Computerized Medical Imaging and Graphics 51 (2016) 11–19

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Multi-modal vertebrae recognition using Transformed Deep Convolution Network

Yunliang Cai [a], Mark Landis [b], David T. Laidley [b], Anat Kornecki [b], Andrea Lum [b], Shuo Li [a,b,*]

[a] Dept. of Medical Biophysics, Schulich School of Medicine and Dentistry, University of Western Ontario, 1151 Richmond St, London, ON, Canada
[b] Dept. of Medical Imaging, Schulich School of Medicine and Dentistry, University of Western Ontario, 1151 Richmond St, London, ON, Canada

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 24 January 2016
Accepted 29 February 2016

Keywords:
Vertebra detection
Vertebra recognition
Deep learning
Convolution network

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 a fully automatic location + naming + pose recognition for routine clinical practice.

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 vertebraes: (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 [1–5], and vertebra pose information 9s 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 (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. Štern et al. [7] utilized the curved spinal geometric structure extracted from both modality. Kelm et al. and Lootus et al. [8,9] used boosting-trained Haar features and SVM-trained Histogram

* Corresponding author at: Dept. of Medical Biophysics, Schulich School of Medicine and Dentistry, University of Western Ontario, 1151 Richmond St, London, ON, Canada.
E-mail address: slishuo@gmail.com (S. Li).

http://dx.doi.org/10.1016/j.compmedimag.2016.02.002
0895-6111/

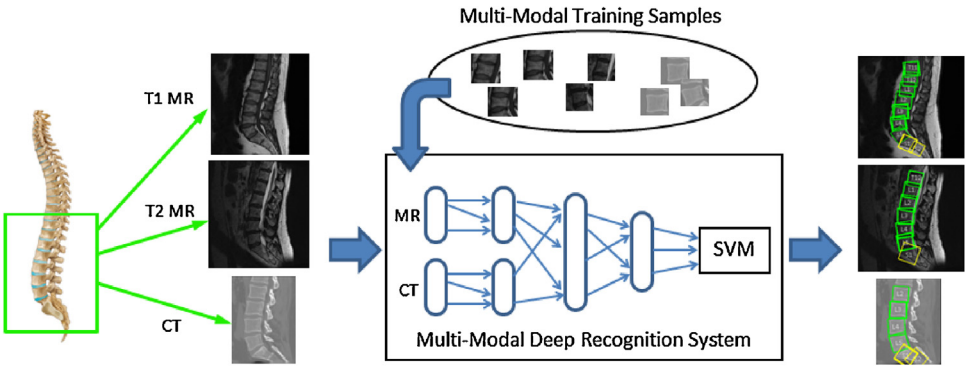

Fig. 1. The multi-modal recognition for lumbar spine imaging. The modalities are uniformly trained and detected in one unified recognition system. In this system, features from different modalities are fused and enhanced by each other via a deep network.

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, [2–5] 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 [3,18]. (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 [11], 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 Fig. 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 Fig. 2. The main ingredients of the system is a novel deep learning model [12] inspired by groupwise registration [13,14] and multi-modal feature fusion [15,16]. 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 automatic 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 Fig. 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) [12]. 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

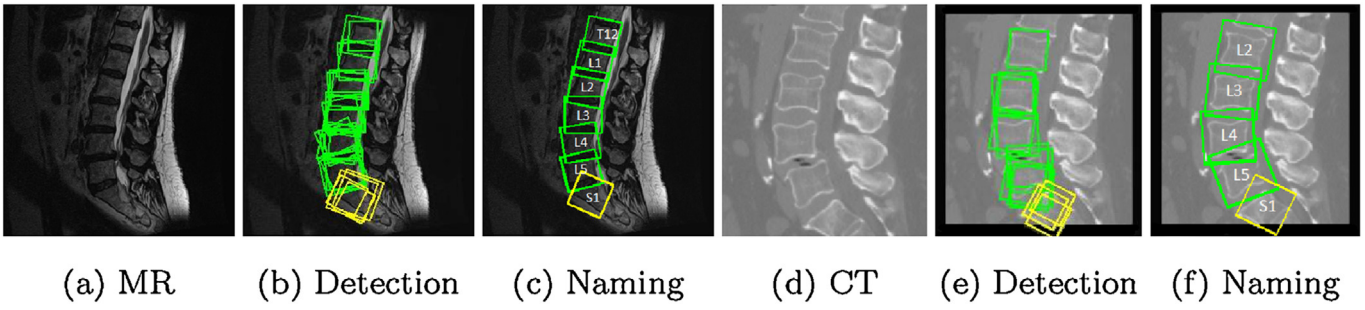

Fig. 2. Examples of detection and naming result for both MR and CT images.