INTERVERTEBRAL DISC DETECTION IN X-RAY IMAGES USING FASTER R-CNN: A DEEP LEARNING APPROACH

Ruhan Sa, William Owens, Raymond Wiegand, Mark Studin, Donald Capoferri, Kenneth Barooha, Alexander Greaux, Robert Rattray, Adam Hutton, John Cintineo, Vipin Chaudhary

*State University of New York (SUNY) at Buffalo
§Spine Metrics, Inc.
‡University of Bridgeport College of Chiropractic
⨿Academy of Chiropractic

ABSTRACT

Automatic identification of specific osseous landmarks on the spinal radiograph can be used to automate calculations for diagnosing ligament instability and injury, which affect 75% of patients injured in motor vehicle accidents and is a precursor for other related diseases. In this work, we propose to use deep learning based object detection method as the first step towards identifying landmark points. The significant breakthrough of deep learning technology has made it a prevailing choice for perception based applications, however, the lack of large annotated training dataset has brought challenges to utilize the technology in medical image processing field. In this work, we try to address this problem by fine-tuning a deep network, in particular Faster-RCNN, a state-of-the-art deep detection network in natural image domain, using small annotated clinical datasets. In the experiment we show that, by using only 81 lateral lumbar X-Ray training images, as shown in Fig. 1, one can achieve satisfactory results. We achieved high average precision of 0.651, evaluated by widely used VOC2007 evaluation metric and the time is reduced to 0.2 second per image, which is significantly faster than traditional sliding window based classification techniques.

Index Terms— intervertebral disc, detection, deep learning, X-Ray

1. INTRODUCTION

Clinical data obtained from radiographic images of the spine is a critical part of care in both the conservative and surgical components of patient care. Huec et al stated the C7 slope has a predictive value of the shape of the cervical spine in the sagittal plane and one-third of the asymptomatic population had cervical kyphosis [1]. The following year, Krej et al provided statistical data of spine shape in a group of healthy young adults in age between 19 and 30 years and reported such statistical analysis should be part and parcel of determining the cut-off level for physiological spinal shape [2]. Back in 2003, Wiegand et al investigated whether a statistical correlation existed between lateral cervical geometry and cervical pathology as identified on neutral anteroposterior (AP) and lateral radiographs within a symptomatic group [3]. According to the above studies, it is possible that there is a correlation between spinal geometrical patterns and spinal symptoms, yet there is very little data collected clinically.

Clinical data sets are generally limited to pathological findings such as arthritis, fracture, infection, dislocation or tumor and in the absence of gross pathology there is very poor correlation to non-specific spine pain clinical symptoms. Since the majority of spine pain patients do not have any correlative pathological findings, it is important to start looking at the mechanics of the spine from a clinical standpoint. Recently, more and more attention is being paid to biomechanical information and data sets in the human spine. This type of biomechanical data includes the pelvic incidence angle, lateral spinal curvatures, vertebral body rotations as well as sagittal curvatures. Historically, laboratory science has embraced biomechanical analysis of the human spine. However, translation into the clinical setting has been slow to evolve, hence the domination of pathological data reporting. Although there are numerous and significant benefits to the diagnosis of biomechanical spinal pathology, it is a time consuming project that often requires additional and specialized training. Taking into consideration the increased demand to see more patients with the clinical complexity of an aging population, one has to consider the practicality of routine manual biomechanical evaluation in conjunction
with traditional pathological evaluation of spinal radiographs. We believe this is one of the main reasons we are not seeing the clinical application of laboratory based assessment in the general health care population hence we hypothesize that automation would lift significant barriers to the adoption of biomechanical analysis from the laboratory to the clinical setting.

Automatic identification of specific osseous landmarks on the spine radiograph can be used to automate calculations needed for diagnosing ligament instability and injury. This serious healthcare condition affects approximately 75% of patients injured in motor vehicle accidents and is a precursor for other spine related diseases. Automation would lead to faster and more accurate diagnosis thereby allowing for faster and more appropriate clinical intervention.

Automatic detection of intervertebral discs is the first step towards identifying the specific osseous landmarks. The accuracy of the identification depends upon robust detection results. Object detection is one of the major research areas in computer vision field due to the challenges brought by various object scales, different image conditions and etc. Recent work on deep learning technology has made significant progress on perception based automation, especially for image based applications. However the lack of annotated training dataset has been a bottleneck for utilizing this technology in medical image analysis field. In this work, we demonstrated the possibilities of using much less data to train a state-of-the-art deep learning network and in the meantime significantly decreasing the processing time compared to previous work in medical image analysis area. In particular, we fine-tuned Faster-RCNN network[4] (which won prestigious object detection challenges, ILSVRC and COCO, in natural image field in 2015) using only 81 lateral lumbar X-Ray images as training data and 11 images as testing data. The average processing time per image is decreased to 0.2 second with a high average precision of 0.65 calculated by VOC2007 evaluation metric [5]. Additionally, we also analyzed the performance of different fine-tuning techniques, which can be utilized for future studies.

2. RELATED WORK

Object detection has been a major research area in computer vision field. Traditionally, hand crafted features, such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) and etc are widely used to train various classifiers. However, compared to recent breakthrough work in deep learning field [6, 7, 8], the performance of traditional methods are far from satisfactory. The main reason is that the traditional features cannot represent the image as well as a deep neural network, such as a deep convolutional network.

Generally, object detection is performed through applying trained classifier on image with sliding window fashion or region proposal based fashion. Sliding windows often give more coverage to the image, however it is often too slow for real-world application due to its large redundancy, especially when one tries to use deep network based classifiers. Much work has been dedicated to accelerate the process [9, 10, 11]. Uijlings et al proposed selective-search, which generates possible object locations by using various techniques to combine local regions into bigger regions[12]. This has been used in many strong object detection algorithms as region of interests extraction steps. Ren et al proposed using a convolutional neural network as region proposal network (RPN)[4] and shared between region proposal and classification task, which significantly reduced the detection time, in the meantime they won the first place for COCO2015 object detection competition with 0.557 mean Average Precision (mAP) for 20 classes by training 80,000 images on ResNet-101 [8]. Furthermore, they achieved up to 0.759 mAP by training on more data. However, in medical image analysis field, it is rare to encounter training dataset of competing volume, which has brought great challenge in our work. Finetuning a pre-existing network is one of the ways to overcome this issue. In this work, we show that by using much less training data we can achieve satisfactory object detection result.

3. FINE-TUNING FASTER-RCNN

Faster-RCNN is an object detection algorithm proposed by Ren et al [4]. The purpose of the study is to generate bounding boxes for each target object and determine the class of the given object. They proposed using Region Proposal Network (RPN) to extract region of interests, where possible object bounding boxes with different scales and width-height-ratios are proposed. Previous methods, such as selective search, often use a separate procedure for region extraction. However, in Faster-RCNN the network only needs to run once for both region of interests proposal and bounding box predictions, which largely reduced the time of detection. This network is generally structured as shown in Fig 2.  *data* layer is training images, *convolutional neural network* is regular deep learning networks. *rois* is the RPN layer, where it constructs 9 different windows with varying scale and width-height-ratios for each pixel in the feature map as potential object boxes. *ROI Pooling layer* takes region of interests and convolutional features as input and generate the bounding box of the objects as well as the corresponding class name. For *convolutional neural network*, different layers and configurations of convolutional neural networks can be added. The author conclude that using deeper networks, such as ResNet-101, would give better results. However, as we have mentioned, in medical image field, it is hard to collect very large annotated training dataset, instead we try to fine-tune the existing network to overcome this drawback.

We explored different fine-tuning configurations, including two stage training methods and four stage training method, which is used by original Faster-RCNN training. We
also explored shallow fine-tuning, which only tunes the last layer of the network, and deep fine-tuning, where we tuned all the layers except for the base convolutional neural network. The details of the experiments as well as the analysis of the results are shown in section 4.

4. EXPERIMENTS

4.1. Training data

Our dataset consists of 92 annotated lateral lumbar X-Ray images. We divide the data into 81 training images and 11 testing images. The size of the image varies largely from 500x600 to 500x1309 pixel.

4.2. Convolutional Neural Network

In this work, we use ZF network [13], which is a 5-layered convolutional neural network as our base network. We initiate the network parameter with Faster-RCNN ZF network using Caffe [14].

4.3. Evaluation metrics

The detection task is judged by precision/recall curve. Detection is considered true or false positives based on the area of overlap with ground truth bounding boxes. Correct detection must have overlap $\alpha$ of more than 50% between predicted bounding box $B_p$ and ground truth bounding box $B_{gt}$ [5], as shown below.

$$\alpha = \frac{B_{gt} \cap B_p}{B_{gt} \cup B_p}$$  \hspace{1cm} (1)

4.4. Hyper-parameter tuning

There are a large number of hyper-parameters in Faster-RCNN, including learning rate, number of iterations, step size for dropping the learning rate, number of layers to fine-tune and etc. To examine the effect each of the parameter has on our dataset, we used two step training: first, fine-tune RPN network alone; second, pass the trained RPN parameter to Faster-RCNN network and fine-tune the rest of the parameters in network.
Fig. 5. Effect of shallow and deep fine-tuning on loss change for Faster-RCNN.

Fig. 6. Effect of shallow and deep fine-tuning VS four step training on loss change.

4.4.1. Learning rates tuning

The first step for training deep network is using small portion of training data and observe the loss convergence. Fig. 3 shows the effects of different learning rates on loss change when fine-tuning last layer of Faster-RCNN on 10 training images. From this image, we see that training with learning rate of 0.001 converges to lower loss value than learning rate of 0.01. As we can see the loss has decreased from around 1.0 to around 0.7.

4.4.2. Shallow tuning and deep tuning

In this section, we use the entire imageset to train the network, but fine-tune different layers and see the performance based on loss change and precision change. Fig. 4 and Fig. 5 show that fine-tuning deeper layer yields better results for both RPN and Faster-RCNN network. Table 1 shows that deeper fine-tuning also gives better precision result.

4.4.3. Four-stage-tuning

In order to examine if combining shallow and deep fine-tuning together would give performance boost, we fine-tuned network using four-stage-tuning, where we first deep tune the network as described in section 4.4.2 and pass the trained parameter to shallow tune, such that the parameter in last layer can have a finer adjustment. However, from Fig. 6 we see that extra stage of tuning does not give a better result.

4.5. Processing time and qualitative result

The average detection time is 0.2 second per image on Geforce GTX 1060 mini GPU. Sample testing results are shown on Fig. 7

5. CONCLUSION

In this work, we used Faster-RCNN object detection method as the first step towards automatically identifying landmarks from spine X-Ray images. Due to the lack of annotated medical images, training deep neural networks can be very challenging. In order to overcome this issue, we show that by fine-tuning a pre-trained deep network on small medical dataset, one can achieve satisfactory results. We also experimented different training techniques and compared the performances on our medical dataset. For future work, we will be focusing on increasing detection precision, as well as locating the landmarks based on detection results.
6. REFERENCES


