best}} + c_2)}{(\mu_{z_i}^2 + \mu_{z_{best}}^2 + c_1)(\sigma_{z_i}^2 + \sigma_{z_{best}}^2 + c_2)}, \tag{3}$$

where z_{best} is the best image, and z_i is one of the others in this case. μ_*, σ_* presents the average and variance of $*$, respectively. $\sigma_{z_i z_{best}}$ is the covariance of z_i

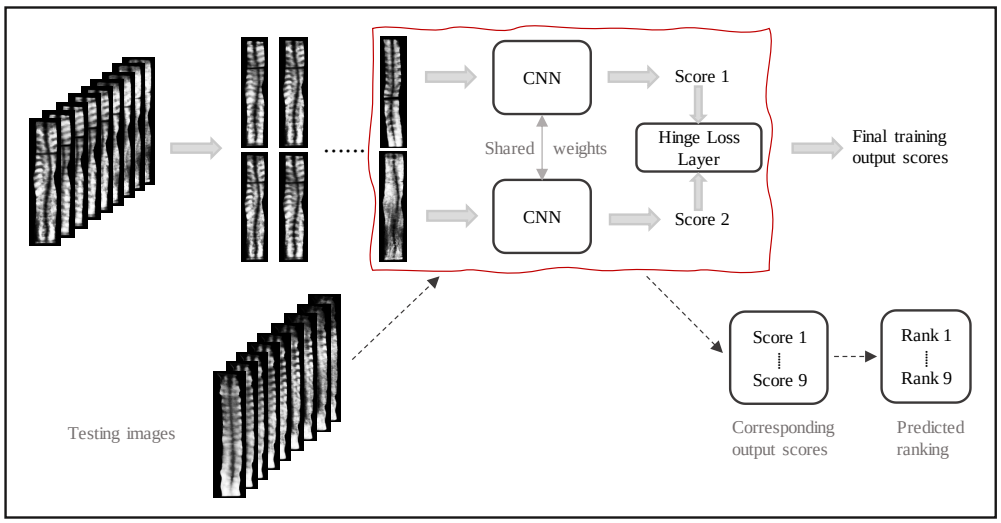

Figure 5: An illustration of the proposed convolutional RankNet for a spine ultrasound image selection process. Each case of images are trained in pairs to get the ranking order. Then, the new data are inputted into one of the two CNNs, and can be ranked based on the prediction scores using the trained model for the selection of good images.

and z_{best}. c_1 and c_2 are two constants to avoid dividing by zero, where $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$. L is the dynamic range of the pixel value, and k_1, k_2 are usually defaulted as 0.01 and 0.03. The reason we adopt SSIM is that, compared with other similarity assessment indexes, SSIM is closer to the human's subjective perception since our reference of the best image is given by the human expert. By learning the SSIM scores, we can obtain a more reasonable ranking order for all the images and their scores are within a range of (0.0, 1.0). The higher output score, the better image. Also, we can rank the images one by one rather than by simply bifurcating them into the best or not.

However, as mentioned in Section II b, it is indistinguishable when the probabilities of two images are very close by using the cross-entropy loss function. Hence, we choose the hinge loss as our loss function, which is widely used in SVM. Given a pair of input images x_i, x_j, we can assume that x_i is ranked higher than x_j, which can be presented as $x_i \rhd x_j$ Burges et al. (2005). The hinge loss function is defined as

$$L(x_i, x_j) = max\{0, m + o_j - o_i\}, \tag{4}$$

where o_i and o_j are the projections of the two branches of the network, and m is the margin. That means that when the gap between the two outputs $o_i - o_j$ is lager than the margin, the loss is zero. Supposing that the system produces N images for each patient, we would set the images of each case as a batch, and there would be a total of $\frac{N^2-N}{2}$ pairs of inputs for each batch. The final ranking hinge loss for a training batch size of H is calculated as

$$L_{rank\_hinge} = \frac{1}{H}\frac{2}{N^2-N}\sum_{h=1}^{H}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ x_i \rhd x_j}}^{N} max(0, m + o_j - o_i). \tag{5}$$

Eventually, we select the good images based on the ranking scores. However, different from other ranking problems, we do not need to rank all the images correctly as we only select the top k images for each case. Furthermore, we only need the top k images to include the best one since they are very similar.

3. RESULTS

We randomly divide the 400 groups of data into training and testing sets of 200:200, and all the data are normalised. With regard to the training images, we preprocess them before training through a 1.5-times contrast enhancement and a 0.8-times colour saturation reduction. As we only compare any pair of images once, there are a total of 36 pairs of images for each batch.

For the proposed convolutional RankNet, due to its output of each branch being within a range of 0.0 to 1.0, the final output is activated using the sigmoid function. Before the output layer, there are 300 dimensions outputs in the fully connected layer. Moreover, in this network, we use the dropout below the flatten layer and the fully connected layer with a keep ratio of 80%. Moreover, the hyperparameter margin in the ranking hinge

function is experimentally set to 0.1, and the network is updated using the Adam optimizer. The initial learning rate of the network is 0.001. We train the model for 200 epochs.

In the testing period, we directly input 9 images for each batch and choose the outputs of one of the branches as the final output. In fact, the outputs of two branches are the same due to the shared weights. Then, based on the outputs, we can easily select the good images.

The experiment is implemented in Python code using Tensorflow and Tensorlayer deep learning libraries, running on a Mars cluster of the University of Technology Sydney ARCLab, which has a 16GB Nvidia Quadro P5000 GPU.

We evaluate the ability of the proposed convolutional RankNet to select spine ultrasound images in three aspects. First, we show how the proposed convolutional RankNet is different using different kinds of labels: the 0 and 1 labels, and the similarity labels. Then, we present the comparison results using different loss functions: cross entropy, and ranking hinge loss functions. Finally, we evaluate the selection performance by making comparisons with different backbones.

The evaluation indices we use are the top-1 and top-3 accuracies and the normal used two indexes in ranking tasks, Normalized Discounted Cumulative Gain (NDCG) Järvelin and Kekäläinen (2000), the Spearman Rank Order Correlation Coefficient (SROCC) Myers et al. (2013) to assess our ranking algorithm. The top-k accuracy signifies whether the labelled best image ranks in the top k position(s). The reason for this is that in those tasks, they mainly concern the ranking orders with respect to the target orders. However, our main purpose using RankNet is not for the whole exact ranking orders, we aims to select the high quality of the ultrasound images to undergo further process. Based on the experiment and evaluated by clinical experts, the most top-3 2D VPI images are recommended high quality images for further assessment. Thus, in the next section, we are investigated to select the top-3 VPI images using the proposed ConvRankNet.

3.1. Comparison of different kinds of labels

The different performances of the binary outcomes and the similarity labels trained by the proposed method are shown in Table I, which are presented by the top-1 and top-3 accuracies with the confidence limits in 95% confidence level. The accuracies by using similarities are significant higher than binary labels. The top-3 accuracy using similarity label is greater than binary outcomes. This obviously indicates that it is difficult to select the good images if the targets are only 0 and 1, meaning 8 of the 9 images in each group have the same target scores. However, for a series of images which need to be ranked correctly, giving them labels based on similarities offers more objectivity than even humans could provide.

3.2. Comparison of Different Loss Functions

The conventional loss function in the traditional RankNet is the cross entropy loss function Burges et al. (2005). However, the samples from the previous tasks do not have great similarities as ours do. Therefore, as shown in Table II, using the hinge loss function in our ranking network is better than using the cross entropy loss function, where the performances are also presented by the top-1 and top-3 accuracies with the confidence limits in 95% confidence level. It reveals that the hinge loss function has more discriminative power for the pairwise similar images in our task.

3.3. Comparison with Different Backbones of the Network

We compare the proposed backbone with the classic pre-trained VGG-16 Simonyan and Zisserman (2014) and DenseNet Huang et al. (2017). However, we cannot use the original models directly as there are only a total of 200 $\times$9 images for training. We optimize the hyperparameters and reduce some layers of the two models, thus obtaining the best accuracies of the pre-trained VGG-16 and DenseNet. This is shown as Table III. Due to our image size being different from that used in the pre-trained VGG-16, the pre-trained layers we use are from conv1_1 to conv3_1. We are able to obtain the best result, and the fully connected layer is also 300 neurons, which is the same as that of the proposed network. For the DenseNet, the growth rate we use is 8, and we only build 3 dense blocks, 2 transition layers, and also 300 neurons' fully connected layer. The 3 dense blocks have 6, 8, and 12 bottlenecks, respectively. Even though it has been proven that the VGG-16 and DenseNet have outstanding performance in many tasks, in terms of our problem, our proposed architecture is more powerful. Then, we compared the NDCG and SROCC indexes of using different backbones, the larger of two values means the better ranking result. It can be seen that the proposed convolutional RankNet also achieves the best ranking result, not only in top 3 images, among three kinds of backbones.

We also show the computational complexities of using different backbones of the convolutional RankNet, which is evaluated by the widely adopted floating-point operations (FLOPs). We also present the number of

Table 1: Comparison Results of Different Kinds of Labels

Label Type	Top-1 Acc (%) ± Confidence limits (%)	Top-3 Acc (%) ± Confidence limits (%)
Binary outcomes	16.60 ± 2.26	36.80 ± 2.67
Similarities (this work)	**44.95 ± 0.86**	**89.80 ± 0.28**

Table 2: Comparison Results of Different Loss Functions

Loss Function	Top-1 Acc (%) ± Confidence limits (%)	Top-3 Acc (%) ± Confidence limits (%)
Cross Entropy	43.15 ± 0.76	88.35 ± 0.52
Hinge (this work)	**44.95 ± 0.86**	**89.80 ± 0.28**

parameters and the training speed, the average number of images trained per second, for reference. The details are shown in Table 3. Compared with the pre-trained VGG-16 backbone, our model performs better with lower computational complexity, higher speed and only using less than 1/13 parameters. For the DenseNet backbone we used, despite fewer FLOPs and the number of parameters than the proposed approach, however, it is trained more than 3.4 times lower and with lower accuracies. Moreover, when we enlarge the DenseNet to more layers the accuracies decline. Therefore, it indicates that the proposed convolutional RankNet is more efficient than the compared backbones.

3.4. Comparison with a Human Expert

As introduced above, there is a total of 21 cases that our proposals are fallen out the top-3 ranking for the all 200 testing cases using our proposed method. We call these the failed matched cases. Following our discussion, we find that there are three main reasons, for this: the AI problem, the quite close similarity problem, and the human expert problem. The AI problem means our approach cannot find a good image at all, the similarity problem means the prediction results are so similar to the target best image that we believe that both of them are acceptable, and the human expert problem means the prediction results using AI are better that the targets given by a human expert. As shown in Table IV, these three situations cause 4, 5 and 12 failed cases, respectively. Therefore, we obtain a **95.5%** top-3 accuracy if only the human problem is accounted for. However, if we only exclude the AI problem, we can achieve a 98% top-3 accuracy. Analogically, we assume that these problems also occur when getting the top-1 ranking, and it is even more severe. Namely, the performance of our proposed ultrasound image selection network based on the qualities ranking goes beyond that of a human expert.

Above all, the proposed convolutional RankNet shows the state-of-the-art performance in the scoliosis ultrasound image selection task. Using the AI technique to select a good image greatly saves time for any subsequent assessment.

4. Discussion

In this section, we analyse and discuss the details of the outputs given by our proposed convolutional RankNet for the mismatched cases. We discover the reasons behind the three potential problems mentioned before: the AI problem, the human expert problem, and the problem involving a high similarity. During the analysis, we first invited two human experts to select the good images in each case blindly. Then, we compared and summarised the results given by the human experts and our results given by the proposed algorithm. The conclusion regarding what kind of problem that each case belongs to is drawn by the principle shown in Table V.

First, we address the AI problem. The proposed network cannot select good images correctly for 4 cases. After analysis, there are two kinds of situations. One is as shown in Fig. 6(a); the imaging quality is bad, and the middle dark lines are all unclear and damaged during the imaging process. Even the best image given by the human expert also does not have a clear middle line. However, the AI gives the higher scores for the two images that have worst middle lines for the compared 4 images. It indicates that the large noises influence the discretion of the AI. For this case, we finally prefer P4

Table 3: Comparison Results of Different Backbones

Backbone	Top-1 Acc (%)	Top-3 Acc (%)	NDCG	SROCC	FLOPs ($\times 10^9$)	No. Param (M)	Images/s
Pre-trained VGG-16 Simonyan and Zisserman (2014)	32.5	80.50	0.9802	0.7578	25.46	58.16	44.44
DenseNet Huang et al. (2017)	39.50	82.50	0.9832	0.7977	**2.41**	**0.27**	16.67
this work	**44.00**	**89.50**	**0.9859**	**0.8307**	9.58	4.43	**57.14**

Table 4: The Comparison Result with the Human Expert

Methods	Total Number of the Mismatched Cases		Top-3 Acc (%) (similar results excluded)	Top-3 Acc (%) (similar results included)
	Number of Failed Cases	Number of Similar Results		
Human Expert	12	5	91.50	94.00
AI (this work)	**4**		**95.50**	**98.00**

Table 5: The Principle of the Conclusion

Our Top 1	Our Top 2 &3	Target	Conclusion
×	–	√	AI problem
√	–	√	Similar result
√	–	×	Human expert problem
×	√	×	Human expert problem

and P5. Another situation is that the middle dark lines are not as sharp and striking as in the other cases shown in Fig. 6(b). Our method also provides a bad suggestion of P3; P3 has the worst quality both in the middle and the end parts of the middle line of these 4 images. However, after discussion, P6 is also accepted, and we finally prefer P5 and P6. In general, it can be seen that the generalization ability of the proposed method is affected by the cases that have bad qualities for all the images, and this would need to be improved in future work.

Secondly, there is the human problem. The given targets are not better than the predicted images after our discussion. This problem is more complex, and there are four main situations. Firstly, the images in some cases are quite similar, so sometimes the human expert may be influenced by subjectivity, as shown in Fig. 7(a). Also, some images have bad qualities, as shown in Fig. 7(b), and at the same time they are so similar that the human expert would find it hard to identify the best image. For some other cases, the human's selection is affected by the noise interference in the surrounding area of the middle lines, such as the marked area in Fig. 7(c). There are also some mistakes when the human expert selects the best images, such as a lack of rib information and the fact that their middle lines are not better than the selection images from the AI, as shown in Fig. 7(d), the good image probability of the target image is only 0.108. Compared with the AI, the human expert has subjectivity and fatigability. Therefore, when facing the images with high similarities or low identifications, he/she may give discrepant labels for different cases.

Thirdly, there is the close similarity problem. There are 5 failed cases where the output results are very similar to the target best image. Although the results given by the AI do not include the labelled best image from the human expert, they are so similar that we believe that both of them are acceptable (see the P5 and P4 shown in Fig. 8(a)). In fact, in some cases there are more than 3 good images, especially in terms of the middle dark line of the spines, such as the 4 images shown in Fig. 8(b). We also calculate the average similarity between the AI proposals and the human expert's recommended image in each case, reaching 0.854, which also indicates that these images are highly similar. Hence, although the target best image is not selected correctly, the prediction results are good enough to be used in our further

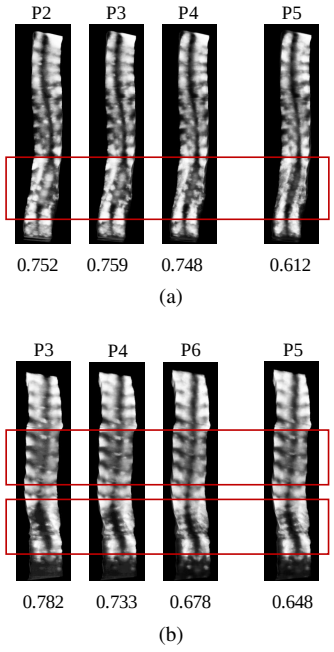

Figure 6: Samples of mismatched cases provided by AI. The serial number beyond the images is named according to their corresponding depths from P0 to P8 for each case. The bottom of the images are the prediction probabilities using our proposed ranking network, and the higher value signifies a better image. Furthermore, for both (a) and (b), the left three images are the best three good images selected by our deep learning approach, and the right hand image is the best image recommended by the human expert. The parts inside the red rectangles are the main differences between the images.

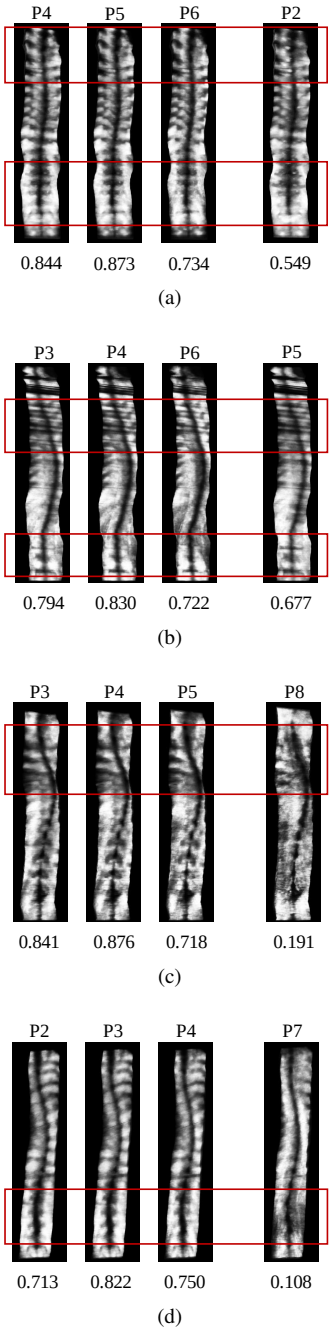

Figure 7: The samples of mismatched cases caused by the human expert. (a), (b), (c) and (d) are 4 situations where the mistakes were caused by the human expert.

tasks.

To sum up, the main impacts on good image selection are the image quality, the noises surrounding the middle dark line, and the close similarity. Therefore, there are two aspects we need to improve in future. On the one hand, the imaging process is supposed to improve so as to obtain a better image quality and to reduce the noise. On the other hand, the generation ability of our method should be enhanced, especially for complex cases.

To select the ultrasound spine images with good qualities for any further scoliosis assessment study, in this paper, we use an AI technique to select them automatically. Since the coronal 2D images produced by a single 3D volume have really high similarities and are difficult to distinguish, we firstly regard this task as a ranking problem. Therefore, we propose a convolutional pairwise learning-to-rank approach to solve the problem, which is a combination of the CNN and the traditional ranking method. Furthermore, we replace the cross-entropy loss function in the conventional RankNet to the hinge loss function, and the result is improved. Surprisingly, the proposed method performs better than the human expert in the selection results. In the future, we will firstly evaluate the proposed network on more data and improve its generation ability.

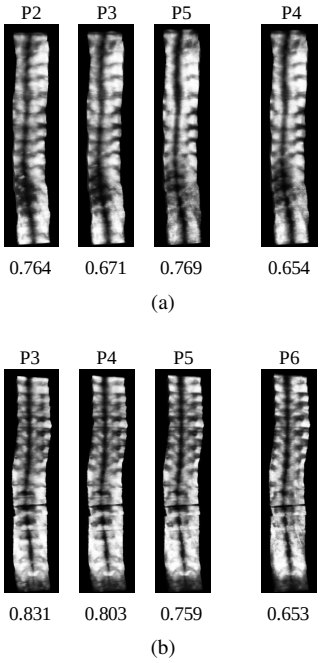

Figure 8: The samples of mismatched cases caused by close similarity.

Conflict of Interest

No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

Acknowledgment

The project is partially supported by Hong Kong Research Grant Council Research Impact Fund (R5017-18).

References

Ahmed, N., Asif, H.M.S., 2019. Ensembling convolutional neural networks for perceptual image quality assessment, in: 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), IEEE. pp. 1–5.

Ben-Shlomo, A., Bartal, G., Mosseri, M., Avraham, B., Leitner, Y., Shabat, S., 2016. Effective dose reduction in spine radiographic imaging by choosing the less radiation-sensitive side of the body. The Spine Journal 16, 558–563.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.N., 2005. Learning to rank using gradient descent, in: Proceedings of the 22nd International Conference on Machine learning (ICML-05), pp. 89–96.

Burges, C.J., 2010. From ranknet to lambdarank to lambdamart: An overview. Learning 11, 81.

Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H., 2007. Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th international conference on Machine learning, ACM. pp. 129–136.

Cheung, C.W.J., Law, S.Y., Zheng, Y.P., 2013. Development of 3-d ultrasound system for assessment of adolescent idiopathic scoliosis (ais): and system validation, in: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, IEEE. pp. 6474–6477.

Cheung, C.W.J., Zhou, G.Q., Law, S.Y., Mak, T.M., Lai, K.L., Zheng, Y.P., 2015. Ultrasound volume projection imaging for assessment of scoliosis. IEEE transactions on medical imaging 34, 1760–1768.

Cobb, J., 1948. Outline for the study of scoliosis. Instr Course Lect AAOS 5, 261–275.

Dunn, J., Henrikson, N.B., Morrison, C.C., Blasi, P.R., Nguyen, M., Lin, J.S., 2018. Screening for adolescent idiopathic scoliosis: evidence report and systematic review for the us preventive services task force. Jama 319, 173–187.

Faria, R., McKenna, C., Wade, R., Yang, H., Woolacott, N., Sculpher, M., 2013. The eos 2d/3d x-ray imaging system: a cost-effectiveness analysis quantifying the health benefits from reduced radiation exposure. European journal of radiology 82, e342–e349.

Geijer, H., Verdonck, B., Beckman, K.W., Andersson, T., Persliden, J., 2003. Digital radiography of scoliosis with a scanning method: radiation dose optimization. European radiology 13, 543–551.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Hui, S.C., Pialasse, J.P., Wong, J.Y., Lam, T.p., Ng, B.K., Cheng, J.C., Chu, W.C., 2016. Radiation dose of digital radiography (dr) versus micro-dose x-ray (eos) on patients with adolescent idiopathic scoliosis: 2016 sosort-irssd "john sevastic award" winner in imaging research. Scoliosis and spinal disorders 11, 46.

Hussain, M., Bird, J.J., Faria, D.R., 2018. A study on cnn transfer learning for image classification, in: UK Workshop on Computational Intelligence, Springer. pp. 191–202.

Hwang, Y.S., Lai, P.L., Tsai, H.Y., Kung, Y.C., Lin, Y.Y., He, R.J., Wu, C.T., 2018. Radiation dose for pediatric scoliosis patients undergoing whole spine radiography: Effect of the radiographic length in an auto-stitching digital radiography system. European journal of radiology 108, 99–106.

Järvelin, K., Kekäläinen, J., 2000. Ir evaluation methods for retrieving highly relevant documents, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM. pp. 41–48.

Jiang, W., Zhou, G., Lai, K.L., Hu, S., Gao, Q., Wang, X., Zheng, Y., 2019. A fast 3-d ultrasound projection imaging method for scoliosis assessment. Mathematical biosciences and engineering: MBE 16, 1067.

Knott, P., Pappo, E., Cameron, M., deMauroy, J.C., Rivard, C., Kotwicki, T., Zaina, F., Wynne, J., Stikeleather, L., Bettany-Saltikov, J., et al., 2014. Sosort 2012 consensus paper: reducing x-ray exposure in pediatric patients with scoliosis. Scoliosis 9, 4.

Konieczny, M.R., Senyurt, H., Krauspe, R., 2012. Epidemiology of adolescent idiopathic scoliosis. Journal of children's orthopaedics 7, 3–9.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90.

Larson, A.N., Schueler, B.A., Dubousset, J., 2019. Radiation in spine deformity: State-of-the-art reviews. Spine deformity 7, 386–394.

Law, M., Ma, W.K., Lau, D., Chan, E., Yip, L., Lam, W., 2016. Cumulative radiation exposure and associated cancer risk estimates for scoliosis patients: impact of repetitive full spine radiography. European journal of radiology 85, 625–628.

Law, M., Ma, W.K., Lau, D., Cheung, K., Ip, J., Yip, L., Lam, W., 2018. Cumulative effective dose and cancer risk for pediatric population in repetitive full spine follow-up imaging: How micro dose

is the eos microdose protocol? European journal of radiology 101, 87–91.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Lee, T.T.Y., Cheung, J.C.W., Law, S.Y., To, M.K.T., Cheung, J.P.Y., Zheng, Y.P., 2019. Analysis of sagittal profile of spine using 3d ultrasound imaging: a phantom study and preliminary subject test. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization , 1–13.

Liu, X., van de Weijer, J., Bagdanov, A.D., 2017. Rankiqa: Learning from rankings for no-reference image quality assessment, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1040–1049.

Luo, T.D., Stans, A.A., Schueler, B.A., Larson, A.N., 2015. Cumulative radiation exposure with eos imaging compared with standard spine radiographs. Spine deformity 3, 144–150.

Ma, N., Volkov, A., Livshits, A., Pietrusinski, P., Hu, H., Bolin, M., 2019. An universal image attractiveness ranking framework, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 657–665.

Myers, J.L., Well, A.D., Lorch Jr, R.F., 2013. Research design and statistical analysis. Routledge.

Po, L.M., Liu, M., Yuen, W.Y., Li, Y., Xu, X., Zhou, C., Wong, P.H., Lau, K.W., Luk, H.T., 2019. A novel patch variance biased convolutional neural network for no-reference image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology 29, 1223–1229.

Popko, J., Kwiatkowski, M., Gałczyk, M., 2018. Scoliosis: Review of diagnosis and treatment. Polish Journal of Applied Sciences 4, 31–35.

Pruijs, J., Hageman, M., Keessen, W., Van Der Meer, R., Van Wieringen, J., 1994. Variation in cobb angle measurements in scoliosis. Skeletal radiology 23, 517–520.

Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K., 2009. Complex wavelet structural similarity: A new image similarity index. IEEE transactions on image processing 18, 2385–2401.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Tian, X., Long, Y., Lv, H., 2018. Relative aesthetic quality ranking, in: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. pp. 2509–2516.

Wang, Z., Bovik, A.C., Sheikh, H.R., Member, S., Simoncelli, E.P., Member, S., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612.

Wang, Z., Simoncelli, E.P., 2005. Translation insensitive image similarity in complex wavelet domain, in: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., IEEE. pp. ii–573.

Wybier, M., Bossard, P., 2013. Musculoskeletal imaging in progress: the eos imaging system. Joint Bone Spine 80, 238–243.

Yan, Q., Gong, D., Zhang, Y., 2018. Two-stream convolutional networks for blind image quality assessment. IEEE Transactions on Image Processing 28, 2200–2211.

Zheng, Y.P., Lee, T.T.Y., Lai, K.K.L., Yip, B.H.K., Zhou, G.Q., Jiang, W.W., Cheung, J.C.W., Wong, M.S., Ng, B.K.W., Cheng, J.C.Y., et al., 2016. A reliability and validity study for scolioscan: a radiation-free scoliosis assessment system using 3d ultrasound imaging. Scoliosis and spinal disorders 11, 13.

Zhou, G.Q., Jiang, W.W., Lai, K.L., Zheng, Y.P., 2017. Automatic measurement of spine curvature on 3-d ultrasound volume projection image with phase features. IEEE transactions on medical imaging 36, 1250–1262.

Zhou, G.Q., Zheng, Y.P., 2015. Assessment of scoliosis using 3-d ultrasound volume projection imaging with automatic spine curvature detection, in: 2015 IEEE International Ultrasonics Symposium (IUS), IEEE. pp. 1–4.