Fast Automatic Vertebrae Detection and Localization in Pathological CT Scans - A Deep Learning Approach

Amin Suzani[1], Alexander Seitel[1], Yuan Liu[1], Sidney Fels[1], and Robert N. Rohling[1,2], and Purang Abolmaesumi[1]

[1] Department of Electrical and Computer Engineering
[2] Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada

Abstract. Automatic detection and localization of vertebrae in medical images are highly sought after techniques for computer-aided diagnosis systems of the spine. However, the presence of spine pathologies and surgical implants, and limited field-of-view of the spine anatomy in these images, make the development of these techniques challenging. This paper presents an automatic method for detection and localization of vertebrae in volumetric computed tomography (CT) scans. The method makes no assumptions about which section of the vertebral column is visible in the image. An efficient approach based on deep feed-forward neural networks is used to predict the location of each vertebra using its contextual information in the image. The method is evaluated on a public data set of 224 arbitrary-field-of-view CT scans of pathological cases and compared to two state-of-the-art methods. Our method can perform vertebrae detection at a rate of 96% with an overall run time of less than 3 seconds. Its fast and comparably accurate detection makes it appealing for clinical diagnosis and therapy applications.

Keywords: Vertebrae localization, vertebrae detection, deep neural networks.

1 Introduction

Automatic vertebrae detection and localization in spinal imaging is a crucial component for image-guided diagnosis, surgical planning, and follow-up assessment of spine disorders such as disc/vertebra degeneration, vertebral fractures, scoliosis, and spinal stenosis. It can also be used for automatic mining of archived clinical data (PACS systems in particular). Furthermore, it can be a pre-processing step for approaches in spine segmentation, multi-modal registration, and statistical shape analysis.

The challenges associated with building an automated system for robust detection and localization of vertebrae in the spine images arise from: 1) restrictions in field-of-view; 2) repetitive nature of the spinal column; 3) high inter-subject variability in spine curvature and shape due to spine disorders and pathologies; and 4) image artifacts caused by metal implants.

N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 678–686, 2015.
DOI: 10.1007/978-3-319-24574-4_81

Several methods have been proposed in the literature for automatic vertebrae detection and localization in Computed Tomography (CT) [8,9,11,16,10] and Magnetic Resonance Imaging (MRI) volumes [10,15,14,1]. Several studies either concentrate on a specific region of the spine, or make assumptions about the visible part of the vertebral column in the image. A few recent studies claim handling arbitrary-field-of-view scans in a fully-automatic system [8,9,11,16]. The methods proposed in [8] and [16] rely on a generative model of shape and/or appearance of vertebrae. As a result, these methods may be challenged with pathological subjects, especially when metal implants produce artifacts in the image. The most promising method reported [9], uses classification forests to probabilistically classify each voxel of the image as being part of a particular vertebra. Based on predicted probabilities, the centroid of these points are obtained using the mean-shift algorithm. Although excellent results are obtained on a challenging data set, this method requires an additional post-processing step for removing false positives. This step adds to the computation time. In fact, this approach [9], which is the fastest method reported to-date, has a computation time of about a minute. Slow computation time can limit the application of automatic methods for image-guided tasks in clinics. A method with a faster computation time, on the order of seconds, has the potential to be used during guided interventions and may broaden the scope of such automatic analysis techniques.

In this work, we aim to find a faster solution to the problem of vertebra localization in general clinical CT scans by using deep neural networks [2]. No assumptions are made about which and how many vertebrae are visible in the images. The computation of image features is adopted from [8,9]. To allow for low computation times, our method does not require computationally expensive post-processing steps. We evaluate our method on a publicly-available data set, and compare it against the methods proposed by Glocker et al. [8,9] that use the same data set.

2 Methods

The vertebrae localization problem is parametrized as a multi-variate non-linear regression. Similar to [8,9] intensity-based features are extracted from voxels in the image. The features represent the short-range contextual information of the reference voxel. The targets of the regression are the relative distances between the center of each vertebral body and the reference voxel. In other words, the vector from the reference voxel to the center of the vertebral body is considered as one target in the regression.

The number of observations in our regression problem is equal to the number of selected voxels from the image. Since the images of our CT data set are labeled with 26 landmarks (26 vertebral bodies), the target of our regression includes 26 three-dimensional vectors for each observation. Therefore, the vertebrae localization problem is parametrized as a multi-variate regression with 500 features and $26 \times 3 = 78$ targets.

For a test image, we first extract intensity-based features from all the voxels. We then use a deep neural network to predict the relative distance vector of

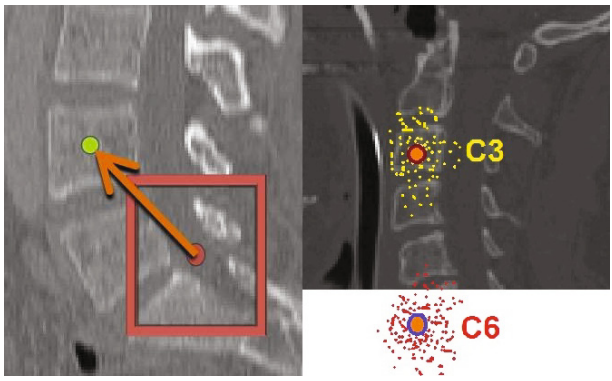

Fig. 1. Left: The vertebrae localization problem is parametrized as a regression problem. The targets of the regression are the vectors from the reference voxel to the center of each vertebral body (The orange arrow). The reference voxel is described by 500 intensity-based features which are extracted from the area around it. The green point is considered as the vote of the red voxel for a specific vertebra. Right: Location of each vertebra is estimated by getting the centroid of the votes of the voxels. This centroid may or may not be located inside the field of view of the image. In this case, C3 is considered as a true positive, and C6 is considered as a true negative.

each vertebral body with respect to the reference voxel. Knowing the location of the reference voxel and the relative distance vector to the center of a specific vertebral body, we can compute the predicted absolute location of the vertebral body on the test image. This absolute location is considered the vote of that voxel for the location of a vertebral body. Note that these votes might be either inside or outside of the field-of-view of the image. Each voxel in the image votes for the location of each vertebral body, so for each specific vertebral body, we compute the centroid of these votes to obtain a single prediction for the location of the vertebral body.

2.1 Point Selection

In our method, each vertebra is localized by aggregating the votes of the points in the image. However, in most of the images there are certain regions, like those in the background, that do not help with the prediction of vertebrae. Disregarding these points increases the accuracy and decreases the computational time. To this end, we use the Canny edge detector [5] and disregard points that are not close to the extracted edges.

2.2 Feature Extraction

The value of each feature is the mean intensity over a three-dimensional cuboid displaced with respect to the reference voxel position. The cuboid dimensions

and the displacement of each feature are chosen randomly. For the reference voxel p, the feature vector $v(p) = (v_1, ..., v_j, ..., v_m)$ is computed as follows:

$$v_j = \frac{1}{|F_{p;j}|} \sum_{q \in F_{p;j}} I(q), \tag{1}$$

where $I(q)$ is the image intensity at position q in the image, and $q \in F_{p;j}$ are the image voxels within the cuboid. $F_{p;j}$ denotes the feature cuboid j displaced in respect to voxel p. Similar features are used in [8,7,6] for object detection and localization. Extracting mean intensity over cuboidal regions can be computed very quickly by using an integral image (introduced in [18]). These features are then used to train a regressor for vertebra localization.

2.3 Deep Neural Network

In recent years, state-of-the-art results have been produced by applying deep learning to various tasks in computer vision [12,17]. In this work, a deep feed-forward neural network with six layers is used for solving the multi-variate non-linear regression problem. The parameters of the network were set as an input layer holding 500 neurons, and four hidden layers with 200, 300, 200, and 150 neurons, followed by 78 neurons in the output layer. The intensity-based features of each selected voxel are given as the input to the network. The network output is the estimated relative distances of the voxel to the center of each vertebral body. These relative distances are then converted to absolute voxel positions in the image. We use a rectifier function, $g_{hidden}(x) = max(0, x)$, to activate the hidden layers, and a linear function, $g_{output}(x) = x$, to activate the output layer. The deep neural network is trained by using layerwise pre-training [3] and then, by fine-tuning the connection weights using conventional backpropagation. Layerwise pre-training involves breaking down the deep network into several shallow networks and training each of them greedily. Stochastic gradient descent (SGD) is used for minimizing the cost function in both pre-training and fine-tuning steps.

2.4 Centroid Estimation

In our approach, we used an adaptive kernel density estimation method [4] to obtain a fast and reliable density function for the voxel votes for each vertebral body. The global maximum of this density function is considered as the predicted location of the centroid of the vertebral body in the image. The main advantage of using this method as opposed to e.g. the popular Gaussian kernel density estimation is its lower sensitivity to outliers and its fast, automatic data-driven bandwidth selection that does not make any assumptions about normality of the input data [4,13].

2.5 Refinement

The predicted vertebra locations are refined by estimating the centroid of only the votes which are close to the predicted location. For each visible vertebra (according to the prediction of the previous step) in the image, the points around itself and its adjacent vertebrae (if present) are aggregated using kernel density estimation. The previously-localized points are then replaced by the points that are obtained from this step.

3 Experiments and Results

The performance of our method is evaluated on a publicly-available data set[1] consisting of 224 spine-focused CT scans of patients with varying types of pathologies. The pathologies include fractures, high-grade scoliosis, and kyphosis. Many cases are post-operative scans where severe image artifacts are caused by surgical implants. Various sections of the spine are visible in different images. The field-of-view is mostly limited to 5-15 vertebrae while the whole spine is visible in only a few cases.

In this work, *detection* records whether or not the input image contains a specific vertebra while *localization* determines the center of a specific vertebra in the image. After estimating the centroid of the votes of the voxels, our system may conclude that a specific vertebra is outside of the field-of-view of the image. If expert annotation confirms that the vertebra is not visible in the image, we consider it as a true negative (TN). Otherwise, if the image contains the vertebra, it will be a false negative (FN). Similarly, if our system concludes that the vertebra is in the field-of-view of the image and the expert annotations confirm, it is considered as a true positive (TP). Otherwise, it will be a false positive (FP). Based on these parameters, the detection rates are evaluated in terms of accuracy, precision, sensitivity, and specificity. Localization error is defined as the Euclidean distance between the estimated centroid of a vertebral body and its expert annotation. The mean and standard deviation of these localization errors are reported for each region and also for the whole vertebral column. Two-fold cross-validation is performed on two non-overlapping sets of volumetric CT scans with 112 images each. The results on data from fold 1 were obtained after training the deep neural network on fold 2, and vice versa. The folds are selected exactly as in [9] to enable comparing the results. All deep network parameters were set the same for this two-fold cross-validation. This approach also mitigates the risk of overfitting, as the training and testing are performed on two independent data sets.

Our final detection and localization results are presented in Table 1. Figure 2 illustrates exemplary localization results. The results show the capability of our method for detecting the vertebrae present in the image.

We compared our results to [8] and [9] and evaluated their methods on the same data set with a two-fold cross validation with exactly the same fold separation. The results of this comparison is summarized in Table 2.

[1] `http://spineweb.digitalimaginggroup.ca/`

Table 1. Detection rates and localization error for different regions of the vertebral column.

	Accuracy	Precision	Sensitivity	Specificity	Mean error	Std
All	96.0%	94.4%	97.2%	95.0%	18.2 mm	11.4 mm
Cervical	96.0%	91.2%	97.8%	95.0%	17.1 mm	8.7 mm
Thoracic	95.1%	93.9%	95.9%	94.5%	17.2 mm	11.8 mm
Lumbar	98.1%	97.5%	99.4%	96.1%	20.3 mm	12.2 mm

Table 2. Comparison of the detection rates and the mean localization error of our method with prior works. The same training and test sets are used in evaluations of all three methods.

	Accuracy	Precision	Sensitivity	Specificity	Mean error	Std
Ours	**96.0%**	**94.4%**	**97.2%**	**95.0%**	18.2 mm	11.4 mm
CF [9]	93.9%	93.7%	92.9%	94.7%	**12.4 mm**	11.2 mm
RF+HMM [8]	-	-	-	-	20.9 mm	20.0 mm

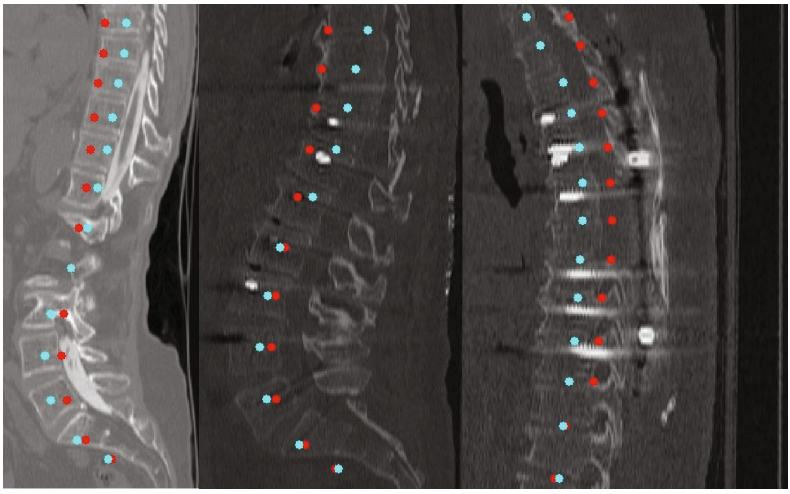

Fig. 2. Visual representation of the refinement step are shown on the mid-sagittal plane of three example images. The localization points before refinement are shown in red while the points after refinement by local centroid estimation are shown in cyan. Refined points have a better representation of the spine curvatures.

For evaluating the influence of Kernel Density Estimation (KDE) for centroid estimation, we repeated the experiments and used the mean and median of the points instead of the maximum of the density function. Using the median of the points instead of KDE reduced the overall accuracy from 96% to 93%, while using the mean of the points reduced it further to 88%.

4 Discussion

We proposed an approach for automatic and simultaneous detection and localization of vertebrae in three-dimensional CT scans using deep neural networks.

We evaluated our algorithm on a large publicly-available data set and compared it against two state-of-the-art methods. We achieved a localization accuracy of 96% which is comparable to results achieved by these methods. Detection and localization of all visible vertebrae can be performed in less then 3 seconds per CT volume which is significantly faster than reported for the methods of comparison that run in the order of minutes.

The key differences between the proposed method and the approaches introduced by Glocker et al. [8,9] are in the choice of the estimator (deep neural networks vs. random forests) and the method used for estimating the vertebra centroid (KDE vs. mean shift clustering). The deep learning approach allowed us to achieve comparably high localization accuracies while reducing the amount of post-processing needed. The influence of KDE in this matter can be regarded relatively low, as even a simple centroid estimation method using the median instead of the maximum of the density function achieves only slightly worse localization accuracies (93% using median vs. 96% using KDE) that are still in the range of the ones reported in [9]. While our results with deep neural networks are highly promising, deep neural networks has a very large number of parameters that need to be optimized preferably on very large data sets. We mitigate this issue in this work by using the greedy layerwise pre-training algorithm [3], which helps with tuning the network parameters layer-by-layer within a much smaller search space.

In [8], a voting framework based on regression forests is used to obtain a rough localization, and then a graphical model based on a hidden Markov model is used for refinement. The results of their method on this public data set is provided in [9]. They have reported lower localization errors on a data set of non-pathological cases. However, the performance of their method degrades significantly in this public pathological data set. A possible reason is that the graphical model cannot accommodate high variations in pathological cases and consequently fails to refine the predictions in this data set. Using a deep learning technique, we can eliminate the need for model-refinement which has relatively high computational cost and did not prove to be robust in pathological cases. The method presented in [9] uses a method based on classification forests that does not require any model-based refinements. While their approach has a lower mean error, our method shows higher accuracy, precision, sensitivity, and specificity. The computational cost of their method is adversely affected by adding extra steps such as a semi-automatic framework to provide dense annotations for the training data, mean shift clustering, and a post-processing step for removing false positives. The algorithms presented in [8] and [9] take about 2 minutes and 1 minute per image, respectively, on a desktop machine. In [8] a joint shape and appearance model in several scales is used to refine the predictions. In [9] centroid density estimates based on vertebra appearance are combined with a shape support term to remove false positives. Our method does not require any post-processing step and runs in less than 3 seconds on a desktop machine.

Combining the advantages of a deep neural network for regression and KDE for final centroid estimation allows us to efficiently detect and localize vertebrae

in CT volumes with high accuracy. The low computational cost of our approach makes it appealing for clinical diagnosis and therapy applications.

Acknowledgement. This work is funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Canadian Institutes of Health Research (CIHR).

References

1. Alomari, R., Corso, J., Chaudhary, V.: Labeling of lumbar discs using both pixel- and object-level features with a two-level probabilistic model. IEEE Transactions on Medical Imaging 30(1), 1–10 (2011)
2. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends® in Machine Learning 2(1), 1–127 (2009)
3. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems 19, 153–168 (2007)
4. Botev, Z., Grotowski, J., Kroese, D., et al.: Kernel density estimation via diffusion. The Annals of Statistics 38(5), 2916–2957 (2010)
5. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 679–698 (1986)
6. Criminisi, A., Robertson, D., Pauly, O., Glocker, B., Konukoglu, E., Shotton, J., Mateus, D., Möller, A., Nekolla, S., Navab, N.: Anatomy detection and localization in 3D medical images. In: Decision Forests for Computer Vision and Medical Image Analysis, pp. 193–209. Springer, Heidelberg (2013)
7. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in CT volumes. In: Proc MICCAI Workshop on Probabilistic Models for Medical Image Analysis, pp. 69–80 (2009)
8. Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E.: Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 590–598. Springer, Heidelberg (2012)
9. Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A.: Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 262–270. Springer, Heidelberg (2013)
10. Kelm, B., Wels, M., Zhou, S., Seifert, S., Suehling, M., Zheng, Y., Comaniciu, D.: Spine detection in CT and MR using iterated marginal space learning. Medical Image Analysis 17(8), 1283–1292 (2013)
11. Klinder, T., Ostermann, J., Ehm, M., Franz, A., Kneser, R., Lorenz, C.: Automated model-based vertebra detection, identification, and segmentation in CT images. Medical Image Analysis 13(3), 471–482 (2009)
12. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. Lim, P., Bagci, U., Bai, L.: Introducing Willmore flow into level set segmentation of spinal vertebrae. IEEE Transactions on Biomedical Engineering 60(1), 115–122 (2013)

14. Oktay, A., Akgul, Y.: Simultaneous localization of lumbar vertebrae and intervertebral discs with SVM-based MRF. IEEE Transactions on Biomedical Engineering 60(9), 2375–2383 (2013)
15. Peng, Z., Zhong, J., Wee, W., Lee, J.: Automated vertebra detection and segmentation from the whole spine MR images. In: 27th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS 2005, pp. 2527–2530. IEEE (2006)
16. Rasoulian, A., Rohling, R., Abolmaesumi, P.: Automatic labeling and segmentation of vertebrae in CT images. In: SPIE Medical Imaging, pp. 903623–903623. International Society for Optics and Photonics (2014)
17. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. In: Advances in Neural Information Processing Systems, pp. 2553–2561 (2013)
18. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)