Automatic Vertebrae Localization from CT Scans using Volumetric Descriptors

Juan Karsten and Ognjen Arandjelović
School of Computer Science
University of St Andrews
Scotland, United Kingdom

Abstract— **The localization and identification of vertebrae in spinal CT images plays an important role in many clinical applications, such as spinal disease diagnosis, surgery planning, and post-surgery assessment. However, automatic vertebrae localization presents numerous challenges due to partial visibility, appearance similarity of different vertebrae, varying data quality, and the presence of pathologies. Most existing methods require prior information on which vertebrae are present in a scan, and perform poorly on pathological cases, making them of little practical value. In this paper we describe three novel types of local information descriptors which are used to build more complex contextual features, and train a random forest classifier. The three features are progressively more complex, systematically addressing a greater number of limitations of the current state of the art.**

I. Introduction

The task of localizing and identifying vertebrae plays an important role in the context of numerous clinical tasks. For example, it is crucial in the diagnosis of spinal problems such as slipped vertebrae, herniated disks, and vertebrae degeneration [9], as well as for a range of surgical procedures such as spinal biopsies, lumbar discectomies, or insertions of pedicle screws [8].

A common physical method of vertebrae localization focuses on the tactile sensing of spinous processes through the skin. [8]. The surgeon usually searches for a specific vertebra, e.g. the first thoracic or the first cervical, and then continues making subsequent ones [8]. Unfortunately, this can be an error-prone and time-consuming process which becomes even more challenging in the presence of significant amounts of fatty tissue [8].

A different type of manual vertebrae localization concerns the localization in electronic images e.g. by marking salient points in X-ray computed tomography (CT) scans with the aid of a computer. However, this process too is time-consuming, takeing up to 15 minutes per patient to localize 11 vertebrae [9]. An additional challenge is posed by the fact that CT scans often capture only a portion of the entire spine [2] which makes it more difficult to identify reliably a reference vertebra (e.g. the first thoracic or cervical, as mentioned before) without prior information of which segment of the spine is visible.

The premise driving the development of automatic methods lies in the expectation that higher consistency, and time and labour efficiency, can be achieved. Nevertheless, this is a not a simple task but one marked with several major challenges. For example, CT data acquired using different scanners exhibits variations in resolution and noise characteristics. Moreover, the field view of scans can vary significantly and is often limited [7], and thus lacks full contextual information. The structure of the spinal column also poses inherent difficulties by its repetitive appearance which makes it challenging to distinguish between different vertebrae. The presence of pathologies, such as severe scoliosis, complicates the task further by making it difficult to impose strong geometric constraints on the inter-vertebral relationships. Lastly, when present, metal surgical implants increase contrast around bone boundary thus adding further confounds [7].

The key contributions of the work described in the present article are three novel types of features used to describe 3D volumes employed in a random forest based automatic vertebrae localization framework, and their evaluation on the largest real-world data set available for public use. In particular, due to computational limitations the existing algorithms in the literature rely on excessively lossy 3D volume descriptors – the simple average of 3D CT scan intensity in the volume – thereby discarding any information on intensity variation or its spatial distribution. The features we describe herein are also compact and thus computationally feasible for use, and address the aforementioned limitations. Specifically, we describe features based on (i) the entropy of intensity, (ii) the maximum entropy intensity histogram, and (iii) the 3D local binary patterns. These respectively capture the amount of intensity variability, the distribution of intensity variability, and intensity variability with spatial information.

II. Method detail

In this section we describe the overall localization framework, as well as three novel cuboid features.

A. Overview

At its core our algorithm employs random forest based classification. As we will elaborate on in detail shortly, the input to this classifier comprises features which combine spatially local CT data descriptors as well as contextual, spatially long range information. The former, being applicable and built from any local descriptors are discussed first, in Section II-C. The three novel types of local CT volume descriptors we propose are detailed in Section II-E. The first of these, based around local entropy seeks to address the key limitation of previous work which, as we highlighted already, discards a significant amount of relevant information for the sake of computational tractability. Our feature retains a higher amount of salient information without compromising on efficiency. The two subsequent features present further enchantments, being able to capture an even greater amount of relevant information while retaining computational feasibility in practice.

B. Dense label generation from sparse manual annotations

Our random forest based approach inherently demands a supervised learning framework. While conceptually simple, this setting poses a serious practical challenge due to the volume of data needing to be labelled. Labelling all CT voxels manually is clearly practically implausible. For this reason we adopt automatic dense label generation from a small, sparse seed set of manual labels. Our dense labelling approach comprises the following steps:

1) Vertebra specific centroid weight $\psi_v(x)$ is computed for all voxels (3D loci within a CT scan) x and all vertebrae v. Given a training image I and the manually annotated location of the vertebra centroid c_v:

$$\psi_v(x) = \exp\left\{-\frac{\|c_v - x\|^2}{h_v}\right\}. \tag{1}$$

Clearly the function produces large values for points x closer to c_v. To distinguish vertebrae and the background, the corresponding background weight is calculated as:

$$\psi_b(x) = 1 - \max_v \psi_v(x). \tag{2}$$

2) Weight distribution for each vertebra and the background is used to compute the corresponding vertebra likelihood:

$$p(l|x) = \frac{\psi(x)}{\sum_{m \in L} \psi_m(x)} \tag{3}$$

3) Hard label value for each voxel is obtained by finding vertebrae with the highest likelihood value from step 2. The training voxel x is thus labelled by the hard label l:

$$l = \arg\max_l p(l|x). \tag{4}$$

This process assigns labels densely to all CT voxels producing sphere like label clusters, as illustrated in Figure 1.

C. The use of contextual information

Capturing local CT scan intensity alone is not sufficient to distinguish between two different vertebrae as different vertebrae have similar appearances. Therefore the use of contextual information spanning a greater spatial range is necessary.

To build a contextual feature v, we consider two cuboid volumes (F_1 and F_2) offset from a CT voxel of interest x. The corresponding contextual feature is computed as the difference of cuboid descriptors $v(x) = d(F_1) - d(F_2)$. Figure 2 illustrates the idea conceptually in two dimensions. Shown are two rectangles (i.e. cuboids in 3D), F_1 and F_2, displaced by d_1 and d_2 relative to voxel x. Local descriptors are computed for the two rectangles and thus the contextual feature based on their difference. Offsets d_1, d_2, and area of box for each feature are randomly chosen before training a random forest tree.

Calculating average intensity can be time consuming especially when dealing with large number of feature and training data. However through the use of the 3D integral image (i.e. the integral volume) representation, the computation can be made dramatically faster [4]; indeed, our implementation adopts this approach.

D. Vertebra centroid estimation

Applying a trained random forest on an input CT scan results in multiple positive labels for each vertebra. From these the best estimate of the vertebra centroid needs to be estimated [1]. We achieve this by employing the well-known mean shift algorithm which operates in an iterative fashion, adjusting the current estimate to be at the point to the average point of neighbourhood density until the local maximum is reached [3]. The mean shift function is defined as below:

$$m(x) = \frac{\sum_{i=1}^{N} K(x - x_i) w(x_i) x_i}{\sum_{i=1}^{N} K(x - x_i) w(x_i)}, \tag{5}$$

where $w(x_i)$ is the weight corresponding to x_i and $K(x - x_i)$ is the value of the kernel function evaluated at $x - x_i$. After calculating $m(x)$, x is updated (shifted) to $m(x)$, and the process repeated convergence. To avoid local minima, multiple starting voxels are considered and the one with highest density chosen, where the density function $q(x)$ is defined as follows:

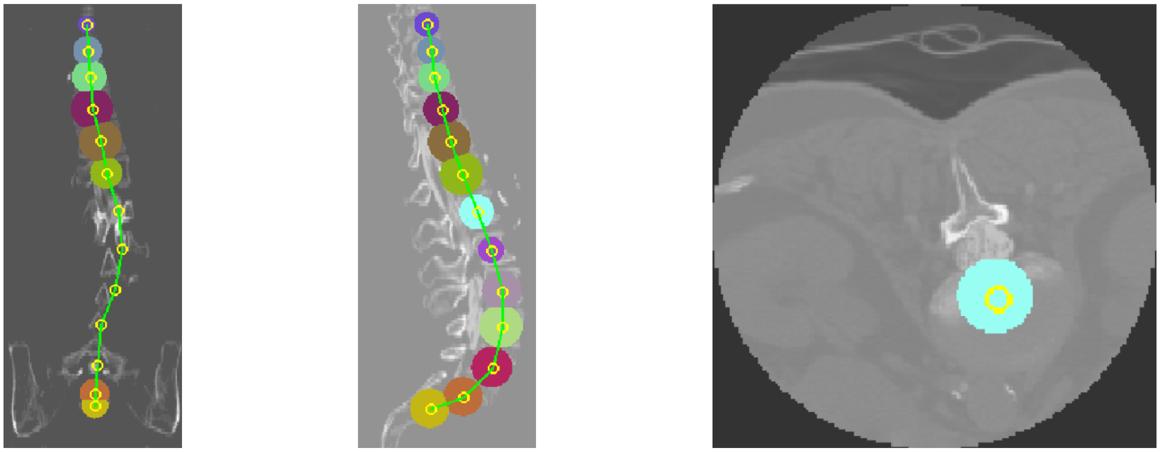
$$q(x) = \sum_{i=1}^{N} K(x - x_i) w(x_i). \tag{6}$$

In this work we used the standard Gaussian kernel, and the weighting function $w(x) = p(v|f(x))$.

E. Local CT volume descriptors

As we noted earlier, the existing state of the art uses an exceedingly simple local information descriptor of a cuboid, in the form of its average value. This choice is largely dictated by the need for computational efficiency, given the size of training data and memory requirements for training a random forest. The limitations of such a simple descriptor are readily apparent. Firstly, no information on the variability of intensity within the cuboid is retained. Moreover, no geometric information corresponding to this variability is captured either. Hence, we describe three novel features which address these limitations while imposing no, or minimal additional memory requirements.

1) Entropy based feature: The first and simplest feature we propose as more descriptive than that used by the current state of the art is based on the entropy of CT scan voxel intensities within a cuboid. In particular, we first create a histogram of the corresponding values (in our experiments we used a four bin histogram) and then compute its entropy. This feature can be seen to capture more in terms of information content within the voxel but no geometric information.



(a)

(b)

(c)

Fig. 1. Sparse manual annotations of vertebrae centroids (yellow circles) and the corresponding dense positive labels (differently coloured spheres, or circles when projected into two dimensions).

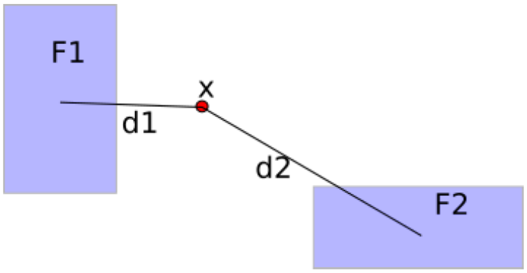


Fig. 2. Conceptual illustration of contextual feature building from local descriptors, in two dimensions.

2) Histogram based feature: Like the previous feature, the second feature we propose herein also does not capture any geometric information but it does retain a greater amount of local intensity content. In particular, instead of ‘collapsing’ the aforementioned histograms into a single value (entropy or, as in previous work, the average value), we use these short histograms as features themselves.

3) 3D local binary pattern based feature: The local binary pattern (LBP) is a feature which has been used with great success in a wide range of 2D image understanding tasks, such as texture analysis and face recognition [5], [10]. The elementary LBP descriptor considers an image patch of size 3×3 pixels. By comparing the values of the 8 neighbouring pixels with the value of the central pixel, the neighbourhood is mapped to a series of binary digits (0 or 1) depending on whether a specific pixel has a smaller value than the central pixel or not. The 8 bit sequence corresponds to an integer in the range $[0, 127]$ and describes the local appearance. The cuboid descriptor we propose here applies the same idea in 3D, i.e. using a neighbourhood of size 26 $(3 \times 3 \times 3 - 1)$ voxels.

III. Evaluation

In this section we describe the experiments we conducted to evaluate the proposed cuboid representations in the context of the wider algorithmic framework we adopted.

TABLE I

Sparse landmarks accompanying image 2804506.

Point	x	y	z
T11_center	72.95	54.6734	240
T12_center	73.9119	51.467	210
L1_center	79.6834	44.7335	177.5
L2_center	82.5692	36.0762	145
L3_center	80.966	30.6253	110
L4_center	68.461	30.946	70
L5_center	54.3528	54.0321	30

A. Data

We used a publicly accessible data set of 224 X-ray computed tomography scans collected and released by Microsoft Research. All images in the corpus are stored as three dimensional metaimages (mhd) commonly used in medical imaging research [6]. The size of images varies between approximately 50 megabytes to 120 megabytes. In total, the corpus takes up around 20.4 gigabytes.

Accompanying each metaimage is a landmark file which contain sparse annotations of vertebrae centroid positions. Table I illustrates the information contained in the file. In the case shown, image #2804506 contains seven vertebrae: 2 thoracic (T11 and T12) and 5 lumbar vertebrae (L1, L2, L3, L4, and L5). The three floating numbers associated with each vertebra specify the coordinates of its centroid, localized manually. Note that most scans do not include the entire vertebrae set.

B. Methodology and parameters

To ensure robust results and lack of experimental bias, we employed the standard five-fold cross-validation protocol. The values of the parameters of the main method used in all experiments are summarized in Table II, while those of the mean shift algorithm used for centroid estimation are shown in Table III.

TABLE II

KEY METHOD PARAMETERS USED FOR OUR EXPERIMENTS.

Parameter	Value
Number of trees	10
Total features	10,000
Candidate features	200
Candidate threshold	10
Minimum samples in node	8
Minimum box offset	0 mm
Maximum box offset	100 mm
Minimum box size	2 mm
Maximum box size	100 mm

TABLE III

MEAN SHIFT PARAMETERS USED FOR OUR EXPERIMENTS.

Parameter	Value
Number of starting points	100
Maximum iteration	40,000
Delta	0.00001
Number of highest density points considered	50,000

C. Results

To obtain baseline performance, we first evaluated our method using the simple, cuboid average features employed by the current state of the art. The results are summarized in Table IV and corroborate the findings reported in the literature: the approach is broadly successful but with much room for improvement left. In particular, note that the correct vertebra identification rate is highly dependent on the vertebra type, being best for cervical, and worst for lumbar and sacral complexes.

Equivalent results obtained using the 3D LBP based descriptor we described in Section II-E are shown in Table V. Consistently worse performance than that achieved with the simple average cuboid representation can be seen throughout the table both in terms of accuracy and precision. Similar trends were noticed for the other two features (detailed statistics are not included due to a lack of space). Though at first sight this highly surprising result is disappointing, we believe that our findings both call for more examination of the reasons behind it, as well as illuminate future work directions. In particular, we believe that the more expressive nature of our features inherently requires a larger training data set. In that sense, our comparison can be interpreted at being unfair towards the novel features.

TABLE IV

SUMMARY OF EXPERIMENTAL RESULTS (STATISTICS OF LOCALIZATION ERROR IN PIXELS AND ID RATE) USING THE SIMPLE, CUBOID AVERAGE FEATURES USE BY THE CURRENT STATE OF THE ART AS REPORTED IN THE LITERATURE.

Vertebrae set	Mean	STD	Median	ID rate
All	21.03	23.43	14.35	74.93%
Cervical	14.73	22.42	10.47	90.18%
Thoracic	22.70	23.66	15.97	70.82%
Lumbar and sacral	23.18	22.99	16.72	69.80%

TABLE V

SUMMARY OF EXPERIMENTAL RESULTS (LOCALIZATION ERROR STATISTICS AND ID RATE) USING OUR 3D LBP BASED FEATURE.

Vertebrae set	Mean	STD	Median	ID rate
All	29.98	41.31	19.02	65.64%
Cervical	22.44	44.81	14.67	81.74%
Thoracic	32.76	41.01	20.57	61.16%
Lumbar and sacral	31.41	38.23	21.09	60.37%

IV. SUMMARY

In this paper we considered the problem of localizing human vertebrae in three dimensional CT scan data. We proposed three new types of local features aimed at overcoming the limitations of those used by the current state of the art – namely, the loss of local appearance and geometric information. Our findings derived from experiments on the largest public data set of CT scans of pathological cases should help guide future research efforts.

REFERENCES

[1] O. Arandjelović and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1:860–867, 2005.

[2] H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J. C. Y. Cheng, and P.-A. Heng. Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. *Medical Image Computing and Computer-Assisted Intervention*, pages 515–522, 2015.

[3] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[4] F. C. Crow. Summed-area tables for texture mapping. *In Proc. Conference on Computer Graphics*, 1:207–212, 1984.

[5] C. Fare and O. Arandjelović. Ancient Roman coin retrieval: a new dataset and a systematic examination of the effects of coin grade. *In Proc. European Conference on Information Retrieval*, pages 410–423, 2017.

[6] T. Glatard, A. Marion, H. Benoit-Cattin, S. Camarasu-Pop, P. Clarysse, R. F. da Silva, G. Forestier, B. Gibaud, C. Lartizien, H. Liebgott, K. Moulin, and D. Friboulet. Multi-modality image simulation with the virtual imaging platform: illustration on cardiac echography and MRI. *IEEE International Symposium on Biomedical Imaging*, pages 98–101, 2012.

[7] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. *Medical Image Computing and Computer-Assisted Intervention*, pages 262–270, 2013.

[8] J. L. Herring and B. M. Dawant. Automatic lumbar vertebral identification using surface-based registration. *Journal of Biomedical Informatics*, 34(2):74–84, 2001.

[9] Z. Peng, J. Zhong, W. Wee, and J.-H. Lee. Automated vertebra detection and segmentation from the whole spine MR images. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2527–2530, 2005.

[10] H. Tang, B. Yin, Y. Sun, and Y. Hu. 3D face recognition using local binary patterns. *Signal Processing*, 93(8):2190–2198, 2013.