Multi-Modality Vertebra Recognition in Arbitrary Views using 3D Deformable Hierarchical Model

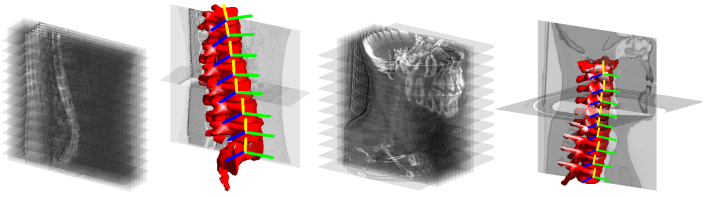
Yunliang Cai, *Member, IEEE*, Said Osman, Manas Sharma, Mark Landis, and Shuo Li*

***Abstract*—Computer-aided diagnosis of spine problems relies on the automatic identification of spine structures in images. The task of automatic vertebra recognition is to identify the global spine and local vertebra structural information such as spine shape, vertebra location and pose. Vertebra recognition is challenging due to the large appearance variations in different image modalities/views and the high geometric distortions in spine shape. Existing vertebra recognitions are usually simplified as vertebrae detections, which mainly focuses on the identification of vertebra locations and labels but cannot support further spine quantitative assessment.In this paper, we propose a vertebra recognition method using 3D deformable hierarchical model (DHM) to achieve cross-modality local vertebra location+pose identification with accurate vertebra labeling, and global 3D spine shape recovery. We recast vertebra recognition as deformable model matching, fitting the input spine images with the 3D DHM via deformations. The 3D model-matching mechanism provides a more comprehensive vertebra location+pose+label simultaneous identification than traditional vertebra location+label detection, and also provides an articulated 3D mesh model for the input spine section. Moreover, DHM can conduct versatile recognition on volume and multi-slice data, even on single slice. Experiments show our method can successfully extract vertebra locations, labels, and poses from multi-slice T1/T2 MR and volume CT, and can reconstruct 3D spine model on different image views such as lumbar, cervical, even whole spine. The resulting vertebra information and the recovered shape can be used for quantitative diagnosis of spine problems and can be easily digitalized and integrated in modern medical PACS systems.**

***Index Terms*—Spine recognition, vertebra detection, vertebra pose estimation, vertebra segmentation**

I. Introduction

AUtomatic spine recognition which supports quantitative measurement is essential in numerous spine related applications in orthopaedics, neurology, and oncology. The task of automatic spine recognition is to extract the set of numerical parameters that can uniquely determine the global structure of the spine and certain local structures of the vertebrae. Currently, spine recognition is often simplified as vertebra detection, which extracts the locations and labels of the vertebrae in input images. Instead, we consider spine recognition as a more general problem: to identify the set of geometric parameters that precisely determine the local and global spine structures. This parametrization obtained from spine recognition provides a unified geometry model that can be shared among spine structures from different modalities, different image views, and different formats. The particular parameters for a spine can be immediately used in any quantitative measurement systems for diagnosis purposes and can be efficiently stored/retrived by medical PACS systems. As shown in Fig. 1, our recognition will extract vertebra locations, poses, and shapes in multiple image modalities, as well as the complete global 3D shape.



(a) MR Lumbar (b) Recognition (c) CT Cervical (d) Recognition

Fig. 1. (a)(c) Input MR/CT slices from different views. (b)(d) Resulting 3D meshes after deformable matching of HDM.

Spine recognition is a challenging problem in spine image analysis. The main difficulties arise from the high variability of image appearance due to modalities differences or shape deformations: 1) Multiple modalities. The image resolution, contrast and appearance for the same spine structure could be very different when it is exposed to MR/CT, or T1/T2 weighted MR images. 2) High repetition. The appearances of vertebrae and intervertebral discs are highly repetitive that mismatching could happen easily. 3) Various poses. The vertebrae sizes and orientations are highly diverse in pathological data that regular detectors such as appearance detectors are insufficient to match all vertebrae. 4) Complex shape composition.

The recognition of the local vertebra structures and global spine shapes are often separately done by spine detection and spine shape matching.

Spine Detection Methods. The existing spine detection methods are often learning-based and focused on one specific image modality [1]–[5]. Long *et al.* [6] and Seifert *et al.* [7] proposed the early studies on vertebrae detection and identification in X-ray images using deformable contour model. Klinder *et al.* [2] used generalized Hough transform model for detecting vertebrae in CT, they extracted the 3D vertebra meshes from their detection-assisted segmentation but did not consider the geometric relations between vertebra meshes. Zhan *et al.* [3] utilized Adaboost method for learning detectors for vertebrae on MR. They proposed a hierarchical probability model of spine and applied it in inferring vertebra

Yunliang Cai is with the Department of Medical Biophysics, University of Western Ontario, London ON, Canada. (e-mail: ycai82@uwo.ca)

Said Osman is with St.Joseph's Health Care London (SJHC), London ON, Canada. (e-mail: sidosman@hotmail.com)

Manas Sharma is with Department of Medical Imaging, University of Western Ontario, London ON, Canada. (e-mail: msharm54@uwo.ca)

Mark Landis is with Victoria Hospital, London Health Sciences Center, London ON, Canada. (e-mail: mark.landis@lhsc.on.ca)

*Shuo Li is with the Digital Imaging Group of London, GE Healthcare, 268 Grosvenor St, London ON, Canada. (e-mail: shuo.li@ge.com)

(*) indicates corresponding author.

Manuscript received Jul 28, 2014; revised December 11, 2014.

locations and labels, but vertebra poses and spine shape are not involved. Roberts *et al.* [8] used random forest for detecting spine in DXA images and depended on active appearance models (AAM) for vertebra identification. Some methods considered detection on both MR and CT scans. [9] utilized the curved spinal geometric structure that can be extracted from both modality. Michael *et al.* [10] used boosting-trained Haar features for detecting vertebra discs and vertebra parts in MR and CT. Lootus *et al.* [11] used a classical SVM-trained Histogram of Oriented Gradients (HOG) features for detecting vertebrae in single MR/CT image. These methods relied on the classifiers that are separately trained on MR and CT. It is also worth noting that, most of supervised learning methods, as pointed out in [12], required dense manual annotations for the vertebrae in startup, i.e., manual annotations for the corners and the center of each vertebrae. For vertebrae/disc labeling, [2]–[5] had very successful labeling on fully or partially scanned image volumes. The local vertebrae labels relied on the identification of some special landmarks detected from multiple image views, i.e., template models for axial view vertebrae [2], annotation of spinal canals [5] or anchor vertebrae [3] in axial views. The complete vertebrae labels in the input images are inferred by a probability inference model, i.e., a graph model [13] [14], Hidden Markov Model (HMM) [4], or hierarchical model [15] [3]. For pose estimation, Pekar *et al*. [1] estimated the vertebra disc orientations using the orientations of detected local line structures. Kelm *et al.* [10] provided vertebra pose estimation in addition to vertebra detection using an iterative candidate searching technique call marginal space learning. These estimation methods exploited the multi-planar detectors to match the correct vertebrae poses.

Despite the successes of detection methods in vertebra locating and labeling, they are still limited in the following ways: 1) Vertebra detection often ignores vertebra shapes and poses, which cannot support further clinical spine diagnosis as shapes and poses are heavily used in quantitative measurement. 2) Learning-based detectors are limited by the image modality of training samples, excessive training data and intense manual annotations are required for handling additional modalities. Moreover, many methods are based on the detection of isotropic 3D image volume (i.e., CT or 3D MR), which cannot be directly applied on multi-slice data such as T1/T2 MR.

Spine Shape Matching. Most spine shape matching methods often employ deformable model based or part-based approaches for estimation/recovery of spine shapes. The model-base shape matching methods are originated from the Active Shape Models (ASM) [16] or Active Appearance Models (AAM) [17] with focus on contour matching. Liu *et al*. [18] modified the ASM model by combining the statistical shape information with the boundary orientedness property, and conducted the contour matching of cervical vertebrae. Markelj *et al*. [19] used gradient amplitudes for matching spine structures in 2D X-ray/fluoroscopic images and 3D-MR CT volumes. Kadoury *et al*. [20] proposed an articulated spine model inference for X-ray images. The inference reconstructed spine structure using Markov Random Field. Since spine is a compositional structure, part-based models such as pictorial models [21] are also studied in spine shape analysis. Lim *et al*. [22] incorporated a set of prior shapes under kernel density estimation, and combined the priors with level set to obtain vertebra segmentation. Zhang *et al*. [23] proposed a part+geometry model for groupwise registration, which can be used to accurately annotate spine images using nonrigid registration with a set of parameterized sample images. Rasoulian *et al*. [24] and Kadoury *et al* [25] represented the change of spine shape and poses as a statistical model, and applied this model in registration-assisted segmentation of CT images. Ibragimov *et al*. [26] utilized the transportation theory for matching the landmarks of vertebrae, and used the matched shape in spine segmentation.

Most existing AAM/ASM series methods are for segmentation purposes. They focus mainly on matching the articulated model boundaries to the image boundaries of the vertebra structures other than identifying the comprehensive quantitative vertebra information. Part-based registration methods focus on the appearance alignment on specific spine sections with reference images. However, this alignment is often restricted to local area and small distortions, it can hardly be applied to full scale matching on arbitrary spine sections.

Hierarchical Model. To handle the limitations of detecting/segmenting anatomical structures by local appearances, a number of hierarchical models were employed [27] [28] [29]. For segmentation, Zhan *et al* [3] found boundary appearance of the anatomical structure can be hierarchically modeled through an iterative global/local clustering. Ma *et al* [28] used the hierarchical coarse-to-fine mechanism to identify the precise edge locations of 3D CT thoracic images, providing accurate 3D segmentation. The HSMOR model proposed by Bagci *et al* [29] applied hierarchical model to general organs detection without using sophisticated optimization techniques.

We propose a comprehensive recognition method that provides simultaneous identification of local and global spine structures. The recognition extracts vertebra locations, labels, and poses from multiple modalities in arbitrary views, with reconstruction of 3D parameterized model for the input spine sections (i.e., lumbar, thoracic, cervical) even whole spine. Our method can work on both volume and multi-slice data, even single slice. This makes it adaptive for different clinical image protocols and significantly reduces the processing time. Our method is implemented by a novel anatomy-inspired Hierarchical Deformable Model (HDM) which simulates the global/local structures of spine to perform deformable matching of spine images.

The HDM follows the anatomic structure of spine, using multiple local compositional models to simulate the local rigid and global non-rigid deformation for the matching of local image structures and global spine structure respectively. Unlike most existing part-based model where local parts are not deformable, vertebra parts in HDM are more flexible. The dimensions and orientations of local vertebra parts can be adjusted to obtain the optimal shape matching. In addition, probability semantic relations are imposed between the local and global components in HDM to resolve the ambiguity of global shape. This provides a more general hierarchical representation for spine structure which has stronger adaptation to shape/appearance variations than existing methods.

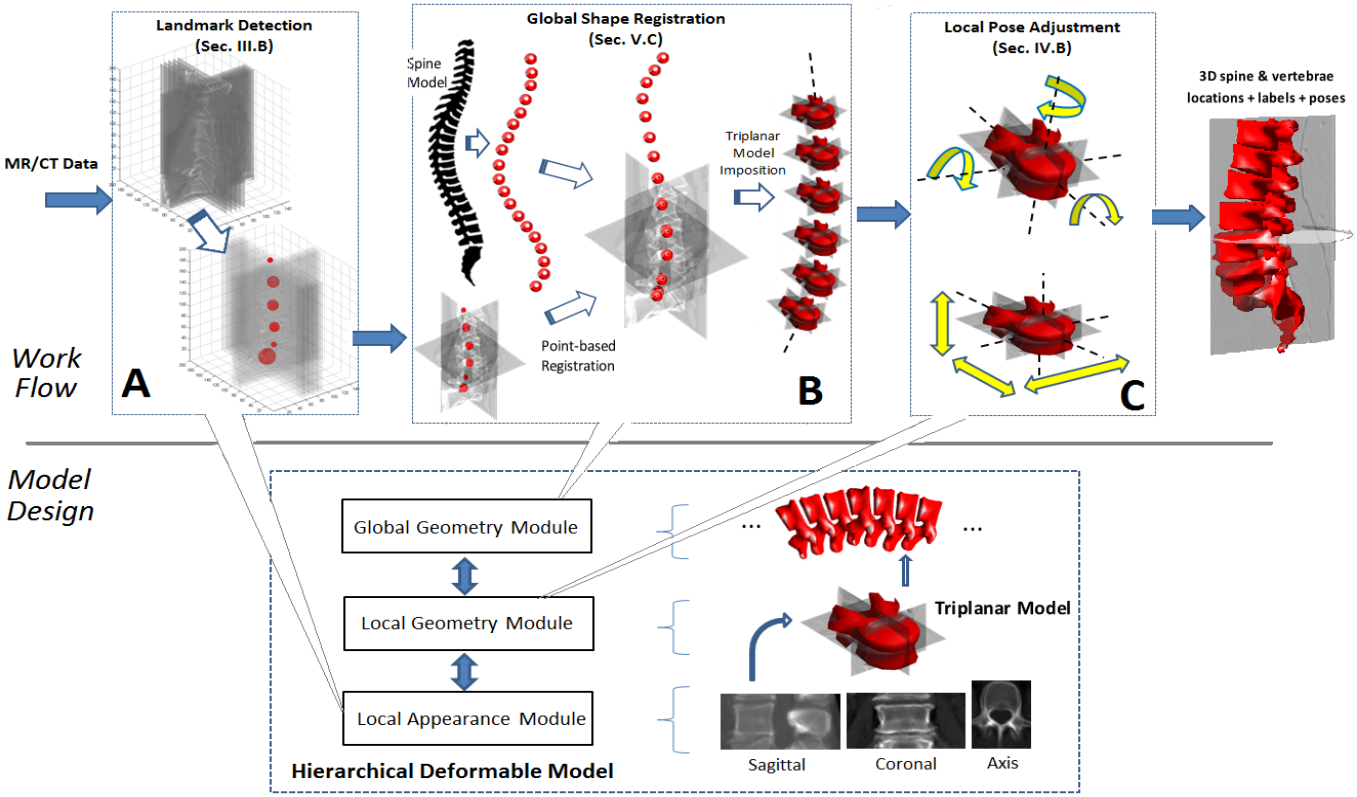


Fig. 2. The overview of the work flow.

II. Outline of the Approach

The overall approach of our method is a three stage recognition approach: landmark detection, global shape registration, and local pose adjustment, as shown in Fig. 2. These stages cover the matching from local to global spine structures. The three stages are individually implemented by the three modules in the hierarchical deformable model: the local appearance module, global geometry module, and local geometry module.

A. Workflow Overview

Our overall workflow can be understood as a three-stage top-down registration. The goal of our registration is 1) to align the global shape of the spine model with the identified vertebra landmarks, and 2) to align the vertebrae poses with the local image structures around the identified landmarks. The overall workflow is decomposed into three steps as shown in Fig. 2, with intermediate results of each step shown in Fig. 3. We describe the steps as follows:

- The landmark detection aims to identify the potential vertebrae locations, as shown in step A of Fig. 2. The detection is done by a feature matching between the input images and the vertebra templates. The feature we used is the cross-modality features introduced in Sec. III. The detection examples of MR and CT are shown in Fig. 3b.
- The global shape registration aims to match the detected landmarks with the global spine model. As presented in step B of Fig. 2, for a given spine model, each vertebra is abstracted as a 3D point so that the registration is done by a point-based alignment which minimizes the landmark-vertebra point distances. The registration is formulated as a coherence point drift process (see Sec. V.C). The aligned spine model provides a coarse identification of vertebrae locations and labels. The 3D vertebra models are then imposed on the model points, as also shown in Fig. 2 step B and the examples of Fig. 3c.
- The local pose adjustment aims to recover the perfect orientation and dimension of each vertebra, as shown in step C of Fig. 2. The adjustment seeks the optimal alignment between the vertebrae and corresponding image structures, which is done by a groupwise registration discussed in Sec. IV.B. The refined vertebrae are combined to construct a full 3D model for the spine, as illustrated in the output of Fig. 2 and the results in Fig. 3d.

The spine model contains intrinsic labels for each vertebra, thus the registration of spine model and vertebrae landmarks immediately provides the vertebra detection and labeling.

B. Basic Models

Triplanar Model. Triplanar model is designed for the joint representation of 3D shape and 2D appearance of a vertebra. Each triplanar vertebra model contains a 3D mesh model embedded with a triplanar template representation (see model design of Fig. 2). The three planar templates in this model are three MR or CT patches describing the MR/CT appearance of a 3D vertebra projected on coronal, sagittal, and axial view respectively. The detection of a vertebra landmark in an image becomes the search of identical matching between the planar templates and the input image. In addition, the 2D warping of a planar template on the associated image plane can be considered as applying a 3D deformation for the 3D vertebra model then projecting the appearance on that image plane.

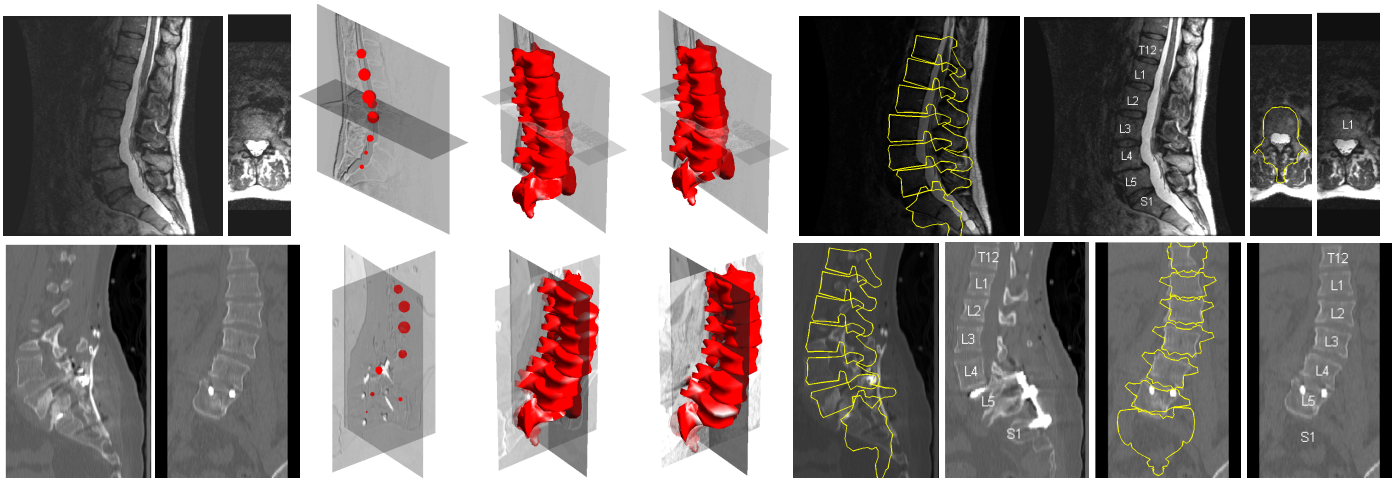


Fig. 3. Intermediate results of different algorithm stages. (a) Examples of original input scans, (b) identified landmarks, (c) results after global shape registration, (d) results after local pose adjustment, (e) sagittal/coronal(axial) silhouettes with vertebrae labeling.

This special 2D/3D appearance projection property will be exploited in pose adjustment in Sec. IV.

3D Spine Model. The spine model used in global shape registration (Fig. 2 step B) is built upon a 3D spine mesh model. The meshes are manually built from CT scans of a healthy spine using the Mimics software [1]. Particularly, each vertebra in 3D spine mesh is individually built and manually assigned to construct the corresponding MR or CT triplanar vertebra model. The meshes assigned and the corresponding triplanar templates are aligned using 3D CAD software. The resulting meshes can be applied to both MR and CT matching as the obtained anatomic structure is independent to image modalities. The 3D spine model and the associated triplanar vertebrae are assigned with different deformation mechanism and are unified in HDM.

Hierarchical Deformable Model. The Hierarchical Deformable Model (HDM) is proposed for the deformable matching of local/global spine structures. We combine the triplanar vertebra models under the organization of a 3D spine, constructing a compositional structure that simulates the spine anatomical deformation mechanism for spine matching. Our model contains three major modules (lower part of Fig. 2):

- The local appearance module contains the planar templates for each vertebra in each view. The templates are described by a cross-modality features which unifies the appearances of both MR and CT. This module conducts vertebra landmark detection using the planar templates.
- The local geometry module contains the triplanar vertebra models composed by the planar templates from the local appearance module. This module conducts pose adjustment for each vertebra via the warping of planar templates. The warping is guided by the groupwise registration (joint-alignment) of the planar templates.
- The global geometry module contains the set of connected triplanar vertebra models organized under the spatial layout of a spine. This module conducts shape registration between detected landmarks and the built-in vertebra models. The registration is implemented by a point-based registration which aligns the abstracted vertebra points with the landmark points.

The advantages of using this hierarchical model includes: 1) Weak supervision. HDM borrows geometric deformation in handling vertebra detection so that training sophisticated vertebra detectors is no longer necessary. This solves the shortcomings of most learning-based method, where a large number of samples and intense manual labeling should be provided in the training stages. 2) Versatile data adaptation. The triplanar representation can perform spine recognition on both specified MR/CT slices and whole image volume. User not only can obtain the recognition result significantly faster, but also can control the progressive spine shape approximation by simply feeding the model with more/less slices. 3) 3D reconstruction as a natural by-product. The spine detection and matching directly correspond to the deformations of a standard 3D spine model, which in turns provides a model-based 3D reconstruction of the spine image.

III. Local Appearance Module for Vertebra Landmark Detection

Local appearance module is for extracting cross-modality features that robustly encode the vertebra appearances in different image modalities. The cross-modality features are obtained by fusing the image features from MR and CT using a multi-model deep network. Using the fused features, the vertebra landmark detection can be performed on both MR and CT data with improved accuracy.

A. Multi-Modal Feature Extraction using Deep Networks

We apply a deep network model for learning and extracting multi-modal image features. Cross-modality features are more reliable than single modality one in vertebra detection. This is because some image structure in one modality will be nearly missing in another modality, i.e., vertebra discs features in MR scans will be missing in CT. Single modality features from either MR or CT will be insufficient to describe the complete vertebra structure. Also, features from different modalities tend to compensate and enhance each other. For example, disc

[1] http://biomedical.materialise.com

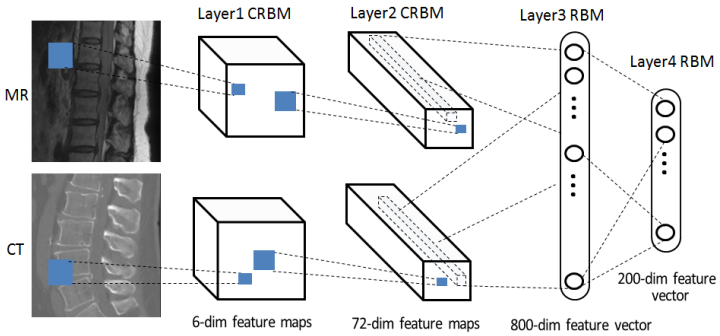


Fig. 4. The multi-modal deep network. Layer 1 and 2 are convolution restricted Boltzmann machine (CRBM) layers for adaptive feature extraction; layer 3 and 4 are RBM layers for feature fusion.

features from MR can help to identify the vertebra locations for CT scans as discs and vertebrae share a lot common image structures (i.e., vertebrae boundaries). This implies that good vertebra features can be learned by fusing image features from all presenting modalities.

Network Design. Our multi-modal deep network design is analogous to those in [30] and [31]. The architecture of the network is presented in Fig. 4. It is slightly different from [30] and [31] that our network is built upon two layers of convolution restricted Boltzmann machine (CRBM) [32] and two layers of restricted Boltzmann machine (RBM) [33]. Through the layer-wise iterative update of convolution filter-banks [32], the CRBMs can adaptively extract significant 2D image features. For each image modality, we deploy a unique set of CRBMs for extracting translation-invariant features from it. Similar to CRBM, through a layer-wise update of their connecting weights, the RBMs can provide a neater representation of the input signals. Unlike single-modal learning, we train the upper layer RBMs by feeding them the extracted lower-layer MR and CT features together. The RBMs will automatically mix the MR and CT features, generating a unified and neater representation for both MR and CT modality. The complete learning process is conducted in an unsupervised fashion.

Feature Fusion. The feature fusion is a unique property of deep network models. The purpose of feature fusion is to combine the common features shared among different modalities and to enhance the feature representation for capturing more image details. Traditional learning-based detection methods often rely on the direct classification of handcrafted features such as SIFT, HOG, or Haar functions. Because of the distinct appearances of original image modalities, the corresponding feature maps of the handcrafted features can be huge diverse. The supervised classifiers will require a large amount of labeled data to resolve this diversity. In contrast to traditional methods, deep network contains multi-layer abstraction and adaptive feature tuning which can automatically combine the higher layers feature representation of different modalities. These combinations also enhance the single modality features extracted from lower layers as the combined features can now represent extra image structures from another modalities. In other words, we can learn a better MR feature by using the CT samples and vice versa. Therefore, our multi-modal deep network requires less training samples than the traditional approaches but still can obtain good discriminative features.

Configurations. The parameters of the deep network are as follows. The layer 1 CRBM consists of a 7x7x6 filter-bank and the layer 2 CRBM consists of a 9x9x12 filter-bank. Each filtering is followed with a 1/2 sub-sampling and probability pooling [32]. The resulting output of layer 2 is a 72-dimensional binary feature maps with each map becomes 1/4 of the original image size. The layer 3 will map each 6x6x72 cube in the layer 2 to a feature vector of 800 dimensions. The last layer 4 will reduce the dimension of the layer 3 vector to 200 dimension. The final deep feature maps is a 200-dimensional maps where each map is of 1/4 the original image size. In other words, every 24x24 image patch is encoded to a 200 descriptor vector through the network. The CRBMs and RBMs are learned through a training set which contain 50 T1 MR slices, 50 T2 MR slices, and 100 CT slices. All slices are sampled near the mid plane in the image volume with the same size of 250x250, as pixel size is 1x1 mm. The MR features and CT features are separatively learned then fused by the RBM layers.

B. Planar Templates Construction

The planar appearance template used in landmark detection are the MR/CT template patches in the triplanar vertebra model. An planar template is generated by taking the mean of a set of training patches from different image slices, then map the mean patches to high dimensional feature map using the deep network. For example, the sagittal template of the lumbar vertebrae in CT is generated by taking the mean appearance of the set of lumbar image patches in the CT training set then covert it to feature maps. Note that the training image patches are aligned beforehand by using the part-based registration model discussed in the next section. In practice, we consider the four different types of template: the lumbar (L), thoracic (T), cervical (C), and sacrum (S). Each template type is generated separately by the same approach. There are in total $4 \times 2 \times 3$ templates that serve for the 4 vertebra types (S,L,T,C), 2 image modalities (MR and CT), and 3 views (coronal, sagittal, and axial). Once the input image modality is identified from image header, the associated templates will be loaded in the triplanar vertebra model to perform the appearance matching.

Templates v.s. Trained Classifiers. The main reason for choosing the template approach is that the invariant appearance of a vertebra pattern is highly unique to other local image appearances and most of the mismatching are caused by pose distortions. It is more efficient to discriminate vertebrae from local image structures using the invariant vertebra appearance along with the pose deformations. Supervised learning-based detection methods like [3] and [4] focus on vertebra appearance only and resolve the pose variations by training their classifiers with extra samples, especially large amount of pose varied samples. The use of planar template with unsupervisedly learned features and pose deformations not only reduces the workload of collecting large amount of medical data with massive manual labeling, but also significantly simplifies the training process while still remains the performance of vertebrae detection.

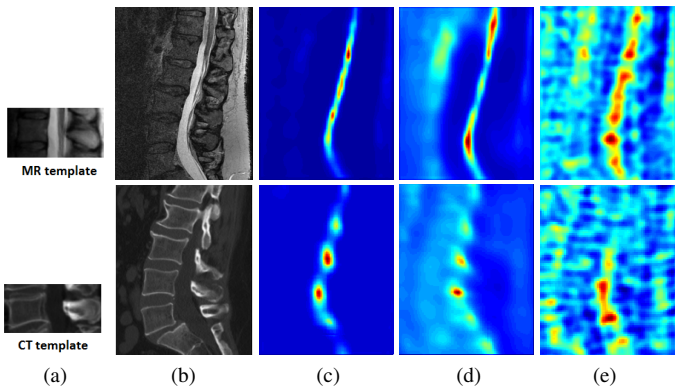


Fig. 5. Comparison of different features in template matching. (a) Planar MR/CT templates, (b) input MR/CT images, (c) the response of multi-modal deep feature defined by Fig. 4, (d) the response of single-modal (MR or CT) deep feature descriptor, and (e) the response of HOG descriptor. The deep features have more sharper and cleaner responses than HOG descriptors.

C. Landmark Detection

The initial landmark detection is done by matching the vertebra appearance templates with the input image. The trained deep network is applied in this detection. According to the input image modality, we substitute the layer 1 and 2 in the deep network with the related MR/CT CRBMs, and with the RBM layers remained unchanged. Following the filtering and pooling process of the deep network, the input image slice is expanded to a set high dimensional feature maps.

Detection. Using the deep-feature descriptor mentioned above, the template matching is done by comparing the L^2 distances of the feature vectors between input image and the appearance template. Suppose f_{dp} is the deep feature descriptor defined by the deep network Fig. 4, for a template $\mathbf{T}$ the matching response on point $\mathbf{p}$ in an input image I is

$$r(\mathbf{p}, \mathbf{T}) = \exp(-\epsilon||f_{dp}(\mathbf{I_p}) - f_{dp}(\mathbf{T})||^2) \quad (1)$$

where $\mathbf{I_p}$ is the image patch centered at $\mathbf{p}$ with the same size of $\mathbf{T}$ and ϵ is a fixed constant.

A comparison of single modality feature, cross-modality feature, and handcrafted feature (HOG descriptor [34]) is shown in Fig. 5. The single modality feature is trained on out deep network without feature fusion (with layer 3 and 4 removed in Fig. 4). The cross-modality feature obtain the best result with sharper response peaks and less noises.

In practice, the template $\mathbf{T}$ is deformed with a set of transforms $\mathcal{G}$ to match different vertebrae with various poses. $\mathcal{G}$ can be a series of 2D rotations and rescalings. The overall appearance matching is presented in the algorithmic form Algorithm 1. After the basic matching for each input scan $\mathbf{I}$, we can synthesize the responses for their corresponding image views and obtain the final 3D responses as shown in Fig. 6. The peak positions are the desired vertebra landmarks. Note if the input is multi-slice data, the obtained peak positions will be attached on the slice planes. Finer positioning can be obtained by using a pose adjustment for each landmarks as discussed in Sec. IV.

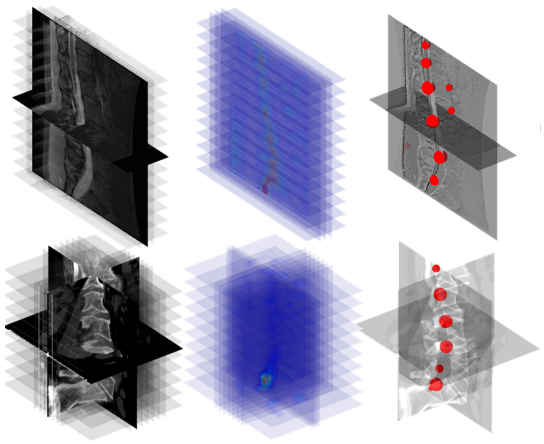


Fig. 6. Example of appearance template matching. Left to right: the input slice samples, the collection of matching responses, and the final detection results. The spheres represent the peaks of the synthesized responses whereas the response magnitudes are illustrated as the sphere sizes.

Input: image scan $\mathbf{I}$, template $\mathbf{T}$, transform set $\mathcal{G}$
Output: matching response r
$r(\mathbf{p}) \leftarrow 0$ for each $\mathbf{p}$;
foreach $G \in \mathcal{G}$ **do**
  $\mathbf{T}' \leftarrow \mathbf{T} \circ G$;
  $r(\mathbf{p}) \leftarrow \max\{r(\mathbf{p}), e^{-\epsilon||f_{dp}(\mathbf{I_p}) - f_{dp}(\mathbf{T}')||^2}\}$ for all $\mathbf{p}$;
end

Algorithm 1: Appearance template matching

IV. Local Geometry Module for Vertebra Pose Adjustment

The goal of local geometry module is to estimate the 3D pose of each triplanar model for the best model-image alignment. We define the pose of a triplanar vertebra model as the orientation and the anisotropic scales of that model in 3D space. The 3D pose can be described by the projections of 3D vertebra on 2D planes, which is equivalent to the planar poses of the intrinsic planar templates from the triplanar model. The optimal 3D poses of the collected triplanar models are obtained by a groupwise registration for the planar templates, and then by the back-projecting the planar poses to 3D.

A. Planar Pose Representation

The planar pose of a triplanar vertebra model is the 2D orientation and scales of its three built-in planar templates. Due to the rigidity of vertebra model, we can define the planar poses as invertible 2D affine transforms. The invertibility of pose transforms implies that any arbitrary planar poses can be generated from one reference pose. This generative assumption can be exploited in estimating the planar poses of each triplanar vertebra model.

Generative Model. For a given image view, i.e., sagittal view, we assume that the vertebra patterns appeared on the input slices are generated by the same appearance template as illustrated in Fig. 7. The reference template is from the sagittal template patch of a triplanar model. According to the repetition of vertebra patterns, we assume that each spine

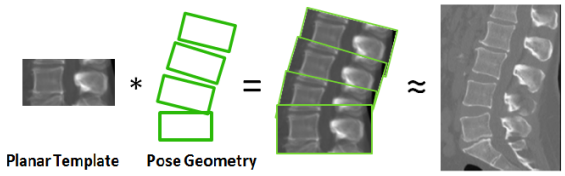


Fig. 7. Illustration of the generative model of vertebra. The spine appearance is generated by the replicas of a vertebra template with a set of pose geometries.

section (sacrum, lumbar, thoracic, and cervical) has its own reference appearance. Different vertebrae in the same section differ only by their poses. As shown in the example of Fig. 7, a template of sagittal CT lumbar generates the repetitive lumbar vertebra patterns in sagittal CT slice with varied poses. Let $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2}$ be the planar template, as d_1, d_2 be the height and width of the template patch. The potential vertebra landmarks p on an arbitrary image slice I satisfies

$$\mathbf{T} \circ G_{\mathbf{p}} = \mathbf{I}_{\mathbf{p}} \tag{2}$$

where $\mathbf{I}_{\mathbf{p}}$ is the local vertebra patch with arbitrary pose (orientation+scale), and $G_{\mathbf{p}}$ is a geometric transform that warps $\mathbf{T}$ to $\mathbf{I}_{\mathbf{p}}$ so that $\mathbf{T}$ 'generates' $\mathbf{I}_{\mathbf{p}}$ via $G_{\mathbf{p}}$. Stitching a set of deformed template replicas together can reconstruct the spine appearance, as shown in the last part of Fig. 7.

Pose Geometry. We define the 2D transforms on planar poses by an algebraic model. According to the generative model, the planar pose of a vertebra at $\mathbf{p}$ is uniquely determined by transform $G_{\mathbf{p}}$. The $G_{\mathbf{p}}$ is defined as an affine transform which can be explicitly formulated as a 3x3 invertible transform matrix. All possible transform matrixes form a Lie group $\mathrm{Aff}(2) \subset \mathrm{GL}(3)$. Using the matrix group form, $G_{\mathbf{p}} \in \mathrm{Aff}(2)$ is formulated as

$$G_{\mathbf{p}} = \mathrm{Exp}(\sum_{k=1}^{6} a_{\mathbf{p}}^{k} E_k) \tag{3}$$

$$\begin{aligned} E_1 &= \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right), \quad E_2 = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right), \quad E_3 = \left(\begin{smallmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right), \\ E_4 &= \left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right), \quad E_5 = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right), \quad E_6 = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right), \end{aligned} \tag{4}$$

where $\mathbf{a}_{\mathbf{p}} = [a_{\mathbf{p}}^1, \ldots, a_{\mathbf{p}}^6]$ and the infinitesimal generators $\{E_k\}$ form the Lie algebra $\mathfrak{aff}(2)$ of $\mathrm{Aff}(2)$. The deformations on planar templates on different views will at last determine a 3D pose of the 3D vertebra model (as later shown in Fig. 10). This allows us to estimate the 3D pose through 2D methods as 2D models can be directly applied on single/multiple slice data or volume data.

The Lie group formulation is adopted by our model because the pose changes between adjacent vertebrae are smooth. Also, the planar deformations are considered to be smooth in local vertebra set and during the deformation process. This formulation agrees with the kinematics model for robotics [35], where the continuous deformations on rigid structure are represented by smooth matrix groups. The pose of a planar template is parameterized by the 6 parameters in vector $\mathbf{a}_{\mathbf{p}}$. The pose of the associated 3D vertebra model are thus controlled by 18 parameters, which correspond to the three degree of freedom on sagittal, axial, and coronal view respectively. We will explicitly define the conversion between the parameterized planar deformation and the 3D rotation in Sec IV-C.

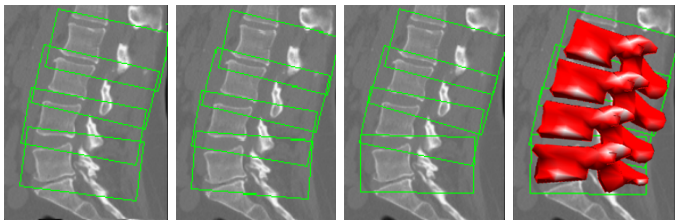


Fig. 8. Example of part-based groupwise registration on single plane. Left to right: the 1st, 5th, and 10th iteration in registration; the corresponding deformed 3D vertebra templates.

B. Planar Pose Adjustment by Groupwise Registration

The recovery of template pose is done by aligning the planar template to the identified landmark in image, as shown in Fig. 8. This can be understood as solving equation (2) for landmark patch $\mathbf{I}_{\mathbf{p}}$ when $G_{\mathbf{p}}$ is unknown. In addition, because of the repetitions of vertebra appearances, the alignment is not only applied on the landmark-template pairs but also the between landmark-landmark pairs. The landmark-landmark alignment help to enhance the regularity of landmark patches and reduce the ambiguity of the landmark identification. We apply a part-based groupwise registration for regularizing the landmark detection using the vertebra repetition. Our registration model is borrowed from [36] which is a variant of the congealing model [37]. Unlike the congealing model, our model is part-based that can be applied on repetitive image patches. This agrees with our assumption that spine is a structure that contains multiple repetitive parts.

Mathematical Formulation. Given a set of sampled scans from the same image view, we have initial identified landmark set $\mathcal{I}$ such that $\mathbf{I}_{\mathbf{p}}$ is a deformable landmark patch observed on $\mathbf{p} \in \mathcal{I}$. $\mathbf{p}$ is represented as the planar coordinates of the image view. Suppose $\mathbf{T}$ is the corresponding template for the identified landmarks, then the alignment is formulated as the minimization of the functional

$$u(G_{\mathbf{p}}) = \phi_{\mathbf{T}}(G_{\mathbf{p}}) + \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \psi(G_{\mathbf{p}}, G_{\mathbf{q}}) \tag{5}$$

where ϕ is a data term and ψ is a smoothness term for neighbor set $\mathcal{N}$. Suppose we let $\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1}$ and $\mathbf{T}$ be vectors in $\mathbb{R}^{d_1 \times d_2}$, ϕ and ψ are then defined as

$$\phi_{\mathbf{T}}(G_{\mathbf{p}}) = ||\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1} - \mathbf{T}||^2, \tag{6}$$

$$\psi(G_{\mathbf{p}}, G_{\mathbf{q}}) = ||\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1} - \mathbf{I}_{\mathbf{q}} \circ G_{\mathbf{q}}^{-1}||^2. \tag{7}$$

The expression of $\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1}$ means the deformed landmark patch $\mathbf{I}_{\mathbf{p}}$ is warped by transform $G_{\mathbf{p}}^{-1}$ to exactly the same size of $\mathbf{T}$. In other words, $\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1}$ and $\mathbf{I}_{\mathbf{q}} \circ G_{\mathbf{q}}^{-1}$ are of the same size such that direct subtraction is possible. To make the registration more robust, we encode the patch by the deep feature descriptor presented in Fig. 4. Each landmark patches become a feature patch written as $f_{dp}(\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1})$. Note that $G_{\mathbf{p}}$ is parameterized by $\mathbf{a}_{\mathbf{p}}$, then we can write

$$\mathbf{F}_{\mathbf{p}}(\mathbf{a}_{\mathbf{p}}) = f_{dp}(\mathbf{I}_{\mathbf{p}} \circ G_{\mathbf{p}}^{-1}).$$

Therefore, for a $\lambda_u > 0$, (5) becomes a functional for variable $\mathbf{a}_{\mathbf{p}}$ as

$$u(\mathbf{a}_{\mathbf{p}}) = \phi_{\mathbf{T}}(\mathbf{a}_{\mathbf{p}}) + \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \psi(\mathbf{a}_{\mathbf{p}}, \mathbf{a}_{\mathbf{q}}) \tag{8}$$

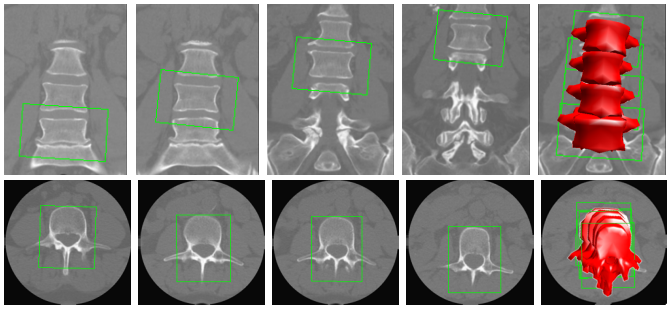


Fig. 9. Examples of part-based groupwise registration across different planes. The results of 10th iteration are shown.

$$\phi_{\mathbf{T}}(\mathbf{a_p}) = ||\mathbf{F_p}(\mathbf{a_p}) - f_{dp}(\mathbf{T})||^2 \tag{9}$$

$$\psi(\mathbf{a_p}, \mathbf{a_q}) = ||\mathbf{F_p}(\mathbf{a_p}) - \mathbf{F_q}(\mathbf{a_q})||^2. \tag{10}$$

The functional u can be solved by Gauss-Newton method, with the alignment being taken over each $\mathbf{p}$ in sampled image scans. The pose change $\triangle\mathbf{a_p}$ is obtained via

$$\triangle\mathbf{a_p} = (J^{\mathrm{T}}J)^{-1}J^{\mathrm{T}}d(\mathbf{a}) \tag{11}$$

$$d(\mathbf{a}) = \mathbf{F_p}(\mathbf{a_p}) - f_{dp}(\mathbf{T}) + \sum_{(\mathbf{q},\mathbf{p})\in\mathcal{N}} (\mathbf{F_p}(\mathbf{a_p}) - \mathbf{F_q}) \tag{12}$$

$$J(\mathbf{a_p}) = \left(\frac{\partial d(\mathbf{a_p})}{\partial \mathbf{a_p^1}}, \ldots, \frac{\partial d(\mathbf{a_p})}{\partial \mathbf{a_p^k}}\right)^{\mathrm{T}} \tag{13}$$

The coefficient vector $\mathbf{a_p}$ is iteratively updated: $\mathbf{a_p} + \triangle\mathbf{a_p}$, leading to the progressive alignment of the vertebra parts. The updated $\mathbf{a_p}$ is substituted back to (3), warping the planar temple and the 3D vertebra. Fig. 8 shows the progressive alignment of landmark patches. Note that the part-based registration can be applied on landmark patches within a single image scan or those distributed among multiple scans. This allows us to conduct the registration on arbitrary image view as some repetitions does not appear in single image. We show the multiple scans examples on coronal view and axial view respectively in Fig. 9. The overall groupwise registration is summarized in Algorithm 2.

Input: $\{\mathbf{I_p}\}$, initial poses $\{\mathbf{a_p}\}$, $\mathbf{T}$, t_{max}
Output: aligned poses $\{\mathbf{a_p^*}\}$
$k \leftarrow 0$, $\mathbf{a_p^{(0)}} \leftarrow \mathbf{a_p}$;
while $t < t_{max}$ **do**
  foreach $\mathbf{I_p}$ **do**
    Compute $\triangle\mathbf{a_p}$ using (11), (12), and (13)) with $\mathbf{T}$;
    $\mathbf{a_p^{(t+1)}} \leftarrow \mathbf{a_p^{(t)}} + \triangle\mathbf{a_p}$;
  end
  $t \leftarrow t + 1$;
end
$\mathbf{a_p^*} \leftarrow \mathbf{a_p^{(t)}}$;

Algorithm 2: Local pose adjustment

C. Estimation of 3D Pose

Through the groupwise registration on different image views, the planar pose of each potential landmark patches are described by the optimal vector $\mathbf{a_p}$ obtained by (8). The next task is to back-project the planar poses from three views (coronal, sagittal, axial) to synthesize a 3D pose so that the dimensions and orientation of a triplanar vertebra model are fully recovered. The estimation problem is similar to that in visual servoing [38], where the 3D transform is recovered by projection 2D planar deformations.

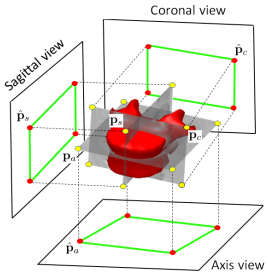


Fig. 10. 3D pose estimation by the planar poses. The planar bounding boxes represent the orientation and sizes of the planar templates. Warping the planar bounding boxes via groupwise registration will change the 3D pose accordingly.

From Aff(2) to SE(3). We assume that the 3D vertebra model is a rigid structure whose pose is defined by a invertible 3D rotation and a translation. This assumption is different from that of the planar template assumption as the planar template is allowed to both rotate and dilate under affine transforms. For a potential landmark which is identified in three image views, we still let $\mathbf{p}$ denote the landmark center for the sake of simplicity. The possible 3D transforms form a Lie group $\mathrm{SE}(3)$ such that $R_{\mathbf{p}} \in \mathrm{SE}(3)$ is a 4x4 transform with a matrix group representation similar to (3):

$$R_{\mathbf{p}} = \mathrm{Exp}(\sum_{k=1}^{6} c_{\mathbf{p}}^k V_k). \tag{14}$$

$$V_1 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, V_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, V_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$
$$V_4 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, V_5 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, V_6 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{15}$$

For a triplanar vertebra model, the three template patches inside the model is rotated and translated according to the pose of the 3D vertebra. In other words, the warped planar landmark obtained from groupwise registration is in fact the projection of the corresponding template patch in the vertebra model. It can be observed from Fig. 10 that the corners of the built-in three template patches define a 3D bounding box for the 3D vertebra model. The template-landmark projection indicates that, the projections of the 3D corners also define the warped bounding boxes of the planar landmarks. Utilizing the 2D-3D projection between corners, we can recover the 3D pose of the vertebra model.

Estimation Formulation. Let $\mathbf{p}_s$ be the 3D corner of the sagittal template patch in the triplanar vertebra model, we consider the orthographic projection P_s such that $P_s\mathbf{p}_s$ is the 2D corner of the warped landmark bounding box on the projected sagittal image plane. Similarly, we can define the projection of sagittal corners, axial corners, and coronal

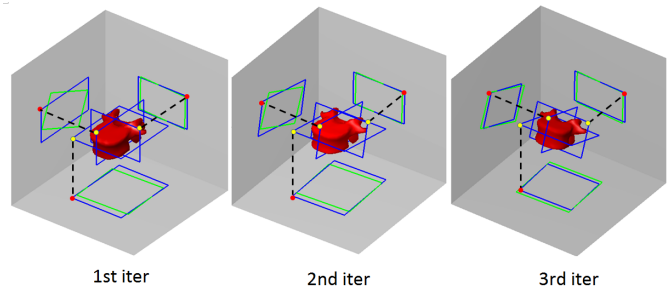


Fig. 11. Example of 3D pose estimation through 2D planar poses. **Green**: the contours of the bounding boxes obtained by groupwise registration. **Blue**: the contours of the built-in template patches and their 2D projections. Left to right show the three iterations of the poses obtained by solving (19).

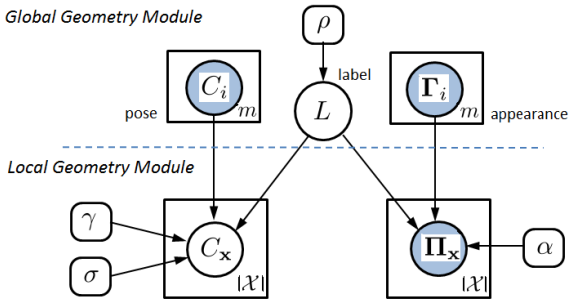


Fig. 12. The graphical probability model representation of the global and local geometry module. The global module acts as priors, generating local module instances. The nodes in dark represent the observed variables.

corners respectively as $\hat{\mathbf{p}}_s$, $\hat{\mathbf{p}}_a$, $\hat{\mathbf{p}}_c$

$$\begin{aligned}\hat{\mathbf{p}}_s &= P_s\mathbf{p}_s\\ \hat{\mathbf{p}}_a &= P_a\mathbf{p}_a\\ \hat{\mathbf{p}}_c &= P_c\mathbf{p}_c.\end{aligned} \tag{16}$$

Note that any 3D corner $\mathbf{p}$ and 2D corner $\hat{\mathbf{p}}$ in above equations are represented in homogeneous coordinates, and projection P is a 4×4 matrix which agrees with our matrix definition in (14). Assume that the initial positions of the 3D corners are $\mathbf{p}_{s0}$, $\mathbf{p}_{a0}$, and $\mathbf{p}_{c0}$ respectively, the relations of the template-landmark projection is now described as

$$\begin{aligned}e(R) = \sum_{i=1}^{4} &||P_sR\mathbf{p}_{s0}^i - \hat{\mathbf{p}}_s^i||^2 + ||P_aR\mathbf{p}_{a0} - \hat{\mathbf{p}}_a^i||^2\\ &+ ||P_cR\mathbf{p}_{c0} - \hat{\mathbf{p}}_c^i||^2\end{aligned} \tag{17}$$

where $\mathbf{p}^1,\ldots,\mathbf{p}^4$ represent the four corners of the template patch. R^* is the desired 3D pose such that

$$R^* = \arg\min_{R\in \mathrm{SE}(3)}\{e(R)\}. \tag{18}$$

Similar to (8), functional (17) is represented in Lie algebra form to enhance the smoothness of transformation. By substituting (14) in (17) we have

$$\begin{aligned}e(\mathbf{c}) = \sum_{i=1}^{4} &||P_s(\mathrm{Exp}(\sum_{k=1}^{6} c^kV_k))\mathbf{p}_{s0}^i - \hat{\mathbf{p}}_s^i||^2\\ &+||P_a(\mathrm{Exp}(\sum_{k=1}^{6} c^kV_k))\mathbf{p}_{a0} - \hat{\mathbf{p}}_a^i||^2\\ &+||P_c(\mathrm{Exp}(\sum_{k=1}^{6} c^kV_k))\mathbf{p}_{c0} - \hat{\mathbf{p}}_c^i||^2.\end{aligned} \tag{19}$$

The optimal $\mathbf{c}^* = \arg\min\{e(\mathbf{c})\}$ is obtained by solving (19) in Gauss-Newton method. The resulting pose updates via the iterative updates of $\mathbf{c}$ are shown in an example in Fig. 11.

V. Global Geometry Module for Spine Shape Registration

Global geometry module performs the global shape registration for the detected vertebra landmarks. This module is built upon the 3D spine model and its shape registration. The 3D spine model is formed by the combined local triplanar vertebrae models and is abstracted as a point-connected global model. Shape registration of the spine model provides a top-down point matching for the potential landmarks. The global module is described by three parts: the unification of triplanar models that forms the global spine, the vertebra parsing that removes outlier landmarks, and the non-rigid deformation of global spine model that conducts the shape registration.

A. Unification of Local Models

Similar to the generative model of local triplanar model, the global spine model can be described as higher order generative model as shown in Fig. 12. We apply a graphical probability model to formulate this higher order generative model. Our formulation is inspired by the generative segmentation model in [39].

Similar to the compositional anatomic structure of spine, the 3D spine model (without point abstraction) in a HDM is composed by the set of 3D triplanar vertebrae models. The ith vertebra in the spine model is the local triplanar geometry model whose pose is C_i and has triplanar appearance $\mathbf{\Gamma}_i$:

$$C_i = (\mathbf{z}_i, \mathbf{c}_i),\ \mathbf{\Gamma}_i = \{\mathbf{T}_i^v\}$$

where $\mathbf{z}_i$ is the vertebra center location in $\mathbb{R}^3$, $\mathbf{c}_i$ represents the 3D pose as in (19), and set $\{\mathbf{T}_i^v\}$ contains the triplanar templates for $v\in\{\mathrm{sagittal}, \mathrm{axial}, \mathrm{coronal}\}$. The whole spine global model which contains m vertebrae can be represented as vertebra set $\{C_1,\ldots,C_m\}$. Suppose through a landmark detection, we observe a set of potential landmarks $\mathcal{X}\subset\mathbb{R}^3$. A detected landmark at $\mathbf{x}\in\mathcal{X}$ whose configuration is described by its pose and appearance

$$C_\mathbf{x} = (\mathbf{x}, \mathbf{c}_\mathbf{x}),\ \mathbf{\Pi}_\mathbf{x} = \{\mathbf{I}_\mathbf{x}^v\}$$

where $\{\mathbf{I}_\mathbf{x}^v\}$ are the deformed planar patches around $\mathbf{x}$ from different image views v.

Suppose each detected landmark $\mathbf{x}$ can only correspond to one triplanar model with one pose. We can assume in the model-landmark matching that, an observed local landmark and the planar image patches around the landmark are considered as being generated from one of the triplanar vertebrae models in the global spine model. We can construct a probability density function about landmark $\mathbf{x}$ following the generative assumption:

$$p(\mathbf{x}) = p(\mathbf{\Pi}_\mathbf{x}, C_\mathbf{x}; \{\mathbf{\Gamma}_i, C_i\}) \tag{20}$$

where $\{C_i\}_{i=1,\ldots,m}$ is the set of vertebrae in the global model. The above formulation implies that the density function will increase when local landmark at $\mathbf{x} \in \mathcal{X}$ is highly correlated to the global model. The correlation can be understood as a deformable matching when deformable model is driven by the force of likelihood. According to our assumption, The driving force of $\mathbf{x}$ is tuned specifically for a single triplanar models in the global model. To explicitly define (20), we first revise it as prior-likelihood form

$$p(\mathbf{\Pi}_\mathbf{x}, C_\mathbf{x}; \{\mathbf{\Gamma}_i, C_i\}) = p(\mathbf{\Pi}_\mathbf{x}, C_\mathbf{x} | \{\mathbf{\Gamma}_i, C_i\}) p(\{\mathbf{\Gamma}_i, C_i\}), \tag{21}$$

where $p(\{\mathbf{\Gamma}_i, C_i\})$ represents the prior initialization of the global model. We then introduce a latent variable (see also Fig. 12): $L : \mathcal{X} \to \{1, \ldots, m\}$ to represent the membership index of vertebra models $\{\mathbf{\Gamma}_i, C_i\})$ from a the global spine, such that the landmark on $\mathbf{x}$ is considered as generated from the $L(\mathbf{x})$-th vertebra model. In addition, L satisfies

$$\begin{aligned} L(\mathbf{z}_{i+1}) &= L(\mathbf{z}) + 1, \quad i = 1, \ldots, m-1, \\ L(\mathbf{z}_{i-1}) &= L(\mathbf{z}) - 1, \quad i = 2, \ldots, m, \end{aligned} \tag{22}$$

where $\{\mathbf{z}_i\}$ belongs to the global model $\{C_1, \ldots, C_m\}$. Given a initialized global model, by our independence assumption the planar appearances and pose of the local landmark at $\mathbf{x}$ are generated by a mixture model

$$\begin{aligned} &p(\mathbf{\Pi}_\mathbf{x}, C_\mathbf{x} | \{\mathbf{\Gamma}_i, C_i\}) = \\ &\qquad \sum_{L=1}^{m} p(L) p(\mathbf{\Pi}_\mathbf{x} | L, \{\mathbf{\Gamma}_i, C_i\}) p(C_\mathbf{x} | L, \{C_i\}), \end{aligned} \tag{23}$$

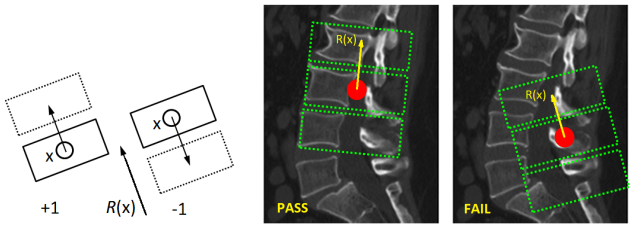
where $p(L)$ is the latent variable describing the association of current landmark and the specific vertebra model.

In the above formulation (23), $p(L)$ can be simply defined as uniform distribution $1/m$ and the deformable matching will reduce to blind registration. Instead of uniform distribution, we define $p(L)$ as confidence of local spatial compatibility where the explicit definition is presented in Sec V-B. The confidence is adaptive according to vertebra type, so that some special landmarks can have stronger attraction to the corresponding vertebra models. We first define the landmark-model correspondence of planar appearances and the correspondence of poses respectively as follows

$$\begin{aligned} p(\mathbf{\Pi}_\mathbf{x} | L, \{\mathbf{\Gamma}_i^v, C_i\}) &= \prod_v p(\mathbf{I}_\mathbf{x}^v | L, \{\mathbf{T}_i^v, \mathbf{a}_i^v\}) \\ &= \prod_v p(\mathbf{I}_\mathbf{x}^v | \mathbf{T}_L^v, \mathbf{a}_L^v) = \frac{1}{Z_\mathbf{\Pi}} \prod_v \exp\left(-\alpha \phi_{\mathbf{T}_L^v}(\mathbf{a}_L^v)\right) \end{aligned} \tag{24}$$

and

$$\begin{aligned} p(C_\mathbf{x} | L, \{C_i\}) &= p(\mathbf{x} | L, \{\mathbf{z}_i, \mathbf{c}_i\}) p(\mathbf{c}_\mathbf{x} | L, \{\mathbf{c}_i\}) \\ &= p(\mathbf{x} | L, \{\mathbf{z}_i\}) \prod_v p(\mathbf{a}_\mathbf{x}^v | L, \{\mathbf{a}_i^v\}) \\ &= p(\mathbf{x} | \mathbf{z}_L) \prod_v \prod_{j \in \mathcal{N}(i) \cup i} p(\mathbf{a}_\mathbf{x}^v | \mathbf{a}_j^v) \\ &= p(\mathbf{x} | \mathbf{z}_L) \times \frac{1}{Z_\mathbf{a}} \prod_v \exp\left(-\gamma \sum_{j \in \mathcal{N}(i) \cup i} \psi(\mathbf{a}_\mathbf{x}^v, \mathbf{a}_j^v)\right) \end{aligned} \tag{25}$$



(a) (b)

Fig. 13. Example of vertebra parsing. (a) Example of $\mathcal{N}_k(\mathbf{x})$ when $k = 1$ and -1 respectively. (b) The parsing verification of a correctly detected landmark (Red), and falsely detected one that fails in the parsing.

where ϕ and ψ are defined respectively in (8). We thus integrate the local geometry alignment (8) into our global model of (21), whereas the local alignment can now contribute to the global deformable matching. The precise definition of $p(\mathbf{x}|\mathbf{z}_L)$ is presented in Sec. V-C.

Since the global spine model is represented as the set of connected triplanar vertebra models, the deformation of the global model corresponds to the collaborated deformations of triplanar models which is highly complicated. The graphical probability model can not only unify the appearance-pose representations among local models, but can also unify the local and global deformation formulations in one neater formulation. The interactions between global spine and local vertebra can be easily accomplished using this formulation.

B. Vertebra Parsing

Vertebra parsing is the verification for removing falsely detected landmarks using the spatial inter-vertebra correspondences. It reduces the ambiguity of spine shape registration. Our parsing model is inspired by the probability parsing of [40] with modifications of our inter-vertebra verifications.

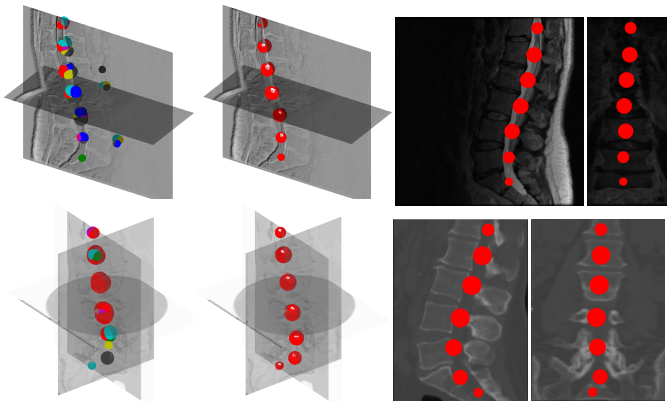
There are two spatial clues that help to verify the correct landmarks: 1) Pose compatibility of adjacent landmark. As spine is a part-connected structure, the correctly matched of a landmark and a vertebra model suggests that the spatial organization of nearby landmarks should also be compatible with the organization of the adjacent vertebra models in the global spine model. We can easily remove the outliers by its spatial compatibility. 2) Anchor vertebrae [3] [5]. The anchor vertebrae are the vertebrae whose appearances are significantly different from the other vertebrae in the spine. These special vertebrae are used to identify the spine section (lumbar or cervical) of the input data. The anchor vertebrae we use are: (a) the S1 for identifying lumbar related sections; (b) the C1+C2 (combined as one model) for identifying cervical related sections.

The overall vertebra parsing is carried out as:

- Identify spine section by anchor vertebrae;
- Remove outliers that are not compatible with local spatial organization.

We formulate the vertebra parsing in the form of graphical model (20). Let $p(L)$ represent the classification of landmark $\mathbf{x} \in \mathcal{X}$ as vertebra $L(\mathbf{x}) \in \{1, \ldots, m\}$. L is considered as a Markov random field (MRF) over $\mathcal{X}$

$$p(L) = \prod_{x \in \mathcal{X}} p(L(\mathbf{x}) | \{L(\mathbf{y})\}) \tag{26}$$



(a) (b) (c)

Fig. 14. Example of vertebra parsing. (a) the detected landmarks are obtained from Algorithm 1 (different colors indicate different poses in $\mathcal{G}$); (b) the landmarks that pass parsing; (c) the correctly identified landmarks in sagittal and coronal views.

where $\{L(\mathbf{y})\}_{\mathbf{y}\in\mathcal{N}_k(\mathbf{x})}$ represents the labels assigned to a modified neighbor set $\mathcal{N}_k(\mathbf{x})$. Unlike traditional definition of neighbor set $\mathcal{N}(\mathbf{x})$, $\mathcal{N}_k(\mathbf{x})$ is defined as

$$\mathcal{N}_k(\mathbf{x}) = \mathcal{N}(\mathbf{x} + kR(\mathbf{x})\mathbf{d}), \quad k \in \{-K, \dots, K\} \tag{27}$$

where $k > 0$ is integer and $\mathbf{d} \in \mathbb{R}^2$ is a unit vector. $\mathcal{N}_k$ is the neighbors of $\mathbf{x}$ obtained by shifting $\mathbf{x}$ with displacement $\mathbf{d}$ oriented by $R(\mathbf{x})$, where $R(\mathbf{x})$ is the planar pose obtained from (18). A toy example of $\mathcal{N}_k$ is shown in Fig. 13. The potential landmark location is confirmed only if its $\mathcal{N}_{\pm 1}$ neighbors are in compatible locations.

The probability of $p(L(\mathbf{x})|\{L(\mathbf{y})\}$ is defined as

$$p(L(\mathbf{x})|\{L(\mathbf{y})\}) = \frac{1}{Z_L}\exp\left(\rho\sum_{k=-K}^{K}\sum_{\mathbf{y}\in\mathcal{N}_k(\mathbf{x})}\delta(L(\mathbf{x})+k, L(\mathbf{y}))\right) \tag{28}$$

where δ is the Kronecker delta function. The definition indicates that $\mathbf{x}$ can be labeled as the $L(\mathbf{x})$-th vertebra only if its shifted position $\mathbf{y}$ is a potential landmark labeled as $L(\mathbf{x})+k$. The parsing of $L(\mathbf{x})$ is then done by maximizing the log-likelihood of (28), which can be simply implemented by linear searching algorithm. A pair of MR and CT vertebra parsing examples are shown in Fig. 14. The outliers are eliminated after parsing, leaving only the correct vertebra landmarks.

C. Registration for Spine Shape

The shape registration of spine model and detected landmarks are implemented by point-based registration. In this registration, each vertebra models in the spine model are abstracted as points to match the identified landmarks obtained from vertebra parsing. The registration is driven by minimizing point pair inter-distances modulated with landmark alignment and appearance matching. The deformation applied in our registration is inspired by coherence point drift (CPD) [41].

As the registration is point-based, we need to define the explicit point correlation. Following the formulation of generative mixture model (23), we now provide the explicit definition of $p(\mathbf{x}|\mathbf{z}_L)$ in (25):

$$p(\mathbf{x}|\mathbf{z}_L) = \frac{1}{(2\pi\sigma^2)^{3/2}}\exp(-\frac{||\mathbf{x}-(\mathbf{z}_L+\mathbf{v}(\mathbf{z}_L))||^2}{2\sigma^2}) \tag{29}$$

where $\mathbf{v}(\mathbf{z}_L)$ is the displacement of the original vertebra location $\mathbf{z}_L$. The likelihood above represents the attraction between $\mathbf{x}$ and $\mathbf{z}_L$, which can be understood as a Gaussian Mixture Model (GMM). A variation for the position $\mathbf{z}_L+\mathbf{v}(\mathbf{z}_L)$ or for the standard deviation σ will cause the likelihood to change accordingly, which implies the move of $\mathbf{z}_L$ towards or away from $\mathbf{x}$. Our goal is then to search for the optimal $\mathbf{v}$ and σ. Consider the negative log likelihood objective function derived from (23)

$$\begin{aligned} Q(\mathbf{v},\sigma) &= -\log\prod_{\mathbf{x}} p(\mathbf{\Pi}_\mathbf{x}, C_\mathbf{x}|\{\mathbf{\Gamma}_i, C_i\}) \\ &= -\sum_{\mathbf{x}}\log\sum_{L} p(L)p(\mathbf{\Pi}_\mathbf{x}|L,\{\mathbf{\Gamma}_i,C_i\})p(C_\mathbf{x}|L,\{C_i\}) \\ &= -\sum_{\mathbf{x}}\log\sum_{L}\Theta(\mathbf{x},L)p(\mathbf{x}|\mathbf{z}_L) \end{aligned} \tag{30}$$

where we denote

$$\Theta(\mathbf{x},L) = p(L)p(\mathbf{\Pi}_\mathbf{x}|L,\{\mathbf{\Gamma}_i,C_i\})\prod_{v}\prod_{j\in\mathcal{N}(i)\cup i} p(\mathbf{a}_\mathbf{x}^v|\mathbf{a}_j^v) \tag{31}$$

and $p(L)$, $p(\mathbf{\Pi}_\mathbf{x}|L,\{\mathbf{\Gamma}_i,C_i\})$, $p(\mathbf{a}_\mathbf{x}^v|\mathbf{a}_j^v)$ are defined in (28), (24), and (25) respectively. Objective function (30) can be minimized using the *expectation-maximization* (EM) algorithm, which iteratively updates the parameters in (30) to solve $\mathbf{v}$ and σ. Using the EM algorithm formulation, from a series of mathematical manipulations [42], we can reformulate (30) as an equivalent minimization objective function

$$Q = -\sum_{\mathbf{x}}\sum_{L} p^{\text{old}}(\mathbf{z}_L|\mathbf{x})\log\left(\Theta^{\text{new}}(\mathbf{x},L)p^{\text{new}}(\mathbf{x}|\mathbf{z}_L)\right) \tag{32}$$

where $\Theta^{\text{new}}(\cdot,\cdot)$ and $p^{\text{new}}(\cdot|\cdot)$ represent the prior and likelihood in the M-step using the newly evaluated parameters. $p^{\text{old}}(\mathbf{z}_L|\mathbf{x})$ is the posterior of the E-step which is evaluated using the old parameters. By minimizing the above function (32) instead of (30) we can obtain the same optimal $\mathbf{v}$ and σ with simpler update steps. From (29) and (25), we will have (32) revised as

$$\begin{aligned} Q(\mathbf{v},\sigma) = \sum_{\mathbf{x}}\sum_{L} p^{\text{old}}(\mathbf{z}_L|\mathbf{x})\Big(&\frac{||\mathbf{x}-\mathbf{z}_L-\mathbf{v}(\mathbf{z}_L)||^2}{2\sigma^2} \\ &-\log\Theta(\mathbf{x},L)\Big) + \frac{3N_p}{2}\log\sigma^2 + \frac{\lambda}{2}\sum_{\mathbf{x}}||\mathbf{v}(\mathbf{z}_L)||_H^2 \end{aligned} \tag{33}$$

$$p^{\text{old}}(\mathbf{z}_L|\mathbf{x}) = \frac{\exp\left(\frac{-||\mathbf{x}-\mathbf{z}_L-\mathbf{v}^{\text{old}}(\mathbf{z}_L)||^2}{2(\sigma^{\text{old}})^2}\right)\Theta(\mathbf{x},L)}{\sum_{\mathbf{y}}\sum_{l}\exp\left(\frac{-||\mathbf{y}-\mathbf{z}_l-\mathbf{v}^{\text{old}}(\mathbf{z}_l)||^2}{2(\sigma^{\text{old}})^2}\right)\Theta(\mathbf{y},l)} \tag{34}$$

where $N_p = \sum_{\mathbf{x}}\sum_{L} p^{\text{old}}(\mathbf{z}_L|\mathbf{x})$, $\lambda > 0$. $||\cdot||_{\mathbb{H}}$ is the norm for a Hilbert space $\mathbb{H}$ which is defined as

$$||\mathbf{v}||_{\mathbb{H}}^2 = \sum_{k=0}^{\infty}\beta_k||D^k\mathbf{v}(\mathbf{x})||^2.$$

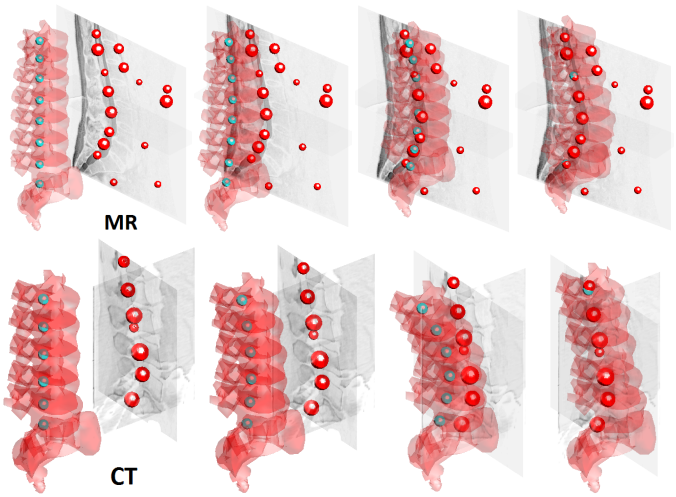


Fig. 15. Example of global shape registration for MR and CT. By using the correspondence matching (33), the registration can be conducted even in a noisy background. The results of 1st, 3rd 5th and 10th iteration are shown.

where $\beta_k = b^{2k}/(k!2^k)$ for a fixed constant $b > 0$, D is a derivative operator that satisfies $D^{2k}\mathbf{v} = \nabla^{2k}\mathbf{v}$ and $D^{2k+1}\mathbf{v} = \nabla(\nabla^{2k}\mathbf{v})$. ∇ is a gradient operator and ∇^2 represents a Laplacian operator over $\mathbb{R}^3$.

To obtain the optimal $\mathbf{v}$, we first fix σ in (33) then from the Euler-Lagrange equation (see [43]) we can have

$$\frac{1}{2\sigma^2\lambda}\sum_{\mathbf{x}}\sum_{L} p^{\text{old}}(\mathbf{z}_L|\mathbf{x})(\mathbf{x}-\mathbf{z}_L-\mathbf{v}(\mathbf{z}_L))\delta(\mathbf{y}-\mathbf{z}_L) = \sum_{k=0}^{\infty}(-1)^k\beta_k D^{2k}\mathbf{v}(\mathbf{y}). \tag{35}$$

The solution of $\mathbf{v}$ is derived from the Green's function of the differential operator on the right side

$$\mathbf{v}(\mathbf{y}) = \frac{1}{2\sigma^2\lambda}\sum_{\mathbf{x}}\sum_{L} p^{\text{old}}(\mathbf{z}_L|\mathbf{x})(\mathbf{x}-\mathbf{z}_L-\mathbf{v}(\mathbf{z}_L))K(\mathbf{y},\mathbf{z}_L) \tag{36}$$

where K is the Green's function. We adopt the definition of [41] and choose K as a Gaussian function $K(\mathbf{y},\mathbf{z}_L) = \exp(-\frac{1}{2b^2}||\mathbf{y}-\mathbf{z}_L||^2)$.

We then fix $\mathbf{v}$, and obtain σ as

$$\sigma^2 = \frac{1}{3N_p}\sum_{\mathbf{x}\in\mathcal{X}}\sum_{L}||\mathbf{x}-\mathbf{z}_L-\mathbf{v}(\mathbf{z}_L)||^2. \tag{37}$$

The updated $\mathbf{v}$ and σ will be substituted to (34), then continues to compute the new $\mathbf{v}$ and σ in next iteration.

An example of iterative global shape registration for MR and CT is shown in Fig. 15. By utilizing the alignment of local geometry model, the global registration is insensitive to erroneously detected landmarks. The overall registration of the global model is presented in the following Algorithm 3.

Input: detected landmarks: $\{(\mathbf{x},\mathbf{c}_\mathbf{x},\mathbf{\Pi}_\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$; global model: $\{(\mathbf{z}_i,\mathbf{c}_i,\mathbf{\Gamma}_i)\}_{i=1,\ldots,m}$; $t_{\max}$
Output: $\{\mathbf{z}_i^*\}_{i=1,\ldots,m}$
Initialize $\sigma^{(0)} \leftarrow \left(\frac{1}{3N_p}\sum_{\mathbf{x}\in\mathcal{X}}\sum_{i=1}^{m}||\mathbf{x}-\mathbf{z}_i||^2\right)^{1/2}$;
Initialize $K_{j,k} \leftarrow \exp(-\frac{1}{2b^2}||\mathbf{z}_j-\mathbf{z}_k||^2)$, $1 \le j,k \le m$;
Initialize $\mathbf{z}_i^{(0)} \leftarrow \mathbf{z}_i$, $t \leftarrow 0$;
while $t < t_{\max}$ **do**
  Compute $p^{\text{old}}(\mathbf{z}_i^{(t)}|\mathbf{x})$ in (34) with $\sigma = \sigma^{(t)}$;
  Compute $\mathbf{v}^{(t+1)}$ in (36) with $p^{\text{old}}(\mathbf{z}_i^{(t)}|\mathbf{x})$, $K_{j,k}$;
  Compute $\sigma^{(t+1)}$ in (37) with $\mathbf{v}^{(t+1)}$;
  $\mathbf{z}_i^{(t+1)} \leftarrow \mathbf{z}_i^{(t)} + \mathbf{v}^{(t+1)}(\mathbf{z}_i^{(t)})$;
  $t \leftarrow t+1$;
end
$\mathbf{z}_i^* \leftarrow \mathbf{z}_i^{(t)}$ for $1 \le i \le m$;

Algorithm 3: Global shape registration

VI. Experiments

We test our vertebra recognition in a variety of spine images that cover most popular image modalities and image views in clinical use. Our method successfully extracts local vertebra information (location+pose+label) and global spine information (3D spine) simultaneously under various input conditions. Our recognition results obtain high accuracy under three criterions include position error, angular error, and dimension error. The performance of our method is also proven by its high successful labeling rate and fast running time.

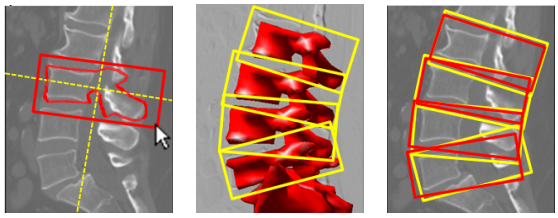
A. Data

Testing Data. We test our method in 55 MR and 85 CT samples. The samples are from three different datasets: (1) Dataset 1: the 30 MR + 30 CT spine samples from SpineWeb[2] which contains lumbar spine images from both healthy cases and patients with minor spondylosis/fracture. (2) Dataset 2: the 30 healthy and pathological CT samples for lumber and thoracic spine selected from the Annotated Spine CT Database [3] which contains different arbitrary CT views. (3) Dataset 3: the 25 MR + 25 CT samples we collected from Ontario area, Canada for evaluations in additional image views and protocols. The data collected covers from lumbar, thoracic, cervical, and whole spine. The modalities of the datasets include T1/T2 MR and CT. All images are resampled to an isotropic resolution of 1mm for vertebra recognition.

Training Data. We first collect 1200 MR+CT image patches from different spine sections/views to construct the initial HDM model. The deep network of local appearance module is trained using 600 MR (including 300 T1 and 300 T2) and 600 CT randomly sampled planar patches from 5 MR and 5 CT volumes in Dataset 3. The planar templates in HDM are then constructed by 30 lumbar patches, 30 thoracic patches from the same training volumes. The training volumes are not involved in subsequent evaluation to avoid the bias. The 3D spine model in HDM is manually built as discussed in Sec. II.B. The HDM contains 24 triplanar vertebra models: S1, L1∼L5, T1∼T12, C3∼C7, and C1+C2 (as one model).

Data Resampling. Our method can work on both volume and multi-slice data. Slice sampling from volume data can be applied to reduce the processing time. The performance of the recognition is insensitive to slice sampling as long as the

[2]http://spineweb.digitalimagegroup.ca
[3]http://research.microsoft.com/en-us/projects/spine/



(a) Pose annotation (b) Recognition (c) Comparison

Fig. 16. Example of accuracy evaluation for sagittal view (e.g., err^{sagittal}). (a) Pose ground truth manually annotated by bounding box (red); (b) the recognized bounding boxes (yellow) (c) evaluation by comparing positions, orientations, and sizes.

selected slices cover the target spine structure. Single slice processing is tested, as a simulation of the practical condition in spine diagnosis, for the MR/CT scans where the spines are with little spatial distortion. For a 3D volume data (i.e., CT volume), we sample slices located in ± 20 mm near the middle slice in sagittal view, and subsample axial view slice with step size 4mm. For a multiple-slice data (i.e., some T1/T2 MR scans), we directly use all the slices from the data.

B. Ground Truth and Evaluation Methods

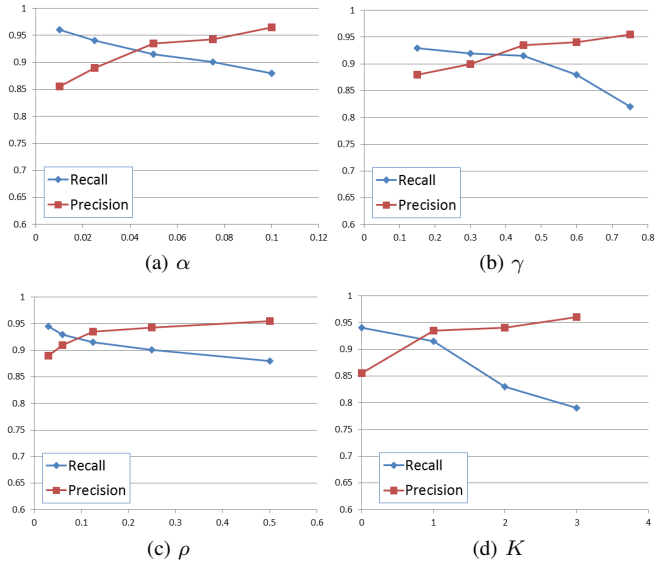
The identification of vertebrae label, locations, and poses are evaluated by the success labeling rate and the pose estimation accuracy respectively. The success labeling rate includes the rate of correct vertebra/non-vertebra landmark classification and the rate of correct labeling out of identified vertebra landmarks, both of which are tested with ground truth labels of landmarks. The pose accuracy is evaluated by comparing the output pose parameters with the ground truth values.

Ground Truth Annotation. The vertebrae labels and poses are manually annotated on each sample in the datasets. Each sample will be annotated in sagittal, axial, and coronal views respectively, using a planar bounding box overlaying on each vertebra. The vertebra label is then assigned to each planar box, obtaining the ground truth vertebra labels. An example for sagittal annotation is demonstrated in Fig. 16. The vertebra locations are set as the centers of the spinal cords, which correspond to the centers of the planar bounding boxes as illustrated in Fig. 16a. The poses are thus denoted as the orientations of the medial axes and the scales of the bounding boxes, for angular and dimension error assessment respectively.

Classification Evaluation. The correct rate of vertebra/non-vertebra classification is evaluated on all detected potential vertebra landmarks, including both the vertebra and non-vertebra ones. The evaluation of classification is presented in the form of precision/recall as discussed in Sec. VI.C (Fig. 17).

Labeling Evaluation. Each correctly identified vertebra landmark is assigned to a specific vertebra label (i.e., L2, T12,...) through the shape registration in Sec. V. The correct labeling rate is then evaluated by comparing the automatic labeling with ground truth annotated labels.

Pose Accuracy Evaluation. For pose estimation accuracy, we directly compare the resulting planar poses from pose adjustment (Sec. IV) with ground truth annotated poses. The planar poses are used for direct measurement with the annotated ground truths to avoid unnecessary computation error



(a) α (b) γ (c) ρ (d) K

Fig. 17. The precisions and recalls of vertebra/non-vertebra classification for parameter α, ρ, γ, and K on all 55 MR and 85 CT scans.

in 2D-3D pose conversion. The poses are evaluated in three perspectives: 1) position error err_{pos}: the differences between obtained vertebra centers and ground truths; 2) angular error err_{ang}: angle differences between the recognized orientations and ground truths; 3) dimension error err_{dim}: size differences between the recognized boxes and ground truths. For each vertebra, the above three types of errors are calculated from three image views:

$$err_{\text{pos}} = err_{\text{pos}}^{\text{sagittal}} + err_{\text{pos}}^{\text{coronal}} + err_{\text{pos}}^{\text{axial}}$$

$$err_{\text{ang}} = err_{\text{ang}}^{\text{sagittal}} + err_{\text{ang}}^{\text{coronal}} + err_{\text{ang}}^{\text{axial}}$$

$$err_{\text{dim}} = err_{\text{dim}}^{\text{sagittal}} + err_{\text{dim}}^{\text{coronal}} + err_{\text{dim}}^{\text{axial}}$$

An example of error measurement in sagittal view is demonstrated in Fig. 16. The result bounding boxes in Fig. 16b will compare with the ground truth boxes as annotated in Fig. 16a. The evaluations in axial and coronal view are done similarly.

C. Results

Environment. The recognition is performed in the Matlab environment on a 2.7 GHz dual core PC with GPU support. Under this computation platform, the HDM model with 24 triplanar vertebra models takes 0.5s average processing time for a 512×512 input slice. A multi-slice MR input with 15 sagittal and 55 axial scans will cause less than 30s. For slice-by-slice full processing of a $512 \times 512 \times 512$ CT volume the processing time will be around 10 min.

Vertebra/Non-vertebra Classification. The vertebra classification rates for of our recognition on all 55 MR and 85 CT samples in datasets are evaluated and shown in Fig. 17. The precision and recall curves are individually obtained in global shape registration (Sec. V) for global module parameters: α in (24), γ in (25), ρ and K in (28). The default choice of parameter values are: $\alpha = 0.05$, $\gamma = 0.45$, $\rho = 0.125$ and $K = 1$. The precision/recall for each parameter is obtained by varying the particular parameter and fixing the

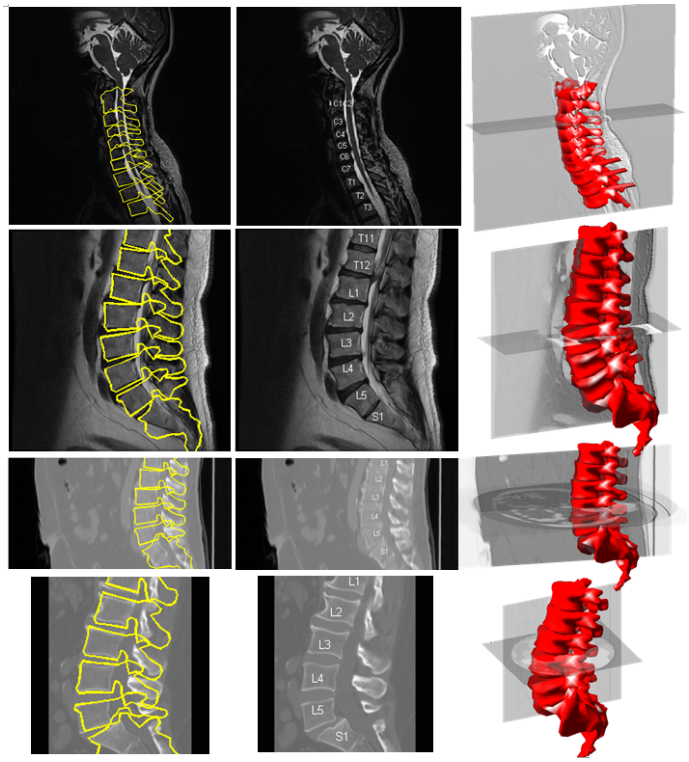


Fig. 18. Examples of MR/CT arbitrary view vertebra recognition in Dataset 1. Left to right: the recognized vertebrae location and poses (represented as 2D silhouettes); vertebrae labels; recovered 3D meshes.

	(%)	Cerv.	Thor.	Lumb.	Avg.
Dataset 1	MR	96.2	96.5	98.5	98.2
	CT	97	-	97.9	97.4
Dataset 2	CT	93.1	92	95.7	93.8
Dataset 3	MR	95.3	95.9	97.9	96.5
	CT	96.1	94.8	98.6	97.1

TABLE I
CORRECT LABELING RATE OF THE IDENTIFIED VERTEBRAE.

other parameters to default values. The labeling is obtained by the registration of spine model and detected landmarks, which is formulated in (33). Using default values, our recognition can obtain recall 91.5% and precision 93.5%. The increase of α, γ, and ρ will enhance the Θ term (31) in (30), making the 3D spine model in HDM collide in shape registration (i.e., >1 points merge to one position). The most significant precision-recall improvements occurs in the curves of K. This proves the effect of vertebra parsing in landmark outliers removal.

Vertebra Labeling. The correct labeling rate is calculated on all successfully identified vertebra landmarks. The labeling results are shown in Table I. Due to the robustness of cross-modality shape registration, the labeling rate is ranging consistently on both MR and CT data from 92% to 98.5%. For spine sections, the correct labeling rates on lumbar or cervical sections are often slightly higher than those in thoracic. This is because the anchor vertebrae C1+C2 and S1 are located in the cervical and lumbar section respectively which enhances the discrimination of the nearby repetitive vertebrae structures in those sections.

Pose Estimation We present the pose estimation results in Table II (Dataset 1), III (Dataset 2), and IV (Dataset 3):

- Fig. 18 and Table II show the example recognition results on Dataset 1. Our method obtain the best average position error 2.54mm in cervical MR images among the tested sections. The error slightly increases in Lumber sections because the larger templates sizes in lumbar parts will introduce larger displacements in registration. The average angular error of identified vertebrae consistently ranges from 2.33° to 3.39°, with maximum standard deviation 2.64° in CT cervical samples. The evaluation of thoracic section for CT samples is not possible due to the lack of thoracic views in the dataset. The mean dimension error in this dataset is around 8.41mm to 9.98mm. Large dimension error occurs in some T1 MR lumbar cases especially those where the S1 vertebrae are only partially captured. The limited coverage of S1 causes large dimension distortion in recognition.

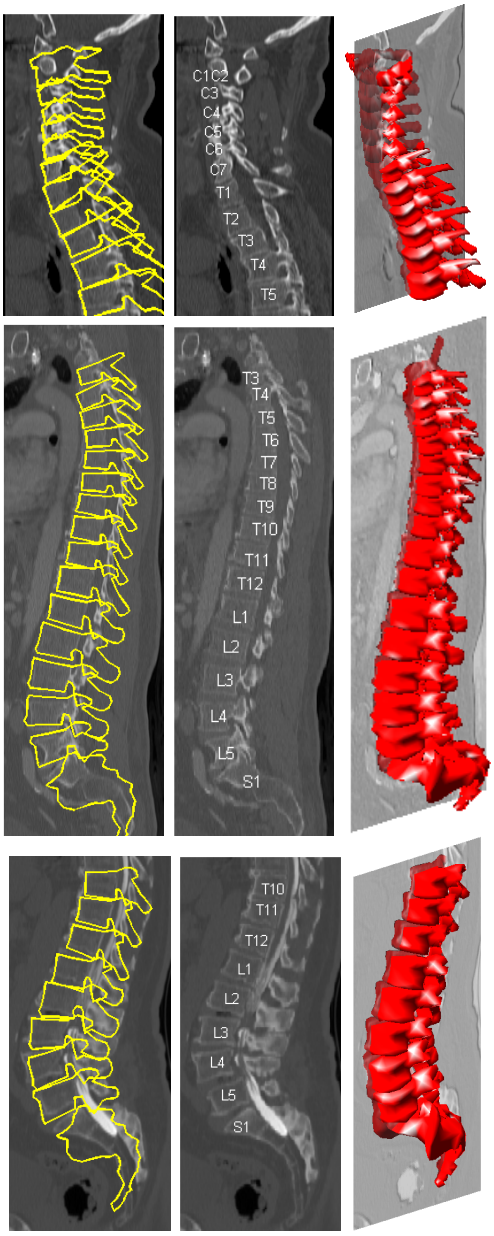


Fig. 19. Examples of CT arbitrary view vertebra recognition in Dataset 2. Left to right: the recognized vertebrae location and poses (represented as 2D silhouettes); vertebrae labels; recovered 3D meshes.

		Position error (mm)				Angular error (deg)				Dimen. error (mm)			
		Cerv.	Thor.	Lumb.	Avg.	Cerv.	Thor.	Lumb.	Avg.	Cerv.	Thor.	Lumb.	Avg.
MR	Mean	2.13	2.49	3.07	2.54	2.33	2.27	2.84	2.57	8.41	8.87	9.82	9.47
	Stdev	2.08	2.16	2.73	2.66	1.86	1.95	2.31	1.97	3.78	4.57	4.94	4.87
	Median	1.92	2.57	2.95	2.56	2.35	2.45	2.21	2.30	8.33	8.76	9.84	9.17
CT	Mean	1.94	-	3.37	3.12	2.53	-	3.39	3.24	8.91	-	9.98	9.66
	Stdev	1.95	-	2.85	2.08	1.64	-	2.57	1.95	3.41	-	4.82	4.47
	Median	2.01	-	3.02	2.91	2.25	-	3.51	2.98	8.74	-	9.68	9.57

TABLE II
EVALUATION OF VERTEBRA RECOGNITION USING 30 MR AND 30 CT SAMPLES IN DATASET 1.

		Position error (mm)				Angular error (deg)				Dimen. error (mm)			
		Cerv.	Thor.	Lumb.	Avg.	Cerv.	Thor.	Lumb.	Avg.	Cerv.	Thor.	Lumb.	Avg.
CT	Mean	2.04	3.38	3.53	3.05	4.30	3.48	3.05	3.63	7.87	8.48	8.60	8.52
	Stdev	1.90	2.80	2.37	2.81	1.41	2.18	1.96	1.77	4.05	4.78	4.82	4.69
	Median	1.93	3.15	3.16	2.89	4.22	3.12	2.91	3.15	7.95	8.80	8.65	8.20

TABLE III
EVALUATION OF VERTEBRA RECOGNITION USING ANNOTATED 30 CT SAMPLES IN DATASET 2 (NO MR SAMPLES INVOLVED).

		Position error (mm)				Angular error (deg)				Dimen. error (mm)			
		Cerv.	Thor.	Lumb.	Avg.	Cerv.	Thor.	Lumb.	Avg.	Cerv.	Thor.	Lumb.	Avg.
MR	Mean	2.85	2.59	2.87	2.69	3.87	3.03	2.80	3.20	8.69	7.99	9.96	8.97
	Stdev	1.80	2.46	2.04	2.42	1.92	1.47	1.84	1.65	3.95	3.12	4.10	3.55
	Median	2.91	2.45	2.80	2.64	3.55	2.65	2.45	2.90	8.10	8.01	10.02	8.57
CT	Mean	2.95	2.97	3.61	3.34	2.17	3.15	2.52	2.87	9.01	8.91	9.32	8.95
	Stdev	2.01	2.33	3.28	2.80	1.35	1.91	2.94	2.52	4.30	4.55	3.89	4.27
	Median	2.98	3.05	3.52	3.44	2.33	2.85	2.49	2.70	8.89	9.05	9.55	9.42

TABLE IV
EVALUATION OF VERTEBRA DETECTION USING 20 MR AND 20 CT SAMPLES (REMOVED WITH TRAINING SAMPLES) IN DATASET 3.

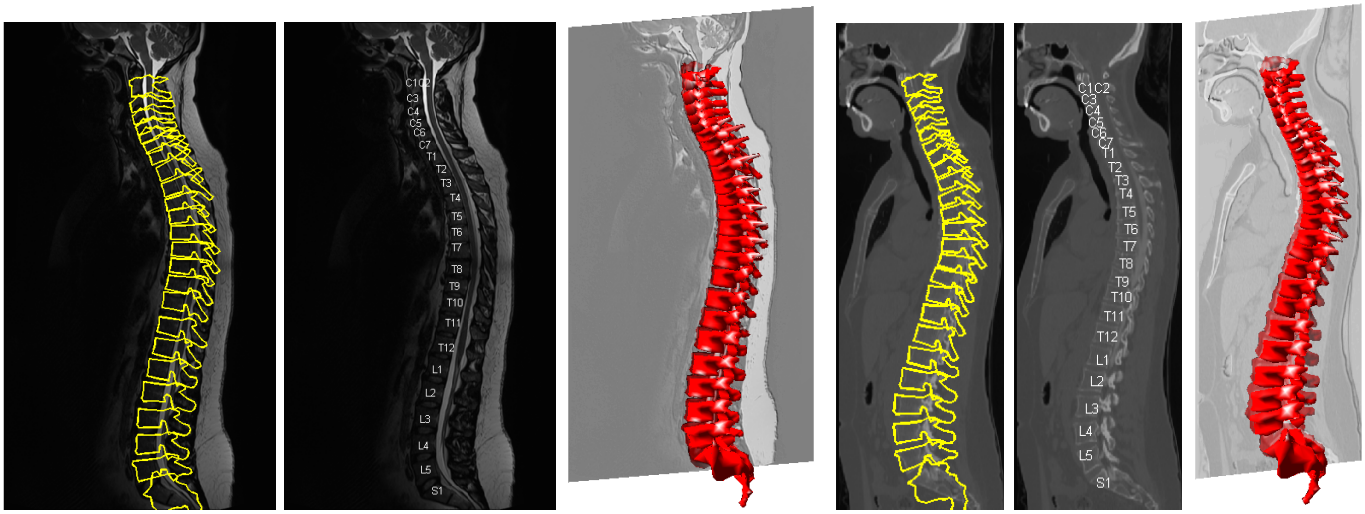


Fig. 20. Examples of MR/CT whole spine recognition in Dataset 3. The recognized vertebrae location and poses (represented as 2D silhouettes), vertebrae labels, and recovered 3D meshes for MR and CT are shown.

- The results of Dataset 2 are shown in Fig. 19 and Table III. The presented CT samples include the cases of larger cervical+thoracic (C+T) and lumbar+thoracic (L+T) coverage. The cervical+thoracic section is identified through the initial detection of C1-C2 landmark analogous to the identification cervical section in Dataset 1. The position error increases to 3.53mm in lumbar due to the higher spine deformation occurred in this dataset. The larger section coverage in the volumes (i.e., more than one spine section) also makes the average standard deviation increase to 2.81mm.
- The whole spine recognition are demonstrated in Fig. 20 using a MR and a CT whole spine scan from Dataset 3 and the overall results on the same dataset are shown in Table IV. Our method can clearly identify the general 3D shape of each vertebra with accurate pose estimation. The position error, angular error, and dimension error all remain low even whole spine scans are involved in this dataset. The highest position and dimensio error occurs in lumbar CT and lumbar MR respectively due to the large template displacements, and the highest angular error happens in cervical MR due to the jittering in the registration of small templates.

Note that Dataset 2 has its own ground truth vertebrae locations (see also [4], [12]) which are different from our annotated locations. Our method prefers the spinal cord center

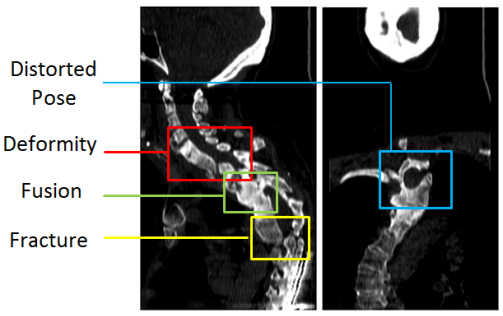
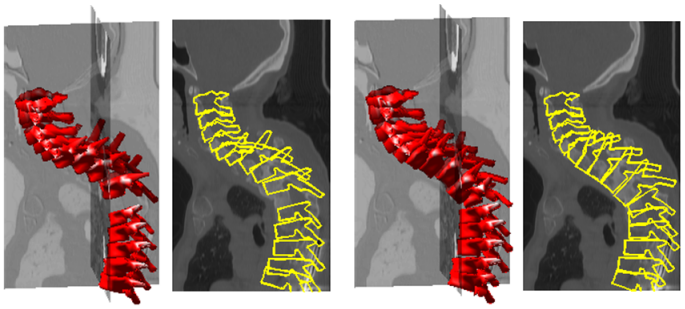


Fig. 21. An example of pathological spine CT in sagittal and coronal views. Typical spine pathological problems such as fracture, fusion, and spinal deformity can be found in this case. Some vertebrae have highly distorted poses, i.e., vertebrae (in blue) have near 90 degree rotations in sagittal plane.



(a) Without articulated transform (b) With articulated transform

Fig. 22. The corresponding recognition results of Fig. 21. Some vertebrae are missing in (a) due to their distorted poses. These missing vertebrae can be recovered by using articulated triplanar poses, as shown in (b).

as the vertebra center while the default locations in Dataset 2 are the centers of vertebrae bodies. We use our annotated ground truths in evaluation instead of the default ones.

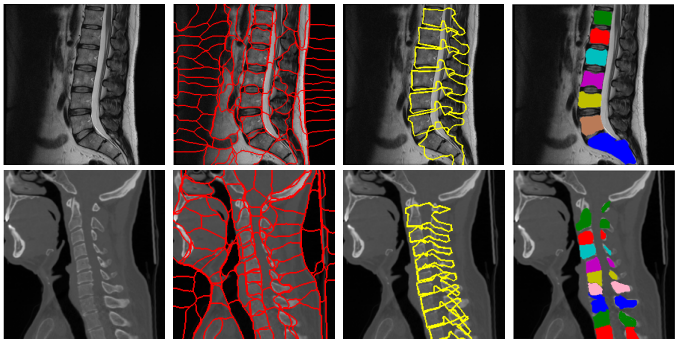
The recognition result proves the robustness of our method in handling different image modalities and arbitrary image views. It also show our method can generate an accurate 3D spine model whose built-in vertebra labeling successfully provides correct identification and labeling for the detected vertebrae landmarks.

VII. Discussion

The proposed recognition method provide a solid basis for more general spine image analysis problems. Extensions can be easily made to cover more extreme cases and provide more analytic functionalities.

Pathological Cases. For some extreme pathological cases, slight extensions on triplanar poses will be needed. We use a typical pathological case in Fig. 21 to demonstrate the extended recognition. As shown in Fig. 22(a), using the default configuration in Sec.VI.C, some vertebrae are missing due to the highly distorted poses. This is because the distorted poses (near 90 degree in sagittal view) can not be reconstructed by the planar projection (shown in Fig. 10 and Fig. 11). A remedy for this is to explicitly transform the triplanar model in sagittal plane, with an articulated rotation of $-\pi/4$ or $\pi/4$, to simulate the distorted poses. Then follow the same estimation scheme of Sec.IV.B and Sec.IV.C to obtain the correct vertebrae poses using the transformed model. The result obtained from the articulated transformed model is shown in Fig. 22(b). The missing vertebrae can now be identified and the poses of other vertebrae are better aligned than the default configuration.

Spine Segmentation The extension from our recognition to segmentation is immediate. We show the MR and CT examples in Fig. 23 for demonstrating the process of cross-modality spine segmentation in different views. The segmentation is done by fitting a set of pre-segmented regions (superpixels) to the silhouette of the 3D spine model on the same image plane. For an input MR/CT slice, we first apply a pre-segmentation as shown in Fig. 23b. The superpixels can be obtained by classical watershed method, or by some popular superpixel methods such as [44]. We then apply our vertebra recognition on the input slice, generating 2D contours for each vertebra (Fig. 23c). The segmentation is immediately obtained by



(a) (b) (c) (d)

Fig. 23. Examples of MR/CT spine segmentation. (a) input slice from MR lumbar and CT cervical part; (b) superpixels (c) the silhouettes (shape contours) used in segmentation; (d) segmentation results.

collecting the superpixels that are surrounded by the contours or have large overlapping area with the corresponding vertebra shape. The recognized vertebrae shapes act as priors to prevent the segmentation from merging unrelated non-vertebrae areas, and resulting segmentation in turns refine the shape contour of the recognized vertebrae. Further generalization to handle more spine analysis problems is possible with the strong shape guidance provided by our recognition method.

VIII. Conclusion

We propose a cross-modality vertebra recognition framework for the identification of local vertebra and global spine information in arbitrary image views. Our recognition can provide simultaneous locations, labels, and poses identification for local vertebrae, and can provide 3D spine reconstruction for specific spine section or even whole spine. The local and global identification are jointly implemented in the Hierarchical Deformable Model (HDM). The local appearance module in HDM conducts the cross-modality feature extraction and provide initial vertebra landmark detection. The global geometry module in HDM continues to match the detected landmarks with the global spine model using point-based registration. The local geometry module in HDM adjust the local poses for all vertebrae, generating the final 3D spine model. Our recognition method successfully extract local and global spine information from different image modalities include T1/T2 MR and CT, and can successfully identify spine structures from lumbar, thoracic, and cervical sections even whole spine.

ACKNOWLEDGEMENT

Computations were performed using the data analytics Cloud at SHARCNET (www.sharcnet.ca) provided through the Southern Ontario Smart Computing Innovation Platform (SOSCIP); the SOSCIP consortium is funded by the Ontario Government and the Federal Economic Development Agency for Southern Ontario. The authors also wish to thank Dr.Jinhui Qin for assistance with the computing environment.

REFERENCES

[1] V. Pekar, D. Bystrov, H. S. Heese, S. P. Dries, S. Schmidt, R. Grewer, C. J. Den Harder, R. C. Bergmans, A. W. Simonetti, and A. M. Van Muiswinkel, "Automated planning of scan geometries in spine mri scans," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*. Springer, 2007, pp. 601–608.

[2] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in ct images," *Medical image analysis*, vol. 13, no. 3, pp. 471–482, 2009.

[3] Y. Zhan, D. Maneesh, M. Harder, and X. S. Zhou, "Robust mr spine detection using hierarchical learning and local articulated model," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. Springer, 2012, pp. 141–148.

[4] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. Springer, 2012, pp. 590–598.

[5] D. Major, J. Hladuvka, F. Schulze, and K. Bühler, "Automated landmarking and labeling of fully and partially scanned spinal columns in ct images," *Medical image analysis*, vol. 17, no. 8, pp. 1151–1163, 2013.

[6] L. R. Long and G. R. Thoma, "Identification and classification of spine vertebrae by automated methods," in *Medical Imaging*. SPIE, 2001, pp. 1478–1489.

[7] S. Seifert, O. Burgert, I. Wächter, R. Dillmann, and U. Spetzger, "Deformable modelling of the cervical spine for neurosugical navigation," in *International Congress Series*, vol. 1268.

[8] M. G. Roberts, T. F. Cootes, and J. E. Adams, "Automatic location of vertebrae on dxa images using random forest regression," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. Springer, 2012, pp. 361–368.

[9] D. Štern, B. Likar, F. Pernuš, and T. Vrtovec, "Automated detection of spinal centrelines, vertebral bodies and intervertebral discs in ct and mr images of lumbar spine," *Physics in medicine and biology*, vol. 55, no. 1, p. 247, 2010.

[10] B. Michael Kelm, M. Wels, S. Kevin Zhou, S. Seifert, M. Suehling, Y. Zheng, and D. Comaniciu, "Spine detection in ct and mr using iterated marginal space learning," *Medical image analysis*, 2012.

[11] M. Lootus, T. Kadir, and A. Zisserman, "Vertebrae detection and labelling in lumbar mr images," in *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*, 2013.

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

[13] S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, and C. Schnörr, "Spine detection and labeling using a part-based graphical model," in *Information Processing in Medical Imaging*. Springer, 2007, pp. 112–133.

[14] R. S. Alomari, J. J. Corso, and V. Chaudhary, "Labeling of lumbar discs using both pixel-and object-level features with a two-level probabilistic model," *IEEE Trans. on Medical Imaging*, vol. 30, no. 1, pp. 1–10, 2011.

[15] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu, "Hierarchical parsing and semantic navigation of full body ct data," in *Medical Imaging*. SPIE, 2009, pp. 1478–1489.

[16] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on PAMI*, vol. 23, no. 6, pp. 681–685, 2001.

[18] J. Liu and J. K. Udupa, "Oriented active shape models," *IEEE Trans. on Medical Imaging*, vol. 28, no. 4, pp. 571–584, 2009.

[19] P. Markelj, D. Tomazevic, F. Pernus, and B. Likar, "Robust gradient-based 3-d/2-d registration of ct and mr to x-ray images," *IEEE Trans. on Medical Imaging*, vol. 27, no. 12, pp. 1704–1714, 2008.

[20] S. Kadoury, H. Labelle, and N. Paragios, "Automatic inference of articulated spine models in ct images using high-order markov random fields," *Medical image analysis*, vol. 15, no. 4, pp. 426–437, 2011.

[21] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[22] P. H. Lim, U. Bagci, and L. Bai, "Introducing willmore flow into level set segmentation of spinal vertebrae," *IEEE Trans. on Biomedical Engineering*, vol. 60, no. 1, pp. 115–122, 2013.

[23] P. Zhang and T. F. Cootes, "Automatic construction of parts+geometry models for initializing groupwise registration," *IEEE Trans. on Medical Imaging*, vol. 31, no. 2, pp. 341–358, 2012.

[24] A. Rasoulian, R. Rohling, and P. Abolmaesumi, "Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model." *IEEE Trans. on Medical Imaging*, vol. 32, no. 10, p. 1890, 2013.

[25] S. Kadoury, H. Labelle, and N. Paragios, "Spine segmentation in medical images using manifold embeddings and higher-order mrfs," *IEEE Trans. on Medical Imaging*, vol. 32, no. 7, pp. 1227–1238, 2013.

[26] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, "Shape representation for efficient landmark-based segmentation in 3d," *IEEE Trans. on Medical Imaging*, vol. 33, no. 4, pp. 861–874, 2014.

[27] Y. Zhan, M. Dewan, and X. S. Zhou, "Cross modality deformable segmentation using hierarchical clustering and learning," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*. Springer, 2009, pp. 1033–1041.

[28] J. Ma, L. Lu, Y. Zhan, X. Zhou, M. Salganicoff, and A. Krishnan, "Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. Springer, 2010, pp. 19–27.

[29] U. Bagci, X. Chen, and J. K. Udupa, "Hierarchical scale-based multiobject recognition of 3-d anatomical structures," *IEEE Trans. on Medical Imaging*, vol. 31, no. 3, pp. 777–789, 2012.

[30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 689–696.

[31] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[32] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 609–616.

[33] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, no. 5786, pp. 504–507, 2006.

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of Computer Vision and Pattern Recognition 2005*, vol. 1. IEEE, 2005, pp. 886–893.

[35] R. M. Murray, Z. Li, and S. S. Sastry, *A Methematical Introduction to Robotics Manipulation*. CRC Press, 1994.

[36] Y. Cai and G. Baciu, "Detecting, grouping, and structure inference for invariant repetitive patterns in images," *IEEE Trans. on Image Processing*, vol. 22, no. 6, pp. 2343–2355, 2013.

[37] E. G. Learned-Miller, "Data driven image models through continuous joint alignment," *IEEE Trans. on PAMI*, vol. 28, no. 2, pp. 236–250, 2006.

[38] T. Drummond and R. Cippolla, "Application of lie algebra to visual servoing," *International Journal of Computer Vision*, vol. 37, no. 1, pp. 21–41, 2000.

[39] M. R. Sabuncu, B. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. on Medical Imaging*, vol. 29, no. 10, pp. 1714–1729, 2010.

[40] D. Ramanan, "Learning to parse images of articulated bodies," in *Advances in Neural Information Processing Systems*, 2006, pp. 1129–1136.

[41] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. on PAMI*, vol. 32, no. 12, pp. 2262–2275, 2010.

[42] C. M. Bishop *et al.*, *Neural Networks for Pattern Recognition*. Oxford Press, 1995.

[43] A. L. Yuille and N. M. Grzywacz, "A mathematical analysis of the motion coherence theory," *International Journal of Computer Vision*, vol. 3, no. 2, pp. 155–175, 1989.

[44] G. Mori, "Guiding model search using segmentation," in *International Conference on Computer Vision*, vol. 2. IEEE, 2005, pp. 1417–1423.