Journal of Digital Imaging (2019) 32:336–348
https://doi.org/10.1007/s10278-018-0140-5

Automatic Vertebrae Localization and Identification by Combining Deep SSAE Contextual Features and Structured Regression Forest

Xuchu Wang[1] · **Suiqiang Zhai**[1] · **Yanmin Niu**[1,2]

Published online: 10 January 2019

Abstract

Automatic vertebrae localization and identification in medical computed tomography (CT) scans is of great value for computer-aided spine diseases diagnosis. In order to overcome the disadvantages of the approaches employing hand-crafted, low-level features and based on field-of-view priori assumption of spine structure, an automatic method is proposed to localize and identify vertebrae by combining deep stacked sparse autoencoder (SSAE) contextual features and structured regression forest (SRF). The method employs SSAE to learn image deep contextual features instead of hand-crafted ones by building larger-range input samples to improve their contextual discrimination ability. In the localization and identification stage, it incorporates the SRF model to achieve whole spine localization, then screens those vertebrae within the image, thus relieves the assumption that the part of spine in the field of image is visible. In the end, the output distribution of SRF and spine CT scans properties are assembled to develop a two-stage progressive refining strategy, where the mean-shift kernel density estimation and Otsu method instead of Markov random field (MRF) are adopted to reduce model complexity and refine vertebrae localization results. Extensive evaluation was performed on a challenging data set of 98 spine CT scans. Compared with the hidden Markov model and the method based on convolutional neural network (CNN), the proposed approach could effectively and automatically locate and identify spinal targets in CT scans, and achieve higher localization accuracy, low model complexity, and no need for any assumptions about visual field in CT scans.

Keywords Vertebrae localization · Stacked sparse autoencoder (SSAE) · Contextual feature · Structured regression forest (SRF) · Kernel density estimation

Introduction

Automatic vertebrae localization and identification is a key step for spine analysis in medical CT scans [1]. It is also pre-order for tasks such as vertebrae segmentation [2, 3], vertebrae fracture detection [4], intervertebral disc labelling [5], and vertebrae shape statistical analysis [3]. In addition, accurate localization and identification of vertebrae can greatly reduce the risk of wrong-level surgery. However, it is highly challenging to automatically localize and identify the vertebrae due to the high similarity of morphological appearance, low contrast between vertebrae and surrounding anatomic structure, spine deformation, and limited field of view of CT scans [6].

To solve this challenging problem, several methods currently have been proposed to localize and identify vertebrae automatically and they can be roughly classified into two types according to whether priori assumptions are taken into account or have constraints about the region of the spinal images. The first type mainly concentrates on specific regions of the spine such as lumbar vertebrae and thoracic regions [5, 7], or depend on the prior knowledge about which part is visible [1, 6, 8–11]. These methods usually achieve high localization accuracy; however, the priori assumptions or constraints will make them less applicable to general cases or pathological images.

The second type focuses on relieving the prior assumptions or the limit of visible part of the vertebral column in the image. Klinder et al. [12] present a comprehensive

✉ Xuchu Wang
seadrift.wang@gmail.com; xcwang@cqu.edu.cn

[1] Key Laboratory of Optoelectronic Technology and Systems of Ministry of Education, College of Optoelectronic Engineering, Chongqing University, Campus A, Room 1105 of Main Building, 174 Shazhen Street, Shapinba District, Chongqing 400040, China

[2] College of Computer and Information Science, Chongqing Normal University, Chongqing 400050, China

solution to copy with the spine image with any field of view while it is a quite complex multi-step process. And Markov random field also is applied to the whole spinal localization by considering the adjacent vertebra information [13]. However, the model assumptions of their approach may make it fail to detect these vertebrae, when larger implants are present. Glocker et al. [14] localized and identified vertebrae in arbitrary field-of-view CT scans with regression forest and hidden Markov model. However, their work makes assumptions about the shape and appearance of vertebrae, which may not be satisfied on pathological or abnormal spinal images. To address the limitations, they further designed a random forest classifier for avoiding explicit parametric modelling of appearance by employing a semi-automatic labelling strategy where sparse centroid annotations are transformed into dense probabilistic labels [15].

One limitation of the above-mentioned methods was that models were trained by hand-crafted feature descriptors such as box features or local intensity features which cannot encode more representative feature of spinal images, so it may fail to handle more complicated cases when spine curvature and pathologies exist. In addition, a six-layer feed-forward neural network is employed to model vertebra localization into a multi-variable nonlinear regression problem [16], and in this method, the input samples are still built based on the hand-crafted box features. Instead of employing low-level hand-crafted features, a joint learning model with convolutional neural network (CNN) is proposed to exploit high-level feature representations by considering both the appearance of vertebrae and the pairwise conditional dependency of neighboring vertebrae [17]. However, this method also incorporates some complex refinement steps to improve the localization accuracy. Besides, a progressive optimization strategy is proposed by combining multiple neural networks [18]. These networks consist of a deep image-to-image network for initializing vertebra locations, a convolutional long short-term memory network for modelling centroid probability map sequence, and a network for further refining and regularizing the landmark positions. Furthermore, a combined method [19] inherently learned to incorporate both the short-range and long-range contextual information by 3D fully convolutional neural network and multi-task bidirectional recurrent neural network in a supervised manner.

Although the state-of-the-art methods have already achieved acceptable performance in 3D spine scans, however, the complex network models generally come at high computational cost. Compared to the above-mentioned methods, we argue that to further improve vertebrae localization and identification results and reduce computational cost, we should (1) exploit the richer contextual high-level features such that we can better capture more discriminative sample features representation; (2) use a structured regression model to localize all vertebrae in CT scans such that we can enforce the structural information of the adjacent pixels in image patches. Firstly, we insist on that low-level hand-crafted features and local patch hardly represent the abundant shape and texture features of CT images, owing to the fact that there are repetitive nature of spine and high inter-subject variability in spine curvature and shape due to spine disorders and pathologies, so it is essential to explore the deep contextual features. Secondly, we also emphasize that the ordinary regression-based methods usually predict the central pixel of each patch separately and ignore the structured information of the adjacent context of an image patch. It is evidential that structured regression strategy could improve the result performance [20–23] since the label space of training samples exhibits an inherently topological structure, which renders the class labels explicitly interdependent.

To this end, we propose a method for automatic vertebrae localization and identification by combining deep stacked sparse autoencoder (SSAE) contextual features and structured regression forest (SRF). The proposed method has the following characteristics: (1) In the feature learning stage, we employ SSAE network to learn inherent discriminating deep contextual features instead of low-level hand-crafted features. (2) In the vertebrae localization and identification stage, we incorporate a structured regression forest by embedding structured information into standard regression forest. Compared with the latter, the structured prediction function maps the input space to structured label space instead of discrete label space that it is likely to promote the localization accuracy because of incorporating contextual topology information. In addition, based on the regression model, we can locate the entire spine and relieve the dependence of prior knowledge about which part is visible in CT scan. (3) In the refined localization stage, we develop a two-stage progressive refining strategy with the mean-shift kernel density estimation and the Otsu method instead of Markov random field (MRF) to reduce model complexity.

Method

As shown in Fig. 1, our proposed approach consists of a training stage and a testing stage. In the training stage, all spine CT images are normalized, then a series of contextual patch sample points are generated on the normalized image and fed to train a SSAE feature learning network. Subsequently, the deep contextual features learned by SSAE are sent to train a SRF. In the testing stage, given a previously unseen CT scan, image testing sample points

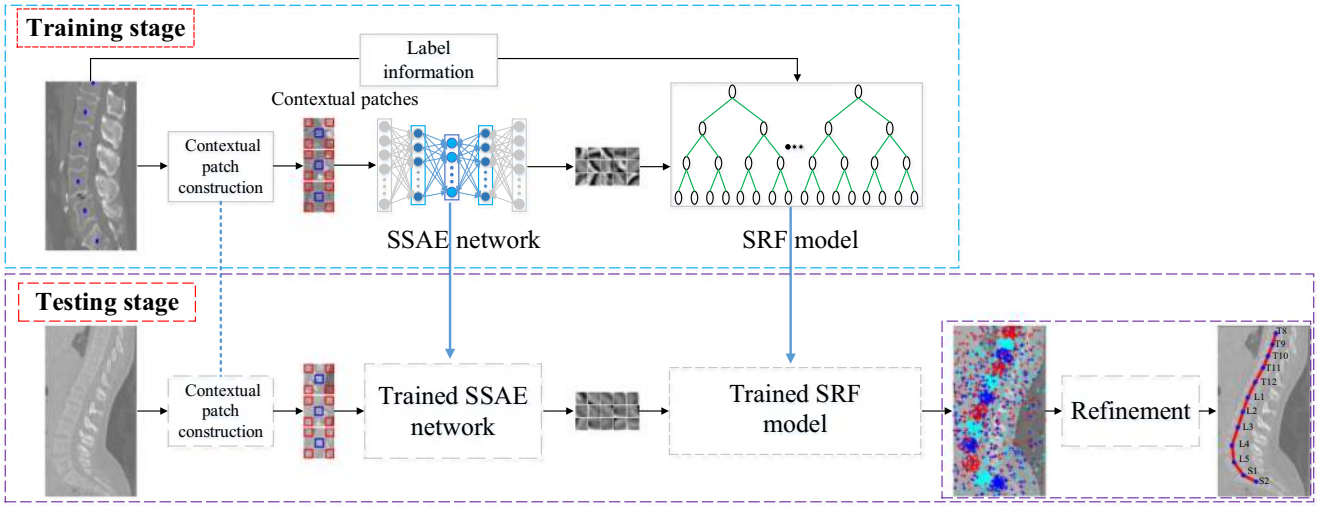

Fig. 1 Proposed flowchart of automatic localization and identification of vertebrae

could cast a series of (probabilistic) structured votes on the positions of all vertebrae. Finally, we employ the proposed two-stage progressive refining strategy to adjust the output of the regression model to improve localization results. The following sections will present the details related to the data preprocessing, the contextual patch construction, the deep contextual feature learning of the SSAE network, the structured regression forest, and the two-stage progressive refining strategy.

Contextual Patch Construction Stage

The proposed contextual patch construction mainly consists of sample point adjustment preprocessing step and contextual patch construction step.

(1) Sample point adjustment preprocessing step. In some spine CT images, the local texture structure of different image patches exists with quite high difference. In general, patch features are highly discriminative when sample points are close to the distinct object regions. However, there are certain regions in most images, like those in the background, that do not benefit the localization and identification of vertebrae. Considering some features such as the structure and texture of image patch around the vertebrae region are highly correlated, these areas would be desirably employed as candidates for sample generation. Based on these samples, the subsequent SSAE network could extract more discriminative feature representation which helps to improve model prediction accuracy. Thus, in this step, we employ the unsupervised Otsu method to divide the obvious difference in gray scale between the vertebrae and surrounding anatomical structures in the CT image, and then morphologically separate images into binary connected blocks. Based on this, sample points around the centroid of the larger connected blocks are selected as candidate points.

(2) Contextual patch construction step. After sample points are randomly selected and adjusted, we develop a contextual patch-constructing strategy to capture image contextual information in a larger-range manner. Figure 2 illuminates these two steps, where green patches represent filtered samples near to the background and blue ones are samples to be reserved in the left column, while four red patches and one blue patch represent contextual patch in the right column. Here, we only exemplify three points of all blue samples as example for illustrating the constructing process of the contextual patch, where there are five sparse patches with the same size in a larger range to capture the contextual information. It can alleviate the influence of the repeating structures in vertebrae regions.

Deep Contextual Feature Learning Using SSAE

Instead of employing low-level hand-crafted features, some supervised CNN-based feature learning methods [16–19] have obtained competitive results for vertebrae localization and identification. In contrast to CNNs that apply a series of convolution-pooling-subsampling operations to learn deep feature representations, SSAE employs a full connection of units for deep feature learning. SSAE contains multiple hidden layers and millions of trainable parameters that enable it to capture highly nonlinear mapping between input and output; thus, recently, it has been widely used in image recognition fields. Some existing results indicate that the architecture of SSAE is essential for achieving better

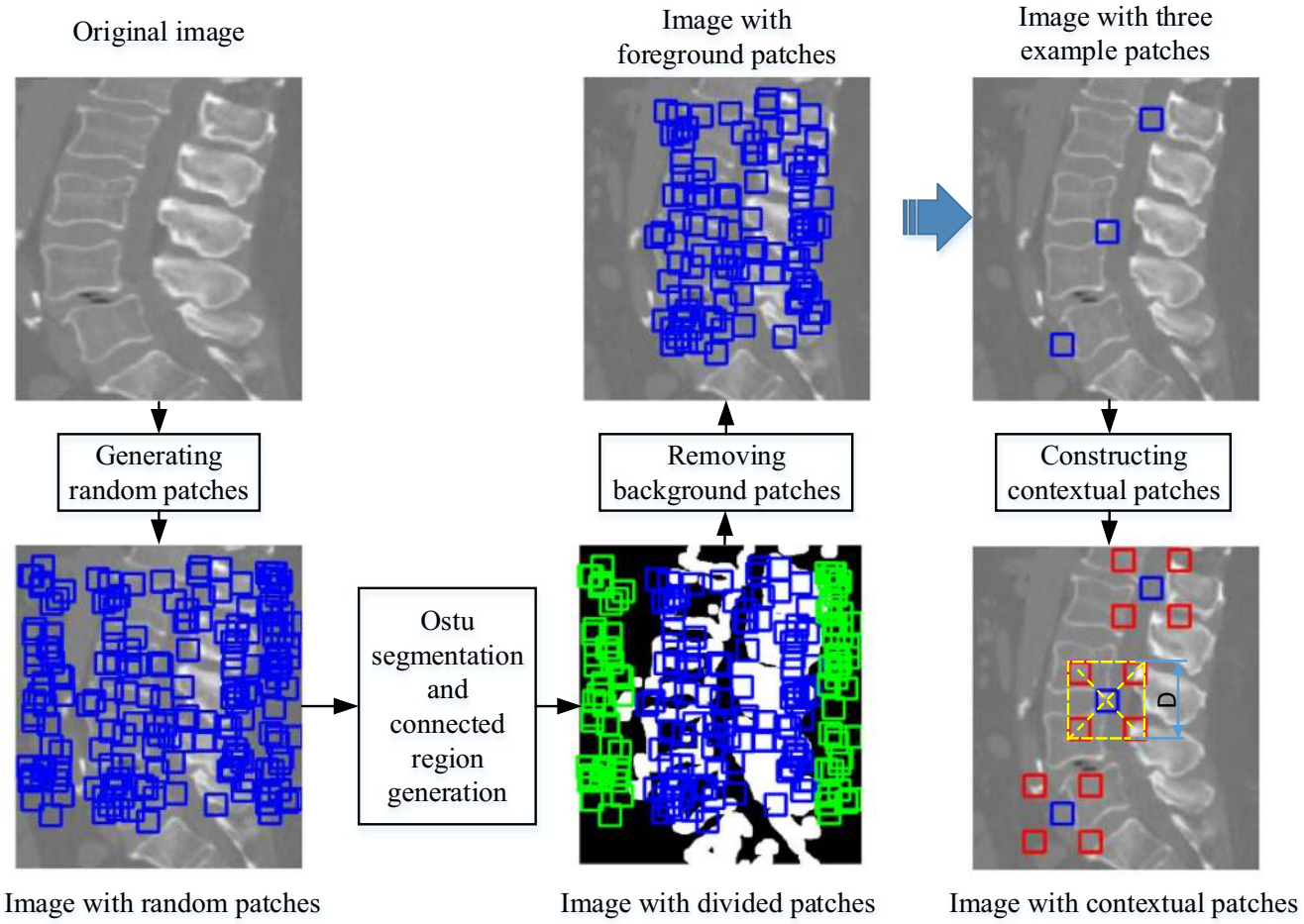

Fig. 2 Diagram of sample point adjustment step and contextual patch construction step

performance in some specific tasks [24], which motivates us to investigate the SSAE-based feature learning for vertebrae localization and identification.

SSAE essentially is a neural network consisting of multiple layers of sparse autoencoders (AEs) in which the outputs of each layer are connected to the inputs of the successive layer [24, 25]. AE is an unsupervised learning algorithm that implements feature encoding by setting the target values to be equal to the inputs. If some specific structures are implied between the input samples, such as some input features are related to each other, then the autoencoder algorithm can capture a useful "hierarchical grouping" or "part-whole decomposition" of the input. In the autoencoder algorithm, the number of hidden units generally is less than the number of the input layer units. But when the number of hidden units is larger (or even larger than the dimension of the input vector), if we impose sparse constraints on the hidden layer to make only a few neurons activated, the network will be forced to learn a compressed representation of the input to discover an interesting structure in the data. Let h_j denote the activation of hidden units j in the autoencoder and T denote the number of training data, and the average activation ratio of hidden unit j could be written

$$\hat{\rho} = \frac{1}{T}\sum_{t=1}^{T} h_j(t). \tag{1}$$

Typically, the average activity is limited to a small value close to zero, that is, the hidden unit's activations must mostly be near to 0. To achieve this, we will choose the following optimization function

$$KL(\rho||\hat{\rho}_j) = \rho log\frac{\rho}{\hat{\rho}} + (1-\rho)log\frac{1-\rho}{1-\hat{\rho}_j}, \tag{2}$$

where ρ is a sparsity parameter, typically a small value close to zero. The formula is sparse parameter (usually a very small value, for example, 0.05). The above functions essentially rely on the penalty term of Kullback-Leibler divergence to enforce activity of hidden unit is close to 0.

Combined with the reconstruction error of the autoencoder algorithm, the overall cost function can be written as

$$J_s(W,b) = \frac{1}{2T}\sum_{t=1}^{T}||\hat{x}_t - \hat{y}_t||_2^2 + \lambda||W||^2 + \beta\sum_{j=1}^{N} KL(\rho||\hat{\rho}_j), \quad (3)$$

where $\hat{x}_t$ denotes the tth contextual patch in the training set; $\hat{y}_t$ is the reconstructed representation with input of $\hat{x}_t$. The second term is a regularization term with weight decay control parameter λ. It aims to avoid overfitting by preventing the magnitude of the neuron weights (W) from increasing dramatically. β is a sparsity parameter on N neurons in the hidden layer. In general, SSAE is greedily layer-wise trained to obtain optimized parameters. Figure 3 shows the SSAE network structure for extracting the proposed contextual patch features. Given an input training image, firstly numerous sample points denoted with blue boxes are generated in 3D CT scans, and then the 2D contextual patches are constructed corresponding to these points. Next, these contextual patches are rearranged into 1D vectors to feed into the SSAE network. As a result, the trained network aims to learn a compressed representation of the input by limiting sparsity to small value close to 0. Compared with AE, SSAE enjoys all the benefits of any deep network of greater expressive power and tends to discover more abstract higher-order features; thus, it is more likely to help the subsequent structured regression forest to localize and identify each vertebra.

Structured Regression Forests

The random forest approach provides a promising perspective for localizing and identifying the vertebrae from spine images [14]. However, the standard random forest usually ignores the structural information of the adjacent context of an image patch since it commonly predicts the central position of each patch separately. For many computer vision problems, the standard model is limited because the label space of training samples exhibits an inherently topological structure, which renders the class labels explicitly interdependent. To overcome this limitation, a simple and effective way is presented to integrate ideas of structured learning into the standard random forest framework for the task of semantic image labeling [20], then the structured random forest (SRF) strategy is extended to apply to boundary detection [23], medical image myocardium delineation [21], and hand detection and hand part labeling [22].

Our structured label construction is proposed to identify every vertebra because of the following consideration. Generally, the human spine contains 26 individual vertebrae, where the regular 24 from the cervical (C1-C7), thoracic (T1-T12), lumbar (L1-L5), and sacrum (S1-S2) vertebrae regions. Therefore, the 3D voxel coordinates of the 26 vertebrae centroids can be defined as the 78 dimensions regressive vector. Let $C = \{c_i\}_{i=1}^{26}$ denote 26 vertebrae centroid coordinates where $c_i = \{x_i, y_i, z_i\}$ represents individual vertebrae voxel coordinates. Given annotated CT scans, the construction process of the proposed structured label for training samples is illustrated in Fig. 4. The details in Fig. 4 are described as follows, firstly we divide the central blue patch of the contextual patch into $\omega \times \omega$ subregion, then we compute its relative displacements to all available vertebrae centroids given by employing $y_p^{(i)(j)} = (c_j - p^{(i)})^T, j = 1, 2, \cdots n$ for each subregion centroid $p^{(i)}$. In this way, each sample will aggregate a structured label vector with capacity of ω^2 in which training sample can be expressed as $\chi = \{x_p, y_p\}$ where x_p denotes SSAE deep contextual feature and $y_p = \{y_p^1, y_p^2, \cdots, y_p^n\}$ is structured label.

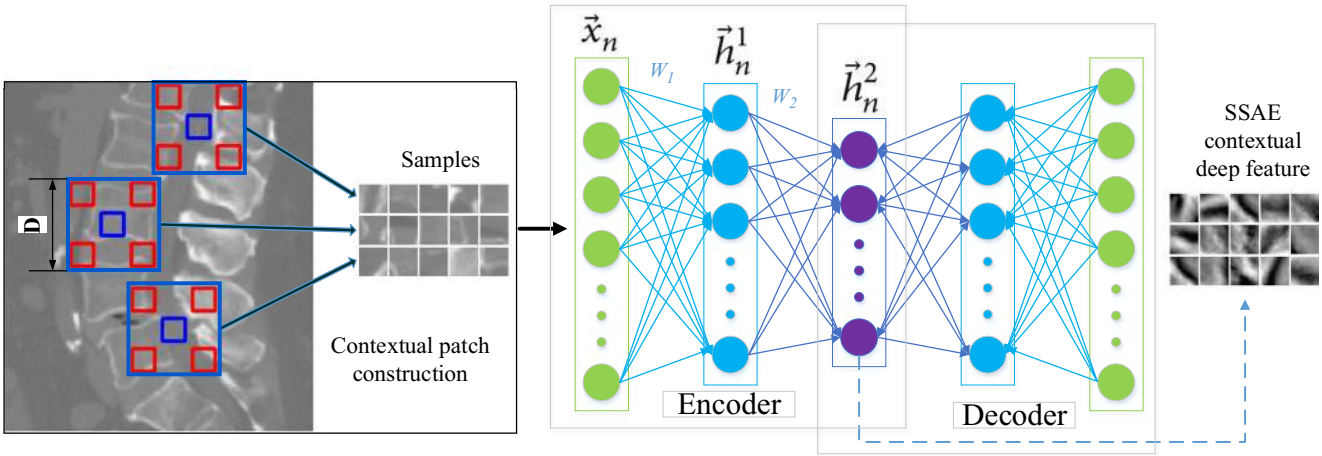

Fig. 3 Two-hidden-layer SSAE network for learning contextual patch features

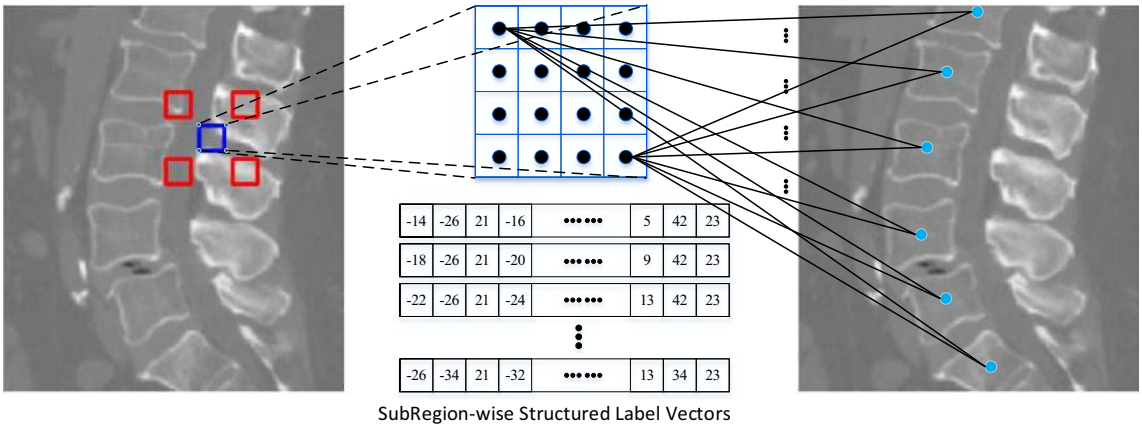

Fig. 4 Structured label construction based on contextual patch

The sample points are obtained from the 3D CT scans to construct contextual feature. For example, given the 3D center coordinates $p = \{p_x, p_y, p_z\}$ of a patch p, we can calculate label $y_p = (c_1 - p)^T, (c_2 - p)^T, \cdots, (c_{26} - p)^T$ corresponding to this point. Although sample label is calculated in 3D space, we only extract 2D image patch at the YZ plane corresponding to this point for constructing the contextual image patch feature. Based on the above strategy, we can achieve the vertebrae localization and identification in 3D CT scans.

The structured label construction provides an adaptive configuration in regression forests (RF) for predicting each vertebra. RF intrinsically is a supervised learning technique for the probabilistic estimation of continuous variables and has demonstrated good performance on the localization of anatomical structures in CT volumes [15]. A forest is an ensemble of several binary regression trees, where each tree t learns its own predictor $p_t(y|\chi)$. Given a training set $S^{(t)} = \{(x_{(i)}, y_{(i)})\}_{i=1}^{n}$ obtained from the SSAE network, training a regression tree is done by recursively subdividing training samples into the left and right child nodes. A local split function is determined at each internal node based on arriving examples. For each tree, the part of training samples $S^{(t)} = \{(x_{(i)}, y_{(i)})\}_{i=1}^{n}$ is randomly selected from the training set to train the tree, and the stopping criterion is used to determine each node whether or not it is a leaf node. For a split node, the splitting aims to optimize and determine parameter of splitting function. Here, the node split function is defined as

$$f_{\theta_m;\theta_\tau(x)} = \begin{cases} 1; \text{ if } \theta_m \cdot x \geq \theta_\tau \\ 0; \text{ otherwise} \end{cases} \tag{4}$$

where θ_m is an M-dimensional binary vector with m term nonzero at random-selected m indexes, and $\theta_\tau \in \mathbf{R}$ is a threshold determining whether the samples is divided into the left or right node. The split is defined by the formula (5), and the training purpose is to determine the optimal parameter by greedy learning.

$$\begin{aligned} SL^{(j)} &= \{(x^{(i)}, y^{(i)} : f_{\theta_m;\theta_\tau(x)} = 1)\}; \\ SR^{(j)} &= \{(x^{(i)}, y^{(i)} : f_{\theta_m;\theta_\tau(x)} = 0)\}. \end{aligned} \tag{5}$$

Here, $SL^{(j)}$ and $SR^{(j)}$ denote left and right child training subset, respectively.

Based on the information theory, the smaller the expectation of the data, the larger the information gain, and the higher the purity of the data, then the objective of optimizing the node splitting parameters can be determined by maximizing the information gain; thus, objective function could is defined

$$I(S_j, \theta) = H(S_j) - \sum\nolimits_{i \in \{L,R\}} \frac{|S_j^i|}{|S_j|} H(S_j^i), \tag{6}$$

where H is a measure of entropy [26]. Splitting parameters are determined by following a random and greedy optimization strategy. That is, selecting m terms in feature vector randomly, we will evaluate the information gain according to a set of different thresholds from a uniform distribution. The parameter pair $\{\theta_m, \theta_\tau\}$ corresponding to the largest information gain is considered the optimal parameter of split function $f_{\theta_m;\theta_\tau}$. That is

$$\theta_j = \arg\max_{\theta \in \Gamma_j} I(S_j, \theta). \tag{7}$$

For the multivariable regression task of the vertebrae localization, the priori distribution model stored in each node can be modeled into 78-dimensional multivariate normal distribution $p(y|x) \triangleq N_{78}(\bar{y}, \sum)$, where $\bar{y}$ corresponds to the mean and $\sum$ is the covariance matrix of all training samples that reaches the node. Based on

multivariate normal distribution, the information gain in Eq. 7 could be rewritten as

$$I(S_j,\theta)=\sum_{x\in S_j}\log(|\sum_{y}(x)|)-\sum_{i\in\{L,R\}}\sum_{x\in S_j^i}\log(|\sum_{y}(x)|). \tag{8}$$

In the normal distribution model, the definition of entropy is proportional to logarithm of the determinant of the covariance matrix, namely differential entropy $H_N = \log\left|\sum\right|$. Plugging this into Eq. 7 fully defines the objective function for regression tree training [14]. Minimizing the trace of the covariance matrix tends to aggregate similar samples in the training set, that is, the variance of the corresponding offset vectors is minimized. And the samples in the leaf node tend to come from similar anatomical regions that lead to spatially consistent clusters of the offset vectors. Following the above strategy, we can achieve the goal of optimizing tree parameters, and the regression forest is the integration of multiple trained trees.

Forest Testing In general, given a previously unseen CT scan, a series of deep SSAE contextual features from image sample points are applied for predicting the positions of all vertebrae based on an empirical distribution at the leaf node. In fact, each sample is pushed through all trained trees. Each sample feature is split recursively into the left and right child nodes until the point reaches a leaf node. The corresponding predictor function stored in a leaf node will give single-predicted offset vector for all vertebrae positions. Compared to standard regression forest, each sample label in SRF is attached to a structured label. Based on the proposed structured labels strategy, we have redesigned the prediction model in leaf nodes. Considering that each training sample contains a structured label with a capacity of $\omega\times\omega$, we design an effective average modelling way to preserve the structural properties in the prediction. That is

$$P^t(y|x)=\frac{1}{N}\sum\nolimits_{j=1}^{N}P^t(y^{(j)}|x), \tag{9}$$

where $P^t(y^{(j)}|x)$ is prior distribution at the leaf node. When given a test point, it will obtain one average structured prediction vector with a capacity of $\omega\times\omega$ instead of a single-prediction vector.

The final probabilistic prediction of the regression forest is determined by simple averaging over tree predictions to obtain the structural result corresponding to the unlabelled sample point. That is

$$P(y|x)=\frac{1}{T}\sum\nolimits_{t=1}^{T}P^t(y|x). \tag{10}$$

Given an unseen CT scan, the structured regression forest model can obtain 26 three-dimensional point clusters corresponding to the all vertebrae, and those clusters will be used for the subsequent refinement localization stage.

Refinement Stage

The First Refinement Stage The posterior of centroids obtained from the regression forest could be directly used to generate the MAP estimate. However, this does not yield very accurate results because of the prediction properties of the structured regression forest. Thus, it is necessary to design a refinement stage to more accurate localizations. Based on the prediction of structured regression forest, we develop a two-stage progressive strategy to refine vertebrae localization by means of the mean-shift kernel density estimation [27] and the Otsu segmentation approach.

Given a series of point clusters from the regression forest, we employ the mean-shift algorithm to estimate the centroids of 26 vertebrae. Mean-shift is a kernel density estimation algorithm which the estimation value is gradually converging along the density gradient direction, and the last convergence locates at local probability density maximum position. Let $x_i(i=0,1,\cdots,n)$ be n samples and $K(x)=c_{k,d}k(\|x^2\|)$ denotes kernel function, and the kernel density of the point x in the space is defined as

$$\hat{f}(x)=\frac{1}{nh^d}\sum\nolimits_{i=1}^{n}K(\frac{x-x_i}{h}), \tag{11}$$

where $c_{k,d}$ is the normalization parameter so that the integral of the kernel function is 1. To obtain the point of maximum density in the point sets, computing derivative of Eq. 12, let $g(x)=-k'(x)$ and $G(x)=c_{k,d}g(\|x^2\|)$, the gradient is computed as follows:

$$\begin{aligned}\hat{\nabla}f_{h,k}(x)&=\frac{2c_{k,d}}{nh^{d+2}}[\sum_{i=1}^{n}g(||\frac{x-x_i}{h}||^2)]\times\\&[\frac{\sum_{i=1}^{n}x_i g(\left\|\frac{x-x_i}{h}\right\|^2)}{\sum_{i=1}^{n}g(\left\|\frac{x-x_i}{h}\right\|^2)}-x].\end{aligned} \tag{12}$$

Assuming that all weights of the sampling points are $1/n$ and the bandwidth matrix is proportional to the unit matrix $H_i=h^2I$, then the iteration formula of the mean-shift can be expressed as

$$x_{i+1}=\frac{\sum_{i=1}^{n}g(\left\|(x-x_i)h^{-1}\right\|^2)x_i}{\sum_{i=1}^{n}g(\left\|(x-x_i)h^{-1}\right\|^2)}. \tag{13}$$

Given the multivariate normal kernel function $G(x)$ and admissible error ϵ, the iterative steps of the mean-shift algorithm for evaluating the vertebral centroid are described as follows:

Step 1. Initialize search sphere area O of the radius h in the each point cluster.

Step 2. Calculate the mean $m_{h,G(x)}$ of all points in the sphere area according to the iterative formula

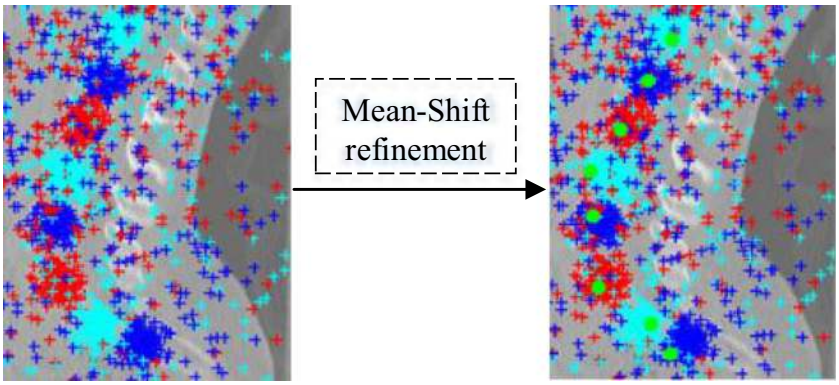

Fig. 5 First stage of refining the vertebrae localization, where the dot clusters represent SRF output and light green dots represent mean-shift estimated results

where the density in $m_{h,G(x)}$ is greater than the density of the sphere center O.

Step 3. Calculate the density difference between the sphere center O and the mean $m_{h,G(x)}$. If $\|m_{h,G(x)} - x\| \leq \varepsilon$, let the loop be end. Otherwise, go to step 4. Here, $\|m_{h,G(x)} - x\|$ is the mean-shift vector toward to the direction of the probability density increases.

Step 4. Assign the mean $m_{h,G(x)}$ and the coordinate position of $m_{h,G(x)}$ to the sphere center O , and execute step 2.

Figure 5 shows the sketch of mean-shift kernel density estimation. For the 26-point clusters obtained from the structured regression forest, this algorithm can determine the density maximum point in turn to obtain the centroids of the 26 vertebrae.

The Second Refinement Stage The mean-shift refinement step is beneficial to locate the approximate positions of the corresponding vertebrae. However, these points may deviate from the centroids of the corresponding vertebrae. We can perform the second refinement stage by segmenting the local image around the predicted point of each vertebra using the local Otsu approach, and then seek local binary connected component. The previously localized points are then replaced by the centroid of the local biggest binary connected component close to the predicted points. The detail is shown in Fig. 6, where the green dots denote the mean-shift estimated point, and the red dots denote the final refined point. It can be seen that most red dots have almost approached to the real centroids of vertebrae.

Experimental Results and Discussion

This proposal algorithm was evaluated on the available data set consisting of 98 spine-focused CT scans that include slightly pathological cases and normal CT scans. Based on vertebrae localization and identification results, we can predict the 3D centroid coordinates of 26 vertebrae in each CT scan. However, some vertebrae are missing in most CT scans, and the proposed algorithm still obtains the localization coordinates of 26 vertebrae without any priori assumption about the visibility of each vertebra in CT scan. Considering the above property, we only leave the localization points in the CT scan visual field as the result of the mean-shift refinement output. In the following we will describe the data sets and experiments in details.

Data Sets

The evaluated data set contains 98 CT scans, including 63 normal scans and 35 abdominal ones. Sixty scans were from SpineWeb data set[1] (43 normal and 17 abnormal) and the rest of the scans were from local data set (20 normal and 18 abnormal). The centroids of all visible vertebrae in both data sets were annotated by two experts as the ground truth. In a few scans, the whole spine is visible, while in most scans, the view is limited to 5∼21 vertebrae. The inter-axial distance varies from 0.5 to 3.5 mm. Considering the complexity of the data set, the spatial resolution is adjusted to 1 mm^3/voxel by bilinear interpolation and orientation is RAI (right anterior inferior) for all images in the pre-processing stage. During the experiment, we split the 98 CT scans into two non-overlapping sets with 49 scans each in which one of sets is obtain by randomly selecting 31 cases from 63 normal CT scans and 18 cases from 35 abnormal CT scans. Based on the above-mentioned way, each set is used once for both training and the remaining set is used for testing. Thus, we can report errors for all 98 scans and a total of 1078 vertebrae.

Experimental Setup and Parameter Setting

In the experiment, a two-fold cross-validation was designed, and the model parameters were separately set in each fold experiment. Considering the acceptable bias of SSAE network after partial parameter reuse, we enforced the same

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

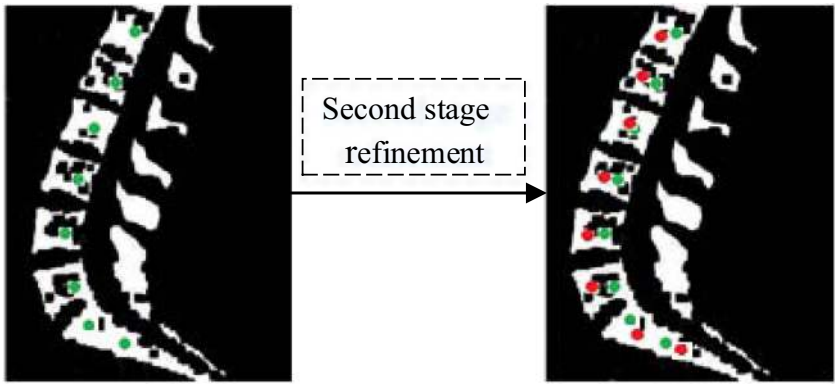

Fig. 6 Second stage of refining the vertebrae localization, where green dots represent mean-shift output and red dots represent the points based on Otsu refinement

settings of the layer number of network and node number of layer for each fold.

The parameters of the proposed algorithm exist in three modules, i.e., the SSAE network, the structured regression forest (SRF), and the two-stage refinement. Firstly, 49 training scans were selected to determine these parameters. In order to investigate the impact of different SSAE architectures and patch sizes on experimental results, four different patch sizes corresponding to different SSAE architectures were designed. In order to reduce the local context redundancy of the image patch in the experimental stage, we rearranged patch into 1D vector by sampling interval of 2. Through an exhaustive search strategy, the desired activation, regularization, and sparsity parameters were set to $\rho = 0.05$, $\lambda = 0.005$, and $\beta = 5.5$. The weights of SSAE are randomly initialized and the training optimization strategy is performed using gradient descent. The detailed parameter configuration and comparison results are summarized in Table 1.

It can be observed from Table 1 that the identification rates are gradually higher with the increase of patch size and the highest result is obtained when patch size is 32×32. In general, SSAE with deeper hidden layer tends to learn high-level feature representation, while more trainable parameters are practically less tunable. Considering the balance between computation cost and identification rate, the number of units in each layer of SSAE is set as 640, 320, and 200, respectively, and patch size is determined as 32×32.

The edge length D of the contextual features is set as 100 according to the input patch size. The SRF parameters are fixed throughout all experiments (forest training (θ_τ is the $0 \sim 1$ uniform distribution of interval 0.0025, m of the θ_m is 5, 40 tree, depth 25)). The bandwidth h of mean-shift is limited to 32. The number of training samples extracted was 53,900, the maximum number of samples for each CT scan was 1 250, and the number of samples for test image was 500.

Results and Discussion

We compared our results with closely related methods RF+HMM [15], J-CNN [17], by employing two evaluation metrics: identification rate and localization error.

Localization Error is defined as the distance (in mm) of each predicted vertebra location from its manual annotation. The results are summarized in Table 2. It is seen from Table 2 that the mean localization errors of our method is about 10.08 mm for all vertebrae. The highest errors are within the thoracic vertebrae region with a median of about 12.45 mm and the cervical region obtains the lowest errors of about 6.56 mm. The reason for the above result is that the visual appearance is more discriminative and is in strong image

Table 1 Comparison results with different input patch sizes and hidden layer depth of SSAE in this study

Patch size		Layer 1	Layer 2	Layer 3	Layer 4	Identification rate(%)
16×16	Input size	320	–	–	–	66.82
	Hidden size	100	–	–	–	
24×24	Input size	720	360	–	–	78.26
	Hidden size	360	150	–	–	
32×32	Input size	1280	640	320	–	82.11
	Hidden size	640	320	200	–	
40×40	Input size	2000	1200	600	320	80.97
	Hidden size	1200	600	320	200	

Table 2 Localization error and identification rate statistics of vertebrae in spine

Vertebrae		SRF localization (mm)			Refinement localization (mm)			Identification	
Region	Counts	Median	Mean	Std	Median	Mean	Std	Correct	Rate
All	1078	12.14	14.41	9.51	11.27	10.08	7.97	886	0.822
Cervical	171	8.74	11.21	9.45	6.56	8.54	7.45	148	0.866
Thoracic	514	14.92	17.52	9.87	12.45	11.69	9.28	392	0.763
Lumbar	349	11.23	13.32	10.21	10.33	9.74	7.33	307	0.880

contrast in the cervical vertebrae region, while partial miss and deformation of the thoracic vertebra. Compared with the thoracic region, the lumbar region with the most samples obtains a lower localization errors of about 10.33 mm.

Identification Error is defined as follows: if the localization error is less than 20 mm corresponding to the vertebra, we call the identification correct. The statistical results of the identification errors are reported in Table 2. The last two columns of Table 2 showed an overall success rate of 82.19%, the highest of 87.97% in the lumbar region and the lowest of 76.26% of the thoracic vertebrae region. This significant difference in the varying regions has the close relationship with the number of samples participating in the training data and the stability of regional structure representation.

Figure 7 showed some typical examples of localization and identification results. Compared with the traditional Box feature [15] and the coarse localization using histogram of oriented gradient (HOG) feature [17] before the refinement step, the proposed method can obtain the coarse location close to centroids of the corresponding vertebrae. Thus, it is likely to replace the complex hidden Markov [15] and the J-CNN [17] with the simple two-stage progressive refining strategy with the mean-shift kernel density estimation and the Otsu method.

Comparison with Related Methods In order to conduct a comprehensive assessment for the proposed method, we employ it to compare with two typical methods of vertebra localization and identification (RF+HMM [15] and J-CNN [17]) by the consistent localization errors and identification errors and results are shown in Figs. 8 and 9, where Fig. 8 is the identification rate for the three methods in individual vertebrae, while Fig. 9 reports the localization errors statistics on each type of vertebrae. In most cases, the proposed method achieved smaller mean errors than the other two methods, while mean errors are between

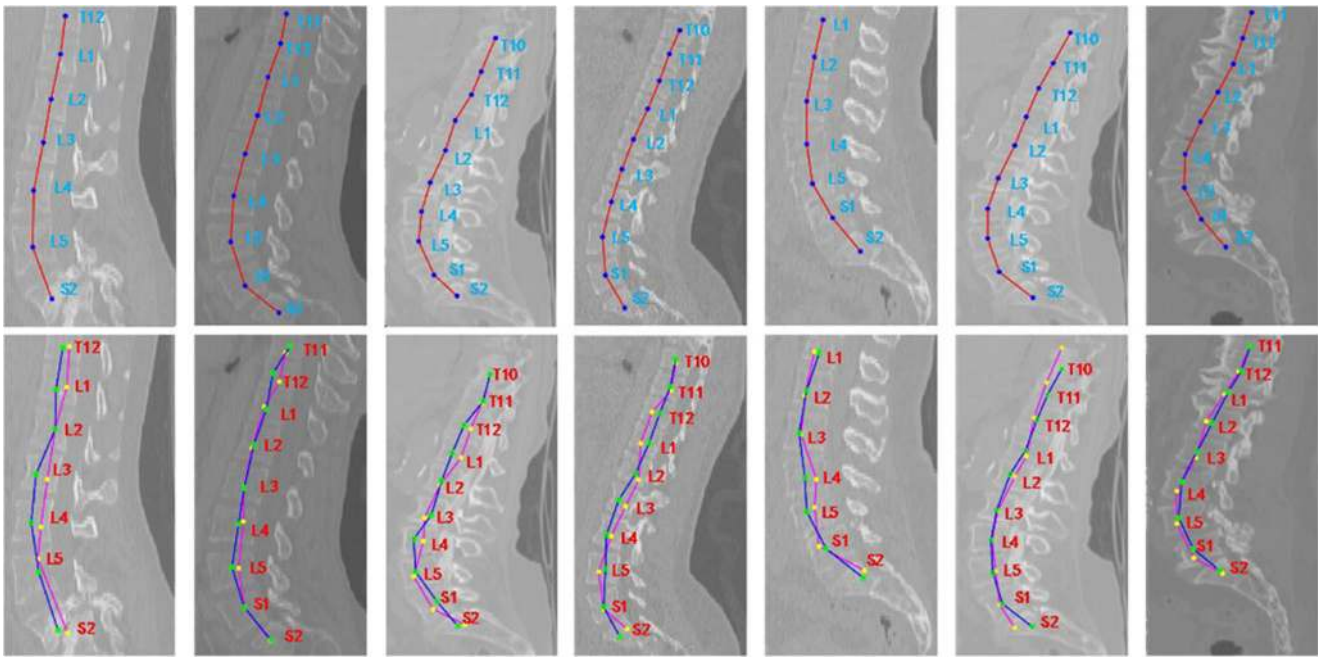

Fig. 7 Experimental results where the first row represents the annotated centroids of vertebrae. The purple lines with yellow dots are the prediction of regression forest and the blue lines with green dots are the output results after refinement step in the second row

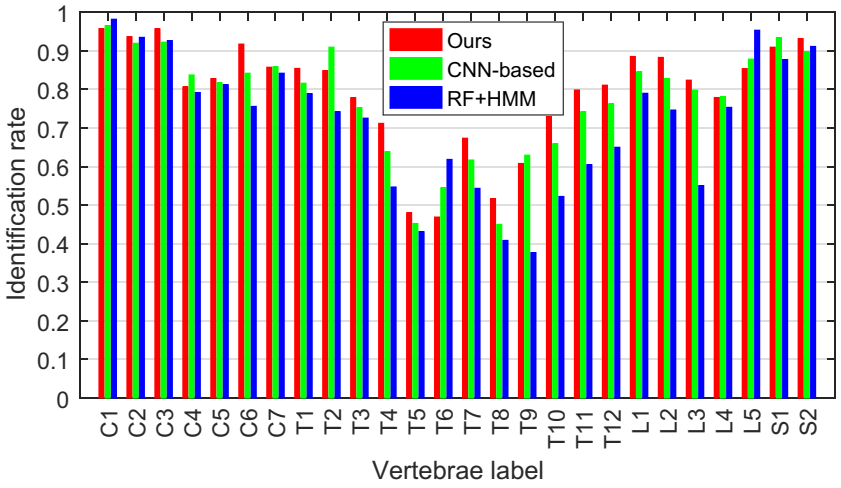

Fig. 8 Identification accuracies on individual vertebra of testing data

RF+HMM and J-CNN in a few cases. We further evaluated the mean localization errors for three methods before the refinement step, and evaluated results show that the proposed method is 14.41 ± 9.51 mm (mean ± std), RF+HMM is 19.05 ± 11.87 mm, and J-CNN is 17.11 ± 13.52 mm. This result demonstrates that the SSAE deep contextual feature has stronger representative ability than traditional Box and HOG features, and structured regression forest has superiority than standard regression forest. In addition, it is shown from Fig. 9 that the resulting accuracies of the three methods in the T3 to T11 section significantly are lower than those in other parts, which is mainly due to spine curvature and pathologies.

It should be noted that the localization accuracies in the thoracic vertebrae region have no obvious advantage in comparison to other regions.This is because of the large similarity of the vertebrae structures, spine curvature, and pathology reasons. However, the proposed method still obtained better results than other methods in this region. In the following work, we will explore more discriminating features with long-range contextual information to improve the localization accuracy of thoracic vertebrae part.

The proposed method has lower computational complexity than RF-HMM and J-CNN. Specifically, for RF-HMM, the three components are time-consuming. Firstly, the vertebra appearance model of HMM needs to compute mean and variance images for per vertebra. And, several iterations of nonlinear registration are performed to increase the sharpness of the mean images. Secondly, joint shape and appearance model of HMM and MAP inference need to be computed via dynamic programming for multiple sampled location candidates. In addition, HMM and RF model parameters also need to be optimized by using the same data set, that is to say, it also costs certain time. So HMM and RF will have a heavier computation cost than SSAE and SRF.

For J-CNN method, it includes three components, namely coarse vertebra candidate localization stage, J-CNN for vertebrae identification stage, and localization refinement with shape regression modelling stage. Firstly, coarse localization component can be divided into 3D

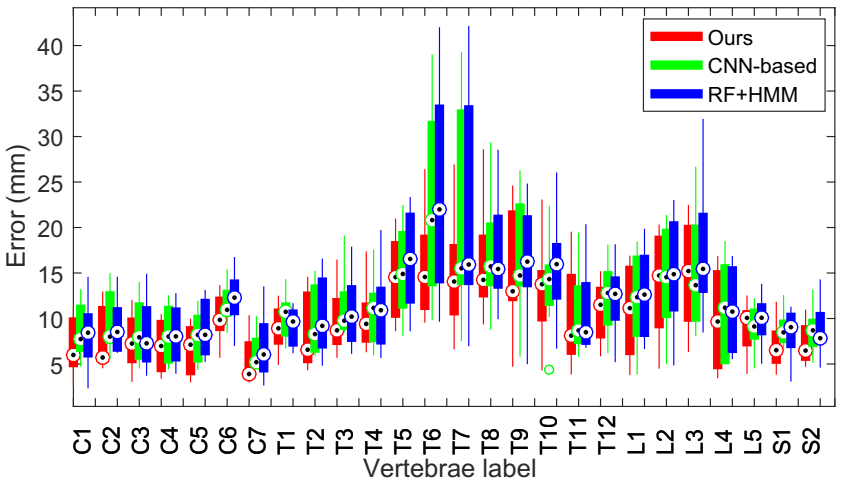

Fig. 9 Localization errors for each type of vertebra

HOG feature construction part and classification forest part, where the classification forest part is relatively time-consuming. Secondly, the vertebrae identification feature is mainly learnt by CNN, whose time cost is comparable to SSAE. Thirdly, the regression model in this method needs to fit a quadratic polynomial curve to refine vertebrae localization. So these three-stage complex models in J-CNN could accumulate remarkable computational complexities.

In contrast to RF-HMM and J-CNN, the proposed method mainly consists of three components (SSAE, SRF, and mean-shift), in which the time-consuming part mainly are SSAE and SRF, while the mean-shift component is simple and fast. In terms of quantization performance, the proposed algorithm runs less than 56 s in one test image while RF-HMM is about 110 s and 87 s to J-CNN. In the test stage, the proposed method employs the mean-shift to simplify the model complexity compared with HMM and CNN, thus speeding up the test time.

This algorithm was implemented mainly by MATLAB 2017R and the interference codes were programmed by Python 3.5 on an Intel i5 3.3Gz CPU, 16G DDR3 memory PC. The algorithm training time was about 299 min, of which the most time-consuming part is SSAE network training, which was about 184 min and the test stage took about 56 s.

Conclusion

This paper has presented a novel automatic approach by combining the deep SSAE contextual features and the structured regression forest (SRF) to achieve vertebrae localization and identification in CT scans. This algorithm does not make any priori assumption about the vision field of input images. The proposed approach utilized deep contextual feature representations learned from the SSAE network instead of low-level hand-crafted features and employed the structured regression forest (SRF) to consolidate structured information between image patches to improve the identification rate. Moreover, we also developed a two-stage progressive refining strategy with the mean-shift kernel density estimation and the Otsu method to further improve performance of vertebrae localization and identification. Experimental results demonstrated that the proposed method achieved better performance than RF+HMM and J-CNN.

In our future work, we will further investigate the influence of hidden neural units of SSAE to all vertebrae localization and identification in CT volumes. In addition, considering that the proposed method achieves better localization and identification on more conventional or slightly pathological data, however, maybe does not adapt well to high variability case. To solve this problem, further investigation will also be carried out w.r.t. highly pathological cases of spine such as high-grade scoliosis and kyphosis.

Acknowledgements The authors would like to sincerely thank the anonymous reviewers for their valuable comments, suggestions, and enlightenment.

Funding Information This research was partially supported by the Basic and Frontier Planning of CQ-CSTC (cstc2016jcyjA0317).

References

1. Huang S-H, Chu Y-H, Lai S-H, Novak CL: Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. IEEE Trans Med Imag 28(10):1595–1605, 2009
2. Ayed IB, Punithakumar K, Minhas R, Joshi R, Garvin GJ: Vertebral body segmentation in MRI via convex relaxation and distribution matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp 520–527
3. Lecron F, Boisvert J, Mahmoudi S, Labelle H, Benjelloun M: Fast 3D spine reconstruction of postoperative patients using a multilevel statistical model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp 446–453
4. Yao J, Burns JE, Munoz H, Summers RM: Detection of vertebral body fractures based on cortical shell unwrapping. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp 509–516
5. Oktay AB, Akgul YS: Localization of the lumbar discs using machine learning and exact probabilistic inference. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2011, pp 158–165
6. Schmidt S, Kappes J, Bergtholdt M, Pekar V, Dries S, Bystrov D, Schnörr C: Spine detection and labeling using a parts-based graphical model. In: Biennial International Conference on Information Processing in Medical Imaging. Springer, 2007, pp 122–133
7. Ma J, Lu L, Zhan Y, Zhou X, Salganicoff M, Krishnan A: Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2010, pp 19–27
8. Kelm BM, Zhou SK, Suehling M, Zheng Y, Wels M, Comaniciu D: Detection of 3D spinal geometry using iterated marginal space learning. In: International MICCAI Workshop on Medical Computer Vision. Springer, 2010, pp 96–105
9. Zhan Y, Maneesh D, Harder M, Zhou XS: Robust MR spine detection using hierarchical learning and local articulated model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp 141–148
10. Zhan Y, Jian B, Maneesh D, Zhou XS: Cross-modality vertebrae localization and labeling using learning-based approaches. In: Spinal Imaging and Image Analysis. Springer, 2015, pp 301–322
11. Forsberg D, Sjöblom E, Sunshine JL: Detection and labeling of vertebrae in MR images using deep learning with clinical annotations as training data. J Digit Imaging 30(4):1–7, 2017

12. Klinder T, Ostermann J, Ehm M, Franz A, Kneser R, Lorenz C: Automated model-based vertebra detection, identification, and segmentation in CT images. Med Image Anal 13(3):471–482, 2009
13. Rak M, Tonnies KD: A learning-free approach to whole spine vertebra localization in MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2016, pp 283–290
14. Glocker B, Feulner J, Criminisi A, Haynor DR, Konukoglu E: Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp 590–598
15. Glocker B, Zikic D, Konukoglu E, Haynor DR, Criminisi A: Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2013, pp 262–270
16. Suzani A, Seitel A, Liu Y, Fels S, Rohling RN, Abolmaesumi P: Fast automatic vertebrae detection and localization in pathological CT scans-a deep learning approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp 678–686
17. Chen H, Shen C, Qin J, Ni D, Shi L, Cheng JC, Heng P-A: Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp 515–522
18. Yang D, Xiong T, Xu D, Zhou SK, Xu Z, Chen M, Park J, Grbic S, Tran TD, Chin SP, et al: Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3DCTvolumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp 498–506
19. Liao H, Mesfin A, Luo J: Joint vertebrae identification and localization in spinal CT images by combining short-and longrange contextual information. IEEE Transactions on Medical Imaging
20. Kontschieder P, Bulo SR, Bischof H, Pelillo M: Structured class-labels in random forests for semantic image labelling. In: International Conference on Computer Vision. 2011, pp 2190–2197
21. Domingos JS, Stebbing RV, Leeson P, Noble JA (2014) Structured random forests for myocardium delineation in 3D echocardiography. Springer International Publishing
22. Zhu X, Jia X, Wong KYK: Structured forests for pixel-level hand detection and hand part labelling. Comput Vis Image Underst 141(C):95–107, 2015
23. Dollar P, Zitnick CL Structured forests for fast edge detection. In: IEEE International conference on computer vision, 2014, pp 1841–1848
24. Zhao G, Wang X, Niu Y, Liwen T, Shaoxiang Z: Segmenting brain tissues from chinese visible human dataset by deep-learned features with stacked autoencoder. Biomed Res Int 2016(6):1–12, 2016
25. Xu J, Xiang L, Liu Q, Gilmore H, Wu J, Tang J, Madabhushi A: Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Trans Med Imaging 35(1):119–130, 2016
26. Criminisi A, Robertson D, Konukoglu E, Shotton J, Pathak S, White S, Siddiqui K: Regression forests for efficient anatomy detection and localization in computed tomography scans. Medical Image Anal 17(8):1293–1303, 2013
27. Comaniciu D, Meer P: Mean-shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619, 2002