Deformable MR Prostate Segmentation via Deep Feature Learning and Sparse Patch Matching

Yanrong Guo[†], Yaozong Gao[†], Dinggang Shen[*], *Senior Member, IEEE*

***Abstract*—Automatic and reliable segmentation of the prostate is an important but difficult task for various clinical applications such as prostate cancer radiotherapy. The main challenges for accurate MR prostate localization lie in two aspects: (1) inhomogeneous and inconsistent appearance around prostate boundary, and (2) the large shape variation across different patients. To tackle these two problems, we propose a new deformable MR prostate segmentation method by unifying deep feature learning with the sparse patch matching. *First,* instead of directly using handcrafted features, we propose to learn the latent feature representation from prostate MR images by the stacked sparse auto-encoder (SSAE). Since the deep learning algorithm learns the feature hierarchy from the data, the learned features are often more concise and effective than the handcrafted features in describing the underlying data. To improve the discriminability of learned features, we further refine the feature representation in a supervised fashion. *Second*, based on the learned features, a sparse patch matching method is proposed to infer a prostate likelihood map by transferring the prostate labels from multiple atlases to the new prostate MR image. *Finally*, a deformable segmentation is used to integrate a sparse shape model with the prostate likelihood map for achieving the final segmentation. The proposed method has been extensively evaluated on the dataset that contains 66 T2-wighted prostate MR images. Experimental results show that the deep-learned features are more effective than the handcrafted features in guiding MR prostate segmentation. Moreover, our method shows superior performance than other state-of-the-art segmentation methods.**

***Index Terms*—MR prostate segmentation, stacked sparse auto-encoder (SSAE), sparse patch matching, deformable model**

I. Introduction

Prostate cancer is the second leading cause of cancer death in American men, behind only lung cancer [1]. As a main imaging modality for clinical inspection of prostate, Magnetic Resonance (MR) imaging provides better soft tissue contrast than ultrasound in a non-invasive way, and has the emerging role in prostate cancer diagnosis and treatment [2, 3]. The accurate localization of the prostate is an important step for assisting the diagnosis and treatment, such as guiding biopsy procedure [2] and radiation therapy [3]. However, the manual segmentation of the prostate is tedious and time-consuming, and also suffers from intra- and inter-observer variability. Therefore, developing automatic and reliable segmentation methods for MR prostate is clinically desirable and an important task.

However, accurate prostate localization in MR images is difficult due to the following two main challenges. First, the appearance patterns vary a lot around the prostate boundary across patients. As we can see from Fig. 1 (a), the image contrasts at different prostate regions, i.e., the anterior, central and posterior regions, change both across different subjects and within each subject. Fig. 1 (b) gives the intensity distributions of prostate and background voxels around the prostate boundary, respectively. As shown in the figure, the intensity distributions highly vary across different patients and do not often follow the Gaussian distribution.

To evaluate the shape difference in our dataset, we adopt the PCA analysis by mapping each high-dimensional shape vector onto a space spanned by the first three principal components. Note that the shape vector is formed by the concatenation of all vertex coordinates, and then linearly aligned to the mean shape before PCA analysis. Fig 2 shows the distribution of 66 prostate shapes, which also indicates the inter-patient shape variation among the shape repository.

A. Related Work

Recently, most studies in T2-weighted MR prostate segmentation focus on two types of methods: multi-atlas-based [4-7] and deformable-model-based [8, 9] segmentation methods. Multi-atlas-based methods are widely used in medical imaging [10-12]. Most research focuses on the design of sophisticated atlas selection or label fusion method. Yan et al [5] proposed a label image constrained atlas selection and label fusion method for prostate MR segmentation. During the atlas selection, label images are used to constrain the manifold projection of intensity images, which can relieve the misleading projection due to other anatomical structures. Ou et al [7] proposed an iterative multi-atlas label fusion method by gradually improving the registration based on the prostate vicinity between the target and atlas images. For deformable-model-based methods, Toth [8] proposed to incorporate different features in the context of AAMs (Active Appearance Models). Besides, with the adoption of the level set, the issue of landmark correspondence can be avoided.

But both types of these methods require careful feature

Manuscript received May 4, 2015. This work was supported in part by NIH grant CA140413.

[†]Y. Guo and [†]Y. Gao are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599 USA (e-mail: yrguo@email.unc.edu, yzgao@cs.unc.edu). †Co-first authors.

[*]D. Shen is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599 USA; and also with Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: dgshen@med.unc.edu). *Corresponding author.

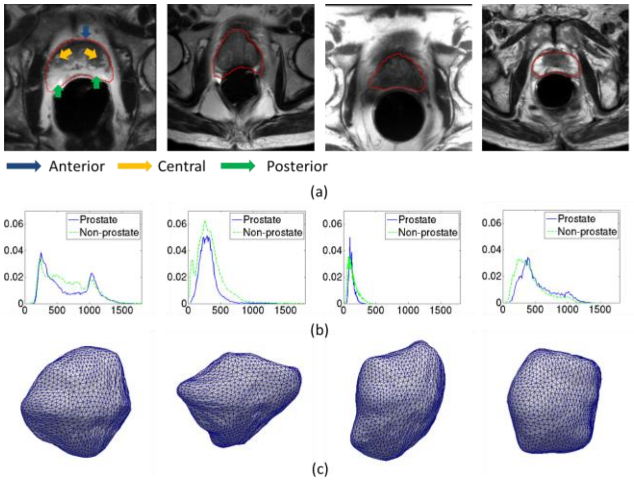


Fig. 1. (a) Typical T2-weighted prostate MR images. Red contours indicate the prostate glands delineated manually by an expert. (b) Intensity distributions of prostate and background voxels around the prostate boundary of (a). (c) The 3D illustrations of prostate surfaces corresponding to each image in (a).

engineering to achieve good performance. The multi-atlas based methods require good features for identifying correspondences between a new testing image and each atlas image [13], while the deformable model relies on discriminative features for separating the target object (e.g., the prostate) from the background [14]. Traditionally, intensity patch is often used as features for the above two methods [15, 16]. However, due to the inhomogeneity of MR images, the simple intensity features often fail in segmentation of MR images with different contrasts and illuminations. To overcome this problem, recent MR prostate segmentation methods started to use features that are specifically designed for vision tasks, such as gradient [17], Haar-like wavelets [18], Histogram of Oriented Gradients (HOG) [19], SIFT [20], Local Binary Patterns (LBP) [21], and variance adaptive SIFT [14]. Compared to simple intensity features, these vision-based features show better invariance to illumination, and also provide some invariance to small rotation. In [22], authors showed that better prostate segmentations could be obtained by using the combination of these features.

One major limitation of the aforementioned handcrafted features is incapable of adapting to data at hand. That means the representation power and effectiveness of these features could vary across different kinds of image data. To deal with this limitation, the learning based feature representation methods [23, 24] are developed to extract latent information, which can be adapted to the data at hand. As one important type of feature learning methods, deep learning recently becomes a hot topic in machine learning [23], computer vision [25], and many other research fields including medical image analysis [26]. Compared with handcrafted features, which need expert knowledge for careful design and also lack sufficient generalization power to different domains, deep learning is able to automatically learn effective feature hierarchies from the data. Therefore, it draws an increasing interest in the research communities. For example, Vincent et al. [27] showed that the features learned by deep belief network and the stacked denoising auto-encoder beat the state-of-the-art handcrafted features for the digit classification problem in the MINST dataset. Farabet et al. [28] proposed to use convolutional network to produce feature representation, which is more powerful in the application of scene labeling than the engineered features, and also achieved the state-of-the-art performance. In the field of medical image analysis, Shin et al. [29] applied the stacked auto-encoders to organ identification in MR images, which shows the potential of deep learning method for application to medical images. In summary, compared with handcrafted features, deep learning has the following advantages: (1) Instead of designing effective features for a new task by trial and error, deep learning largely saves researchers' time by automating this process. Also, it is capable to exploit the complex feature patterns, which the manual feature engineering is not good at. (2) Unlike the handcrafted features, which are usually shallow in representation due to the difficulty of designing abstract high-level features, deep learning is able to learn the feature hierarchy in a layer-by-layer manner, by first learning the low-level features and then recursively building more comprehensive high-level features based on the previously learned low-level features. (3) When unsupervised pre-training is combined with supervised fine-tuning, the deep-learned features can be optimized for a certain task, such as segmentation, thus boosting the final performance.

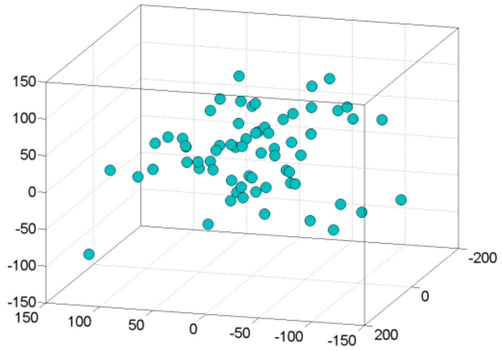


Fig. 2. The prostate shape distribution obtained from the PCA analysis.

B. *Our Contribution*

Motivated by the above factors, we propose to learn the hierarchical feature representation from MR prostate images by deep feature learning. These learned features are further integrated in a sparse patch matching framework to find the corresponding patches in the atlas images for label propagation. Finally, a deformable model is adopted to segment the prostate by combining the shape prior with the prostate likelihood map derived from sparse patch matching. The main contribution of our method lies in threefold:

- Instead of using handcrafted features, we propose to learn the latent feature representation from prostate MR images by the stacked sparse auto-encoder (SSAE) [30, 31], which includes an unsupervised pre-training step and also a task-related fine-tuning step.
- By using deep-learned features for measuring inter-patch

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2508280, IEEE Transactions on Medical Imaging

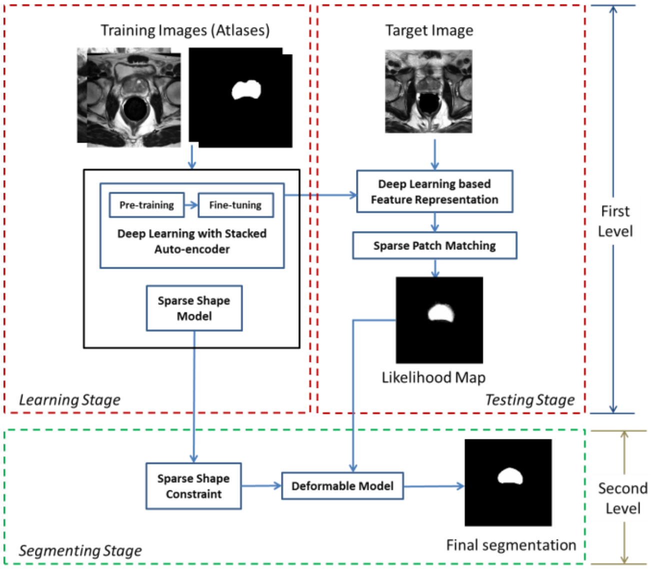


Fig. 3. The schematic description of our proposed segmentation framework.

similarity, a sparse patch matching method is proposed for finding the corresponding patches in the atlas images and then transferring their prostate labels from atlases to the new prostate image.

- A deformable model is adopted to further enforce a sparse shape constraint during segmentation, which aims to cope with the large variation existing in prostate shape space.

The proposed method has been extensively evaluated on the T2-weighted MR prostate image dataset, which contains 66 3D images. The manual prostate segmentations are provided by a radiation oncologist for the evaluation purpose. Experimental results show that the sparse patch matching with deep-learned features achieve better segmentation accuracy than using the handcrafted features, as well as the simple intensity features. Besides, compared to other state-of-the-art prostate segmentation methods, our method obtains competitive segmentation accuracy.

C. *Brief Outline of Our method*

The proposed MR prostate segmentation framework is composed of two levels (Fig. 3). The first level (two upper panels of Fig. 3) learns the deep feature representation and then applies sparse patch matching with the deep-learned features for deriving the prostate likelihood map. Based on the produced likelihood map, the second level (lower panel of Fig. 3) consists of a deformable model by enforcing the shape prior during the evolution of prostate segmentation.

The rest of the paper is organized as follows. In Section II, we present the stacked sparse auto-encoder for feature learning and the sparse patch matching framework for deriving the prostate likelihood map in the first level. Section III elaborates both the deformable model and the sparse shape model in the second level. Section IV evaluates the proposed segmentation method on the T2-weighted prostate MR dataset. Finally, conclusive remarks are presented in Section V.

II. First Level: Learning Deep Feature Representation and Sparse Patch Matching

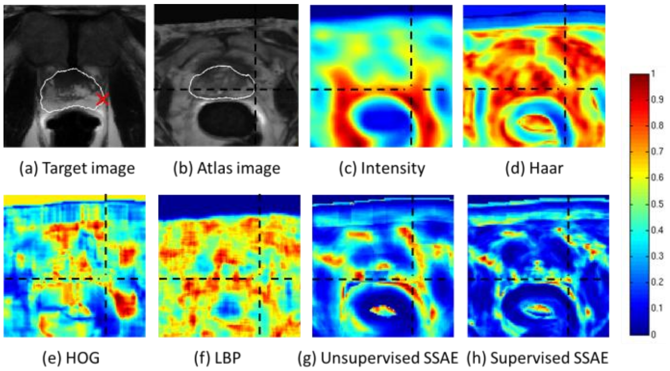
The goal of this level is to learn a latent feature representation for MR prostate images, and then use them to infer a likelihood map of prostate gland for a new image. To achieve this goal, two main stages (i.e., learning stage and testing stage) are conducted as illustrated in the two upper panels of Fig. 3. First, in the learning stage, the intrinsic feature hierarchy from MR prostate image patches is learned by using a deep learning framework, namely the stacked sparse auto-encoder (SSAE). Then, in the testing stage, each image patch from both atlas and target images is first represented by the features learned from the SSAE network. Then, these features are integrated into a sparse patch matching method for estimating the prostate likelihood map by transferring the label information from atlas images to the target image.

The organization of this section is as follows. In Section II.A, we first investigate the limitation of handcrafted features in MR prostate segmentation, and give our motivation of adopting deep learning features. Afterwards, we introduce the feature learning method in Section II.B, and the sparse patch matching in Section II.C, respectively.

A. *The Limitation of Handcrafted Features in MR Prostate Segmentation*

Since our sparse patch matching method belongs to multi-atlas based segmentation methods, in the following, we will illustrate the importance of features in such context. As briefly mentioned in the Introduction, good features in multi-atlas based segmentation should identify the correct correspondences between the target image and the atlas images. In computer vision, various handcrafted features, such as Haar features [18], HOG features [20] and Local Binary Patterns [21], have been proposed in different applications, with promising results such as in object detection of natural images. However, these features are not suitable for MR prostate images, as they are not invariant to both the inhomogeneity of MR images and the appearance variations of prostate gland.

To describe and compare the effectiveness of different features for identifying correspondences in two images, Fig. 4 shows a typical example by computing the similarity maps between one point (shown as red cross in Fig. 4(a)) in the target image (Fig. 4(a)) and all points in an aligned atlas image (Fig. 4(b)). The white contours in (a) and (b) show the prostate boundaries, and the black dashed cross in Fig. 4 (b) indicates the correct correspondence of the red-cross target point in the atlas image. The effectiveness of features can be reflected by the similarity map. If features are distinctive for correspondence detection, the similarity computed by using these features would be high for correct correspondences and low for incorrect correspondences. Fig. 4(c-f) shows the similarity maps computed using different handcrafted features, such as intensity patch features, Haar features, HOG features and LBP features, respectively. It is clear that none of these features could capture correct correspondence, as the similarity between the corresponding voxels indicated by the red crosses


(a) Target image (b) Atlas image (c) Intensity (d) Haar

(e) HOG (f) LBP (g) Unsupervised SSAE (h) Supervised SSAE

Fig. 4. The similarity maps computed between a reference voxel (red cross) in the target image (a) and all voxels in the atlas image (b) by the four handcrafted feature representations, i.e., intensity (c), Haar (d), HOG (e) and LBP (f), as well as the two deep learning feature representations, namely unsupervised SSAE (g) and the supervised SSAE (h). White contours indicate the prostate boundaries, and the black dashed crosses indicate the ground-truth point in (b), which is corresponding to the red cross in (a).

is low, compared to that of nearby voxels. This shows that the existing handcrafted features are insufficient in multi-atlas based segmentation for the MR prostate.

To relieve the limitation of handcrafted features, it is necessary to learn discriminant features adaptive to MR prostate images. To demonstrate the effectiveness of deep learning features, Fig. 4 (g) and (h) provide the similarity maps computed using the two kinds of deep learning features obtained by our proposed *unsupervised* and *supervised* stacked sparse auto-encoder (SSAE), respectively. Compared to similarity maps of handcrafted features, it is clear that the correct correspondence can be better identified with the deep learning features, especially for the supervised SSAE. In the following section, we will elaborate how these features could be adaptively learned from MR prostate images by SSAE.

B. Stacked Sparse Auto-Encoder (SSAE) for Learning the Latent Feature Representation

As illustrated in the previous section, it is necessary to learn the feature representation adaptive to the data, thus alleviating the need of labor-intensive feature engineering. To achieve this purpose, we introduce stacked sparse auto-encoder (SSAE) as a way to learn the latent feature representation from a collection of training prostate image patches. Stacked sparse auto-encoder is a deep learning architecture, which consists of basic feature learning layers, i.e., sparse auto-encoders (SAE). It is built by layer-wise stacking of sparse auto-encoders (Fig. 7). In the following paragraphs, we first introduce the auto-encoder as a basic feature learning algorithm. Then, we explain sparse auto-encoder, which imposes sparsity constraint for learning the robust shallow feature representations. Finally, we elaborate how to learn deep feature hierarchy by stacking multiple sparse auto-encoders layer-wisely.

1) Basic Auto-Encoder

Serving as the fundamental component for SSAE, the basic auto-encoder (AE) trains a feed-forward non-linear neural network, which contains three layers, i.e., input layer, hidden layer, and output layer, as illustrated in Fig. 5. Each layer is represented by a number of nodes. Blue nodes on the left and right sides of Fig. 5 indicate the input and output layers, respectively, and green nodes indicate the hidden layer. Nodes in the two neighboring layers are fully connected, which means that each node in the previous layer can contribute to any node in the next layer. Basically, AE consists of two steps, namely encoding and decoding. In the encoding step, AE encodes the input vector into a concise representation through connections between input and hidden layers. In the decoding step, AE tries to reconstruct the input vector from the encoded feature representation in the hidden layer. The goal of AE is to find a concise representation of input data, which could be used for the purpose of best reconstruction. Since we are interested in the representation of image patches, in this application the input to AE is an image patch, which is concatenated as a vector. In the training stage, given a set of training patches $\boldsymbol{X} = \{\boldsymbol{x}_i \in \mathbb{R}^L, i = 1, \ldots, N\}$, where N and L are the number and the dimension of training patches, respectively, AE automatically learns the weights of all connections in the network by minimizing the reconstruction error in Eq. (1).

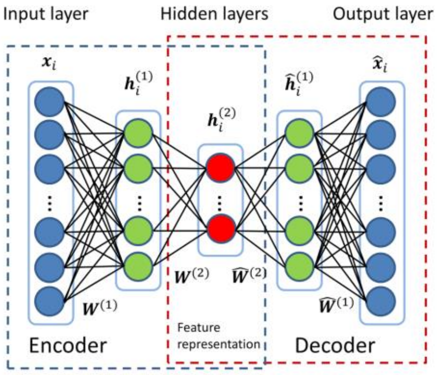


Fig. 5. Construction of the basic AE.

$$\operatorname{argmin}_{\boldsymbol{W},\boldsymbol{b},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{b}}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \left(\widehat{\boldsymbol{W}}\left(\sigma(\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b})\right) + \widehat{\boldsymbol{b}}\right) \right\|_2^2 \quad (1)$$

where $\boldsymbol{W}, \boldsymbol{b}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}$ are the parameters in the AE network, and $\sigma(\boldsymbol{a}) = (1 + \exp(-\boldsymbol{a}))^{-1}$. Given an input vector $\boldsymbol{x}_i$, AE first encodes it into the concise representation $\boldsymbol{h}_i = \sigma(\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b})$, where $\boldsymbol{h}_i$ is the responses of $\boldsymbol{x}_i$ at the hidden nodes, and the dimension of $\boldsymbol{h}$ equals to the number of nodes in the hidden layer. In the next step, AE tries to decode the original input from the encoded representation, i.e., with $\widehat{\boldsymbol{W}}\boldsymbol{h}_i + \widehat{\boldsymbol{b}}$. To learn effective features for the input training patches, AE requires that the dimension of the hidden layer is less than that of the input layer. Otherwise, the minimization of Eq. (1) would lead to trivial solutions, e.g., identity transformation. Studies [32] have also shown that the basic AE learns very similar features as PCA.

Once the weights $\{\boldsymbol{W}, \boldsymbol{b}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}\}$ have been learned through the training patches, in the testing stage AE could efficiently obtain a concise feature representation for a new image patch $\boldsymbol{x}_{\text{new}}$ by a forward passing step, i.e., $\boldsymbol{h}_{\text{new}} = \sigma(\boldsymbol{W}\boldsymbol{x}_{\text{new}} + \boldsymbol{b})$.

2) *Sparse Auto-Encoder*

Rather than limiting the dimension of hidden layer (i.e., feature representation), an alternative could be imposing regularization on the hidden layer. Sparse auto-encoder (SAE) falls into this category. Instead of requiring the dimension of hidden layer less than that of the input layer, SAE imposes sparsity regularization on the responses of hidden nodes (i.e., $\boldsymbol{h}$) to avoid the problem of trivial solutions suffered by the basic AE. Specifically, SAE enforces the average response of each hidden node over the training set to be infinitesimal, i.e., $\rho^j = \sum_{i=1}^{N} \boldsymbol{h}_i^j \approx \rho$, where $\boldsymbol{h}_i^j$ is the response of the i-th training input at hidden node j, and ρ is a very small constant. In this way, to balance both the reconstruction power and the sparsity of the hidden layer, only a few useful hidden nodes could have responses for each input, thus forcing the SAE network to learn sparse feature representation of the training data. Mathematically, we can extend Eq. (1) to derive the objective function of SAE by adding a sparsity constraint term shown below:

$$\operatorname{argmin}_{\boldsymbol{W},\boldsymbol{b},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{b}}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \left(\widehat{\boldsymbol{W}} \left(\sigma(\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b}) \right) + \widehat{\boldsymbol{b}} \right) \right\|_2^2 + \delta \sum_{j=1}^{M} KL(\rho|\rho^j) \qquad (2)$$

$$KL(\rho|\rho^j) = \rho \log \frac{\rho}{\rho^j} + (1-\rho) \log \frac{1-\rho}{1-\rho^j}$$

where δ is a parameter to balance between reconstruction and sparsity terms, and M is the number of hidden nodes. $KL(\rho|\rho^j)$ is the Kullback-Leibler divergence between two Bernoulli distributions with probability ρ and ρ^j. As we can see, the sparsity term is minimized only when ρ^j is close to ρ for every hidden node j. Since ρ is set to be a small constant, minimizing Eq. (2) could lead to the sparse responses of hidden nodes, hence the sparsity of learned feature representation.

3) *Stacked Sparse Auto-Encoder*

By using SAE, we can learn the low-level features (such as Gabor-like features as shown in Fig. 6) from the original data (MR image patches). However, low-level features are not enough due to large appearance variations of the MR prostate. It is necessary to learn abstract high-level features, which could also be invariant to the inhomogeneity of MR images. Motivated by the human perception, which constitutes a deep network to describe concepts in a hierarchical way using multiple levels of abstraction, we recursively apply SAE to learn more abstract/high-level features based on the features learned from the low-level. This multi-layer SAE model is referred to as a stacked sparse auto-encoder (SSAE), which stacks multiple SAEs on top of each other for building deep hierarchies.

Fig. 7 shows a typical SSAE with R stacked SAEs. Let $\boldsymbol{W}^{(r)}$, $\boldsymbol{b}^{(r)}$, $\widehat{\boldsymbol{W}}^{(r)}$ and $\widehat{\boldsymbol{b}}^{(r)}$ denote the connection weights and intercepts between the input layer and hidden layer, and between the hidden layer and output layer in the r-th SAE, respectively. In the encoding part of the SSAE, the input vector $\boldsymbol{x}_i$ is first encoded by the first SAE for obtaining the low-level representation $\boldsymbol{h}_i^{(1)}$, i.e., $\boldsymbol{h}_i^{(1)} = \sigma(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})$. Then, the low-level representation $\boldsymbol{h}_i^{(1)}$ of the first SAE is considered as the input vector to the next SAE, which encodes it into higher level representation $\boldsymbol{h}_i^{(2)}$, i.e., $\boldsymbol{h}_i^{(2)} = \sigma(\boldsymbol{W}^{(2)}\boldsymbol{h}_i^{(1)} + \boldsymbol{b}^{(2)})$. Generally, the r-th level representation $\boldsymbol{h}_i^{(r)}$ can be obtained by a recursive encoding procedure $\boldsymbol{h}_i^{(r)} = \sigma(\boldsymbol{W}^{(r)}\boldsymbol{h}_i^{(r-1)} + \boldsymbol{b}^{(r)})$ with $\boldsymbol{h}_i^{(0)} = \boldsymbol{x}_i$. Similarly, the decoding step of SSAE recursively reconstructs the input of each SAE. In this example, SSAE first reconstructs the low-level representation $\widehat{\boldsymbol{h}}_i^{(r-1)}$ from the high-level representation $\widehat{\boldsymbol{h}}_i^{(r)}$, i.e., $\widehat{\boldsymbol{h}}_i^{(r-1)} = \widehat{\boldsymbol{W}}^{(r)}\widehat{\boldsymbol{h}}_i^{(r)} + \widehat{\boldsymbol{b}}^{(r)}$ with $\widehat{\boldsymbol{h}}_i^{(R)} = \boldsymbol{h}_i^{(R)}$ for $r = R, \ldots, 2$. Then, using the reconstructed low-level representation $\widehat{\boldsymbol{h}}_i^{(1)}$, the original input vector could be estimated, i.e., $\widehat{\boldsymbol{x}}_i = \widehat{\boldsymbol{W}}^{(1)}\widehat{\boldsymbol{h}}_i^{(1)} + \widehat{\boldsymbol{b}}^{(1)}$.

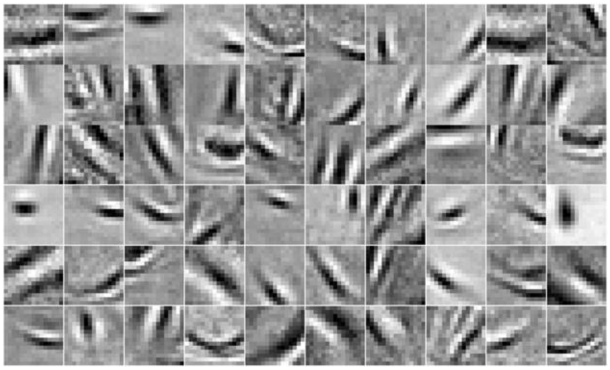


Fig. 6. The low-level feature representation learned from the SAE. Here, we reshape each row in $\boldsymbol{W}$ into the size of image patch, and only visualize its first slice as an image filter.

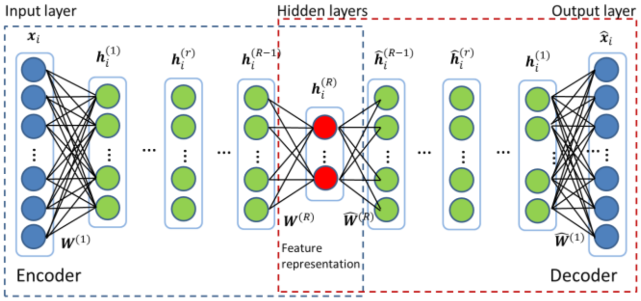
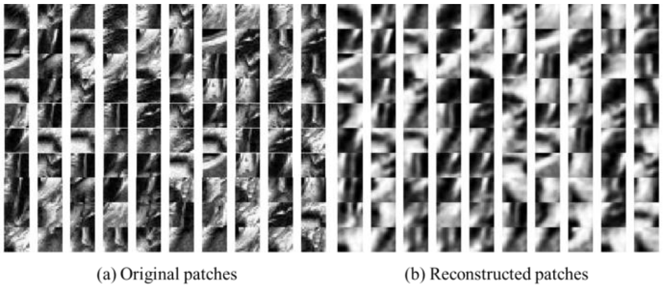

Input layer
Hidden layers
Output layer
$\boldsymbol{x}_i$
$\boldsymbol{h}_i^{(1)}$
$\boldsymbol{h}_i^{(r)}$
$\boldsymbol{h}_i^{(R-1)}$
$\boldsymbol{h}_i^{(R)}$
$\widehat{\boldsymbol{h}}_i^{(R-1)}$
$\widehat{\boldsymbol{h}}_i^{(r)}$
$\widehat{\boldsymbol{h}}_i^{(1)}$
$\widehat{\boldsymbol{x}}_i$
$\boldsymbol{W}^{(1)}$
$\boldsymbol{W}^{(R)}$
$\widehat{\boldsymbol{W}}^{(R)}$
$\widehat{\boldsymbol{W}}^{(1)}$
Feature representation
Encoder
Decoder

Fig. 7. Construction of the *unsupervised* SSAE with R stacked SAEs.

After stacking multiple SAEs together by feeding the output layer from the low-level SAE as the input layer of a high-level SAE, SSAE is able to extract more useful and general high-level features. In the optimization of SSAE, this deep architecture is first pre-trained in an unsupervised layer-wise manner and then fine-tuned by back propagation. Since the aforementioned SSAE network is trained based only on the original image patches, without using the supervised label information, it is denoted as the *unsupervised SSAE*. Fig. 8 shows some typical prostate image patches and their reconstructions by the unsupervised SSAE with $R = 4$.



(a) Original patches (b) Reconstructed patches

Fig. 8. Typical prostate image patches (a) and their reconstructions (b) by using the unsupervised SSAE with four stacked SAEs.

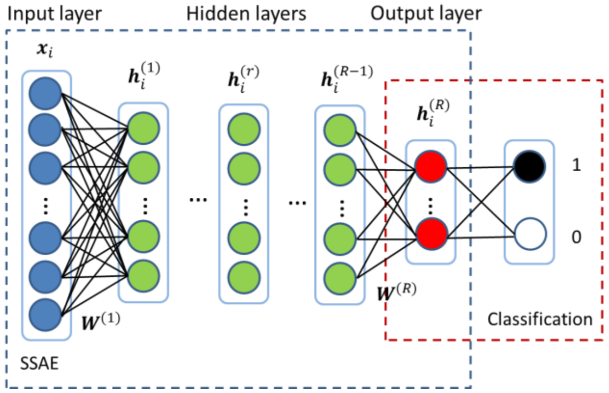
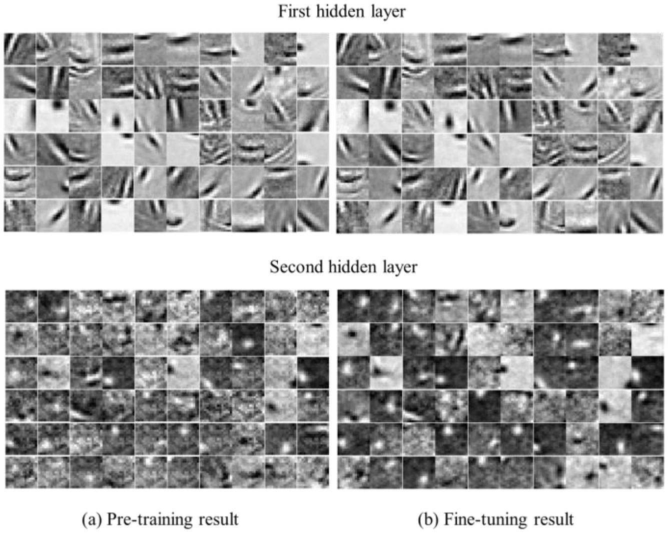


Fig. 9. Construction of the supervised SSAE with a classification layer, which fine-tunes the SSAE with respect to the task of voxel-wise classification between prostate (label = 1) and background (label = 0).

However, since the unsupervised SSAE trains the whole network on the unlabeled data, the high-level features learned from unsupervised SSAE are only data-adaptive, that is, not necessarily discriminative to separate prostate and background voxels. To make the learned feature representation discriminative [33, 34], the supervised fine-tuning is often adopted by stacking another classification output layer on the top of the encoding part of the SSAE, as shown in red dashed box of Fig. 9. This top layer is used to predict the label likelihood of the input data $\boldsymbol{x}_i$ by using the features learned from the most high-level representation $\boldsymbol{h}_i^{(R)}$. The number of nodes in the classification output layer equals to the number of labels (i.e., “1” denotes prostate, and “0” denotes background). Using the optimized parameters from the pre-training of SSAE as initialization, the entire neural network (Fig. 9) can be further fine-tuned by back-propagation to maximize the classification performance. This tuning step is referred to as the supervised fine-tuning, in contrast with the unsupervised fine-tuning mentioned before. Accordingly, the entire deep network is referred to as the *supervised SSAE*. Fig. 10 gives a visual illustration of typical feature representations of the first and second hidden layers learned by a four-layer supervised SSAE based on the visualization method in [35]. Here, Figs. 10 (a) and (b) show the visualization of 60 units obtained from the first and second hidden layers under unsupervised pre-training (with unlabeled image patches) and supervised fine-tuning (with labeled image patches), respectively. It can be seen that higher hidden layer tends to be more affected by the classification layer we introduced.



First hidden layer

Second hidden layer

(a) Pre-training result (b) Fine-tuning result

Fig. 10. Visualization of typical feature representations of the first hidden layer (first row) and second hidden layer (second row) for the unsupervised pre-training (a) and supervised fine-tuning (b), respectively.

After learning all the parameters $\{\boldsymbol{W}^{(r)}, \widehat{\boldsymbol{W}}^{(r)}, \boldsymbol{b}^{(r)}, \widehat{\boldsymbol{b}}^{(r)}\}$ of SSAE ($r = 1, \ldots, R$), where R denotes the number of stacked SAEs, the high-level representations of a new image patch $\boldsymbol{x}_{\text{new}}$ can be efficiently obtained by a recursive forwarding pass, i.e., $\boldsymbol{h}_{\text{new}}^{r} = \sigma\left(\boldsymbol{W}^{(r)}\boldsymbol{h}_{\text{new}}^{(r-1)} + \boldsymbol{b}^{(r)}\right)$ with $\boldsymbol{h}_{\text{new}}^{0} = \boldsymbol{x}_{\text{new}}$ for $r = 1, \ldots, R$. The final high-level representation $\boldsymbol{h}_{\text{new}}^{R}$ will be used as features to guide the sparse patch matching (Section II.C), and propagate labels from atlas images to the target image for estimating the prostate likelihood map.

C. Sparse Patch Matching with the Deep Learning Features

Before sparse patch matching, all atlas images are registered to the target image. This registration includes two steps. First, linear registration is applied for initial alignment, with the guidance from the landmarks automatically-detected around the prostate region [36]. Then, the free-form deformation (FFD) [37] is further adopted to the linearly aligned images for deformable registration.

After learning the SSAE networks (either in unsupervised or supervised manner), each new image patch in the testing stage can be encoded as a high-level feature vector (i.e., the last hidden layer of the SSAE). These features can be fed into a segmentation framework for labeling voxels as either prostate or background. As one of the popular segmentation frameworks, multi-atlas based segmentation demonstrates its effectiveness on dealing with image variations in different applications [38, 39]. However, traditionally the multi-atlas based segmentation adopts only the intensity or handcrafted features for measuring the similarity between different local patches, or computing the weights of different patches during label propagation. Since MR prostate images exhibit large structural and appearance variations, we propose to incorporate the deep learning features, instead of the conventional handcrafted features, into the multi-atlas segmentation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2508280, IEEE Transactions on Medical Imaging

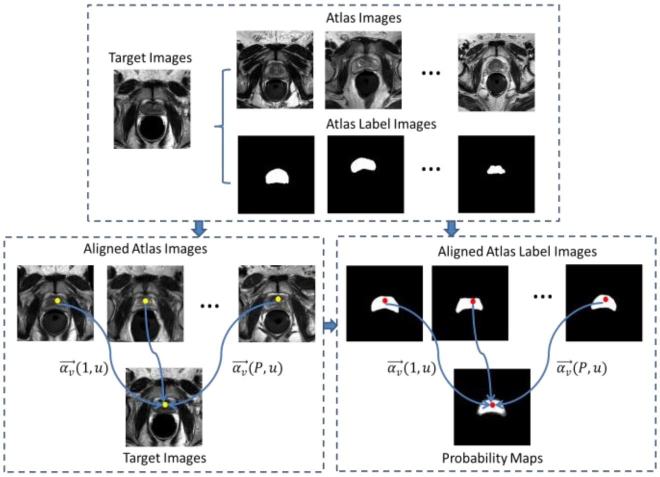

Atlas Images
Target Images
...
Atlas Label Images
...
Aligned Atlas Images
...
$\overrightarrow{\alpha_v}(1,u)$
$\overrightarrow{\alpha_v}(P,u)$
Target Images
Aligned Atlas Label Images
...
$\overrightarrow{\alpha_v}(1,u)$
$\overrightarrow{\alpha_v}(P,u)$
Probability Maps

Fig. 11. The schematic description of sparse patch matching

framework. As the features extracted by the deep learning methods are usually more robust and discriminative, the performance of multi-atlas segmentation can be improved at the same time. Fig. 11 gives the general description of our multi-atlas based method, called sparse patch matching. In this method, instead of computing pair-wise intensity similarity as a matching weight, we propose to select only a small number of similar atlas patches by sparse representation, which is more robust to outliers. In the following, we give the detailed description for our sparse patch matching method.

In order to estimate the prostate likelihood map of a target image I_s, we first align all the atlas images $\{I_p, p = 1, \dots, P\}$ and their label maps $\{G_p, p = 1, \dots, P\}$ onto the target image I_s. Then, to determine the prostate likelihood of a particular voxel v in the target image I_s, we first extract the image patch centered at voxel v from the target image, and then all image patches within a certain searching neighborhood $\mathbb{N}(v)$ across all the aligned atlas images.

Next, the deep learned feature representations for those extracted intensity patches are obtained through the encoding procedure of the learned SSAE as introduced in Section II.B. Denote $\boldsymbol{f}_s(v)$ as the deep learning features for the target image patch at point v, and denote $\boldsymbol{A}_v$ as the feature matrix resulted from column-wise combination of deep learning features of atlas patches, i.e., $\boldsymbol{A}_v = [\boldsymbol{f}_p(u) | p = 1, \dots, P; u \in \mathbb{N}(v)]$. To estimate the prostate likelihood $Q_s(v)$ for voxel v in the target image I_s, we linearly combine the label of each voxel $u \in \mathbb{N}(v)$ from each atlas image I_p with a weighting vector $\boldsymbol{\alpha}_v = [\alpha_v(p,u)]_{u=1,\dots,|\mathbb{N}(v)|; p=1,\dots,P}$ as follows:

$$Q_s(v) = \frac{\sum_{p=1}^{P} \sum_{u \in \mathbb{N}(v)} \alpha_v(p,u) \times G_p(u)}{\sum_{p=1}^{P} \sum_{u \in \mathbb{N}(v)} \alpha_v(p,u)} \tag{3}$$

According to Eq. (3), it is easy to see that the robustness and accuracy of prostate likelihood estimation depends on how well the weighting vector $\boldsymbol{\alpha}_v$ is determined. In the literature, different weight estimation methods have been proposed [40, 41]. Most multi-atlas based segmentation methods directly compute $\boldsymbol{\alpha}_v$ as the pair-wise similarity between intensity patches, such as using the Euclidean distance. In our method, we compute the weighting vector $\boldsymbol{\alpha}_v$ different from the previous methods in respect to the following two aspects. *First*, instead of using the intensity or handcrafted features, the high-level features are learned from the deep learning architecture. *Second*, with the help of recently proposed sparse representation method [4], we enforce sparsity constraint upon the weighting vector $\boldsymbol{\alpha}_v$. In this way, we seek for the best representation of the target patch using a limited set of similar atlas patches. Mathematically, the optimization of $\boldsymbol{\alpha}_v$ can be formulated as the sparse representation problem below:

$$\begin{gathered}\boldsymbol{\alpha}_v = \arg\min_{\boldsymbol{\alpha}_v} \frac{1}{2} \|\boldsymbol{f}_s(v) - \boldsymbol{A}_v \boldsymbol{\alpha}_v\|_2^2 + \eta \|\boldsymbol{\alpha}_v\|_1 \\ s.t. \boldsymbol{\alpha}_v \geq 0\end{gathered} \tag{4}$$

The first term is the data fitting term, which measures the difference between the target feature vector $\boldsymbol{f}_s(v)$ and the linearly combined feature representation $\boldsymbol{A}_v \boldsymbol{\alpha}_v$ from all atlas image patches. The second term is the sparsity term, which attributes to the sparsity property of the weighting vector $\boldsymbol{\alpha}_v$. η controls the strength of sparsity constraint on the weighting vector $\boldsymbol{\alpha}_v$. If η is larger, the number of non-zero elements in $\boldsymbol{\alpha}_v$ will be smaller. In this way, only a few patches in patch dictionary $\boldsymbol{A}_v$ will be selected to reconstruct the target features $\boldsymbol{f}_s(v)$ in a non-parameter fashion, thus reducing the risk of including those misleading atlas patches in the likelihood estimation.

Based on the derived weighting vector $\boldsymbol{\alpha}_v$, the prostate likelihood $Q_s(v)$ for a target point v can be estimated by Eq. (4). Since the weighting vector $\boldsymbol{\alpha}_v$ is sparse, the prostate likelihood $Q_s(v)$ is finally determined by the linear combination of labels corresponding to atlas patches with non-zero elements in vector $\boldsymbol{\alpha}_v$. After estimating the prostate likelihood for all voxels in the target image I_s, a likelihood map Q_s is generated, which can be used to robustly locate the prostate region (as shown in Fig. 11). Usually, a simple thresholding or level set method [42, 43] can be applied to binarize the likelihood map for segmentation. However, since each voxel in the target image is independently estimated in the multi-atlas segmentation method, the final segmentation could be weird as no shape prior is considered. In order to robustly and accurately estimate the final prostate region from the prostate likelihood map, it is necessary to take into account the prostate shape prior during the segmentation.

III. Second Level: Data and Shape Driven Deformable Model

The main purpose of this section is to segment the prostate region based on the prostate likelihood map estimated in the previous section. The likelihood map can be used in two aspects for deformable model construction. First, the initialization of deformable model can be easily built by thresholding the likelihood map. In this way, the limitation of model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2508280, IEEE Transactions on Medical Imaging

initialization problem in the traditional deformable segmentation can be naturally relieved. Second, the likelihood map can be used as the appearance force to drive the evolution of deformable model. Besides, in order to deal with the large inter-patient shape variation, we propose to use sparse shape prior for deformable model regularization. In the following paragraphs, we first introduce the sparse shape composition as a non-parametric shape modeling method. Then, we present the optimization of our deformable model by jointly considering both shape and appearance information. Finally, the proposed deformable segmentation method is summarized.

A. Shape Prior by Sparse Shape Composition

Here, our deformable model is represented by a 3D surface, which is composed of K vertices $\{\boldsymbol{d}_k | k = 1, \ldots, K\}$. After concatenating these K vertices $\{\boldsymbol{d}_k | k = 1, \ldots, K\}$ into a vector $\boldsymbol{d}$, each deformable model can be represented as a shape vector with length of $3 \cdot K$. Let $\boldsymbol{D}$ denotes a large shape dictionary that includes prostate shape vectors of all training subjects. Each column of shape dictionary $\boldsymbol{D}$ corresponds to the shape vector of one subject. The shape dictionary can be used as a shape prior to constrain the deformable model in a learned shape space. Instead of assuming the Gaussian distribution of shapes and then simply using PCA for shape modeling as in the Active Shape Model [44], we adopt a recently proposed method, named sparse shape composition [45], for shape modeling. In the sparse shape composition, the shapes are sparsely represented by shape instances in the shape dictionary without the need of Gaussian assumption. Specifically, given a new shape vector $\boldsymbol{d}$ and shape dictionary $\boldsymbol{D}$, sparse shape composition method reconstructs shape vector $\boldsymbol{d}$ as the sparse representation of shape dictionary $\boldsymbol{D}$ by minimizing the following objective function:

$$(\boldsymbol{\varepsilon}, \psi) = \arg\min_{\boldsymbol{\varepsilon},\psi} \|\psi(\boldsymbol{d}) - \boldsymbol{D}\boldsymbol{\varepsilon}\|_2^2 + \mu\|\boldsymbol{\varepsilon}\|_1 \qquad (5)$$

where $\psi(\boldsymbol{d})$ denotes the target shape $\boldsymbol{d}$ that is affine aligned onto the mean shape space of shape dictionary $\boldsymbol{D}$. $\boldsymbol{\varepsilon}$ indicates the sparse coefficient for the linear shape combination. Once $(\boldsymbol{\varepsilon}, \psi)$ are estimated by Eq. (5), the regularized shape can be computed by $\psi^{-1}(\boldsymbol{D}\boldsymbol{\varepsilon})$, where ψ^{-1} is the inverse affine transform of ψ.

B. Optimization of Deformable Model Method

For each target image, the segmentation task is formulated as the deformable model optimization problem. During the optimization procedure, each vertex of deformable model d_k is driven iteratively by the information from both prostate likelihood map and shape model until converged at the prostate boundaries. Mathematically, the evolution of the deformable model can be formulated as the minimization of an energy function, which contains a data energy E_{data} and a shape energy E_{shape} as in Eq. (6):

$$E = E_{\text{data}} + \lambda E_{\text{shape}} \qquad (6)$$

The data term E_{data} is used to attract the 3D surface towards the object boundary based on the likelihood map. Specifically, each vertex $\boldsymbol{d}_k$ is driven by the force related to the gradient vector of prostate likelihood map. Denote $\overrightarrow{\nabla Q}_s(\boldsymbol{d}_k)$ as the gradient vector at vertex $\boldsymbol{d}_k$ in the prostate likelihood map, and $\vec{n}_s(\boldsymbol{d}_k)$ as the normal vector on the vertex $\boldsymbol{d}_k$ of surface. When vertex $\boldsymbol{d}_k$ deforms exactly to the prostate boundary and also its normal direction aligns with the gradient direction of prostate boundary, the local matching term $\langle \overrightarrow{\nabla Q}_s(\boldsymbol{d}_k), \vec{n}_s(\boldsymbol{d}_k) \rangle$ is maximized. In this case, we formulate to minimize the data energy E_{data} as:

$$E_{\text{data}} = -\sum_k \langle \overrightarrow{\nabla Q}_s(\boldsymbol{d}_k), \vec{n}_s(\boldsymbol{d}_k) \rangle \qquad (7)$$

Since all the vertices on the deformable model are jointly evolved during the deformation, the matching of the deformable model with prostate boundary will be robust to possible incorrect likelihood on some vertices, as well as inconsistency between neighboring vertices.

The shape term E_{shape} is used to encode the geometric property of prostate shape based on the estimated coefficient $\boldsymbol{\varepsilon}$ and the transformation φ in Eq. (5). Specifically, the shape term is formulated as below:

$$E_{\text{shape}} = \|\boldsymbol{d} - \psi^{-1}(\boldsymbol{D}\boldsymbol{\varepsilon})\|_2^2 + \beta \sum_k \left\| \boldsymbol{d}_k - \frac{\sum_{d_j \in \mathbb{N}(\boldsymbol{d}_k)} \boldsymbol{d}_k}{\sum_{d_j \in \mathbb{N}(\boldsymbol{d}_k)} \boldsymbol{1}} \right\|_2^2 \qquad (8)$$

where the first term constrains the deformed shape $\boldsymbol{d}$ to be close to the regularized shape $\psi^{-1}(\boldsymbol{D}\varepsilon)$ by the sparse shape composition, and the second term imposes the smoothness constraint on shape, which prevents large deviations between each vertex $\boldsymbol{d}_k$ and the center of its neighboring vertices $d_j \in \mathbb{N}(\boldsymbol{d}_k)$.

By combining Eq. (7) and Eq. (8) into Eq. (6), the vertices on the deformable model are iteratively driven towards the prostate boundary while constraining the shape in a non-parametric shape space.

C. Summary of the Proposed Deformable Segmentation Method

Generally, the overall optimization of deformable model can be summarized as an expectation-maximization (EM) algorithm, which minimizes the data energy and shape energy alternatively. Given the target image I_s, the initial deformable surface $\boldsymbol{d}^0$ is estimated by solving the sparse learning problem in Eq. (5). Then the "M" step and "E" step are alternatively executed as follows. Based on the likelihood map generated from stacked sparse auto-encoder and sparse patch matching, in the "M" step, the deformable model $\boldsymbol{d}^t$ is first evolved to minimize the data energy function E_{data} (Eq. (7)). Then, in the "E" step, the parameters $(\boldsymbol{\varepsilon}^t, \psi^t)$ is estimated for the shape refinement by solving Eq. (5), and the deformable model $\boldsymbol{d}^t$ is further refined by minimizing the shape energy (Eq. (8)) with the computed parameters $(\boldsymbol{\varepsilon}^t, \psi^t)$. After T iterations of the above EM step, the output shape $\boldsymbol{d}^T$ is converted to a binary label map G_s, which gives the final segmentation result of the target image I_s.

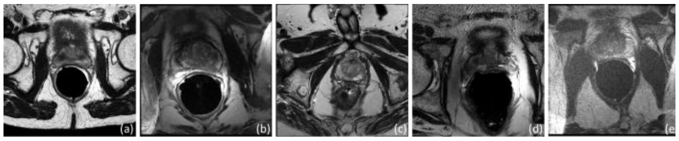


Fig. 12. Five typical T2-weighted MR prostate images acquired from different scanners, showing large variations of both prostate appearance and shape, especially for the cases with or without using the endorectal coils.

IV. Experiments and Analysis

A. *Materials and Parameter Settings*

We evaluate our method on the dataset, which includes 66 T2-weighted MR images from the University of Chicago Hospital. The images are acquired with 1.5T magnetic field strength from different patients under different MR image scanners (34 images from Philips Medical Systems and 32 images from GE Medical Systems). Under this situation, the difficulty for the segmentation task increases since both shape and appearance differences are large. In Fig. 12, images (b) and (e) were acquired from a GE MRI scanner, while the other three were acquired from a Philips MRI scanner. As shown in Fig. 12, image (c) was obtained without the endorectal coil. It has different prostate shape with other four images acquired with the endorectal coil. Besides, the prostate appearance suffers from the inhomogeneity (as in (b) and (d)) and noises (as in (d) and (e)), which further produce large variations. The image dimension and spacing are different from image to image. For example, the image dimension varies from $256 \times 256 \times 28$ to $512 \times 512 \times 30$. The image spacing varies from $0.49 \times 0.49 \times 3$ mm to $0.56 \times 0.56 \times 5$ mm. The manual delineation of the prostate in each image is provided by a clinical expert as the ground truth for quantitative evaluation. As the preprocessing of the dataset, the bias field correction [46] and histogram matching [47] are applied to each image successively. We adopted the two-fold cross-validation. Specifically, in each case, the images of one fold are used for training the models, while the images of other fold are used for testing the performance.

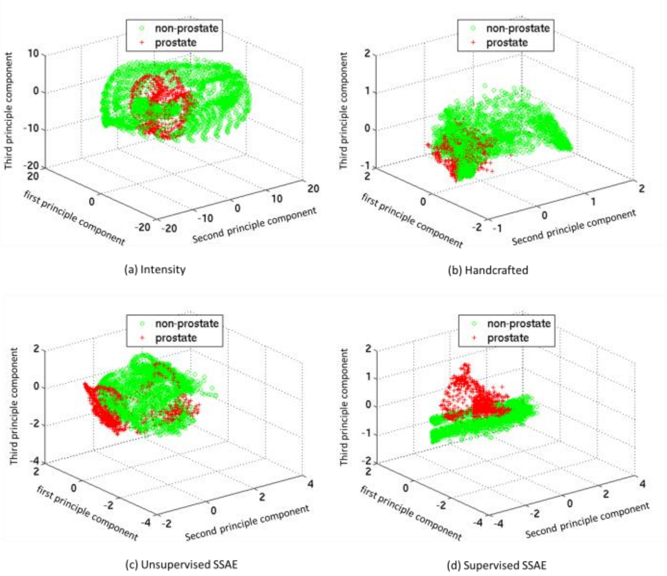
The parameters for deep feature learning are listed below. The patch size is $15 \times 15 \times 9$. The number of layers in SSAE framework is 4. The number of nodes in each layer of SSAE is 800, 400, 200, 100, respectively. Thus, the deep learning features have the dimensionality of 100. The target activation ρ for the hidden units is 0.15. The sparsity penalty β is 0.1. The Deep Learning Toolbox [48] is used for training our SSAE framework.

The searching neighborhood $\mathbb{N}(v)$ is defined as the $7 \times 7 \times 7$ neighborhood centered at voxel v. For sparse patch matching, the parameter η in Eq. (4), which controls the sparsity constraint, is 0.001. For the final deformable model segmentation, the parameter λ in Eq. (6), which weights the shape energy during deformation, is 0.5, and the parameter μ in Eq. (5) is 0.5.

Table I
Definition of Evaluation Measurement

Dice ratio	$\frac{2 \cdot V(S \cap F)}{V(S) + V(F)}$
Precision	$\frac{V(S \cap F)}{V(F)}$
Hausdorff distance	$\max\big(H(e_S, e_F), H(e_S, e_F)\big),$ $H(e_S, e_F) = \max_{d_i \in e_S}\left\{\min_{d_j \in e_F} \text{dist}(d_i, d_j)\right\}$
Average surface distance	$\frac{1}{2}\left(\frac{\sum_{d_i \in e_S} \min_{d_j \in e_F} \text{dist}(d_i, d_j)}{\sum_{d_i \in e_S} 1} + \frac{\sum_{d_j \in e_F} \min_{d_i \in e_S} \text{dist}(d_j, d_i)}{\sum_{d_j \in e_F} 1}\right)$

S: ground truth segmentation; F: automatic segmentation; V: volume size;
e_S: surfaces of ground-truth segmentation;
e_F: surface of automatic segmentation;
d_i: vertices on the surfaces e_S; d_j: vertices on the surfaces e_F;
$\text{dist}(d_i, d_j)$: Euclidean distance between vertices d_i and d_j



(a) Intensity
(b) Handcrafted
(c) Unsupervised SSAE
(d) Supervised SSAE

Fig. 13. Distributions of voxel samples by using four types of features: (a) intensity, (b) handcrafted, (c) unsupervised SSAE, and (d) supervised SSAE. Red crosses and green circles denote prostate and non-prostate voxel samples, respectively.

B. *Evaluation Criteria*

Given the ground-truth segmentation S and the automatic segmentation F, the segmentation performance is evaluated by four metrics: Dice ratio, precision, Hausdorff distance, and average surface distance. Dice ratio and precision measure the overlap between two segmentations. Hausdorff distance and average surface distance measures the boundary distance between two surfaces of segmentation. Detailed definitions are given as Table I, where V indicates the volume size, e_S and e_F are the surfaces for the ground-truth and automatic segmentations, respectively, and $\text{dist}(d_i, d_j)$ denotes the Euclidean distance between vertices d_i and d_j.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2508280, IEEE Transactions on Medical Imaging

TABLE II
MEAN AND STANDARD DEVIATION OF QUANTITATIVE RESULTS FOR MR PROSTATE SEGMENTATION WITH DIFFERENT FEATURE REPRESENTATIONS. BEST PERFORMANCE IS INDICATED BY BOLD FONT.

Method	Intensity	Haar	HOG	LBP	Handcraft	Unsupervised SSAE	Supervised SSAE
Dice (%)	85.3±6.2	85.6±4.9	85.7±4.9	85.5±4.3	85.9±4.5	86.7±4.4	**87.1±4.2**
	(1.1e-04)	(2.2e-05)	(6.1e-05)	(5.5e-05)	(7.0e-06)	(2.1e-01)	(NA)
Precision (%)	85.1±8.3	85.9±8.5	85.3±8.7	83.7±7.7	87.3±7.4	**87.3±7.3**	87.1±7.3
	(1.9e-03)	(3.63e-02)	(4.5e-03)	(1.3e-06)	(7.0e-01)	(7.3e-01)	(NA)
Hausdorff	8.68±4.24	8.50±2.86	8.51±2.69	8.59±2.38	8.55±2.91	8.65±2.69	**8.12±2.89**
	(1.8e-01)	(1.9e-01)	(1.6e-01)	(1.1e-01)	(1.5e-01)	(6.3e-02)	(NA)
ASD	1.87±0.93	1.76±0.52	1.74±0.50	1.75±0.44	1.77±0.54	1.68±0.49	**1.66±0.49**
	(6.0e-03)	(8.9e-03)	(3.1e-02)	(2.5e-02)	(5.0e-04)	(5.8e-01)	(NA)

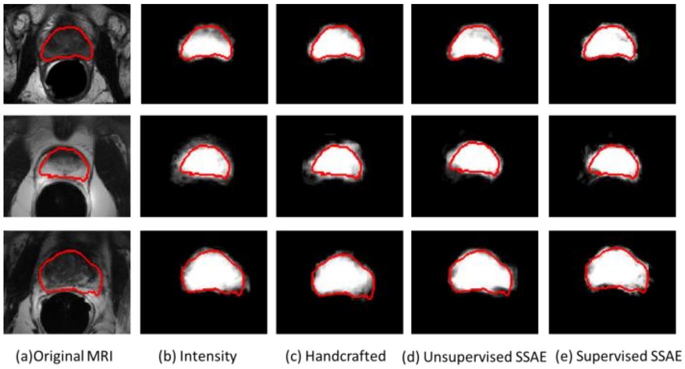


Fig. 14. (a) Typical slices of T2 MR images with manual segmentations. The likelihood maps produced by sparse patch matching with four feature representations: (b) intensity patch, (c) handcrafted, (d) unsupervised SSAE, and (e) supervised SSAE. Red contours indicate the manual ground-truth segmentations.

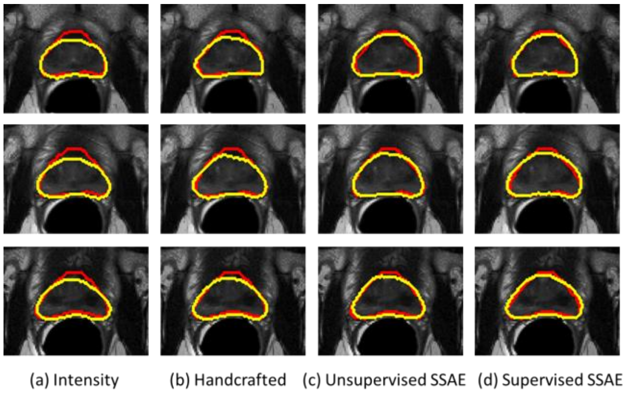


Fig. 15. Typical prostate segmentation results of the same patients produced by four different feature representations: (a) intensity, (b) handcrafted, (c) unsupervised SSAE, and (d) supervised SSAE. Three rows show the results for three different slices of the same patient, respectively. Red contours indicate the manual ground-truth segmentations, and yellow contours indicate the automatic segmentations.

C. Experiment Results

1) Evaluation of the Performance of First Level

Inspired by paper [49], we plot the PCA-projected features to show the effectiveness of different features in separating voxels from different classes (e.g., the prostate class and the non-prostate class). After mapping each feature vector to the subspace spanned by the first three principal components, the effective features would 1) cluster the voxels with the same class label as close as possible and 2) separate the voxels with different class labels as far as possible. First, we demonstrate the discrimination power of our deep learning features in Fig. 13, by visualizing the PCA-projected feature distributions of different feature representations, i.e., intensity patch (Fig. 13 (a)), handcrafted (Fig. 13 (b)), features learned by unsupervised SSAE (Fig. 13 (c)), and features learned by supervised SSAE (Fig. 13 (d)). Specifically, for the case of handcrafted features, we include three commonly used features, i.e., Haar [50], HoG [47] and LBP [21]. The same patch size is used for computing all features under comparison. It can be seen that the deep learning features from supervised SSAE have better clustering results in the projected feature space, and thus better discriminative power than other two predefined features (i.e., intensity, and handcrafted), as well as deep learning features by unsupervised SSAE. The superior performance of supervised SSAE over the unsupervised SSAE indicates the necessity of utilizing label information to improve the discrimination power of learned features.

Next, we evaluate the segmentation accuracy of different feature representations in the context of sparse patch matching. Table II lists the quantitative results (Dice ratio, precision, Hausdorff distance, and average surface distance) for all feature representations. The p-values (computed with paired t-test at 5% significance level), comparing the supervised SSAE with all other methods, are provided below each quantitative result. It can be observed that our supervised SSAE method significantly outperforms all the intensity and handcrafted feature methods. According to the paired t-test at 5% significance level, both our proposed method (unsupervised and supervised SSAE) outperformed the rest of competing method, but the supervised SSAE is not statistically superior to the unsupervised SSAE.

Fig. 14 further shows the typical likelihood maps estimated by four different feature representations for three different patients. It can be observed that the features learned from supervised SAE can better capture the prostate boundary, especially on the anterior and right posterior parts of the prostate. Fig. 15 shows some typical segmentation results obtained by the sparse label matching method with four different feature representations, respectively. Similarly, the proposed method (i.e., supervised SSAE) achieves the best segmentation, especially on the anterior parts of the prostate, which demonstrates the effectiveness of our proposed method. Moreover, Fig. 16 gives the typical prostate segmentation results of different patients produced by four different feature representations, respectively. 3D visualization of the segmentation result has been added below each segmentation result shown in 2D. For each 3D visualization, the red surface indicates automatic segmentation results with different features, such as intensity, handcrafted, unsupervised SSAE and supervised SSAE, respectively. The transparent grey surfaces indicate the ground-truth segmentations. Our proposed supervised SSAE method improves the segmentation accuracy on both the anterior and posterior parts of the prostates.

2) *Evaluation of the Performance of Second Level*

In this section, we further evaluate our deformable model to show its effectiveness. The comparison methods contain three different deformable model based methods. The first one is the conventional Active Shape Model (ASM). The second one uses intensity features for multi-atlas label fusion, and then finalizes the segmentation by adopting a deformable model on the produced likelihood map, similar to our proposed method. The second method follows the same procedure as the first one except using the handcrafted features, such as Haar, HOG, and LBP, instead of intensity patch for multi-atlas label fusion. Table III shows the segmentation results of intensity, handcrafted and supervised SSAE with/without deformable model and the p-value (with paired t-test at 5% significance level) between the supervised SSAE with deformable model and all other methods. According to the paired t-test at 5% significance level on Dice ratio, our proposed deformable model is statistically the best among all the competing methods. Specifically, our proposed supervised SAE outperforms the ASM, the intensity based deformable model, and the handcrafted based deformable model by 10.7%, 2.1% and 1.6%, respectively. Besides, it can be seen that, after adopting the second level of deformable segmentation, the segmentation accuracy can be further improved for all the comparing methods.

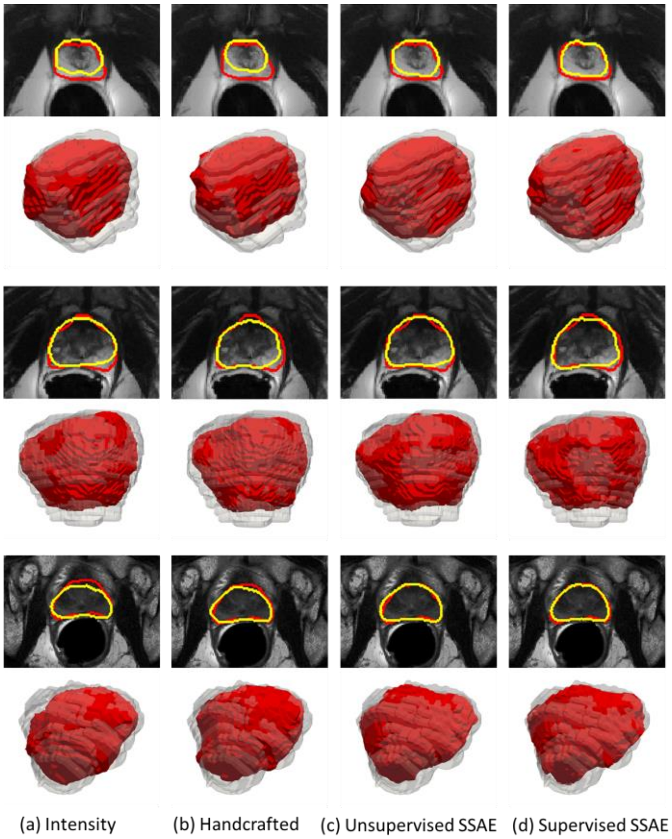


Fig. 16. Typical prostate segmentation results of three different patients produced by four different feature representations: (a) intensity, (b) handcrafted, (c) unsupervised SSAE, and (d) supervised SSAE. Three odd rows show the results for three different patients, respectively. Red contours indicate the manual ground-truth segmentations, and yellow contours indicate the automatic segmentations. Three even rows show the 3D visualization of the segmentation results corresponding to the images above. For each 3D visualization, the red surfaces indicate the automatic segmentation results using different features, such as intensity, handcrafted, unsupervised SSAE and supervised SSAE, respectively. The transparent grey surfaces indicate the ground-truth segmentations.

TABLE III
MEAN AND STANDARD DEVIATION OF QUANTITATIVE RESULTS FOR THE SEGMENTATIONS OBTAINED BY SUPERVISED SSAE WITH/WITHOUT USING DEFORMABLE MODEL. BEST PERFORMANCE IS INDICATED BY BOLD FONT. * DENOTES THE STATISTICALLY BEST PERFORMANCE AMONG ALL THE METHODS WITH/WITHOUT DEFORMABLE MODEL (ACCORDING TO THE PAIRED T-TEST AT 5% SIGNIFICANT LEVEL).

Method	ASM	Intensity		Handcrafted		Supervised SSAE	
		w/o Deformable model	w/ Deformable model	w/o Deformable model	w/ Deformable model	w/o Deformable model	w/ Deformable model*
Dice (%)	78.4±9.7	85.3±6.2	86.0±4.3	85.9±4.5	86.4±4.4	87.1±4.2	**87.8±4.0***
	(2.7e-11)	(3.0e-06)	(7.3e-10)	(7.7e-08)	(5.1e-06)	(2.6e-03)	(NA)
Precision (%)	71.9±13.8	85.1±8.3	89.3±7.4	87.3±7.4	**92.3±7.3**	87.1±7.3	91.6±6.5
	(1.3e-21)	(8.0e-17)	(4.5e-07)	(1.4e-11)	(7.7e-02)	(5.7e-20)	(NA)
Hausdorff	11.50±5.48	8.68±4.24	7.72±2.90	8.55±2.91	7.97±2.92	8.12±2.89	**7.43±2.82***
	(8.5e-07)	(3.6e-04)	(2.5e-02)	(3.8e-05)	(1.2e-04)	(7.4e-03)	(NA)
ASD	3.12±1.71	1.87±0.93	1.78±0.55	1.77±0.54	1.71±0.50	1.66±0.49	**1.59±0.51***
	(9.4e-10)	(6.4e-04)	(2.0e-07)	(2.0e-05)	(1.3e-03)	(1.6e-02)	(NA)

D. Discussion

1) The Issues of Deep Learning Method

The key advantage of deep learning methods is learning good features automatically from data and avoiding using the manually-designed feature extractors, which often require high engineering skills. SAE differs from AE and PCA in the aspect that it imposes sparsity on the mapped features (i.e., responses of hidden nodes), thus avoiding the problem of trivial solutions when the dimensionality of hidden features is more than that of the input features. By stacking SAE together, SSAE is able to learn the hierarchical feature representation, similar to other deep learning models. Besides, we use unsupervised initialization in the pre-training stage, which prevents the later supervised training from falling into the bad local minimum. This also contributes to good performance of our method.

However, the issue of small dataset *vs.* large number of variables arises during the training of the deep network. To prevent the potential issue raised from the limited number of training samples, two strategies are adopted in our method. *First*, we pre-trained the deep network in a layer-by-layer manner [31, 51], which can learn a hierarchy of feature representation one layer a time. Specifically, in the training of each layer, the features learned from previous layer are feed into the next layer. The first three layers consist of 320,000, 80,000, 20,000 parameters, respectively. In our experiment, totally 396,000 training samples were used, which should be sufficient for this lay-wise pre-training step. *Second*, in the fine-tuning stage, the entire deep network is refined only by several iterations, thus better preventing the overfitting issue.

To further relieve the possible overfitting issues, we can also use the idea of transfer learning [52-55]. Specifically, in the unsupervised pre-training step, we can borrow MR images from other body parts (e.g., heart) to initialize our deep network, thus capturing more general MR image appearance. We believe that this initialization could benefit the fine-tuning step and thus overcome the small sample problem. Note that similar strategies have been widely used in the field of computer vision and machine learning [31, 51-56].

2) The Issue of Deformable Segmentation Method

According to Eq. (7), the data term of deformable model is driven by the gradient of the prostate likelihood map. One potential issue may happen if evolving the deformable model according to this data term. That is, the gradient will be zero if the initial shape is a bit away from the boundary. We proposed two strategies to address this potential issue. First, we obtained the initial prostate shape by thresholding the probability map, which makes the initialization *not* far away from the boundary. Second, the deformation is regularized by the shape model. Thus, even some model vertices cannot find the boundary in the capture range, the shape model can still pull them towards the boundary as long as other vertices have been deformed to the boundary. That is, shape regularization makes all vertices deform as a whole, thus addressing the capture range issue.

3) The Computational Time

For our algorithm, the computational time mainly contains three parts: 1) registration part; 2) multi-atlas label fusion part; 3) deformable segmentation part. For registration, the run-time for each affince and non-linear registration is about 20 seconds and four minutes, respectively. For multi-atlas label fusion, the computational time is about 45 minutes, which is the major computational cost of our method. This mainly is due to the individual labeling for a large amount of voxels in each subject image. Currently, we implement multi-atlas label fusion in MATLAB, which is time-consuming for the loop job of sequentially labeling each voxel. For improving the efficiency of our algorithm, one possible solution is to implement the whole label fusion step by using the C++ language. In this way, we expect the computational time to be reduced to 10 minutes for the entire label fusion part. As for the deformable model segmentation part, the computational time is less than one minute.

V. Conclusion

In this paper, we present an automatic segmentation algorithm for T2 MR prostate images. To address the challenges of making the feature representation robust to large appearance variations of the prostate, we propose to extract the deep learning features by the SSAE framework. Then, the learned features are used under sparse patch matching framework to estimate the prostate likelihood map of the target image. To further relieve the impact of large shape variation in the prostate shape repository, a deformable model is driven toward the prostate boundary under the guidance from the estimated prostate likelihood map and sparse shape prior. The proposed method is extensively evaluated on the data set containing 66 prostate MR images. By comparing with several state-of-the-art MR prostate segmentation methods, our method demonstrates the superior performance regarding to the segmentation accuracy.

References

[1] *Prostate Cancer*. Available: http://www.cancer.org/acs/groups/cid/documents/webcontent/003134-pdf.pdf

[2] K. M. Pondman, J. J. Fütterer, B. ten Haken, L. J. Schultze Kool, J. A. Witjes, T. Hambrock, *et al.*, "MR-guided biopsy of the prostate: an overview of techniques and a systematic review," *European Urology,* vol. 54, pp. 517-527, 2008.

[3] H. Hricak, L. Wang, D. C. Wei, F. V. Coakley, O. Akin, V. E. Reuter, *et al.*, "The role of preoperative endorectal magnetic resonance imaging in the decision regarding whether to preserve or resect neurovascular bundles during radical retropubic prostatectomy," *Cancer,* vol. 100, pp. 2655-2663, 2004.

[4] S. Liao, Y. Gao, Y. Shi, A. Yousuf, I. Karademir, A. Oto*, et al.*, "Automatic prostate MR image segmentation with sparse label propagation and domain-specific manifold regularization," in *Information Processing in Medical Imaging*. vol. 7917, J. Gee, S. Joshi, K. Pohl, W. Wells, and L. Zöllei, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 511-523.

[5] P. Yan, Y. Cao, Y. Yuan, B. Turkbey, and P. L. Choyke, "Label Image Constrained Multiatlas Selection," *Cybernetics, IEEE Transactions on,* vol. 45, pp. 1158-1168, 2015.

[6] Y. Cao, X. Li, and P. Yan, "Multi-atlas Based Image Selection with Label Image Constraint," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, 2012, pp. 311-316.

[7] Y. Ou, J. Doshi, G. Erus, and C. Davatzikos, "Multi-Atlas Segmentation of the Prostate: A Zooming Process with Robust Registration and Atlas Selection," *PROMISE12*, 2012.

[8] R. Toth and A. Madabhushi, "Multifeature landmark-free active appearance models: application to prostate MRI segmentation," *Medical Imaging, IEEE Transactions on,* vol. 31, pp. 1638-1650, 2012.

[9] R. Toth, B. N. Bloch, E. M. Genega, N. M. Rofsky, R. E. Lenkinski, M. A. Rosen, *et al.*, "Accurate prostate volume estimation using multifeature active shape models on T2-weighted MRI," *Academic radiology,* vol. 18, pp. 745-754, 2011.

[10] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *Medical Imaging, IEEE Transactions on,* vol. 29, pp. 1714-1729, 2010.

[11] Y. Jin, Y. Shi, L. Zhan, B. A. Gutman, G. I. de Zubicaray, K. L. McMahon, *et al.*, "Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics," *NeuroImage,* vol. 100, pp. 75-90, 10/15/ 2014.

[12] Y. Jin, Y. Shi, L. Zhan, and T. P. M., "Automated multi-atlas labeling of the fornix and its integrity in alzheimer's disease," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, 2015, pp. 140-143.

[13] Y. Jin, C.-Y. Wee, F. Shi, K.-H. Thung, D. Ni, P.-T. Yap, *et al.*, "Identification of infants at high-risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks," *Human Brain Mapping,* vol. 36, pp. 4880-4896, 2015.

[14] M. Yang, X. Li, B. Turkbey, P. L. Choyke, and P. Yan, "Prostate Segmentation in MR Images Using Discriminant Boundary Features," *Biomedical Engineering, IEEE Transactions on,* vol. 60, pp. 479-488, 2013.

[15] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *Medical Imaging, IEEE Transactions on,* vol. 29, pp. 2000-2008, 2010.

[16] S. S. Chandra, J. A. Dowling, S. Kai-Kai, P. Raniga, J. P. W. Pluim, P. B. Greer, *et al.*, "Patient specific prostate segmentation in 3-D magnetic resonance images," *Medical Imaging, IEEE Transactions on,* vol. 31, pp. 1955-1964, 2012.

[17] K. Somkantha, N. Theera-Umpon, and S. Auephanwiriyakul, "Boundary detection in medical images using edge following algorithm based on intensity gradient and texture gradient features," *Biomedical Engineering, IEEE Transactions on,* vol. 58, pp. 567-573, 2011.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition,Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-511-I-518.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893 vol. 1.

[20] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, 2004.

[21] T. Ojala, M. Pietik äinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition,* vol. 29, pp. 51-59, 1996.

[22] S. Liao, Y. Gao, J. Lian, and D. Shen, "Sparse Patch-Based Label Propagation for Accurate Prostate Localization in CT Images," *Medical Imaging, IEEE Transactions on,* vol. 32, pp. 419-434, 2013.

[23] Y. Bengio, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 1798-1828, 2013.

[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009.

[25] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 35, pp. 1915-1929, 2013.

[26] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search Methods," *Image Processing, IEEE Transactions on,* vol. 21, pp. 968-982, 2012.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371-3408, 2010.

[28] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Scene parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12),* 2012, pp. 575-582.

[29] S. Hoo-Chang, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 35, pp. 1930-1943, 2013.

[30] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, 2008, pp. 1096-1103.

[31] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, 2006, pp. 153-160.

[32] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Workshop on Unsupervised and Transfer Learning*, 2012.

[33] S. R. Bulo` and P. Kontschieder, "Neural decision forests for semantic image labelling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 81-88.

[34] H.-I. Suk and D. Shen, "Deep learning-based feature representation for AD/MCI classification," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. vol. 8150, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 583-590.

[35] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," presented at the Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, 2009.

[36] Y. Zhan, X. Zhou, Z. Peng, and A. Krishnan, "Active scheduling of organ detection and segmentation in whole-body medical images," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*. vol. 5241, D. Metaxas, L. Axel, G. Fichtinger, and G. Sz ékely, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 313-321.

[37] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, *et al.*, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine,* vol. 98, pp. 278-284, 6// 2010.

[38] M. Jorge Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, *et al.*, "STEPS: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis,* vol. 17, pp. 671-684, 2013.

[39] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on,* vol. 23, pp. 903-921, 2004.

[40] A. J. Asman and B. A. Landman, "Non-local statistical label fusion for multi-atlas segmentation," *Medical Image Analysis,* vol. 17, pp. 194-208, 2013.

[41] S. Nouranian, S. S. Mahdavi, I. Spadinger, W. Morris, S. Salcudean, and P. Abolmaesumi, "An automatic multi-atlas segmentation of the prostate in transrectal ultrasound images using pairwise atlas shape similarity," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. vol. 8150, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 173-180.

[42] T. F. Chan and L. A. Vese, "Active contours without edges," *Image Processing, IEEE Transactions on,* vol. 10, pp. 266-277, 2001.

[43] M. Kim, G. Wu, W. Li, L. Wang, Y.-D. Son, Z.-H. Cho, *et al.*, "Automatic hippocampus segmentation of 7.0 Tesla MR images by

combining multiple atlases and auto-context models," *NeuroImage,* vol. 83, pp. 335-345, 2013.

[44] 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, pp. 38-59, 1995.

[45] S. Zhang, Y. Zhan, and D. N. Metaxas, "Deformable segmentation via sparse representation and dictionary learning," *Medical Image Analysis,* vol. 16, pp. 1385-1396, 2012.

[46] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *Medical Imaging, IEEE Transactions on,* vol. 17, pp. 87-97, 1998.

[47] Y. Gao, S. Liao, and D. Shen, "Prostate segmentation by sparse representation based classification," *Medical Physics,* vol. 39, pp. 6372-6387, 2012.

[48] https://github.com/rasmusbergpalm/DeepLearnToolbox.

[49] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science,* vol. 313, pp. 504-507, 2006.

[50] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 11, pp. 674-693, 1989.

[51] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation,* vol. 18, pp. 1527-1554, 2006.

[52] P. Sinno Jialin and Y. Qiang, "A Survey on Transfer Learning," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 22, pp. 1345-1359, 2010.

[53] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, "Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols," *Medical Imaging, IEEE Transactions on,* vol. 34, pp. 1018-1030, 2015.

[54] T. Heimann, P. Mountney, M. John, and R. Ionasec, "Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data," *Medical Image Analysis,* vol. 18, pp. 1320-1328, 2014.

[55] Y. Sawada and K. Kozuka, "Transfer learning method using multi-prediction deep Boltzmann machines for a small scale dataset," in *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on,* 2015, pp. 110-113.

[56] M. Wang, W. Li, D. Liu, B. Ni, J. Shen, and S. Yan, "Facilitating Image Search With a Scalable and Compact Semantic Mapping," *Cybernetics, IEEE Transactions on,* vol. 45, pp. 1561-1574, 2015.