Multi-Modal Vertebrae Recognition using Transformed Deep Convolution Network

Yunliang Cai\textsuperscript{a}, Mark Landis\textsuperscript{b}, David T. Laidley\textsuperscript{b}, Anat Kornecki\textsuperscript{b}, Andrea Lum\textsuperscript{b}, Shuo Li\textsuperscript{a,1,*}

\textsuperscript{a}Dept. of Medical Biophysics, Schalich School of Medicine and Dentistry, University of Western Ontario, 1151 Richmond St, London ON, Canada
\textsuperscript{b}Dept. of Medical Imaging, Schalich School of Medicine and Dentistry, University of Western Ontario, 1151 Richmond St, London ON, Canada

Abstract

Automatic vertebra recognition, including the identification of vertebra locations and naming in multiple image modalities, are highly demanded in spinal clinical diagnoses where large amount of imaging data from various of modalities are frequently and interchangeably used. However, the recognition is challenging due to the variations of MR/CT appearances or shape/pose of the vertebrae. In this paper, we propose a method for multi-modal vertebra recognition using a novel deep learning architecture called Transformed Deep Convolution Network (TDCN). This new architecture can unsupervisely fuse image features from different modalities and automatically rectify the pose of vertebra. The fusion of MR and CT image features improves the discriminativity of feature representation and enhances the invariance of the vertebra pattern, which allows us to automatically process images from different contrast, resolution, protocols, even with different sizes and orientations. The feature fusion and pose rectification are naturally incorporated in a multi-layer deep learning network. Experiment results show that our method outperforms existing detection methods and provides an fully automatic location+naming+pose recognition for routine clinical practice.

Keywords: Vertebra detection, vertebra recognition, deep learning,
1. Introduction

Magnetic resonance imaging (MR) and computed tomography (CT) are two main imaging methods that are intensively and interchangeably used by spine physicians. The longitudinal/differential diagnoses today are often conducted in large MR/CT dataset which makes manual identification of vertebrae a tedious and time-consuming task. Automatic locate-and-name system of spine MR/CT images which supports quantitative measurement is thus highly demanded for orthopaedics, neurology, and oncology.

Automatic vertebra recognition, particularly the identification of vertebra location, naming, and pose (orientation+scale), is a challenging problem in spine image analysis. The main difficulty arises from the high variability of image appearance due to image modalities or shape deformations of the vertebrae: 1) Vertebra is difficult to detect due to imaging 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) Vertebra is difficult to automatically name. The vertebrae and intervertebral discs are lack of unique characteristic features that automatic naming could fail easily. 3) Vertebra pose is difficult to estimate. The poses of vertebrae are highly diverse and little stable features can be used for pose estimation. Except for the local pose and appearance problems, the global geometry of spine is often difficult to recover in some medical situations, i.e., spine deformity and scoliosis. The reconstruction of global spine geometry from limited CT/MR slices can be ill-posed and requires sophisticated learning algorithms.

Most current spine detection methods focus on identification of vertebra locations or labels in particular one particular image modality [1, 2, 3, 4, 5], and vertebra pose information is seldom obtained in the same method. 1) For vertebra localization, learning-based detectors were employed for handling specified image modalities, they were proven to work on CT (generalized Hough) [2], MR
Figure 1: The multi-modal recognition for lumbar spine imaging. The modalities are uniformly trained and detected in one unified recognition system. In this system, features from different modalities are fused and enhanced by each other via a deep network.

(Adaboost) [3], or DXA images (random forest) [6]. Their training and testing were performed on the chosen image protocol only. Some detection methods claimed they can work on both MR and CT. [7] utilized the curved spinal geometric structure extracted from both modality. [8, 9] used boosting-trained Haar features and SVM-trained Histogram of Oriented Gradients (HOG) respectively. However, these cross-modality methods often required the separated training for MR and CT, and thus the separated testing for the two modalities too. 2) For vertebrae naming, [2, 3, 4, 5] had successful labeling on fully or partially scanned image volumes. Their methods relied on the identification of some special landmarks detected from multiple image views, i.e., axial view templates [2], spinal canals [5] or anchor vertebrae [3], while the exact labels are inferred by a probability inference model, i.e., a graph model [10], Hidden Markov Model (HMM) [4], or hierarchical model [3] [11]. 3) Besides the detection and naming, vertebral pose is critical information in orthopedics. Pose estimation was used [1, 8, 5] for extracting the 3D structure of the spines. These estimation methods exploited the multi-planar detector to match the correct vertebrae poses, but can not directly used in a single slice input. In addition, most of the training-based methods, as pointed out in [12], required dense manual annotations for ground truth labels, i.e., annotations of all the corners and the center for each vertebrae. This makes the training-based method not convenient to use.
To overcome these limitations, we uniquely propose a unified framework using *Transformed Deep Convolution Network* (TDCN) to provide automatic cross modality vertebrae location, naming, and pose estimation. As presented in Figure 1, our system is a learning-based recognition system which contains a multi-step training stage and an efficient testing stage. The example results on MR and CT are shown Figure 2. The main ingredients of the system is a novel deep learning model [13] inspired by groupwise registration [14] [15] and multi-modal feature fusion [16] [17]. We have the following contributions in this paper:

- **Vertebra Recognition.** The location, name, and pose (scale+orientation) of each vertebra are identified simultaneously. Further spine shape analysis, i.e., spondylolysis analysis, is then possible basing on the recognition results.

- **Multi-Modal Feature Learning.** The vertebra features are jointly learned and fused from both MR and CT. This enhances the features discrimination and improves the classification of vertebra/non-vertebra.

- **Invariant Representation.** In the training and recognition stage, the sampled of detected vertebrae are automatically aligned, generating transform-invariant feature representations or rectified final poses respectively.

- **Simple Annotation.** Thanks to the invariant representation, our method only requires single-clicking for each vertebrae in ground truth annotation while other methods [8, 5, 9] require four clicks or more.
2. The Transformed Deep Convolution Network

The Transformed Deep Convolution Network (TDCN) is a novel deep network structure, which can automatically extract the best representative and invariant features for MR/CT. It employs MR-CT feature fusion to enhance the feature discriminativity, and applies alignment transforms for input data to generate invariant representation. This resolves the modality and pose variation problems in vertebra recognition. The overall structure of TDCN presented in Figure 3. The two major components in TDCN: the feature learning unit and the multi-modal transformed appearance learning unit are presented in details as follows.

2.1. Low Level Feature Learning Unit

The low-level feature learning unit is for unsupervised learning adaptive features that best represent the training MR/CT samples. The feature learning is implemented by layers of Convolution Restricted Boltzmann Machines (CRBM) [13]. CRBM is a multi-output filtering system which can adaptively update its filter weights to obtain the best approximative feature maps for the training samples. The learned features can reveal some unique micro-structures in MR or CT, while manual features like SIFT or HOG cannot adapt to these micro-differences.
As shown in Figure 3, there are two CRBM layers for feature extraction over MR patches and two for the extraction over CT patches. Each CRBM will accept a set of visible layers $v$ as input, generating binary hidden layers $h$. In the training stage, the CRBM continuously tunes the weight coefficients $W_k$ for each of the $k \in \{1, \ldots, K\}$ filter, in order to minimize the following functional

$$E(v, h) = -\sum_{k=1}^{K} h_k \cdot (\bar{W}_k * v) - \sum_{k=1}^{K} h_k b_k \sum_{i,j} h_{i,j}^k - c \sum_{i,j} v_{i,j}$$  

where $b_k$ and $c$ are adjustable bias, $i, j$ will range in $[1, N_V - N_W + 1]$ for a $N_V \times N_V$ input layer $v = [v_{i,j}]$ and a $N_W \times N_W$ filter $W_k$. Operator $\cdot$ is an element-wise product of two matrixes, while $*$ is a convolution. The hidden layer generated by $W_k$ is a $N_H \times N_H$ map $h_k = [h_{i,j}^k]$ with $N_H = N_V - N_W$.

Two CRBMs are stacked together (level 1,2 in Figure 3), providing two level feature extraction. Each CRBM will first unsupervisely train itself until (1) is minimized, then connected its output(input) to the higher(lower) level CRBM. Let level 1 has $K_1$ filters and level 2 has $K_2$ filters, the output of the two-level feature learning will be $K_1 \times K_2$ feature maps. For a given 2D filter $W$ and bias $b$, the output of the CRBM can be defined by:

$$f(v) = \sigma(\bar{W} * v + b)$$  

where $\sigma$ the element-wise logistic sigmoid function and $v$ is in its vectorized form. Suppose $f_1$ is the output of level 1 CRBM in Figure 4, for MR-CRBM (or CT-CRBM), we let $f_1$ be defined in (2) by substituting the corresponding $W$ with the $K_1$ filters’ coefficients, i.e., $\{W_{1MR}^{1}, \ldots, W_{K1}^{1MR}\}$ (or $\{W_{1CT}^{1}, \ldots, W_{K1}^{1CT}\}$). Similarly, the level 2 CRBM in Figure 4 output mapping $f_2$ is defined by substituting $W$ in (2) with the corresponding $K_2$ filters’ coefficients.

### 2.2. Multi-Modal Transformed Appearance Learning Unit

Multi-modal feature fusion is important in classification as the missing of some features in one modality will be compensated by the others, i.e., the missing vertebra discs in CT can be compensated by its MR counterparts. Also, the features are expected to be pose invariant, as geometric invariance can enhance
the discrimination of vertebra/non-vertebra. The multi-modal transformed appearance learning unit is a set of deep network layers specifically for fusing multi-modal feature and transforming input samples to achieve invariant representations. With the fusion of low-level features from different modalities, the feature representation of vertebra can be enhanced to obtain a better discrimination.

**MR/CT Feature Fusion using RBM.** The feature maps extracted by CRBM are sent to two Restricted Boltzmann Machines (RBM) for multi-modal feature fusion. RBM is a simpler version of CRBM, which works with vectorized visible and hidden layers other than 2D layers in CRBM. It also can reduce the dimensionality of the input by setting a shorter hidden layer size. Suppose now $v = [v_i] \in \mathbb{R}^{N_V}$ is the visible layer, which can be either MR or CT features, and $h = [h_j] \in \mathbb{R}^{N_H}$ is the hidden layer, the particular functional RBM minimizes is:

$$E(v, h) = - \sum_{i,j} v_i W_{ij} h_j - \sum_j b_j h_j - \sum_i c_i v_i \tag{3}$$

where $W_{ij}$ now is a $N_V \times N_H$ matrix and $b_j$, $c_i$ are adjustable bias. As shown in Figure 4, the MR/CT feature maps obtained from the previous low-level feature learning unit are vectorized and combined as the visible layer of the RBM, generating the MR-CT fused feature vector as the hidden layer of the RBM. The fusion process is performed specifically by the level 3 RBM in TDCN (see the level 3 layer in Figure 3). The result of level 3 RBM is a dimension-reduced unified feature vector that jointly describe the vertebra patterns in MR and CT.

Similar to (2), we can define the output mapping of RBM as nonlinear
activation function:

\[ f(v) = \sigma(W^T v + b) \]  (4)

where \( v, b \) are vectors other than 2D maps in (2). The mapping of level 3 RBM in Figure 3, particularly \( f_3 \), is thus constructed by substituting (4) with the learned weight matrix and bias from training, i.e., \( W^3 \) and \( b^3 \). \( f_4 \) for the level 4 RBM can be obtained in the same way.

The multi-modal fusion has a number of advantages that cannot be provided by single-modal feature learning (i.e., features learned only from level 1, 2 while bypassing level 3 and 4 in TCDN): 1) MR (T1,T2) and CT features can be compensated for each other in the fused feature vectors; 2) Shape features, which are rich in both modalities, are enhanced in the fused feature vector; 3) The fused feature vector can be used for identifying vertebra structures across MR and CT using only one unified classifier.

**Pose Rectification by Congealing.** Pose Rectification is for unifying the different orientation and sizes (scales) of vertebrae in a spine scan, so that the resulting learned feature can be more representative for vertebra. The vertebra appearance can be considered as invariant such that any vertebra patch can be assumed as be generated by warping the standard invariant patch with certain rotation and rescaling. This warping uniquely describes the 2D geometric transform from a standard model to a sampled data, which we describe as the pose of the vertebra. The pose rectification is performed by congealing.

Congealing is a groupwise registration for an image set. It is employed for aligning the 2D feature maps obtained from the level 3 RBM. Suppose \( v \) is an arbitrary vertebra image patch (MR or CT) with free-form size/orientation, inverse warping \( G^{-1} \) deforms \( v \) to a \( m \times m \) regular patch, denoted as \( v \circ G^{-1} \). Let \( \{v^1, \ldots, v^n\} \) be the collection patches, the corresponding warped versions are represented as vector \( \{v^1 \circ G_1^{-1}, \ldots, v^n \circ G_n^{-1}\} \). We require \( G = G(r, s) \) be a similarity transform described by the rotation \( r \) and scaling \( s \). The congealing
for the patch collection can be formulated as the minimization of energy:
\[
C(r, s) = \sum_{i,j=1, i \neq j}^n ||v^i \circ G_i^{-1}(r_i, s_i) - v^j \circ G_j^{-1}||^2.
\] (5)

As \( v \) can be either a MR or a CT patch, we cannot directly align these two types of patches in (5) unless they have a unified representation. This representation can be obtained by the feature fusion of RBM (Figure 4) where the fused feature mapping allows us to align images from different modalities in the same congealing model. By using the CRBM mapping \( f_1, f_2 \) and the RBM mapping \( f_3 \) obtained by activation (2) and (4) respectively, (5) can be then revised as
\[
C(r, s) = \sum_{i,j=1, i \neq j}^n ||f_3(f_2(f_1(v^i \circ G_i^{-1}(r_i, s_i)))) - f_3(f_2(f_1(v^j \circ G_j^{-1})))||^2.
\] (6)

where →\( \) is the vectorization. \( f_1, f_2 \) will extract the low-level MR/CT features of patch \( v \) and \( f_3 \) will map them into a unified feature vector. The deep congealing (6) not only can align patches in either MR or CT as the original congealing does (5) but also can align the MR and CT patches in the same patch set.

Figure 5 illustrate a toy example of congealing over CT+MR images by minimizing (6). The vertebra patches in the three images are first mapped to unified feature maps using \( f_1, f_2 \) and \( f_3 \). The deep congealing (6) is then applied, obtaining the aligned patches. Note that the bounding boxes of the patches under congealing can reveal the correct poses of the corresponding vertebrae. Therefore, from the congealing operation, we can rectify the correct poses of the vertebrae.
3. The Training of Multi-Modal Recognition System

The training process of the TDCN system is presented in Figure 6. The process starts with the annotation of sample patches in original scans. It then trains TCDN using the selected samples, generating invariant vertebra features. The features are applied in training a SVM, obtaining the desired vertebra classifier.

**One-Click Sample Annotation.** The training samples (positive/negative) are collected by simple clicking operations in image slices. For positive samples, the user only need to click the vertebra center in a training image slice regardless of the size and orientation of the vertebra. This significantly simplifies the sample collection process in other methods (i.e., [8] [9]) where at least four clicks are needed for each vertebra. A coarse labeling of vertebra type: T/L or S vertebra is required after clicking but no further specific labeling is needed. This also simplifies the previous training-based system which detailed labeling is often required. For negative samples, the non-vertebra patches can be simply obtained by random sampling or clicking. The ratio of the number of vertebra to that of non-vertebra can be set around 1/10 to 1/8.

**TDCN Feature Learning.** The selected training samples are served as the input of the TDCN (Figure 3) according to their original modalities. The TDCN then starts the layer-wise learning via the optimization of CRBM energy function (1) and RBM energy (3). Particularly, the learning starts from the lowest layer, updating its parameters until optimized. The parameters of the layer are then freezed, generating output and triggering the learning of the upper CRBM/RBM layer. The layer-wise update continues until all layers are optimized.
Vertebra Classifier Learning. Thanks to the unified feature representation of MR and CT patches, we can train the vertebra/non-vertebra classification for both modalities simultaneously in a SVM. In our system, a standard classification SVM is trained for three classes according to the TDCN trained features: T/L vertebra, S vertebra, and non-vertebra.

After the complete training process, the trained TDCN+SVM can be directly applied to the blind recognition for both MR and CT patches. With a patch sampling process (i.e., sliding window), the trained system can then be used for vertebra recognition for arbitrary MR/CT slices.

4. The Vertebra Recognition

We directly apply the trained multi-modal recognition system for full automatic vertebra recognition on arbitrary spine image. The overall recognition process is shown in Figure 7.

Vertebra Detection. As shown in the first step of Figure 7, to simulate the various poses of the vertebrae, we first rotate or rescale the input MR/CT image, generating a set of articulately transformed images. Regular patches (i.e., 51 × 51) are then randomly sampled from the images, and are sent as input to the trained TDCN and SVM. The locations that generates positive SVM output are marked as new vertebrae positions. The detection result is represented as a set of bounding boxes as green boxes for L/T vertebrae and yellow for S vertebrae in Figure 8(a). A vertebra pose is now described by the box size and orientation.

Pose Post-Rectification. After obtaining the positive results from SVM (2nd-last part of Figure 7), we can rectify the poses of the bounding boxes.
Figure 8: Intermediate results of the recognition process. (a) The vertebra detection results; (b) Pose post-rectification by congealing (three iterations are shown); (c) Naming by label propagation from L5 and S1 along spinal direction (red).

using the same mechanism of deep congealing (6). This can be done by the optimization (6) with the positive patch set substituted as the input variables. The resulting aligned bounding boxes not only recover the correct poses of the detected vertebrae, but can also join the duplicated bounding boxes near a location into one box. Figure 8(b) illustrate the process post-rectification of vertebrae poses using deep congealing.

**Vertebra Naming.** The specific vertebrae names can be obtained using the pose-rectified bounding boxes (last part of Figure 7). Since the vertebra orientation has been correctly recovered by congealing, one can easily link up the separated bounding boxes in pairwise to form a spine structure using their orientations. The orientation-based linking also help to avoid the connecting with unrelated outliers bounding boxes from falsely detected locations. As shown in Figure 8(c), when the bounding boxes are linked, we can easily identify the connecting part of T/L and S vertebra along the linking directions. The vertebra L5 and S1 can then be identified from the connecting part. Starts with L5 and S1, by propagating the specific labels along the linked bounding boxes, we can obtain all vertebra labels for the detected vertebrae. The naming of MR and CT can be done simultaneously using the same work flow.
5. Experiments

Our method is tested on a cross modality MR+CT dataset which contains 60 MR volumes (T1 and T2 included) and 90 CT volumes from subjects with different pathologies (i.e., fracture, spondylolisthesis). Particularly, 30 pairs of MR-CT lumbar volumes are from Spineweb 1, 50 CT volumes (lumbar and cervical) are from MS Annotated Spine CT Database 2, and the rest volumes (lumbar and whole spine) are from our supportive hospitals.

5.1. Experiment Setting

Our TDCN system contains 2 CRBM layers and 2 RBM layers. The level 1 layer CRBM has $7 \times 7 \times 6$ filter banks for both MR and CT modality, it will accept image patches with size $51 \times 51$. The level 2 CRBM has $9 \times 9 \times 10$ filter banks. The resulting feature maps of the low-level feature learning unit are 60-dim feature maps. For the RBM layers in the transformed appearance learning unit, the output of level 3 RBM is a 800-dim feature vector, and level 4 RBM is a 200-dim feature vector. The level 4 feature vectors are then used to train the SVM for vertebra/non-vertebra classification.

For TDCN training, a set of 1150 patches (with size $51 \times 51$) are sampled from 6 selected MR and 4 CT volumes of our dataset. Particularly, 115 patches from the set are manually selected and labeled as L/T vertebra patches by our one-click annotation. The rest patches are non-vertebra patches randomly sampled near the vertebra areas of the same image slices. The proposed multimodal recognition system is then trained using the collected patch set. The groundtruth of the testing data is obtained by following the same selection process.

Our slice selection follows the standard radiology protocol of spine physician, which reflects the common clinical scenario in spine diagnosis. Unlike the groundtruth labeling of the spine segmentation problems where manual labels

---

1http://spineweb.digitalimaginggroup.ca
2http://research.microsoft.com/en-us/projects/spine/
in different slices will affect the final evaluation, slice selection in our test cannot affect vertebra location selection/detection. The slice selection and vertebra groundtruth labeling are in separated manual processes, and a degree of distortion is acceptable thanks to our robust detection. The vertebra location of a selected slice only depends on the success of extracting some vertebra features such as vertebra contours or vertebra discs. With our TDCN-trained feature extraction, the repetitive vertebra features from different sagittal slices can be reliably identified, and thus the location of the vertebra can still be accurately recovered.

5.2. Recognition Experiments

Testing is conducted by 110 (1-2 slices per volume) MR and CT sagittal slices from the 90 MR-CT volumes excluding all training volumes. The overall naming accuracy we obtained is 98.1% (MR) and 96% (CT) with error less than 3.81mm.

Recognition Test. The qualitative example results are presented in Figure 9, Figure 10, and Figure 11 respectively, which show the versatility of our method under different imaging environments. Figure 9 demonstrates the overall cross-modality detection in Spineweb dataset. Figure 10 shows examples of the detection and naming for the MS Spine database. The results in Figure 11 illustrates our successful identification of location+naming+pose for all vertebrae in whole spine scans.

Our results provide a comprehensive numerical profile for the spine, which parameterizes its shape and structure identity. The numerical profile in turns, refine the detection results so that non-vertebra structures near the spine can be excluded. The geometric parametrization not only enables an intuitive visualization of the spine curvature information, but also allows interactive deformations to incorporate further spine movements. With the parameterized model, our vertebra detection results can be immediately used for spine registration, pathology measurement, or bio-mechanical simulation proposes.

The quantitative results of the recognition test are evaluated by the standard
error of the detected vertebrae locations to that of the ground truth locations. The evaluations on specific vertebrae are shown in Table 1, while overall comparative results are presented in Table 2. We can have naming accuracy 95% above for all data from both MR and CT. The naming errors occurred in the results are due to the mis-identification of L5 and S1 in some slices. The detection errors are constantly low (around 3 mm) thanks to the post pose-rectification which aligns the detected patches to refine their locations. Slight increase of detection errors are observed in S1, as it cannot be aligned by its neighbor vertebrae due to its unique appearance. As shown in Table 4, our detection results are competitive to the state-of-the-art detection methods for CT [4] and MR [18].

Additional quantitative comparative results are present in Table 3, where we extend the lumbar detection to thoracic and cervical spine detection using
Table 1: Recognition errors (detection and naming) of specific lumbar vertebrae on both MR and CT of the testing dataset (SpineWeb).

<table>
<thead>
<tr>
<th>Vertebra</th>
<th>S1</th>
<th>L5</th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>T12</th>
<th>T11</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR Detect. Err. (mm)</td>
<td>5.12</td>
<td>3.23</td>
<td>2.29</td>
<td>3.61</td>
<td>2.98</td>
<td>3.1</td>
<td>3.05</td>
<td>2.87</td>
</tr>
<tr>
<td>Naming Acc. (%)</td>
<td>99.1</td>
<td>98.4</td>
<td>98.2</td>
<td>97.4</td>
<td>95.1</td>
<td>98.3</td>
<td>97.1</td>
<td>98.3</td>
</tr>
<tr>
<td>CT Detect. Err. (mm)</td>
<td>6.05</td>
<td>4.23</td>
<td>2.7</td>
<td>2.86</td>
<td>3.28</td>
<td>3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naming Acc. (%)</td>
<td>98.9</td>
<td>97.1</td>
<td>97.9</td>
<td>96.5</td>
<td>95.0</td>
<td>96.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Detection errors (mm) of recognition our method and those in [18] [12]. Average error values for all vertebrae are presented in the table. Our method has the best labeling success rate and detection accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Mean (MR)</th>
<th>Std (MR)</th>
<th>Mean (CT)</th>
<th>Std (CT)</th>
<th>Label Acc (MR/CT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>3.23</td>
<td>2.09</td>
<td>3.81</td>
<td>2.98</td>
<td>98.1%/96%</td>
</tr>
<tr>
<td>Baseline</td>
<td>3.1 ([18])</td>
<td>2.1 ([18])</td>
<td>10.6 ([12])</td>
<td>16.9 ([12])</td>
<td>97.8%/86% ([18]/[12])</td>
</tr>
</tbody>
</table>

the MS Annotated Spine CT Database. This provides a more detailed regional comparison with the methods of [4] and [12]. As Table shows, our method has the best accuracy and robustness. Our error values are much smaller than those in [4] [12]. Note that [4] [12] consider the 3D detection scenario, which involves more spatial deformations/distortions and thus have larger error values.

Instead of attempting to tackle the hard 3D problems, our 2D detection provides a reliable solution that can manage most of the detection tasks in common diagnosis cases. The reliability makes our method practically useful in most spine applications.

Besides lumbar scans, whole spine MR and CT slices are tested as presented in Figure 11 to show the generality of our method. Our system successfully recognize all vertebrae with C1, C2 excluded due to their special appearances. In these examples, the same trained parameters as previous tests are used.

**Classifier Performance.** To more accurately understand the role of vertebra classifier in our TDCN+SVM recognition system, we specifically evaluate the performance of the SVM classifier in vertebra/non-vertebra classification.
Figure 10: Examples of vertebrae detection on CT slices from MS Annotated Spine Dataset. In this test, the vertebra C1-C2 are jointly treated as a special anchor vertebra similar to S1 for the naming process of cervical scans.

tests. As shown in Table 4, the vertebra classifier has high sensitivity (93.8% best) and high specificity (92.1% best). The testing set is formed by 1000 randomly sampled patches from the same slice set of recognition test, with 300 positive and 700 negative patches. In addition, each patch centered at the sampled location is cropped with five different random poses, which leads to 5000 testing patch set in total. The results (see Table 4) are obtained and sorted according to the poses. Our system constantly obtains high sensitivity (around 90%) and specificity (around 93%) in all poses, which makes it possible to apply to detect the whole spine in the image slice. The best classification (sens. 92.1%, spec. 93.4%) is achieved in the pose with rotation 8 degree clockwise and resizing by factor of 0.92. This can be observed in the detection results of Figure 8 and Figure 9, even without the pose post-rectifications, we can still rely on the SVM-classifier to obtain general vertebra localizations with tolerable errors.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Med</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td>Lumbar</td>
<td>5.4</td>
<td>9.7</td>
<td>11.2</td>
</tr>
<tr>
<td>Thoracic</td>
<td>5.5</td>
<td>9.9</td>
<td>10.8</td>
</tr>
<tr>
<td>Cervical</td>
<td>6.5</td>
<td>8.2</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Table 3: Comparative detection results (error mm) for multiple spine regions (MS Annotated Spine Database). Our 2D detection provide a reliable and accurate solution for the recognition task.

Figure 11: Recognition results on whole spine MR and CT scans. The complete spine shape and vertebra labels are jointly identified.

6. Discussion

2D v.s. 3D Methods. There are a number of 3D vertebra detection methods [4] [12] [11] [19] that can provide vertebra identification in 3D space, particularly for CT images. However, as reported in [12] and [14], 3D detection is still very challenging especially in some extreme pathological cases. Instead of competing the performance in 3D spine recognition where large detection error could occur, the proposed method in this paper focuses on the 2D spine recognition and provides a fast, stable and highly accurate solution for vertebra detection. The practical advantage of applying 2D detection/recognition is that
Table 4: The vertebra/non-vertebra classification results in patch testing set.

<table>
<thead>
<tr>
<th>Angle (deg.)</th>
<th>-5</th>
<th>12</th>
<th>18</th>
<th>-14</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale</td>
<td>1.1</td>
<td>0.98</td>
<td>0.95</td>
<td>1.32</td>
<td>0.92</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>90.3</td>
<td>89.7</td>
<td>88.5</td>
<td>91.0</td>
<td>92.1</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>92.6</td>
<td>90.4</td>
<td>92.6</td>
<td>91.3</td>
<td>93.8</td>
</tr>
</tbody>
</table>

it is fully compatible with the current clinical scenario in spine diagnosis, where 2D slice is still favored by spine physicians and the slices widespread in most of existing medical reports and diagnostic documents. A reliable 2D detection tool is much easier to be adopted considering these practical situations. Also, as reported in Table 2 and Table 3, the proposed 2D approach has better accuracy than the state-of-the-art 3D methods. This makes it more applicable to the measurement of spine pathological cases such as spondylosis and spondylolysis. For implementation, the 2D approach is faster than the that proposed in [19], as only single scan is required in the 2D detection.

**Limitations.** As discussed above, the proposed method is a 2D recognition approach such that a potential limitation is that the 2D image plane might not be sufficient to capture an irregular 3D spine shape. Some vertebrae could be missing or deformed in the image and thus cannot be correctly detected. A remedy for this problem is to extend the single plane detection to multiple planes, where the adjacent slices of the sampled image will be involved in detection. To detect the adjacent slices, one can simply apply the same detection. The missing vertebrae can then be identified through the same process. The duplicately detected vertebrae from different slices can be merged by their coordinates, and are linked together to form a single spine model similar to that described in Sec.4. An illustrative example is shown in Figure 12. In this example, the T1, T2, and T3 are missing in the single plane detection of the originally sampled image. By picking the adjacent slice from the same image volume Figure 12(b), the missing vertebrae are identified with C7 vertebra duplicately detected. We then combine the detection result of both images and merge the duplicated ver-
Figure 12: The missing vertebra problem of irregular spine in an image. The missing vertebra (T1-T3) of figure (a) can be detected in additional slice (b). The combined results (c) provide the complete vertebra detection for the original scan.

tebrae. The combined result can be processed by following the same vertebra naming procedure in Sec.4.

7. Conclusion

In this paper, we proposed a multi-modal vertebra recognition framework using transformed deep convolution network (TDCN). TDCN automatically extract modal adaptive, high discriminative, and pose invariant features for recognition. Using the TDCN-based recognition system, we can simultaneously identify the locations, labels, and poses of vertebra structures in both MR and CT. The system has successfully passed the tests on multi-modal datasets for lumbar and whole spine scans with high accuracy and stability.


