Improving the Performance of V-Net Architecture for Volumetric Medical Image Segmentation by Implementing a Gradient Pre-Processor

Elvis Chun-Sing Chui1, Lawrence Chun-Man Lau2, Ka-Bon Kwok3, Xin Ye1, Lik-Hang Hung2, Kwoon-Ho Chow4, Ronald Man-Yeung Wong2, Edmond Wing-Fung Yau4, Sheung-Wai Law5, Wing-Hoi Cheung* and Patrick Shu-Hang Yung*

1Department of Orthopedics and Traumatology, The Chinese University of Hong Kong, Hong Kong
2Department of Orthopedics and Traumatology, Prince of Wales Hospital, Hong Kong
3Department of Orthopedics and Traumatology, Alice Ho Miu Ling Nethersole Hospital, Hong Kong
4Koin 3D Technology (Medical) Limited Company, Science Park, Hong Kong

Abstract

Applications of deep learning models and Convolutional Neural Network (CNN) have been achieving good performance in 3D medical image analysis. In this study, we present a fast and efficient 3D femur segmentation method based on V-net with a gradient pre-processor. Instead of feeding CT data into a standard V-net model, our proposed model is fed with gradient data based on CT scans using a gradient pre-processor, which forces the network to learn from the gradient field. Adopting an objective function based on dice similarity coefficient, the imbalance between the numbers of femur voxels against that of background could be addressed. A dataset of lower limb CT data with 60 samples was trained and tested on a pure V-net model and the proposed V-net model separately. Experimental results show that our proposed method could achieve better segmentation results (1.4% improvement in dice similarity coefficient) and a higher robustness as compared with the pure V-net model, which allows faster training speed and higher segmentation accuracy.

Keywords: Deep Learning; Convolutional Neural Network; Volumetric Segmentation; V-net; Gradient pre-processor; Femur

Introduction

Accurate segmentation of 3D bone models from CT scanning images is essential for Computer Aided Orthopedic Surgical (CAOS) planning and 3D printing assisted. However, manual segmentation of bone models from CT scans is laborious and subjective with unstable accuracy [1], which heavily depends on operators’ experience. Thus, automatic segmentation methods with reliable accuracy are in high demand in orthopedic clinical practice.

Over the past few years, lots of automatic segmentation methods have been proposed to meet these challenges, such as atlas guided methods, level sets-based methods and machine learning based methods [2]. Although great progress is achieved, a rapid and accurate segmentation for lower limb has still not yet been developed.

In recent years, deep learning techniques, particularly Convolutional Neural Networks (CNN), have been used for automatic segmentation from medical images [3]. One of the well-known CNN architectures is U-net, which was proposed by Ronneberger et al. [4]. The U-net has combined an equal amount of up sampling and down sampling layers, which are connected by a skip connection that concatenate features from the contracting and expanding paths. During the training process, the U-net can process input images in one forward pass, which can obtain results in a direct segmentation map and enable the perception of full context of the image. Due to the amazing advantages of U-net, it has been widely used in 2D medical analysis. Çiçek et al. [5] has further applied U-net for 3D segmentation by feeding the U-net with 2D annotated slices from one volume. Milletari et al. [6] has proposed a 3D-variant of U-net architecture called V-net, to perform 3D image segmentation using 3D convolutional layers with a novel objective function directly based on the dice similarity coefficient during training.

V-net architecture comprises of compression layers and expansion layers. Compression layers are responsive for reducing the size of the image data presented as input and increasing the receptive...
field of the features, while the expansion layers are responsible for extracting features and expanding the spatial support of the lower resolution feature maps in order to output a two-channel volumetric segmentation. V-net has been successfully and widely adopted in 3D medical image segmentation due to its fast and accurate properties. Zhou Z et al. [7] has been successful in conducting segmentation of knee joint structures based on V-net architecture.

This study is describing how we apply a Gradient Pre-Processor (GPP) on the input data of V-net architecture. The V-net model can learn from the gradient processed CT images which facilitates training acceleration and accuracy improvement. 60 femur CT scans were applied for algorithm training and performance evaluation testing by comparing between our proposed method and standard V-net method.

**Method**

**V-net architecture**

V-net consists of an equal amount of compression and expansion layers (Figure 1). The compression layers were responsive for reducing the size of the input data and increasing the receptive field of the features by taking 3D convolution kernel with appropriate stride, while the expansion layers were focused on increasing the data size by projecting each input voxel to a bigger region through the 3D de-convolution kernel. A skip connection between compression and expansion layers was used to concatenate features from the compression and expansion paths, and this allowed V-net to take into account the full context of the image [8]. In this way, the fine-grained details of the image would be held for improving the quality of final contour prediction, which would facilitate the convergence of the model. In the compression path, the data was processed through the compressing stages with its resolution reduced and its channels doubled, which was achieved through performing $2 \times 2 \times 2$ convolution with stride 2 (Figure 1).

At the bottom of the V-net architecture, the input image with $512 \times 64 \times 64$ voxels were compressed into a $64 \times 4 \times 4$ feature representation. The magnitude of its channel was increased to 256. The images reached the innermost layer of the architecture and its whole anatomy of interest was perceived at once.

In expansion path, $2 \times 2 \times 2$ de-convolution with stride 2 was adopted to extract features and expand the spatial support of the lower resolution feature maps, combining with detailed information provided by skip connection, to output a volumetric segmentation with the same size as the input data with size of $512 \times 64 \times 64$. In order to output a volume in the same size as the input volume, convolutions with $1 \times 1 \times 1$ kernel size was applied in the very last convolutional layer to convert the input to a probabilistic segmentation of the foreground and background. All the convolutions and de-convolutions were followed by PReLU nonlinearities [6].

**Gradient pre-processor (GPP)**

Our proposed method trained end-to-end to segment 3D femurs from the lower limb CT scans based on the V-net model. The only difference between our proposed model and the standard V-net model is that the gradient result of CT images was fed into the V-net architecture by employing the GPP instead of raw CT images.

The GPP can be written as:

$$\begin{vmatrix}
\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} \\
\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}
\end{vmatrix}$$

where $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial f}{\partial z}$ and are gradient in the x, y, z direction respectively [6].

The GPP in our model, utilizing the gradient magnitude image filter class from simple ITK to compute the magnitude of the image gradient [9], forced the V-net model to learn in a compact way that the unwanted information in the input data was ignored. As shown in Figure 2b, the results obtained by the GPP has an obviously larger background area than raw image (Figure 2a), thus representing the

![Figure 1: Schematic representation of V-net architecture.](image-url)
Data preparation

In this study, 60 sets of angiographic CT images were used with the ages of subjects ranging from 50 to 70 years old and 1:1 male/female distribution. Since we aimed at studying the anthropometry of normal adult knee, CT images of the patients suffering from osteoarthritis, fracture deformity or other diseases affecting the anthropometry were excluded from the study. These images were first undergone manual segmentation by using MIMICS software. The segmented volumes were then exported as DICOM files, which would be further encoded into NPY files [10]. During the GPP process, the raw CT volumes were undergone gradient computation, and similarly, were output as NPY files. Since the forward and backward pass of the volumes in their original size would require more memory than available, all NPY files, including the segmented volume, raw volume and its corresponding gradient results, were down-sampled to $512 \times 64 \times 64$ voxels in each dimension using bilinear interpolation. Figure 2 shows three 2D slices images from our dataset, representing raw input (Figure 2a), gradient representation of raw input (Figure 2b), and their annotated segmentation (Figure 2c). Before feeding these 3D data into V-net model, normalization was applied to these data to obtain distribution ranging from 0 to 1 [11], and thus, all the data fed to V-net had the same scale.

Segmentation evaluation metrics

A standard measurement method is required for evaluating the performance of the deep learning model. The discrepancy between the deep learning model’s prediction output and its annotated segmentation were therefore captured. Dice Similarity Coefficient (DSC) was used to evaluate the similarity between the predicted segmentation output by V-net model and its associated annotated segmentation [12]. The dice similarity coefficient with magnitude of 1 reflects a perfect match between prediction and target, with this sum of “discrepancy” termed as loss function being zero. The Dice Similarity Coefficient (DSC) between two binary volumes can be written as:

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}$$

where X and Y are the annotated and predicted segmentation volumes, respectively [6].

The train loss of deep learning architecture can be written as:
where the training model was optimized by finding the set of parameters (weights and bias in V-net architecture) that minimize it [11].

**Training and testing**

40 femur CT scans were randomly selected for algorithm training and the remaining data (20 CT scans) were used for deep learning model testing. In order to evaluate the performance of our proposed method, two sets of training and testing were conducted. The first set was conducted by using a pure standard V-net model with image data input (Figure 2a); another set was conducted by using a V-net model with a GPP, while the gradient pre-processed raw image data (Figure 2b) was used as the input for the V-net model.

Our models were implemented in python environment based on the TensorFlow library [11]. All the training and testing were performed at a workstation equipped with a NVIDIA GeForce GTX 1080 Ti GPU and 11 GB of memory. The models were trained with a mini-batch size of 2 due to the limited capacity of GPU memory. An optimizer adaptive moment estimation with initial learning rate 0.0001 was applied. The models were trained for 10000 iterations. The training process took 26 h, and it took about 10 sec to process 20 sets of CT images segmentation with size of 512 × 64 × 64.

**Results**

**Training results**

The learning process of the proposed and standard model were evaluated in term of dice loss which was used as measurement of segmentation error. Figure 3 shows the relationship between the dice loss and the number of iterations during the training process. There was no notable difference between the two models. Both curves were in the same shape and overlay each other. Our proposed model approached a final dice loss with magnitude of 0.0010, while the standard model got a final magnitude of 0.0015. The difference between two models in term of final dice loss is small. During the training process, these two models have no significant difference in performance. In strict sense, the proposed model gains a lower dice loss with a magnitude of 0.001, which may imply that our proposed model accelerates training process of V-net.

**Segmentation accuracy**

During the testing process, the trained models on the 20 femur CT volumes were validated. DSC values, average Hausdorff Distance (Avg. HD), and Volume Similarity (VS) were measured. Our proposed model had a better average DSC with magnitude of 0.944 ± 0.011 than that of the standard model with magnitude of 0.931 ± 0.014. A 1.4% improvement was observed.
Figure 4a describes the DSC obtained from testing in 20 CT volumes. The DSC values from our proposed method were more concentrated in a higher magnitude than that of the standard V-net model. As stated in 2.3, dice similarity coefficient is a measurement of the similarity between two samples [12], and a higher value indicates a more accurate segmentation. Thus, our proposed model achieved a better performance on segmentation accuracy.

Figure 4b is the box plot of the Avg. HD results, showing that the standard V-net model had an Avg. HD of 0.162 ± 0.068, while the V-net+GPP had a lower Avg. HD of 0.121 ± 0.057. To examine the volume similarity between the predicted segmentation and annotated segmentation, VS results were included in our study, as shown in following equation [13].

\[
FS = 2 \frac{|S \cap T|}{|S \cup T|}
\]

\[
\text{Avg. HD} = \frac{1}{n} \sum_{i=1}^{n} HD_i
\]

\[
\text{Avg. HD} = \frac{1}{n} \sum_{i=1}^{n} HD_i
\]

where S represents the predicted segmentation volume and T stands for annotated segmentation volume. When S matches perfectly in an ideal case, VS will gain a zero value. As shown in Figure 4c, our proposed model had a lower average VS than that obtained from the standard V-net model. The median of our proposed model was closer to zero compared with V-net model. This implies that our proposed model provided a better segmentation accuracy than the standard V-net model.

Visualization results

Visualization of the segmented 2D slices were compared. Figure 5 shows that segmentation by our proposed method had a clear binary boundary than that of V-net model, where the femur head in our proposed model was more obvious and smoother. The result echoed that our proposed model could provide a better segmentation performance than using a standard V-net model. The GPP was valuable in improving the performance of V-net architecture.

GPP proposed in this study transformed the raw input data into gradient fields which measured the directional changes in the intensity in an image. In the gradient fields, the pixels with the largest gradient values in the direction of the gradient became the boundary between certain parts. Surfaces and edges could be extracted in the direction perpendicular to the gradient direction, which is critical in image segmentation and other image analysis issues [14]. In this study, instead of learning from raw images, our proposed model effectively learned to extract shapes of femur from the gradient fields that provided more useful shape-related information about boundary between different organs, thus earning a performance improvement.

Conclusion

In this study, we proposed a convenient method for improving the performance of V-net on 3D femur segmentation, i.e., the first study investigating the implementation of a gradient pre-processor to the V-net model. By adopting the gradient pre-processor, our proposed method learned quickly and the V-net model could achieve a better accuracy (1.4% DSC improvement) on volumetric medical image segmentation. Future works will be conducted on other 3D organ segmentation.

Acknowledgment

This study was financially supported by the Innovation and Technology Fund from the Government of the Hong Kong Special Administrative Region of the People’s Republic of China (Project reference number: ITS/417/16).

References