Direct Automated Quantitative Measurement of Spine via Cascade Amplifier Regression Network

Shumao Pang[1], Stephanie Leung[2,3], Ilanit Ben Nachum[2,3], Qianjin Feng[1(✉)], and Shuo Li[2,3(✉)]

[1] Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
qianjinfeng08@gmail.com

[2] Department of Medical Imaging, Western University, London, ON, Canada
slishuo@gmail.com

[3] Digital Imaging Group of London, London, ON, Canada

Abstract. Automated quantitative measurement of the spine (i.e., multiple indices estimation of heights, widths, areas, and so on for the vertebral body and disc) is of the utmost importance in clinical spinal disease diagnoses, such as osteoporosis, intervertebral disc degeneration, and lumbar disc herniation, yet still an unprecedented challenge due to the variety of spine structure and the high dimensionality of indices to be estimated. In this paper, we propose a novel cascade amplifier regression network (CARN), which includes the CARN architecture and local shape-constrained manifold regularization (LSCMR) loss function, to achieve accurate direct automated multiple indices estimation. The CARN architecture is composed of a cascade amplifier network (CAN) for expressive feature embedding and a linear regression model for multiple indices estimation. The CAN consists of cascade amplifier units (AUs), which are used for selective feature reuse by stimulating effective feature and suppressing redundant feature during propagating feature map between adjacent layers, thus an expressive feature embedding is obtained. During training, the LSCMR is utilized to alleviate overfitting and generate realistic estimation by learning the multiple indices distribution. Experiments on MR images of 195 subjects show that the proposed CARN achieves impressive performance with mean absolute errors of 1.2496 ± 1.0624 mm, 1.2887 ± 1.0992 mm, and 1.2692 ± 1.0811 mm for estimation of 15 heights of discs, 15 heights of vertebral bodies, and total indices respectively. The proposed method has great potential in clinical spinal disease diagnoses.

Keywords: Spine · Deep learning · Manifold regularization
Disc height measurement · Vertebral body height measurement

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 940–948, 2018.
https://doi.org/10.1007/978-3-030-00934-2_104

1 Introduction

The quantitative measurement of the spine (i.e., multiple indices estimation of heights, widths, areas, and so on for the vertebral body and disc) plays a significant role in clinical spinal disease diagnoses, such as osteoporosis, intervertebral disc degeneration, and lumbar disc herniation. Specifically, the vertebral body height (VBH) and intervertebral disc height (IDH) (as shown in Fig. 1) are the most valuable indices for the quantitative measurement of the spine. The VBHs are correlated with the bone strength, which is of great significance to the vertebral fracture risk assessment for the osteoporotic patients [1,2]. Furthermore, the IDH reduction is associated with the intervertebral disc degeneration [3,4] and lumbar disc herniation [5].

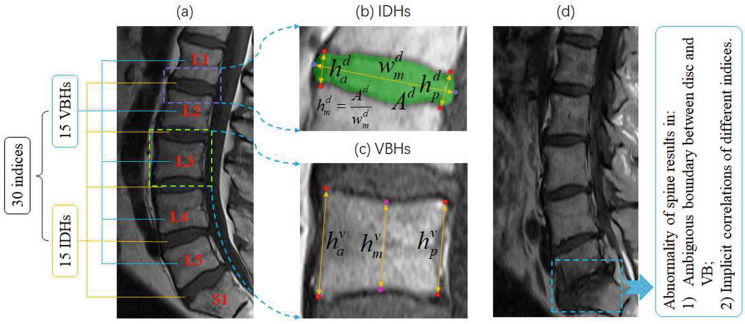

Fig. 1. (a) Illustration of 30 indices to be estimated; (b) Three heights for each disc (i.e., anterior IDH h_a^d, middle IDH h_m^d, and posterior IDH h_p^d); (c) Three heights for each vertebral body (i.e., anterior VBH h_a^v, middle VBH h_m^v, and posterior VBH h_p^v), where A^d denotes the disc area; (d) Ambiguous boundary between disc and VB and implicit correlations of different indices due to spinal abnormality.

Automated quantitative measurement of the spine is of significant clinical importance because it is reliable, time-saving, reproducible, and has higher consistency compared with manual quantitative measurement, which is usually obtained by manually detecting landmarks of the intervertebral disc (ID) and vertebral body (VB) from MR image [5,6].

Direct automated quantitative measurement of the spine is an exceedingly intractable task due to the following challenges: (1) The high dimensionality of estimated indices (as shown in Fig. 1(a)), which leads to difficulty in expressive feature embedding for such complex regression problem. (2) The excessive ambiguity of the boundary between VB and ID for abnormal spine (as shown in Fig. 1(d)), which increases intractability of expressive feature embedding. (3) Implicit correlations between different estimated indices (as shown in Fig. 1(d), the heights of the abnormal disc and the heights of adjacent VB are correlated because disc abnormality leads to simultaneous changes of IDH and the adjacent

VBH), which is difficult to be captured. (4) Insufficient labelled data (as shown in Fig. 1(d)), which possibly results in overfitting.

In recent years, an increasing number of approaches emerged in the direct quantitative measurement of other organs (e.g., heart) [7,8]. Although these methods achieved promising performance in the quantification of the cardiac image, they are incapable of achieving quantitative measurement of the spine because they suffer from the following limitations. (1) Lack of expressive feature representation. Traditional convolutional neural network (CNN) [9] is incapable of generating an expressive feature for multiple indices estimation because CNN possibly loses effective feature due to the lack of an explicit structure for feature reuse. (2) Incapability of learning the estimated indices distribution, which will lead to unreasonable estimation and overfitting.

In this study, we propose a cascade amplifier regression network (CARN), which includes the CARN architecture and local shape-constrained manifold regularization (LSCMR) loss function, for quantitative measurement of the spine from MR images. The CARN architecture is comprised of a cascade amplifier network (CAN) for expressive feature embedding and a linear regression model for multiple indices estimation. In CAN, amplifier unit (AU) is used for selective feature reuse between adjacent layers. As shown in Fig. 2(b), the effective feature of the anterior layer is stimulated while the redundant feature is suppressed, thus generating the selected feature, which is reused in posterior layer by a concatenation operator. CAN reuses multi-level features selectively for representing complex spine, thus an expressive feature embedding is obtained. During training, the high dimensional indices can be embedded in a low dimensional manifold due to the correlations between these indices. LSCMR is employed to restrict the output of the CARN to the target output manifold. As a result, the distribution of the estimated indices is close to the real distribution, which reduces the impact of outliers and alleviates overfitting. Combining the expressive feature embedding produced by CAN with LSCMR, a simple linear regression model, i.e., fully connected network, is sufficient to produce accurate estimation results.

The main contributions of the study are three-fold. (1) To the best of our knowledge, it is the first time to achieve automated quantitative measurement of the spine, which will provide a more reliable metric for the clinical diagnosis of spinal diseases. (2) The proposed CAN provides an expressive feature map for automated quantitative measurement of the spine. (3) Overfitting is alleviated by LSCMR, which utilizes the local shape of the target output manifold to restrict the estimated indices to being close to the manifold, thus a realistic estimation of indices is obtained.

2 Cascade Amplifier Regression Network

The CARN employs the CARN architecture and LSCMR loss function to achieve accurate quantitative measurement of the spine. The CARN architecture is composed of the CAN for expressive feature embedding and the linear regression model for multiple indices estimation. As shown in Fig. 2, in CAN, AU is used

for selective feature reuse between the adjacent layers by a gate, multiplier, adder and concatenate operator. In AU, the effective feature map is stimulated while the redundant feature map is suppressed. CAN provides expressive feature embedding via reusing multi-level features selectively. The linear regression model in CARN is a fully connected network without non-linear activation. During training, overfitting is alleviated by LSCMR, which is employed to oblige the output of CARN to lie on the target output manifold expressed by local linear representation [10], i.e., a sample on the manifold can be approximately represented as a linear combination of several nearest neighbors from the manifold. Local linear representation captures the local shape of the manifold, therefore, the distribution of estimated indices is close to the real distribution and the indices estimated by CARN are realistic.

2.1 Mathematical Formulation

Automated quantitative measurement of the spine is described as a multi-output regression problem. Given a training dataset $T=\{x_i, y_i\}_{i=1}^{N}$, we aim to train a multi-output regression model (i.e., the CARN) to learn the mapping $f : x \in R^{h\times w} \rightarrow y \in R^d$, where x_i and y_i denote the MR image and the corresponding multiple indices respectively, and N is the number of training samples. CARN should learn an effective feature and a reliable regressor simultaneously.

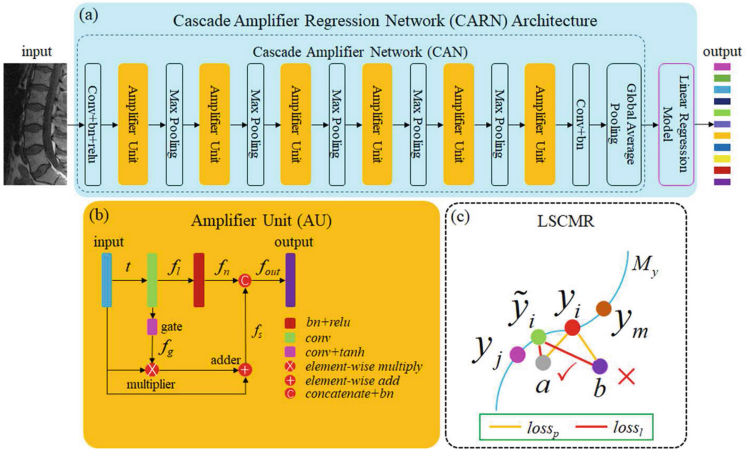

Fig. 2. (a) Overview of CARN architecture, including CAN for expressive feature embedding and a linear regression model for multiple indices estimation. (b) AU for selective feature reuse between adjacent layers. (c) LSCMR for obtaining realistic estimation and alleviating overfitting.

2.2 CARN Architecture

The CARN architecture is comprised of the CAN for expressive feature embedding and the linear regression model for multiple indices estimation.

CAN for Expressive Feature Embedding. The CAN consists of six AUs, two convolutional layers, five max pooling layers, and a global average pooling layer as shown in Fig. 2(a). AU is designed for selective feature reuse between adjacent layers. During feature selection, the selected feature is obtained by amplifying the input feature of AU using an amplifier, whose amplification factor is learned automatically (details in Section Feature Selection Mechanism). The effective low-level feature is stimulated and concatenated by the high-level feature while the redundant low-level feature is suppressed. The selective feature reuse is achieved by CAN level by level; then the multi-level selective reused feature generates an expressive feature embedding. The first convolutional layer with a 7×7 kernel size and stride of 2 reduces the resolution of feature maps from 512×256 to 256×128, while the last convolutional layer with a 1×1 kernel size and stride of 1 linearly combines the feature maps for information integration. The max pooling with a 2×2 kernel size and a stride of 2 is used to provide translation invariance to the internal representation. The global average pooling layer is utilized to reduce the dimensionality of feature maps.

The most crucial component of CAN is AU (as demonstrated in Fig. 2(b)), which is composed of a gate for controlling information propagation between adjacent layers, a convolutional layer with a 3×3 kernel size and stride of 1 for extracting a linear feature map, which is used to control the gate, a batch normalization layer with relu activation for producing non-linear feature map, a multiplier, an adder, and a concatenation operator with batch normalization for combining the selected feature map and non-linear feature map. The input t of AU goes through a convolutional layer and produces the linear feature map $f_l(t) = w_l * t + b_l$ for guiding feature selection, where w_l and b_l are the convolution kernel weight and bias of the convolutional layer respectively, and $*$ is the convolutional operator. Then the $f_l(t)$ flows into two paths. One path consists of batch normalization and relu activation, which is analogous to the traditional CNN to generate non-linear feature map $f_n(t) = relu(bn(f_l(t)))$, where bn and $relu$ denote the batch normalization and relu activation respectively. The other path is a gate composed of a convolutional layer and tanh activation, which generates output $f_g(t) = tanh(w_g * f_l(t) + b_g)$, where w_g and b_g are the convolution kernel weight and bias in the gate respectively, for selecting feature map. The output of the gate flows into a multiplier followed by an adder, and generates the selected feature:

$$f_s(t) = t \odot f_g(t) + t = t \odot (f_g(t) + 1) \tag{1}$$

where $\odot$ denotes the element-wise multiplication. Finally, the f_n and f_s are concatenated along the channel axis and normalized by the batch normalization layer to generate a output feature map $f_{out}(t) = bn(f_n(t) \oplus f_s(t))$, where $\oplus$ denotes the concatenation operator.

Feature Selection Mechanism. In Eq. 1, the value of each pixel in the selected feature map f_s is obtained by multiplying an amplification factor with the corresponding value in the input feature map t. The amplification factor $[f_g(t)+1]$ ranges from 0 to 2; substantially, the selected feature map f_s is equivalent to stimulating or suppressing the input feature map via an amplifier. When the amplification factor is less than 1, the input feature map is suppressed, vice versa. If the amplification factor is 1, the input feature map is directly propagated to the output, which is analogous to the denseNet [11].

Linear Regression Model for Multiple Indices Estimation. The linear regression model is a fully connected layer. The output of the linear regression model is: $f(x_i) = w_o h(x_i) + b_o$, where $h(x_i)$ is the output of the global average pooling (i.e., the feature embedding) as shown in Fig. 2(a), and w_o and b_o are the weights matrix and bias of the linear regression respectively.

2.3 Local Shape-Constrained Manifold Regularization Loss Function

The loss function is divided into two parts, including preliminary loss $loss_p$ and LSCMR loss $loss_m$. The preliminary loss is designed to minimize the distance between the estimation of indices and the ground truth, while the LSCMR loss is aimed at alleviating overfitting and generating realistic results by obliging the output of CARN to lie on the target output manifold using local linear representation. The total loss function is defined as follows:

$$loss_t(w) = loss_p(w) + \lambda_l loss_l(w) \tag{2}$$

where the λ_l is a scaling factor controlling the relative importance of the LSCMR loss. The preliminary loss function is defined as follows:

$$loss_p(w) = \frac{1}{N \times d} \sum_{i=1}^{N} \|y_i - f(x_i)\|_1 + \lambda_p \sum_i \|w_i\|_2 \tag{3}$$

where the first term is the mean absolute error (MAE) of the regression model; the second term is the l_2 norm regularization for the trainable weight w_i in CARN; λ_p is a hyper-parameter.

By using only the preliminary loss function, unreasonable multiple indices estimation may be obtained because the estimated result is possible to be out of their real distribution. For instance, as shown in Fig. 2(c), y_i, y_j, and y_m are the target outputs of samples. The points a and b are two possible estimations of y_i. The distances between the two estimations (the points a and b) and the target output y_i are the same, i.e., they have an identical preliminary loss. However, the loss of point a should be smaller than the point b as a is much closer to the local shape of the output space than b. Hence, a is a better estimation of y_i than b.

LSCMR is proposed to achieve a realistic and accurate estimation of multiple indices. Inspired by [12], y_i lies on a manifold M_y with an inherent dimension

smaller than d as the elements of y_i are correlated. The manifold M_y is spanned by $\{y_i\}_{i=1}^{N}$. We introduce the local linear representation, i.e., a sample on manifold M_y can be approximately represented as a linear combination of several nearest neighbors from M_y [10]. A sample y_i on M_y is locally linearly represented as:

$$\begin{aligned} y_i &= \sum_{j=1}^{k} y_j \alpha_j + \varepsilon \approx \sum_{j=1}^{k} y_j \alpha_j = \tilde{y}_i \\ s.t.\, \|\varepsilon\| &< \tau, \sum_{j=1}^{k} \alpha_j = 1, \alpha_j \geq 0, y_j \in N(y_i) \end{aligned} \tag{4}$$

where ε is the reconstruction error and τ is a small non-negative constant. $N(y_i)$ denotes the k-nearest neighbors of y_i on M_y and α_j is the reconstruction coefficient, which is calculated by LAE [13]. As shown in Fig. 2(c), $\tilde{y}_i$ is the local linear representation of y_i using its k-nearest neighbors (here k is equal to 2) y_j and y_m. The local linear representation of y_i reflects the local manifold shape. If the predicted indices is close to $\tilde{y}_i$, it will be near the manifold M_y. Therefore, the LSCMR loss is defined as:

$$loss_l\,(w) = \frac{1}{N \times d} \sum_{i=1}^{N} \|f\,(x_i) - \tilde{y}_i\|_1 \tag{5}$$

Using the $loss_l$, the prediction of y_i is restricted to being close to the manifold M_y, thus a more realistic result is obtained (e.g., the model generate the point a as the estimation of y_i instead of point b in Fig. 2(c)).

3 Experimental Results

Dataset. The dataset consists of 195 midsagittal spine MR images from 195 patients. The pixel spacings range from 0.4688 mm/pixel to 0.7813 mm/pixel. Images are resampled to 0.4688 mm/pixel and the ground truth values are obtained manually in this space. In our experiments, two landmarks, i.e., the left-top corner of the L1 VB and the left-bottom corner of the L5 VB, are manually marked for each image to provide reference for ROI cropping, in which five VBs, including L1, L2, L3, L4 and L5, and five IDs under them are enclosed. The cropped images are resized to 512×256.

Experimental Configurations. The network is implemented by Tensorflow. Four group experiments under different configurations, including CARN-$loss_p$, CNN-$loss_p$, CNN-$loss_t$, and CARN-$loss_t$, are used to validate the effectiveness of our proposed method. In CNN-$loss_p$ and CNN-$loss_t$, AU is replaced with a traditional convolutional layer, in which the output feature channels are the same as AU; the -$loss_p$ and -$loss_t$ denote the loss function defined in Eqs. 3 and 2 respectively used in the model.

Overall Performance. As shown in the last column of Table 1, the proposed CARN achieves low error for automated quantitative measurement of the spine,

with MAE of 1.2496 ± 1.0624 mm, 1.2887 ± 1.0992 mm, and 1.2692 ± 1.0811 mm for IDHs, VBHs, and total indices respectively. These errors are small referring to the maximums of IDHs (20.9203 mm) and VBHs (36.7140 mm) in our dataset.

CAN and LSCMR Effectiveness. Combining CAN and LSCMR, the performance improved by 2.44%, 1.16%, and 1.80% for IDHs, VBHs, and total indices estimation respectively, which is clearly demonstrated by comparing the third and last columns of Table 1. Using CAN without LSCMR, the MAE decreased by 0.21%, 0.49%, and 0.36% for IDHs, VBHs, and total indices estimation respectively, as shown in the second and third columns of Table 1. These results indicate that the CAN improves the performance for total indices estimation, especially for VBHs. This results from the fact that CAN generates expressive feature embedding although pathological changes in the disc can reduce the intensity of the adjacent VB and lead to ambiguity in the boundary. Using LSCMR without CAN, the performance improved by 2.14%, 1.03% for IDHs, and total indices estimation respectively, as shown in the third and fourth columns of Table 1.

LSCMR alleviates overfitting as shown in Table 1, in which CARN-$loss_t$ and CNN-$loss_t$ have high training errors (0.8591 mm vs 0.5024 mm, 0.9059 mm vs 0.5224 mm) but low test errors (1.2692 mm vs 1.2878 mm, 1.2791 mm vs 1.2924 mm) for total indices estimation compared with CARN-$loss_p$ and CNN-$loss_p$.

Table 1. Performance of CARN in terms of MAE under different configurations for IDH (mm), VBH (mm), and total indices (mm) estimation. MAE is illustrated in each cell. Best results are bolded for each row.

Method		CARN-$loss_p$	CNN-$loss_p$	CNN-$loss_t$	CARN-$loss_t$
IDH	Train	0.4633 ± 0.4706	0.4920 ± 0.4574	0.8689 ± 0.7417	0.8265 ± 0.7012
	Test	1.2782 ± 1.1173	1.2809 ± 1.1172	1.2535 ± 1.0754	$\mathbf{1.2496 \pm 1.0624}$
VBH	Train	0.5414 ± 0.5846	0.5528 ± 0.5615	0.9429 ± 0.8383	0.8916 ± 0.8004
	Test	1.2974 ± 1.0922	1.3038 ± 1.1154	1.3047 ± 1.1215	$\mathbf{1.2887 \pm 1.0992}$
Total	Train	0.5024 ± 0.5321	0.5224 ± 0.5130	0.9059 ± 0.7923	0.8591 ± 0.7531
	Test	1.2878 ± 1.1049	1.2924 ± 1.1163	1.2791 ± 1.0990	$\mathbf{1.2692 \pm 1.0811}$

4 Conclusions

We have proposed a multi-output regression network CARN for automated quantitative measurement of spine. By taking advantage of expressive feature extracted from CAN, and employing LSCMR for alleviating overfitting, CARN is capable of achieving promising accuracy for all indices estimation.

Acknowledgements. This work was supported by China Scholarship Council (No. 201708440350), National Natural Science Foundation of China (No. U1501256), and Science and Technology Project of Guangdong Province (No. 2015B010131011).

References

1. McCloskey, E., Johansson, H., Oden, A., Kanis, J.A.: Fracture risk assessment. Clin. Biochem. **45**(12), 887–893 (2012)
2. Tatoń, G., Rokita, E., Korkosz, M., Wróbel, A.: The ratio of anterior and posterior vertebral heights reinforces the utility of DXA in assessment of vertebrae strength. Calcif. Tissue Int. **95**(2), 112–121 (2014)
3. Jarman, J.P., Arpinar, V.E., Baruah, D., Klein, A.P., Maiman, D.J., Muftuler, L.T.: Intervertebral disc height loss demonstrates the threshold of major pathological changes during degeneration. Eur. Spine J. **24**(9), 1944–1950 (2014)
4. Salamat, S., Hutchings, J., Kwong, C., Magnussen, J., Hancock, M.J.: The relationship between quantitative measures of disc height and disc signal intensity with Pfirrmann score of disc degeneration. SpringerPlus **5**(1), 829 (2016)
5. Tunset, A., Kjaer, P., Chreiteh, S.S., Jensen, T.S.: A method for quantitative measurement of lumbar intervertebral disc structures: an intra- and inter-rater agreement and reliability study. Chiropr. Man. Ther. **21**(1), 26 (2013)
6. Videman, T., Battié, M.C., Gibbons, L.E., Gill, K.: Aging changes in lumbar discs and vertebrae and their interaction: a 15-year follow-up study. Spine J. **14**(3), 469–478 (2014)
7. Zhen, X., Zhang, H., Islam, A., Bhaduri, M., Chan, I., Li, S.: Direct and simultaneous estimation of cardiac four chamber volumes by multioutput sparse regression. Med. Image Anal. **36**, 184–196 (2017)
8. Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J., Li, S.: Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 276–284. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_32
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
10. Pang, S., et al.: Hippocampus segmentation based on local linear mapping. Sci. Rep. **7**, 45501 (2017)
11. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017
12. Liu, G., Lin, Z., Yu, Y.: Multi-output regression on the output manifold. Pattern Recognit. **42**(11), 2737–2743 (2009)
13. Liu, W., He, J., Chang, S.F.: Large graph construction for scalable semi-supervised learning. In: Proceedings of the 27th International Conference on Machine Learning (ICML-2010), pp. 679–686 (2010)