Semi-Automatic Segmentation of Prostate in CT Images via Coupled Feature Representation and Spatial-Constrained Transductive Lasso

Yinghuan Shi, Yaozong Gao, Shu Liao, Daoqiang Zhang, Yang Gao, and Dinggang Shen

Abstract—Conventional learning-based methods for segmenting prostate in CT images ignore the relations among the low-level features by assuming all these features are independent. Also, their feature selection steps usually neglect the image appearance changes in different local regions of CT images. To this end, we present a novel semi-automatic learning-based prostate segmentation method in this article. For segmenting the prostate in a certain treatment image, the radiation oncologist will be first asked to take a few seconds to manually specify the first and last slices of the prostate. Then, prostate is segmented with the following two steps: (i) Estimation of 3D prostate-likelihood map to predict the likelihood of each voxel being prostate by employing the coupled feature representation, and the proposed Spatial-COnstrained Transductive LassO (SCOTO); (ii) Multi-atlases based label fusion to generate the final segmentation result by using the prostate shape information obtained from both planning and previous treatment images. The major contribution of the proposed method mainly includes: (i) incorporating radiation oncologist's manual specification to aid segmentation, (ii) adopting coupled features to relax previous assumption of feature independency for voxel representation, and (iii) developing SCOTO for joint feature selection across different local regions. The experimental result shows that the proposed method outperforms the state-of-the-art methods in a real-world prostate CT dataset, consisting of 24 patients with totally 330 images, all of which were manually delineated by the radiation oncologist for performance evaluation. Moreover, our method is also clinically feasible, since the segmentation performance can be improved by just requiring the radiation oncologist to spend only a few seconds for manual specification of ending slices in the current treatment CT image.

Index Terms—Prostate segmentation, feature representation, feature selection, label fusion

◆

1 INTRODUCTION

It is estimated that prostate cancer possibly caused 233,000 new cases and 29,480 deaths for U.S. male in 2014, according to the data reported from National Cancer Institute.[1] Prostate cancer is thus regarded as one of the most leading reasons for cancer-caused death. Recently, CT image guided radiotherapy for prostate cancer treatment has raised lots of research interests, due to its ability in better guiding the delivery of radiation to the prostate [26].

During the CT image guided radiotherapy, a sequence of CT scans will be acquired from a patient in the planning and treatment days (typically from 20 to 40 days). A CT scan acquired in the planning day is called the planning image, and the scans acquired in the subsequent treatment days are called the treatment images. Since the locations of prostate might vary in CT scans across different treatment days, the key issue is to accurately determine the boundary of prostate in the images acquired from each treatment day, which is currently often done by the radiation oncologist in a slice-by-slice manner. However, this manual segmentation procedure is time-consuming, which normally takes up to 20 minutes for each treatment image. Most importantly, manual segmentation results could be inconsistent across different treatment days due to inter- and intra-operator variability.

Previous studies [13], [11], [27] show that the major challenging issues to accurately segment prostate in the CT images include: (i) the boundary between prostate region and background (non-prostate) region is usually unclear due to the low contrast in the CT images, e.g., Figs. 1a and 1b are the same CT prostate images without and with manually delineated prostate boundary, respectively. (ii) The locations of the prostate regions scanned from the same patient at different treatment days are often different due to the irregular and unpredictable prostate motion, e.g., in Fig. 1c where the cyan and magenta contours denote the manual segmentations of the two bone-aligned CT images scanned from two different treatment days of the same patient. We can observe the large prostate motion even after aligning the two scans based on their bony structures. This

---

[1] Prostate Cancer, report from National Cancer Institute. http://www.cancer.gov/cancertopics/types/prostate

- Y. Shi and Yang Gao are with the State Key Laboratory for Novel Software Technology, Nanjing University, China. E-mail: yinghuan.shi@gmail.com, gaoy@nju.edu.cn.
- Yaozong Gao is with the Department of Radiology and BRIC, UNC-Chapel Hill. E-mail: yzgao@cs.unc.edu.
- S. Liao is with the Siemens Medical Solution. E-mail: liaoshu.cse@gmail.com.
- D. Zhang is with the Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, China. E-mail: daoqiangz@gmail.com.
- D. Shen is with the Department of Radiology and BRIC, UNC-Chapel Hill, and the Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea. E-mail: dgshen@med.unc.edu.

Manuscript received 18 Sept. 2014; revised 13 Feb. 2015; accepted 6 Apr. 2015. Date of publication 19 Apr. 2015; date of current version 7 Oct. 2015.
Recommended for acceptance by D. Xu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2424869

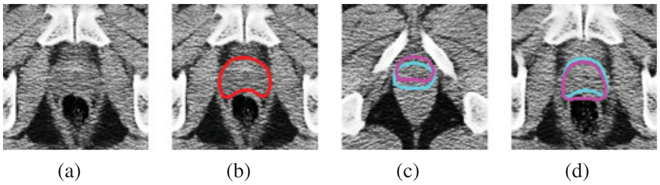


Fig. 1. Challenges in prostate CT segmentation. (a, b) Low contrast in CT image (without and with manual segmentation); (c) Large prostate motion, (d) large shape change, relative to the bones, even after bone-based alignment for the two CT images.

indicates the large prostate motion relative to the bones. (iii) The prostate shape varies at different treatment days, e.g., in Fig. 1d where the cyan and magenta contours denote the manual segmentations at the same slice index of two bone-aligned CT images from different treatment days of the same patient, which indicates the large shape change.

In recent years, several prostate segmentation methods [8], [13], [11], [27], [23], [16] have been developed for CT image guided radiotherapy, with the common goal of segmenting prostate in the current treatment image by borrowing the knowledge learned from the planning and previous treatment images. In this article, we propose a novel semi-automatic learning-based prostate segmentation method for CT image guided radiotherapy. Previous learning-based methods [23], [16], [36], [24] often conduct the feature selection and the subsequent prostate-likelihood estimation jointly for all voxels in the image. However, different local regions may prefer choosing different features to better discriminate between their respective prostate and non-prostate voxels, as indicated by a typical example shown in Fig. 2. In this example, we extracted features for three different local regions, and then apply Lasso (a supervised feature selection technique as introduced in [39]) for selection of their respective features. From the results shown in Fig. 2, we can see that the selected features from three local regions are completely different, demonstrating the necessity of selecting the respective features for each local region. In this article, we design a novel local learning strategy: partition each 2D slice into several non-overlapping local blocks, and then select the respective local features to predict the prostate-likelihood for each local block. This will be achieved by our proposed Spatial-COnstrained Transductive LassO (SCOTO) and support vector regression (SVR), respectively. Compared with the global learning strategy [23], [16], [36], the local learning strategy can help preserve the local information, which guides us to predict the more accurate prostate boundary.

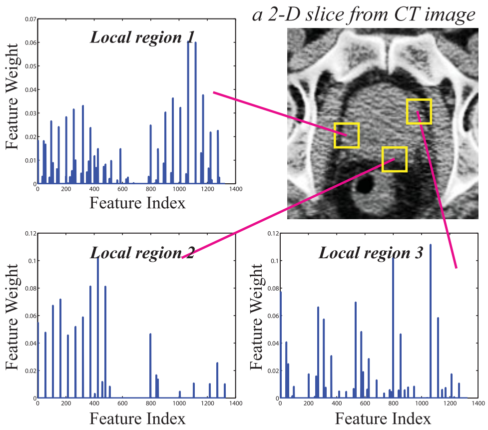


Fig. 2. A typical example to illustrate that the three different local regions prefer choosing different features. We adopt Lasso as feature selection method, and the features include LBP, Haar wavelet, and HoG, which will be discussed in the following section.

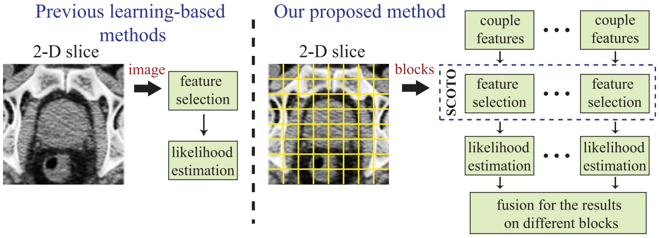


Fig. 3. The major differences between the previous learning-based methods and our proposed method. Our proposed method employs the coupled feature representation, and adopts a local strategy for feature selection and prostate-likelihood estimation.

Also, previous methods [23], [16], [36] often assume all the extracted features are independent during the prostate segmentation. However, it is often believed that different features extracted from both prostate and non-prostate regions are actually not isolated, but connected with each others in some way. Therefore, the independency assumption for the features used in the previous studies might be too strict. In this article, we consider to relax independency assumption by introducing the coupled feature (CF) representation as inspired by [42], [38]. The major difference between the previous learning-based methods and our proposed method can be observed from Fig. 3.

Note that, in our method, before automatic segmentation of prostate from the current treatment CT image, radiation oncologist needs to spend a few seconds to specify the first and last slices of prostate in the CT image. By spending this little manual time, the segmentation results can be significantly improved, compared with the fully automatic methods [23], [26]. The contributions of our proposed method can be summarized into the following three folds:

- A novel semi-automatic prostate segmentation method in CT images is proposed. For segmenting the current treatment image, the patient-specific information is learned from previous planning and treatment images. This information together with radiation oncologist's manual specification of two ending slices are used for guiding the accurate segmentation. It is excepted that the prostate segmentation performance can be further boosted by incorporating manual specification as reference, thus finally achieving better treatment.
- We first attempt to introduce the coupled feature representation to the prostate segmentation, in order to relax the feature independency assumption as used in the previous studies. With the help of coupled features, the original feature (OF) matrix can be complemented by analyzing the intra- and inter- coupled interactions. The experimental results also validate the efficacy of coupled features, compared with original features, for better feature

representation by capturing the coupled interaction information.

- A novel joint feature selection method, called SCOTO, is proposed, which is theoretically formulated as a joint convex optimization problem. SCOTO can successfully select the discriminative features jointly for different local regions (blocks) to guide better prostate-likelihood estimation. Also, a feasible iterative projected gradient descent based strategy is utilized for solving SCOTO. Moreover, we systematically compared SCOTO with several state-of-the-art feature selection methods for prostate segmentation.

A preliminary version of this work was presented in [37]. The preliminary version is extended in this work to (1) employ the coupled feature representation aiming to model both the intra- and inter- coupled interactions among the original features, (2) mathematically prove the joint convexity of the proposed SCOTO, which is the requirement of employing the alternating optimization method. Also, we substantially extend the experimental validation of the proposed method, which includes (3) evaluating the efficacy of coupled features, (4) evaluating the efficacy of the multi-atlases based label fusion, (5) investigating the influence of inaccurate manual specification, (6) analyzing the computational complexity of SCOTO, and (7) adding a discussion section to fully investigate several related issues.

The rest of this article is organized as follows. In Section 2, we briefly introduce the related works for prostate segmentation developed in recent years. The framework and image preprocessing of the proposed method are described in Section 3 and Section 4, respectively. In Section 5, we introduce the usage of coupled feature representation, and propose our new feature selection algorithm, SCOTO, for prostate-likelihood estimation. In Section 6, the multi-atlases based label fusion for final segmentation is discussed. Experimental results and performance comparisons with the competing methods are presented in Section 7. Several related issues are also discussed in Section 8. Finally, we conclude this article by summarizing the proposed method and also discussing the obtained results in Section 9.

2 RELATED WORKS

The previous methods on prostate segmentation for CT image guided radiotherapy can be roughly categorized into three classes: deformable-model-based, registration-based, and learning-based methods.

In deformable-model-based methods [8], [13], the prostate shapes learned from the planning and previous treatment images are first used to initialize the deformable model [35], and then specific optimization strategies are developed to guide prostate segmentation. Chen et al. [8] simultaneously segmented prostate and rectum by imposing the anatomical constraints during the model deformation procedure. Feng et al. [13] segmented the prostate by using deformable model, which integrates the gradient profile features and the probability distribution function features. The deformable-model-based methods can combine the information from both prostate shape and image appearance for guiding prostate segmentation. However, good performance largely requires good quality of deformable model initialization. In the case where large prostate motion happens, obtaining a good initialization for deformable model becomes a very challenging task.

In registration-based methods [9], [11], [26], [27], the planning and previous treatment images are warped onto the current treatment image, and then their respective segmentation images are similarly warped and further combined (by label fusion) to obtain the final segmentation of the current treatment image. Chen et al. [9] designed a strategy that aligns the planning image to treatment image by mesh-less point set modeling and 3D non-rigid registration. Davis et al. [11] segmented the prostate by combining large deformation image registration with a bowel gas segmentation and deflation algorithm. Liao and Shen [26] proposed to use anatomical feature selection and an online updating mechanism to integrate both population and patient-specific information, in order to guide the accurate prostate registration. Also, Liao et al. [27] combined the patch-based feature representation and hierarchical sparse label propagation in order to automatically localize the prostate. Several experimental results have demonstrated the robustness and effectiveness of the registration-based methods for prostate segmentation. However, the segmentation accuracy is limited in the case with inconsistent image appearance changes in CT images [16].

The last class of methods, learning-based methods [23], [36], [16], which attracted lots of interests recently, first use the machine learning techniques to predict the prostate-likelihood map, and then segment the prostate in the estimated prostate-likelihood map. In Li et al.'s method [23] and Gao et al.'s [16], prostate segmentation is first formulated as a prostate-likelihood estimation problem (AdaBoost [14] in [23], sparse representation-based classification [44] in [16]) using visual features (e.g., histogram of oriented gradients (HoG) [10] and auto-context features [41]), and then on the obtained likelihood map, the off-the-shelf segmentation techniques (e.g., levelset) are adopted to segment the prostate. Shi et al. [36] presented a method by modeling the prostate-likelihood estimation problem as a transductive learning task. Our proposed method belongs to this class of learning-based segmentation methods.

It is worth noting that, besides segmenting prostate from CT images, other segmentation methods are also developed for segmenting prostate from other imaging modalities such as magnetic resonance (MR) [17], [21] and ultrasound [47], [48] images. However, these methods cannot be directly borrowed to segment prostate from CT images due to the aforementioned challenges.

3 FRAMEWORK

As a learning-based segmentation method, our proposed method mainly consists of two steps: (i) prostate-likelihood estimation step and (ii) multi-atlases based label fusion step.

In the *prostate-likelihood estimation step*: firstly, all previous and current treatment images are rigidly aligned to the planning image of the same patient based on the pelvic bone structures, for removing the whole-body patient motion that is irrelevant to prostate segmentation. Then, we

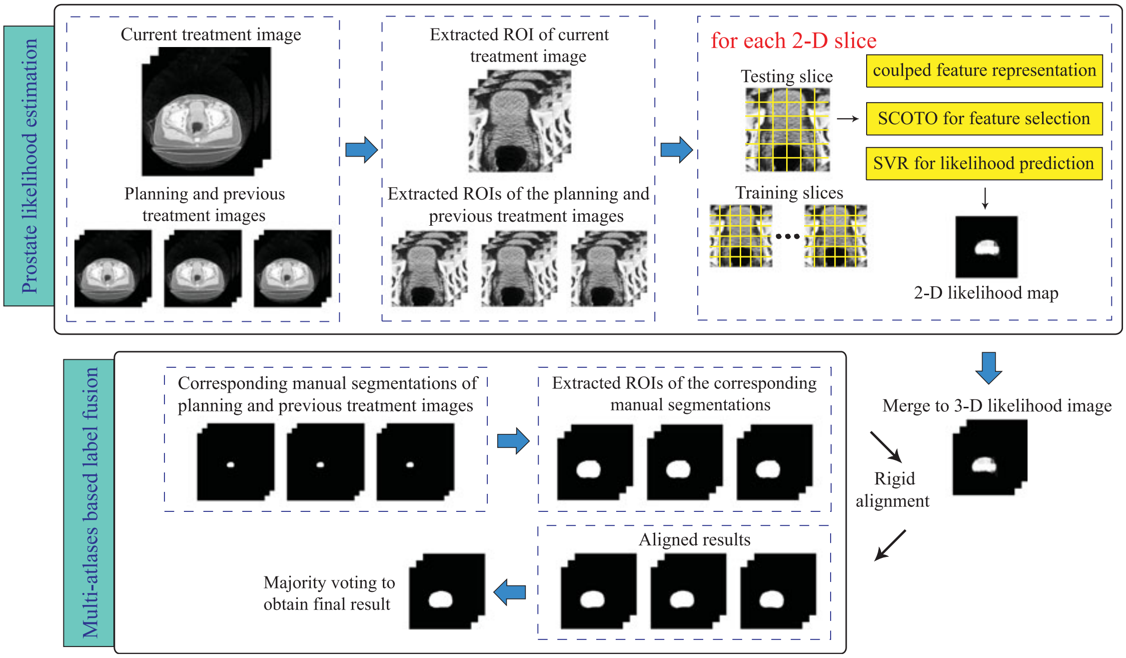


Fig. 4. The flowchart of the proposed prostate segmentation method.

extract the ROI regions according to the prostate center in the planning image.

Secondly, for the current treatment image, radiation oncologist is asked to specify the first and last slices of the prostate. By combining the voxels in the specified slices of the current treatment image with the voxels sampled from the planning and previous treatment images, we can extract 2D low-level features (LBP [34], HoG [10], and Haar wavelet [31]) for all these voxels separately from their original CT images. Then, each 2D slice will be partitioned into several non-overlapping blocks. The coupled feature representation is employed to transform the low-level original feature space into the coupled feature space for each block, the proposed SCOTO is applied for joint feature selection of all blocks upon the coupled feature representation, and SVR is further adopted to predict the 2D prostate-likelihood map for all the voxels in the current slice.

Finally, the predicted 2D prostate-likelihood map of each individual slice will be merged into a 3D prostate-likelihood map according to the order of their original slices.

In *multi-atlases based label fusion step*, to make full use of prostate shape information, those manually segmented prostate images in both planning and previous treatment images of the same patient will be rigidly aligned to the estimated 3D prostate-likelihood map of the current treatment image. Then, majority voting will be applied to fuse the labels from different aligned images, and obtain the final segmentation result. The framework of the proposed method is shown in Fig. 4.

4 IMAGE PREPROCESSING

Before discussing the details of the prostate-likelihood estimation step and multi-atlases based label fusion step, we will first introduce the notations used in the following parts, and present the implementation details of the alignment, ROI extraction, manual interaction, feature representation, as well as strategy for sampling the training voxels.

For each patient, we have one planning image, several previous treatment images with their respective manual prostate segmentations by radiation oncologist offline, and also the treatment image scanned in the current treatment day, which needs to be segmented by the proposed method. The planning image and its corresponding manual segmentation result are denoted as $\mathbf{I}_\mathbf{p}$ and $\mathbf{G}_\mathbf{p}$, respectively. The nth treatment image, which is the current treatment image, is denoted as $\mathbf{I}_n$. The previous treatment images and their corresponding manual segmentations are denoted as $\mathbf{I}_1, \ldots, \mathbf{I}_{n-1}$ and $\mathbf{G}_1, \ldots, \mathbf{G}_{n-1}$, respectively. Also, the final 3D prostate-likelihood map and its segmentation result for the current treatment image $\mathbf{I}_n$ by adopting the proposed method are denoted as $\mathbf{M}_n$ and $\mathbf{S}_n$, respectively.

4.1 Alignment to the Planning Image

During the scanning process of CT image guided radiotherapy at different treatment days, the whole-body rigid motion is often inevitable, which is irrelevant to the prostate motion. To eliminate the influence, we utilize the pelvic bone structure based alignment, which has been already validated in [26]. Detailed steps include: *pelvic bone segmentation*, and *rigid alignment*. For pelvic bone segmentation, since the pelvic bone structure of the same patient has only the rigid motion during the CT image guided radiotherapy, and also it has higher image intensity, which can be easily distinguished from other surrounding anatomical structures, we adopt a thresholding technique to first segment the pelvic bone structure as reference for the subsequent alignment, which is similar to [11]. Specifically, rigid alignment algorithm implemented in FLIRT toolkit [20] is used to perform rigid registration on the segmented pelvic bones. That is, each treatment image $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$ will be rigidly aligned to the planning image $(\mathbf{I}_\mathbf{p})$ with respect to their segmented pelvic bone structures.

4.2 ROI Extraction

Since the prostates are normally located only in a small central part of the CT images, the ROI extraction, aiming to extract the central part by excluding the irrelevant and redundant background voxels, is useful to alleviate the computational burden and also important for improving the final segmentation accuracy. For each patient, the mass center of the prostate in the planning image $\mathbf{I_p}$ is first calculated as the patient-specific prostate center, and then a 3D region centered at the patient-specific center is extracted. Specifically, in this article, the extracted ROI size is $140 \times 140 \times 60$, which is large enough to cover the whole prostate volume (i.e., the normal size of the prostate is usually from $30 \times 40 \times 30$ mm^3 to $40 \times 50 \times 70$ mm^3).

4.3 Simple Manual Interaction

The proposed semi-automatic segmentation method intends to borrow the manual interaction from the experienced radiation oncologist to improve the segmentation performance. Instead of asking radiation oncologist for multiple interactions during the segmentation, we only ask for manual specification of the first and last slices of the prostate along the z-axis. It is noteworthy that the radiation oncologist's manual specification is carried after ROI extraction. Therefore, it will take only less than 10 seconds for this manual step. In the experiments, we will also show that the segmentation results can be largely improved by asking radiation oncologist to spend such a little interaction time, which is also clinically feasible.

4.4 Patch-Based Original Feature Representation

For the original feature representation, three different kinds of low-level features are extracted from 2D slice, which include nine-dimensional histogram of oriented gradient [10], 30-dimensional local binary pattern (LBP) [34] and 14-dimensional multi-resolution Haar wavelet [31]. Typically, HOG is calculated within 3×3 cell blocks with nine histogram bins similar to [10]. LBP is calculated with the radius value 2 and the neighboring voxel number 8. Haar is calculated by convolving the 14 multi-resolution wavelet basis functions with the input image similar to [31]. The window size for feature extraction is empirically set to 21×21. To better incorporate the neighbouring information, we adopt a patch-based feature representation. Specifically, for each voxel, we extract the above three features at every voxel within a $k \times k$ patch for representing the central voxel. The feature vector of each voxel consists of 53 features ($9 + 30 + 14 = 53$ features) extracted from all voxels within its local patch. In this article, k is set to 5, so a 1,325 $(53 \times 5 \times 5)$ dimensional low-level original feature vector is used to represent each voxel before the coupled feature representation.

4.5 Training Voxels Sampling

Because the corresponding manual segmentations are available for both planning and previous treatment images, voxels with the known labels (prostate or background) can be sampled as the training voxels to aid the estimation of prostate-likelihood map for the current treatment image. Since the confusing voxels are frequently lying on the boundary of the prostate region, it is reasonable to sample relatively more voxels around the boundary. That is, the boundary voxels will have higher probability to be sampled, as illustrated in Fig. 5. Similar strategy for sampling the training voxels was also used in literature [27].

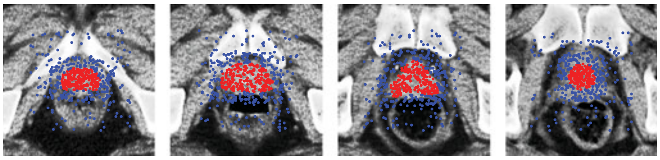


Fig. 5. Typical examples for illustrating the sampling of the training voxels, with the red points denoting the prostate voxels and the blue points denoting the background voxels.

5 PROSTATE-LIKELIHOOD ESTIMATION

We first discuss the details of our coupled feature representation, and then formulate the spatial-constrained transductive feature selection task as the proposed SCOTO. Also, we present a feasible optimization strategy for solving SCOTO, whose joint convexity is theoretically guaranteed. Finally, SVR is employed to estimate the prostate-likelihood for each block (voxel-wise).

5.1 Notations

For each slice, we first partition the slice into non-overlapping $N_x \times N_y$ blocks as shown in Fig. 4. Then for the ith block, we use $l_i \in \mathbb{R}$ and $u_i \in \mathbb{R}$ to denote the number of training and testing voxels, respectively. Note that the training voxels come from the sampled voxels, whose locations are in the current block within the slices $[s_c - 1, s_c + 1]$ of training images, where s_c is the current slice index of testing voxels in z-axis. The reason for this design is that training voxels in adjacent slices have similar distribution in feature space, which guarantees sampling of enough voxels, especially on the base and apex slices of prostate. $N \in \mathbb{R}$ $(N = N_x \times N_y)$ denotes the total number of blocks in the current slice. $\mathbf{y}_i \in \mathbb{R}^{l_i + u_i}$ and $\mathbf{Z}_i \in \mathbb{R}^{(l_i+u_i)\times d}$ are used to denote the ground-truth label and original feature matrix for all the training and testing voxels, respectively. Without loss of generality, all the training voxels are listed before the testing voxels in both $\mathbf{y}_i$ and $\mathbf{Z}_i$. d means the number of low-level original features (i.e., 1,325 in the article). It is noteworthy that the labels of testing voxels in $\mathbf{y}_i$ are set to 0. Also in $\mathbf{y}_i$, the labels of training voxels are set to 1 if they belong to the prostate, and set to 0 if they belong to the background.

5.2 Coupled Feature Representation

Basically, it is always believed that different voxels of both prostate and non-prostate regions are not isolated, but anatomically connected with each other. Thus, the low-level original features extracted from these voxels are also not independent, but related with each other in a certain way. However, previous learning-based prostate segmentation methods simply use the low-level original features without considering their relations. In this article, we first attempt to introduce the coupled feature representation to the prostate segmentation inspired by [42], [38].

For the coupled feature representation, aimed to enhance the representative ability, it usually integrates two types of relational information inherent in the original features: intra-coupled interaction with correlations between the features and their own powers, and inter-coupled interaction with correlations between different powers of different features. Thus, the coupled features can represent the voxels in a more meaningful way by capturing both the linear and nonlinear combination of the low-level original features.

Coupled feature representation can be considered as a procedure with an input as the original feature matrix and an output as the learned coupled feature matrix. Specifically, for each original feature matrix $\mathbf{Z}_i \in \mathbb{R}^{(l_i+u_i)\times d}$ (extracted from the ith block), we will perform the coupled feature representation to obtain the corresponding coupled feature matrix $\mathbf{F}_i \in \mathbb{R}^{(l_i+u_i)\times dE}$, where $E \in \mathbb{R}$ is a number to denote the maximal power (typically $3 \leq E \leq 10$). Please note that all the following steps, i.e., SCOTO and multi-atlases based label fusion, are conducted on the obtained coupled feature matrix $\mathbf{F}_i$, instead of original feature matrix $\mathbf{Z}_i$, which is different to our preliminary work [37].

Matrix expansion. It is always excepted that increasing the dimensionality of features can boost the representative ability [25], [42]. Here, each feature can be extended to new features, by calculating its power with different orders, to complement the feature matrix, aiming to incorporate both linear and nonlinear information. Formally, we denote $z_i^{k,j}$ as the numerical value of kth voxel at jth feature in original feature matrix $\mathbf{Z}_i$, and then map them to an expanded feature space (taking $E = 2$ as an example) as below:

$$\left[\left\langle z_i^{k,1}\right\rangle^1, \left\langle z_i^{k,1}\right\rangle^2, \left\langle z_i^{k,2}\right\rangle^1, \left\langle z_i^{k,2}\right\rangle^2, \ldots, \left\langle z_i^{k,d}\right\rangle^1, \left\langle z_i^{k,d}\right\rangle^2\right]^\top$$

where $\langle z_i^{k,j}\rangle^e$ indicates the eth $(1 \leq e \leq E)$ power of $z_i^{k,j}$ and in this case $e \in \{1, 2\}$.

Intra-coupled interaction. By employing the aforementioned matrix expansion, to exploit the relations of same feature with different powers, we first define an intra-coupled interaction, which considers the correlations between the jth feature and its own powers as follows:

$$\mathbf{C}_a(j) = \begin{pmatrix} \theta_{11}^j & \theta_{12}^j & \cdots & \theta_{1E}^j \\ \theta_{21}^j & \theta_{22}^j & \cdots & \theta_{2E}^j \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{E1}^j & \theta_{E2}^j & \cdots & \theta_{EE}^j \end{pmatrix},$$

where θ_{e_1,e_2}^j denotes a Pearson's correlation coefficient between $\langle \mathbf{Z}_i^{:,j}\rangle^{e_1}$ and $\langle \mathbf{Z}_i^{:,j}\rangle^{e_2}$, since we know that Pearson's correlation is powerful in measuring the agreement of different sets or vectors [15]. Here, $e_1 = \{1, 2, \ldots, E\}$, $e_2 = \{1, 2, \ldots, E\}$, and $\langle \mathbf{Z}_i^{:,j}\rangle^e$ indicates the column vector of the eth power for all the voxels at jth feature in the expanded feature space.

Inter-coupled interaction. Besides the intra-coupled interaction, to exploit the relations of different features with different powers, we also define an inter-coupled interaction to capture the correlations between the jth feature and the powers of other features as follows:

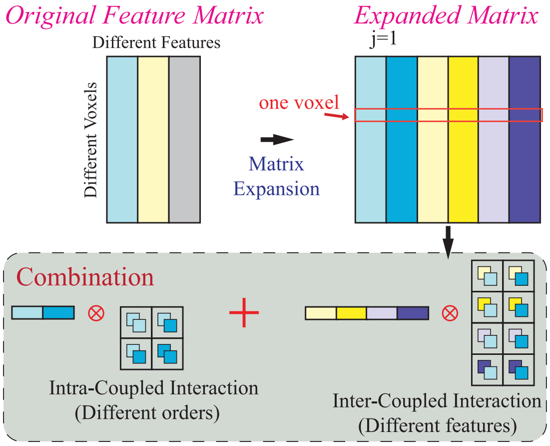


Fig. 6. An illustration (three features, $E = 2$, $j = 1$) of calculating the intra- and inter- coupled interactions. Matrix expansion is firstly applied to calculate each feature up to its second-order power (light blue, yellow, and violet are first-order power, and dark colors are corresponding second-order power), then intra- and inter- coupled interaction matrixes are calculated, and finally we combine them for coupled feature representation. Here, we take one voxel as an example.

$$\mathbf{C}_b(j) = \begin{pmatrix} \sigma_{11}^{v_{j(1)}} \cdots \sigma_{1E}^{v_{j(1)}} & \cdots & \sigma_{11}^{v_{j(d-1)}} \cdots \sigma_{1,E}^{v_{j(d-1)}} \\ \sigma_{21}^{v_{j(1)}} \cdots \sigma_{2E}^{v_{j(1)}} & \cdots & \sigma_{21}^{v_{j(d-1)}} \cdots \sigma_{2E}^{v_{j(d-1)}} \\ \vdots & \ddots & \vdots \\ \sigma_{E1}^{v_{j(1)}} \cdots \sigma_{EE}^{v_{j(1)}} & \cdots & \sigma_{E1}^{v_{j(d-1)}} \cdots \sigma_{EE}^{v_{j(d-1)}} \end{pmatrix},$$

where $\mathbf{v}_j \in \mathbb{R}^{d-1} = [1, \ldots, j-1, j+1, \cdots, d]^\top$ denotes the feature indexes (except current j) with $v_{j(c)}$ as the cth entry of $\mathbf{v}_j$. $\sigma_{e_1,e_2}^{v_{j(c)}}$ is a Pearson's correlation coefficient between $\langle \mathbf{Z}_i^{:,j}\rangle^{e_1}$ and $\langle \mathbf{Z}_i^{:,v_{j(c)}}\rangle^{e_2}$. Fig. 6 gives a typical illustration of calculating the intra- and inter- coupled interactions. Please note that, to take advantage of information from testing voxels, we use both the training and testing voxels in coupled feature representation within an unsupervised manner.

Combination. Here, we respectively denote the intra- and inter-coupled representation of the jth feature for the kth voxels as (*containing jth feature and its powers*)

$$\mathcal{Z}_a(k) = \left[\left\langle z_i^{k,j}\right\rangle^1, \ldots, \left\langle z_i^{k,j}\right\rangle^E\right],$$

and (*excluding jth feature and its powers*)

$$\mathcal{Z}_b(k) = \left[\left\langle z_i^{k,v_{j(1)}}\right\rangle^1, \ldots, \left\langle z_i^{k,v_{j(1)}}\right\rangle^E, \ldots, \left\langle z_i^{k,v_{j(d-1)}}\right\rangle^1, \ldots, \left\langle z_i^{k,v_{j(d-1)}}\right\rangle^E\right].$$

Also, to prevent the possible over-fitting, a parameter $\mathbf{w} \in \mathbb{R}^E$ is introduced to assign the lower weight to the higher order representation as:

$$\mathbf{w} = [1/(1!), 1/(2!), \ldots, 1/(E!)].$$

Thus, by introducing $\mathbf{w}$, we can obtain the normalized intra- and inter- coupled representation as $\widehat{\mathcal{Z}_a(k)} = \mathcal{Z}_a(k) \odot \mathbf{w}$ and $\widehat{\mathcal{Z}_b(k)} = \mathcal{Z}_b(k) \odot \overbrace{[\mathbf{w}, \mathbf{w}, \ldots, \mathbf{w}]}^{d-1}$, respectively, where $\odot$ denotes Hadamard product.

We know that intra- and inter-coupled interactions $\mathbf{C}_a(j)$ and $\mathbf{C}_b(j)$ capture the coupled relations within and between jth feature, respectively. Thus, it is excepted that, a feature, with the larger Pearson's correlation coefficient value in either $\mathbf{C}_a(j)$ or $\mathbf{C}_b(j)$ (indicating it is strongly coupled with current $j\text{th}$ feature), should have larger contribution to the respective intra- and inter- coupled representation of $j\text{th}$ feature. Mathematically, the linear combination can be adopted to integrate $\mathbf{C}_a(j)$ and $\mathbf{C}_b(j)$, aiming to obtain the coupled feature representations of the jth feature for the kth voxel, as follows:

$$\mathbf{u}^k(j) = \widehat{\mathcal{Z}_a(k)} \otimes [\mathbf{C}_a(j)]^\top + \widehat{\mathcal{Z}_b(k)} \otimes [\mathbf{C}_b(j)]^\top,$$

where $\otimes$ denotes matrix multiplication. Thus, the final coupled feature representation for the kth voxel can be represented as follows:

$$\mathbf{u}^k = [\mathbf{u}^k(1), \mathbf{u}^k(2), \ldots, \mathbf{u}^k(d)]^\top \in \mathbb{R}^{E\cdot d}. \tag{1}$$

Lastly, by combining all the $\mathbf{u}^k$ $(k = 1, \ldots, l_i + u_i)$, we can obtain the coupled feature matrix $\mathbf{F}_i$ for the ith block.

5.3 SCOTO: Problem Formulation

For each 2D slice, our goal is to estimate the prostate-likelihood for each voxel in the current slice. Since our coupled feature representation $\mathbf{F}_i$ for each voxel is a high dimensional vector ($\mathbb{R}^{dE}$, $d = 1{,}325$, $3 \leq E \leq 7$), feature selection (by SCOTO in this article) is very critical to remove the redundant features.

To consider the prediction error per block, the usage of unlabeled (testing) voxels, as well as the neighbouring spatial smoothness during the feature selection, we mathematically formulate the objective function of SCOTO as follows:

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_N} \Bigg\{ \sum_{i=1}^{N} \Big[\|\mathbf{J}_i(\mathbf{y}_i - \mathbf{F}_i\boldsymbol{\beta}_i)\|_2^2 + \lambda_S \|\boldsymbol{\beta}_i\|_1 + \frac{\lambda_L}{(l_i+u_i)^2} \boldsymbol{\beta}_i^\top \mathbf{F}_i^\top \mathcal{L}_i \mathbf{F}_i \boldsymbol{\beta}_i \Big] + \frac{\lambda_E}{|H(i)|} \sum_{i=1}^{N} \sum_{j\in H(i)} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2 \Bigg\}, \tag{2}$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ $(\boldsymbol{\beta}_i \in \mathbb{R}^{dE})$ are the parameters to learn, which indicates the weights of individual features for each block. Please note that, although we extend the feature spaces from the original features [37] to the coupled features, SCOTO still works for both the original and coupled features, except we should notice the dimensionality of the feature matrix $\mathbf{F}_i$ and the parameter $\boldsymbol{\beta}_i$ are different when using the original and coupled features.

Also in Eq. (2), $\lambda_S, \lambda_L, \lambda_E \in \mathbb{R}$ are the parameters to control the corresponding terms in Eq. (2), and we will introduce the method for automatic parameter selection in later discussion. $\mathbf{J}_i \in \mathbb{R}^{(l_i+u_i)\times(l_i+u_i)}$ is used to indicate the training voxels since the testing voxels have no contribution on the first term. It is a diagonal matrix defined as

$$\mathbf{J}_i = \mathrm{diag}\Big[\overbrace{1/l_i, \ldots, 1/l_i}^{l_i}, \overbrace{0, \ldots, 0}^{u_i} \Big].$$

$\mathcal{L}_i \in \mathbb{R}^{(l_i+u_i)\times(l_i+u_i)}$ is the graph Laplacian with the same definition as that in literature [2]. $H(i)$ denotes the neighbors of the ith block. Since we use the four-neighbor connection in this article, $|H(i)|$ (the cardinality of $H(i)$) is set to 4 as a constant.

In Eq. (2), the first term with three sub-terms focuses on each individual block: the first sub-term indicates the prediction error, the second sub-term imposes the sparsity constraint with $L1$ norm for feature selection, and the third sub-term is the graph Laplacian imposing the manifold assumption on both the training and testing voxels. In this way, testing voxels without any label information can also be well utilized in the feature selection. The second term of Eq. (2) is the smoothness term on the neighboring blocks, by assuming that the neighboring blocks should choose similar features.

5.4 SCOTO: Optimization

Because the feature weight vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ are jointly convex (see the following theoretical analysis), the alternating optimization method [30] can be employed to solve Eq. (2), which sequentially solves $\boldsymbol{\beta}_i$ with other variables $\boldsymbol{\beta}_j$ $(1 \leq j \leq N, j \neq i)$ fixed. Therefore, the optimization for Eq. (2) can be divided into several alternating sub-problems. Formally, for each $\boldsymbol{\beta}_i$, when fixing the other parameters $\boldsymbol{\beta}_j$, we have the following sub-problem:

$$\min_{\boldsymbol{\beta}_i} \Bigg\{ \|\mathbf{J}_i(\mathbf{y}_i - \mathbf{F}_i\boldsymbol{\beta}_i)\|_2^2 + \lambda_S \|\boldsymbol{\beta}_i\|_1 + \frac{\lambda_L}{(l_i+u_i)^2} \boldsymbol{\beta}_i^\top \mathbf{F}_i^\top \mathcal{L}_i \mathbf{F}_i \boldsymbol{\beta}_i + \frac{\lambda_E}{|H(i)|} \sum_{j\in H(i)} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2 \Bigg\}. \tag{3}$$

The problem of Eq. (3) is convex but not smooth for the feature weight vector $\boldsymbol{\beta}_i$, so the closed-form solution cannot be reached. For solving Eq. (3), we consider adopting the iterative projected gradient descent method [1] for optimization, which separates the sub-problem into the smoothness term and the non-smoothness term, and solves them iteratively until convergence. Therefore, the Eq. (3) can be separated into the smoothness term

$$\mathcal{S}(\boldsymbol{\beta}_i) = \|\mathbf{J}_i(\mathbf{y}_i - \mathbf{F}_i\boldsymbol{\beta}_i)\|_2^2 + \frac{\lambda_L}{(l_i+u_i)^2} \boldsymbol{\beta}_i^\top \mathbf{F}_i^\top \mathcal{L}_i \mathbf{F}_i \boldsymbol{\beta}_i + \frac{\lambda_E}{|H(i)|} \sum_{j\in H(i)} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2,$$

and the non-smoothness term

$$\mathcal{N}(\boldsymbol{\beta}_i) = \lambda_S \|\boldsymbol{\beta}_i\|_1.$$

The iterative projected gradient descent method contains two steps in each iteration. For the kth iteration, *in the first step*, we compute $\boldsymbol{\alpha}_i^{(k)}$ using Newton gradient descent method as

$$\boldsymbol{\alpha}_i^{(k)} = \boldsymbol{\beta}_i^{(k)} - \gamma_k \left[\frac{\partial^2 \mathcal{S}(\boldsymbol{\beta}_i^{(k)})}{\partial \boldsymbol{\beta}_i^{(k)}} \right]^{-1} \frac{\partial \mathcal{S}(\boldsymbol{\beta}_i^{(k)})}{\partial \boldsymbol{\beta}_i^{(k)}}, \tag{4}$$

where $\gamma_k \in \mathbb{R}$ is the step size which can be automatically determined by the back-tracking line search for each iteration. $\frac{\partial \mathcal{S}(\boldsymbol{\beta}_i^{(k)})}{\partial \boldsymbol{\beta}_i^{(k)}}$ and $\frac{\partial^2 \mathcal{S}(\boldsymbol{\beta}_i^{(k)})}{\partial \boldsymbol{\beta}_i^{(k)}}$ denote the first and second order derivatives of $\mathcal{S}(\boldsymbol{\beta}_i^{(k)})$ at $\boldsymbol{\beta}_i^{(k)}$, respectively, which can be calculated as follows:

$$\frac{\partial \mathcal{S}(\boldsymbol{\beta}_i^{(k)})}{\partial \boldsymbol{\beta}_i^{(k)}} = \mathbf{F}_i^\top \left[\hat{\mathbf{J}}_i \mathbf{y}_i + \mathbf{Q}_i \boldsymbol{\beta}_i^{(k)}\right] + \frac{\lambda_E}{|H(i)|} \sum_{j \in H(i)} \left(\boldsymbol{\beta}_i^{(k)} - \boldsymbol{\beta}_j^{(k)}\right),$$

and

$$\frac{\partial^2 \mathcal{S}(\boldsymbol{\beta}_i^{(k)})}{\partial \boldsymbol{\beta}_i^{(k)}} = \mathbf{F}_i^\top \mathbf{Q}_i + \lambda_E \mathbf{I},$$

where $\hat{\mathbf{J}}_i = \mathbf{J}_i \mathbf{J}_i \in \mathbb{R}^{(l_i+u_i)\times(l_i+u_i)}$. For simplicity representation, $\mathbf{Q}_i \in \mathbb{R}^{(l_i+u_i)\times dE}$ is defined as

$$\mathbf{Q}_i = \left[\frac{\lambda_L}{(l_i+u_i)^2}\mathcal{L}_i - \hat{\mathbf{J}}_i\right]\mathbf{F}_i. \tag{5}$$

In the second step, with the obtained $\boldsymbol{\alpha}^{(k)}$, we update $\boldsymbol{\beta}_i^{(k)}$ to $\boldsymbol{\beta}_i^{(k+1)}$ by solving the following $L1$ regularized regression problem:

$$\boldsymbol{\beta}_i^{(k+1)} = \arg\min_{\boldsymbol{\beta}_i} \left\|\boldsymbol{\beta}_i - \boldsymbol{\alpha}_i^{(k)}\right\|_2^2 + \mathcal{N}(\boldsymbol{\beta}_i). \tag{6}$$

The Eq. (6) can be efficiently solved by many existing algorithms, e.g., LARS [12].

The algorithm summary for SCOTO can be referred to Algorithm 1. The inputs are the coupled feature matrices and the corresponding labels of individual blocks, and the outputs are the corresponding feature weights for each block.

Algorithm 1. SCOTO

```
Input: the coupled feature matrices $\mathbf{F}_1,\ldots,\mathbf{F}_N$, and the
       corresponding labels $\mathbf{y}_1,\ldots,\mathbf{y}_N$.
Output: feature weights $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_N$.
 1: for $i \leftarrow 1,\ldots,N$ do
 2:     $\mathcal{L}_i \leftarrow$ calculate the graph Laplacian [2].
 3:     $\boldsymbol{\beta}_i^{(1)} \leftarrow$ solving Lasso on $\mathbf{F}_i$ and $\mathbf{y}_i$ for initialization.
 4: end for
 5: $k \leftarrow 1$
 6: while not converge do
 7:     for $i \leftarrow 1,\ldots,N$ do
 8:         $\boldsymbol{\alpha}_i^{(k)} \leftarrow$ solving Eq. (4).
 9:         $\boldsymbol{\beta}_i^{(k+1)} \leftarrow$ solving Eq. (6).
10:     end for
11: end while
```

After using SCOTO for feature selection, for the ith block, the features, which correspond to the entries in $\boldsymbol{\beta}_i$ with the value larger than 0, will be selected. So we can finally obtain the new feature matrices $\mathbf{F}_i'$ $(i = 1,\ldots,N)$ by selecting the columns in $\mathbf{F}_i$ corresponding to the selected features.

For the computational complexity of SCOTO, taking coupled features as an example, please note that in Algorithm 1, Line.1-Line.4 are for parameter initialization part for each block (without any loop), whose complexity is greatly lower than that of Line.6-Line.11 (iteration part). Thus, we can obtain the complexity of SCOTO by only analyzing the iteration part. Specifically, Eq. (4) mainly consists of matrix inversion with complexity as $O((l_i+u_i)^3)$, calculating $\mathbf{Q}_i$ with $O(dE(l_i+u_i)^2)$, 1st order derivative with $O(d^2E^2(l_i+u_i))$, and second order derivative with $O(d^2E^2)$. Eq. (6) adopts soft-thresholding strategy with complexity as $O(dE)$. Also, we know that in practice $d \gg l_i+u_i$. Thus, we can obtain the total complexity of SCOTO as $O(d^2E^2(l_i+u_i)Nk)$ for coupled features, and $O(d^2(l_i+u_i)Nk)$ for original features, respectively. In addition, $(l_i+u_i)N$ actually equals to the total number of voxels in each slice. Moreover, the iterative projected gradient descent method [1] can mathematically converge in function value as $O(1/k^2)$, which is relatively fast to compute, and widely adopted in related optimization problems. Furthermore, we illustrate the convergence of using SCOTO for the top, central, and bottom slices of Patient 1 as an example in Fig. 7.

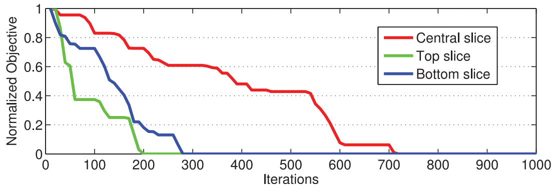


Fig. 7. An illustration for convergence of SCOTO (PA1).

5.5 Theoretical Analysis: Joint Convexity

The requirement of employing the alternating optimization method for solving Eq. (2) is that all the unknown variables $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_N$ should be joint convex. We first conceptively give the separate definitions of the convex (see Lemma 1) and joint convex (see Lemma 2).

Lemma 1 (Boyd and Vandenberghe [3]). *A function $f(x)$ is said to be convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$,*

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

Lemma 2 (Boyd and Vandenberghe [3]). *If $f(x,y) : \mathbb{R}_x^{d_1} \times \mathbb{R}_y^{d_2}$ is convex (as a function of $z = (x,y)$; this is called joint convexity) and the function*

$$g(x) = \inf_y f(x,y)$$

is proper ($> \infty$) everywhere and is finite at least at one point, then g is convex.

Lemma 3 (Boyd and Vandenberghe [3]). *If H function $f_1(x)$, $f_2(x),\ldots, f_H(x)$ are all convex functions, then $f_1(x) + f_2(x) + \cdots + f_H(x)$ are convex.*

Lemma 3 indicates the property that the linear combination of several convex functions is also convex. Here, the Eq. (2) can be normally transformed into the following four sub-terms:

$$\begin{cases} \mathcal{T}_1 = \sum_{i=1}^{N} \|\mathbf{J}_i(\mathbf{y}_i - \mathbf{F}_i\boldsymbol{\beta}_i)\|_2^2, \\ \mathcal{T}_2 = \sum_{i=1}^{N} \|\boldsymbol{\beta}_i\|_1, \\ \mathcal{T}_3 = \sum_{i=1}^{N} \boldsymbol{\beta}_i^\top \mathbf{F}_i^\top \mathcal{L}_i \mathbf{F}_i \boldsymbol{\beta}_i, \\ \mathcal{T}_4 = \sum_{i=1}^{N} \sum_{j\in H(i)} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2. \end{cases}$$

Our sketch for joint convexity proof is to individually validate that all the four sub-terms (i.e., $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, $\mathcal{T}_4$) are joint convex. Thus, we can also consequently obtain that the Eq. (2) is joint convex using Lemma 3.

Proposition 1. *Both the sub-terms $\mathcal{T}_1$ and $\mathcal{T}_2$ are joint convex.*

Proof. The proof is trivial by employing mathematical induction (M.I.). □

Proposition 2. *The sub-term $\mathcal{T}_3$ is joint convex.*

Proof. According to the literature (He et al. [18]),

$$\boldsymbol{\beta}^\top \mathbf{F}^\top \mathcal{L} \mathbf{F} \boldsymbol{\beta} = \frac{1}{2}\sum_{m,n} [(\mathbf{x}_m - \mathbf{x}_n)^\top \boldsymbol{\beta}]^2 \mathbf{W}_{m,n},$$

where $\mathbf{x}_m \in \mathbb{R}^d$ and $\mathbf{x}_n \in \mathbb{R}^d$ are the mth and nth feature vector, respectively. $\mathbf{W}_{m,n}$ is the pair-wise similarity of $\mathbf{x}_m$ and $\mathbf{x}_n$. Taking two variables $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as an example, for $\forall\lambda$ $(0 \le \lambda \le 1)$, $\forall\boldsymbol{\beta}_1^A, \boldsymbol{\beta}_1^B \in \Omega(\boldsymbol{\beta}_1)$ (with $\Omega(x)$ as the domain of x), and $\forall\boldsymbol{\beta}_2^A, \boldsymbol{\beta}_2^B \in \Omega(\boldsymbol{\beta}_2)$, we can let $\mathcal{X}_{m,n} = \mathbf{x}_m - \mathbf{x}_n$, and have

$$\begin{aligned} &\mathcal{T}_3\Big(\lambda\boldsymbol{\beta}_1^A + (1-\lambda)\boldsymbol{\beta}_1^B, \lambda\boldsymbol{\beta}_2^A + (1-\lambda)\boldsymbol{\beta}_2^B\Big) \\ &= \frac{1}{2}\sum_{m,n} \Big[\mathcal{X}_{m,n}^\top\big(\lambda\boldsymbol{\beta}_1^A + (1-\lambda)\boldsymbol{\beta}_1^B\big)\Big]^2 \mathbf{W}_{m,n} \\ &\quad + \frac{1}{2}\sum_{m,n} \Big[\mathcal{X}_{m,n}^\top\big(\lambda\boldsymbol{\beta}_2^A + (1-\lambda)\boldsymbol{\beta}_2^B\big)\Big]^2 \mathbf{W}_{m,n} \\ &= \frac{1}{2}\sum_{m,n} \Big[\lambda\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_1^A + (1-\lambda)\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_1^B\Big]^2 \mathbf{W}_{m,n} \\ &\quad + \frac{1}{2}\sum_{m,n} \Big[\lambda\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_2^A + (1-\lambda)\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_2^B\Big]^2 \mathbf{W}_{m,n}. \end{aligned} \tag{7}$$

$$\begin{aligned} &\lambda\mathcal{T}_3\big(\boldsymbol{\beta}_1^A, \boldsymbol{\beta}_2^A\big) + (1-\lambda)\mathcal{T}_3\big(\boldsymbol{\beta}_1^B, \boldsymbol{\beta}_2^B\big) \\ &= \frac{1}{2}\sum_{m,n} \Big[\lambda\big(\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_1^A\big)^2 + (1-\lambda)\big(\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_1^B\big)^2\Big] \mathbf{W}_{m,n} \\ &\quad + \frac{1}{2}\sum_{m,n} \Big[\lambda\big(\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_2^A\big)^2 + (1-\lambda)\big(\mathcal{X}_{m,n}^\top\boldsymbol{\beta}_2^B\big)^2\Big] \mathbf{W}_{m,n}. \end{aligned} \tag{8}$$

Moreover, we can obtain that

$$\begin{aligned} &\text{Eq.(8)-Eq.(7)} \\ &= \frac{1}{2}\sum_{m,n} (1-\lambda)\big[\mathcal{X}_{m,n}^\top\big(\boldsymbol{\beta}_1^A - \boldsymbol{\beta}_1^B\big)\big]^2 \mathbf{W}_{m,n} \\ &\quad + \frac{1}{2}\sum_{m,n} (1-\lambda)\big[\mathcal{X}_{m,n}^\top\big(\boldsymbol{\beta}_2^A - \boldsymbol{\beta}_2^B\big)\big]^2 \mathbf{W}_{m,n} \ge 0. \end{aligned}$$

For the case with all ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,..., $\boldsymbol{\beta}_N$), the property still holds using the M.I. Therefore, we can conclude that the sub-term $\mathcal{T}_3$ is joint convex. □

Proposition 3. *The sub-term $\mathcal{T}_4$ is joint convex.*

Proof. Taking two variable $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as an example, for $\forall\lambda$ $(0 \le \lambda \le 1)$, $\forall\boldsymbol{\beta}_1^A, \boldsymbol{\beta}_1^B \in \Omega(\boldsymbol{\beta}_1)$, and $\forall\boldsymbol{\beta}_2^A, \boldsymbol{\beta}_2^B \in \Omega(\boldsymbol{\beta}_2)$, we have

$$\begin{aligned} &\mathcal{T}_4\Big(\lambda\boldsymbol{\beta}_1^A + (1-\lambda)\boldsymbol{\beta}_1^B, \lambda\boldsymbol{\beta}_2^A + (1-\lambda)\boldsymbol{\beta}_2^B\Big) \\ &= \sum_{1,j}\sum_{1,j} \|\lambda\boldsymbol{\beta}_1^A + (1-\lambda)\boldsymbol{\beta}_1^B - \lambda\boldsymbol{\beta}_{1,j}^A - (1-\lambda)\boldsymbol{\beta}_{1,j}^B\|_2^2 \\ &\quad + \sum_{2,j}\sum_{2,j} \|\lambda\boldsymbol{\beta}_2^A + (1-\lambda)\boldsymbol{\beta}_2^B - \lambda\boldsymbol{\beta}_{2,j}^A - (1-\lambda)\boldsymbol{\beta}_{2,j}^B\|_2^2. \end{aligned} \tag{9}$$

$$\begin{aligned} &\lambda\mathcal{T}_4\big(\boldsymbol{\beta}_1^A, \boldsymbol{\beta}_2^A\big) + (1-\lambda)\mathcal{T}_3\big(\boldsymbol{\beta}_1^B, \boldsymbol{\beta}_2^B\big) \\ &= \lambda\sum_{1,j} \big\|\big(\boldsymbol{\beta}_1^A - \boldsymbol{\beta}_{1,j}^A\big)\big\|_2^2 + (1-\lambda)\sum_{1,j} \big\|\big(\boldsymbol{\beta}_1^B - \boldsymbol{\beta}_{1,j}^B\big)\big\|_2^2 \\ &\quad + \lambda\sum_{2,j} \big\|\big(\boldsymbol{\beta}_2^A - \boldsymbol{\beta}_{2,j}^A\big)\big\|_2^2 + (1-\lambda)\sum_{2,j} \big\|\big(\boldsymbol{\beta}_2^B - \boldsymbol{\beta}_{2,j}^B\big)\big\|_2^2. \end{aligned} \tag{10}$$

Moreover, we can obtain that

$$\begin{aligned} &\text{Eq.(10)-Eq.(9)} \\ &= \sum_{1,j}\sum_{1,j} \lambda(1-\lambda)\big\|\lambda\boldsymbol{\beta}_1^A + \lambda\boldsymbol{\beta}_1^B - \lambda\boldsymbol{\beta}_{1,j}^A - \lambda\boldsymbol{\beta}_{1,j}^B\big\|_2^2 \\ &\quad + \sum_{2,j}\sum_{2,j} \lambda(1-\lambda)\big\|\lambda\boldsymbol{\beta}_2^A + \lambda\boldsymbol{\beta}_2^B - \lambda\boldsymbol{\beta}_{2,j}^A - \lambda\boldsymbol{\beta}_{2,j}^B\big\|_2^2 \ge 0. \end{aligned}$$

For the case with all ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,..., $\boldsymbol{\beta}_N$), the property still holds using the M.I. Thus, it mathematically validates the sub-term $\mathcal{T}_4$ is joint convex. □

Theorem 1. *The objective function of Eq. (2) is convex.*

Proof. According to Lemma 2, and Proposition 1-3, the Theorem 1 can be proved. □

5.6 Prostate-Likelihood Estimation

After applying SCOTO for joint feature selection across different blocks, we can obtain the new feature matrices $\mathbf{F}_i'(i = 1, \ldots, N)$ by removing the unselected features. In order to further estimate the prostate-likelihood for each block (voxel-wise), for each individual block, we apply SVR, which is a conventional regression method with a supervised manner, to predict the prostate-likelihood for all the voxels in each block. Specifically, SVR model is first trained by the training voxels in $\mathbf{F}_i'$ as well as the available labels in $\mathbf{y}_i$, and then preformed over the u_i testing voxels on the ith block for prediction of prostate. All the predicted likelihood will be finally normalized into $[0, 1]$. It is noteworthy that we will first obtain 2D prostate-likelihood maps slice by slice, and then merge all those 2D results to get the final 3D prostate-likelihood map according to the ordinary order, which is denoted as $\mathbf{M}_n$.

6 MULTI-ATLASES BASED LABEL FUSION

Here, we introduce the technical details of the multi-atlases based label fusion, which generates the final binary segmentation results according to the prostate-likelihood map (continuous value) obtained in the previous step. Although the prostates scanned at different days usually have large irregular motion, non-linear transformation, and even appearance changing, the patient-specific prostate shape information can still be well used to guide final segmentation, since the shape differences among the same patient are generally smaller than that among different patients, in terms of the volume size, general shape, etc. To make full use of all the shape information from the planning and previous treatment images for guiding the segmentation, we adopt the multi-atlases based label fusion with the following steps:

- First step is rigid alignment. All previous binary segmentation results (the ground truth of the training images manually delineated by radiation oncologist) $\mathbf{G}_1, \ldots, \mathbf{G}_{n-1}$ and $\mathbf{G_p}$ will be rigidly aligned to the estimated prostate-likelihood map $\mathbf{M}_n$ (after using coupled features for representation and also SCOTO for feature selection) by using the mutual-information based similarity metric with Powell's optimization strategy [27], which is implemented by FLIRT toolkit [20]. Thus, the output of first step are the aligned results of previous training images, in which the value of each voxel is of the range [0, 1] with the higher value indicating the higher likelihood to be prostate voxel.
- Second step is voxel-wise majority voting. After the rigid alignment, we average all the obtained aligned results and further conduct the voxel-wise majority voting for each voxel of current treatment image (i.e., testing image). Specifically, if the averaged likelihood of a voxel is larger than 0.5, it will be classified as prostate voxel, otherwise as non-prostate voxel. Thus, the output of second step is the final binary segmentation result denoted as $\mathbf{S}_n$.

The flowchart of the above steps in our whole segmentation method can be referred to the *multi-atlases based label fusion* in Fig. 4. The multi-atlases based label fusion is a simple yet effective schema for final binary prostate segmentation. In the experimental evaluation, we will validate the advantages of the multi-atlases based label fusion step in our whole segmentation method.

7 EXPERIMENTAL RESULTS

Now we present the qualitative and quantitative evaluation to validate the advantages of the proposed method. After the discussion on the CT dataset, experimental setting, and parameter selection mechanism, we first investigate the effectiveness of the coupled feature representation by comparing it with the method only using the low-level original features [37]. Then, we evaluate the proposed SCOTO and compare it with several popular feature selection methods. Also, we extensively present the efficacy of the multi-atlases based label fusion as the second step in our whole method. Moreover, we will also discuss the performance under two particular cases, i.e., (1) when inaccurate manual specification happens, and (2) when patients are with large irregular prostate motion. Finally, we compare the proposed method with several state-of-the-art methods, and also report additional quantitative results.

Please note that, to better distinguish the previous method (published in [37]) using the original features (OF) and the current novel method using the coupled features, we denote the previous method as OF+SCOTO, and the current novel method as CF+SCOTO, in the following sections.

7.1 Dataset and Settings

We systematically evaluated the proposed method on a CT prostate dataset consisting of 24 different patients with totally 330 images. Here, each patient has at least nine images (nine CT scans) obtained from one planning day and several treatment days. The original resolution of each CT image is $512 \times 512 \times 60$, with in-plane voxel size of $0.98 \times 0.98\ \text{mm}^2$ and the inter-slice thickness of 3 mm.

All CT images of patients are manually segmented by an experienced radiation oncologist, which we used as ground-truth for quantitative evaluation in the following experiments. For each patient, the first three images (i.e., the planning image and the first two treatment images) are used as training images, from which the training voxels are sampled, along with the available segmentation ground-truths. As for the clinical feasibility of our setting, first we need to clarify that, for each patient, the number of scans to be segmented is way more than three images (usually 20-40 scans per patient), so manual segmentation of the first three scans can still be considered as a minor effort for radiation oncologist.

For quantitative evaluation, we employ four common evaluation metrics: the Dice ratio, the true positive fraction (TPF), the average surface distance (ASD), and the centroid distance (CD), which are also popularly used in the related literature [26], [27], [16]. Specifically, the Dice ratio between two binary images A and B can be calculated as $2|A \cap B|/(|A| + |B|)$. The TPF indicates that the percentage of corrected predicted prostate voxels in the manually segmented prostate regions. The ASD is adopted to discretely measure the euclidean distance between the surfaces of the manual segmentation result and the predicted segmentation result. The CD means the euclidean distance between the central locations of the manual segmentation result and the predicted segmentation result. Since prostate CT-images are 3D, the CD along three directions, including the lateral (x-axis), anterior-posterior (y-axis), and superior-inferior (z-axis) directions, need to be calculated. Please note that, in the superior-inferior (z-axis) direction, the CD is calculated as three times of the obtained value since the inter-slice voxel size is 3 mm, which is approximately three times of that in the x-axis and y-axis.

7.2 Parameter Selection Mechanism

For SCOTO as a feature selection tool, the parameters λ_S, λ_L, λ_E are automatically selected by grid searching with leave-one-out cross-validation on three patient-specific training images independently. The parameters λ_S, λ_L, λ_E, which can obtain the lowest mean square error (MSE) value, are selected and used for the feature selection on the testing images. Specifically, λ_S, λ_L and λ_E are selected from $(10^{-9}, 10^{-8}, \ldots, 10^{-1})$. Using the aforementioned mechanism, we experimentally found $\lambda_S = 10^{-6}$, $\lambda_L = 10^{-8}$, $\lambda_E = 10^{-3}$ (for OF+SCOTO), and $\lambda_S = 10^{-4}$, $\lambda_L = 10^{-7}$, $\lambda_E = 10^{-2}$ (for CF+SCOTO, $E = 4$) to be best. As for the running time affected by the grid searching, for CF+SCOTO, we

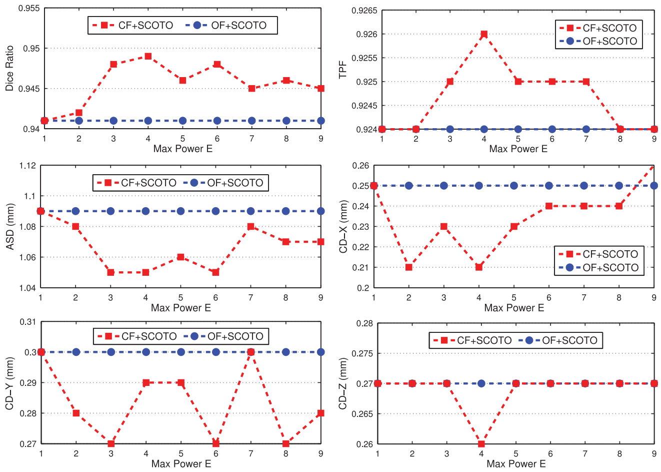


Fig. 8. Comparison of Dice Ratio, TPF, ASD, CD on 24 patients between OF+SCOTO and CF+SCOTO by using different maximum power number E.

spend a couple of days to search the best parameters according to the previous selected parameters in OF+SCOTO, which is used to accelerate the searching time. It is worth noting that searching the optimal parameters only applies to the training, which will not affect our testing time.

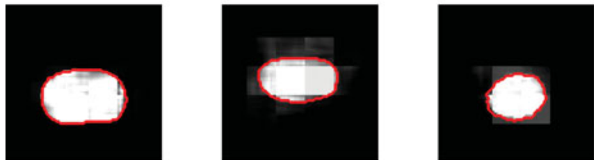
For parameters in prostate-likelihood regression, we implement the SVR by LIBSVM toolbox [5], in which all the parameters (e.g., c, g) are set to the default ones.

For the size of blocks, we empirically set to 10 (note that the slice size after ROI extraction is 140×140). Choosing either too large or too small block size is impracticable. Too large block size will ignore the variations of appearance along the prostate boundary, while too small block size will increase the unnecessary computational burden. However, it is still an open problem on how to automatically choose the best block size, which we will study in our future work.

7.3 Efficacy of Coupled Features

Here, we evaluate the efficacy of coupled feature representation for prostate segmentation. In our previous work [37], only low-level original features are used; in our current work, by introducing the coupled features, we use both original features and coupled features. Please note that the subsequent feature selection (using SCOTO) and the final segmentation (using multi-atlases based label fusion) are same during the comparison. Specifically, we investigate the segmentation performance by varying the maximum power number E (the only parameter in coupled features [42]) from 1 to 9. $E = 1$ indicates the case that we only use the original features (without feature matrix expansion). The larger E means the more complex feature representation is employed. We compare the results of Dice ratio, TPF, ASD and CD (x-/y-/z-axis) between the coupled features (CF+SCOTO) and original features (OF+SCOTO) on 24 patients, respectively (see Fig. 8). By varying different E, the results indicate that CF+SCOTO is better than OF+SCOTO in terms of higher Dice ratio, TPF, and lower ASD, CD. In addition, we can notice that the CDs in z-axis are almost same between CF+SCOTO and OF+SCOTO. The reason might be that we asked the radiation oncologist to provide the first and last slice indexes of prostate region in z-axis, thus both methods (i.e., CF+SCOTO and OF+SCOTO) can use the same information. In the following experiments, we choose $E = 4$ as the tradeoff between the computational burden and segmentation performance. When $E = 4$, CF+SCOTO statistically outperforms OF+SCOTO in terms of Dice ratio with p-value as $5.14e-3$.

(a) Example 1 (OF) (b) Example 2 (OF) (c) Example 3 (OF)



(d) Example 1 (CF) (e) Example 2 (CF) (f) Example 3 (CF)

Fig. 9. Three typical prostate-likelihood maps showing that OF+SCOTO cannot achieve good results, but CF+SCOTO can. Red curves indicate manual delineations.

Also, we illustrate three typical prostate-likelihood maps where OF+SCOTO cannot achieve good performance in our pervious work [37] (see Fig. 9, with red contour as the manual delineation by radiation oncologist). By using the coupled features, we can observe that the prostate-likelihood maps of CF+SCOTO are much better than that of OF+SCOTO.

Moreover, to show the intrinsic relations among features, we report the normalized selection percentages by using SCOTO for three kinds of features (i.e., HoG, LBP, and Haar wavelet) in Table 3. Specifically, here we illustrate the selection percentages by averaging the feature selection results of each slice for 24 patients. In Table 3, $\mathcal{P}_{\text{HoG}}$ indicates the percentage of original HoG features in all the selected features (similar as $\mathcal{P}_{\text{LBP}}$ and $\mathcal{P}_{\text{Haar}}$), while $\mathcal{P}_{\text{HoG-c}}$ indicates the percentage of coupled HoG features within expanded matrix in all the selected features (similar as $\mathcal{P}_{\text{LBP-c}}$ and $\mathcal{P}_{\text{Haar-c}}$). We can observe that, by relaxing the feature independency assumption, certain coupled features will be further selected by replacing the original features. Also, it is worth noting that, for LBP features, which are seldom selected using the original feature representation, the selection percentage will be greatly increased when using coupled feature representation.

7.4 Evaluation on the SCOTO

We now evaluate the performance of SCOTO on the feature selection task for both original features and coupled features, respectively. Specifically, SCOTO belongs to the block-level feature selection methods, which simultaneously selects the features within different local regions. Thus, to validate if the block-level feature selection is better than the slice-level feature selection (i.e. the methods that select the features per slice globally), as well as to validate if the manifold assumption constraint can benefit the final segmentation results, several related feature selection

TABLE 1
Comparison of Experimental Results among Different Feature Selection Methods, with the Best Results Marked by Bold Font

Level	**Method**	**Dice** (mean ± std)	**TPF** (mean ± std)	**ASD** (mean ± std) (mm)	**CD** (mean ± std) (x/y/z) (mm)
Slice-Level (OF)	Lasso-S	0.874 ± 0.083	0.869 ± 0.107	1.41 ± 0.83	0.71 ± 0.56 / 0.80 ± 0.61 / 0.67 ± 0.53
	tLasso-S	0.917 ± 0.053	0.899 ± 0.084	1.32 ± 0.78	0.54 ± 0.37 / 0.50 ± 0.38 / 0.40 ± 0.33
Block-Level (OF)	mRMR	0.893 ± 0.033	0.912 ± 0.047	1.38 ± 0.74	0.50 ± 0.34 / 0.72 ± 0.41 / 0.36 ± 0.33
	Lasso-B	0.922 ± 0.039	0.909 ± 0.042	1.35 ± 0.81	0.47 ± 0.39 / 0.47 ± 0.37 / 0.33 ± 0.34
	tLasso-B	0.932 ± 0.036	0.919 ± 0.040	$1.17 \pm \mathbf{0.68}$	$0.37 \pm \mathbf{0.17}$ / 0.41 ± 0.35 / 0.32 ± 0.33
	Fused Lasso	0.928 ± 0.047	0.906 ± 0.043	1.28 ± 0.80	0.34 ± 0.37 / 0.42 ± 0.38 / 0.34 ± 0.51
	OF+SCOTO	$\mathbf{0.941} \pm \mathbf{0.030}$	$\mathbf{0.924} \pm \mathbf{0.037}$	$\mathbf{1.09} \pm 0.74$	$\mathbf{0.25} \pm 0.18$ / $\mathbf{0.30} \pm \mathbf{0.22}$ / $\mathbf{0.27} \pm \mathbf{0.29}$
Slice-Level (CF)	Lasso-S	0.885 ± 0.091	0.887 ± 0.109	1.36 ± 0.81	0.74 ± 0.53 / 0.81 ± 0.63 / 0.63 ± 0.49
	tLasso-S	0.921 ± 0.064	0.903 ± 0.073	1.32 ± 0.79	0.54 ± 0.36 / 0.51 ± 0.34 / 0.40 ± 0.32
Block-Level (CF)	mRMR	0.899 ± 0.031	0.915 ± 0.046	1.35 ± 0.72	0.46 ± 0.33 / 0.70 ± 0.43 / $0.32 \pm \mathbf{0.31}$
	Lasso-B	0.929 ± 0.038	$0.910 \pm \mathbf{0.039}$	1.32 ± 0.80	0.45 ± 0.36 / 0.46 ± 0.37 / 0.32 ± 0.35
	tLasso-B	0.937 ± 0.030	0.919 ± 0.042	1.12 ± 0.66	$0.32 \pm \mathbf{0.18}$ / 0.38 ± 0.30 / 0.31 ± 0.33
	Fused Lasso	0.936 ± 0.048	0.908 ± 0.043	1.25 ± 0.80	0.29 ± 0.33 / 0.40 ± 0.37 / 0.35 ± 0.54
	CF+SCOTO	$\mathbf{0.949} \pm \mathbf{0.028}$	$\mathbf{0.926} \pm 0.041$	$\mathbf{1.05} \pm \mathbf{0.63}$	$\mathbf{0.23} \pm \mathbf{0.18}$ / $\mathbf{0.27} \pm \mathbf{0.22}$ / $\mathbf{0.27} \pm 0.32$

methods are used for extensively comparison, which include Lasso-S and tLasso-S (applying Lasso [39] and tLasso [36] on slice-level feature selection, respectively), Lasso-B and tLasso-B (applying Lasso and tLasso on block-level feature selection, respectively), mRMR [32] and Fused Lasso [40] (applying the respective method on block-level feature selection). For all these methods, the parameters are experimentally set using leave-one-out cross-validation. It is noteworthy that the same multi-atlases based label fusion is adopted for all the methods.

Table 1 lists the segmentation accuracies obtained by different feature selection schemes on all the 24 patients with 330 CT images by using original features and coupled features, respectively, and the best results across different methods are marked by the bold fonts. We found the proposed SCOTO can achieve superior performance over the related methods. Specifically, we also found that (i) the block-level methods are better than the slice-level ones, which validates our assumption that different local regions prefer choosing different features; (ii) Manifold assumption constraint is useful for improving the results, i.e., by comparing tLasso-S with Lasso-S, and tLasso-B with Lasso-B; (iii) Spatial-constraint smoothness term leads to better results, i.e., by comparing SCOTO with tLasso-B; (iv) Coupled features can help boost the performance, i.e., by comparing the using original and coupled features, respectively.

Also, compared with these feature selection methods (Lasso-S, tLasso-S, mRMR, Lasso-B, tLasso-B, and Fused Lasso) on the Dice ratios, our proposed OF+SCOTO and CF+SCOTO are statistically better than their results through calculating the corresponding p-values (see Table 2). Fig. 10 presents the mean Dice ratio among these different feature selection methods per patient, respectively, which demonstrates the advantages of SCOTO on feature selection task.

TABLE 2
The p-Values (the Dice Ratio) of OF+SCOTO and CF+SCOTO Compared with Related Feature Selection Methods: Lasso-S, tLasso-S, mRMR, Lasso-B, tLasso-B and Fused Lasso by Using Both the Original and Coupled Features, Respectively

Using original features (compared with OF+SCOTO)					
Lasso-S	tLasso-S	mRMR	Lasso-B	tLasso-B	Fused Lasso
3.43e-16	1.47e-10	3.86e-31	2.10e-13	7.51e-6	5.08e-9
Using coupled features (compared with CF+SCOTO)					
Lasso-S	tLasso-S	mRMR	Lasso-B	tLasso-B	Fused Lasso
9.51e-15	2.04e-12	5.28e-22	8.42e-16	6.27e-3	1.19e-4

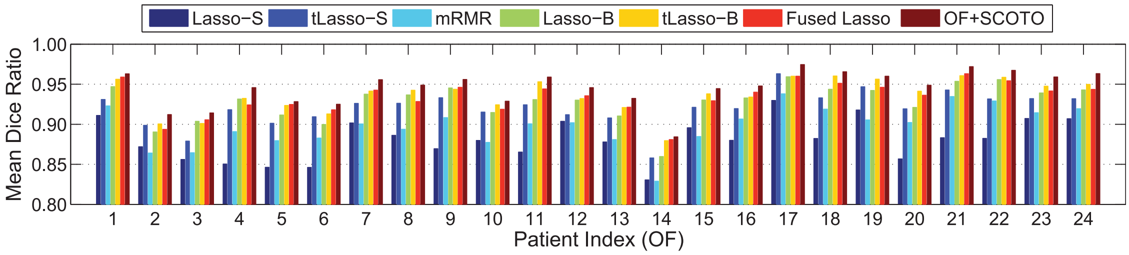
Since the automatic approach for selecting the best block size of SCOTO is not available, we investigate the segmentation performance with respect to the change of predefined block size (see Fig. 11). It can be observed that, by setting the block size from 8 × 8 to 14 × 14, the overall Dice Ratios for CF+SCOTO and OF+SCOTO are all above 0.94, and also remain relatively stable. When setting block size with a small value (i.e., 6 × 6), the performance is poor since it cannot well handle prostate motion problem. In this case, due to large prostate motion, especially in the prostate boundary part, the training and testing voxels will come from very different prostate regions, which makes training voxels not able to fully guide feature selection for the testing voxels. When setting the block size with a large value (i.e., 16 × 16), the prostate region will be covered by a very small number of blocks, and the same block will contain different prostate regions. In this case, the results will be similar as performing Lasso on each slice globally.

7.5 Evaluation on the Multi-Atlases Based Label Fusion

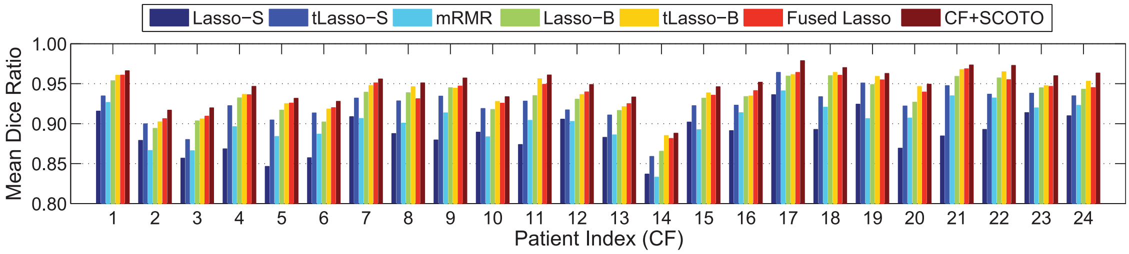
To evaluate the multi-atlases based label fusion, we adopt two related methods, OTSU's segmentation [33] and level-set segmentation [22] for comparison, both of which can also generate the final prostate segmentation results on the prostate-likelihood maps. In our multi-atlases based label fusion, we first apply the original features (or

TABLE 3
Normalized Selection Percentages for Three Kinds of Features

Using original features (OF+SCOTO)					
$\mathcal{P}_{\text{HoG}}$	$\mathcal{P}_{\text{LBP}}$	$\mathcal{P}_{\text{Haar}}$	$\mathcal{P}_{\text{HoG-c}}$	$\mathcal{P}_{\text{LBP-c}}$	$\mathcal{P}_{\text{Haar-c}}$
47.9%	23.2%	28.9%	N/A	N/A	N/A
Using coupled features (CF+SCOTO)					
$\mathcal{P}_{\text{HoG}}$	$\mathcal{P}_{\text{LBP}}$	$\mathcal{P}_{\text{Haar}}$	$\mathcal{P}_{\text{HoG-c}}$	$\mathcal{P}_{\text{LBP-c}}$	$\mathcal{P}_{\text{Haar-c}}$
14.7%	21.5%	5.3%	23.4%	29.5%	5.6%



(a) Using the original features (OF)



(b) Using the coupled features (CF)

Fig. 10. Comparison of mean Dice Ratio among different feature selection methods per patient using OF and CF, respectively.

coupled features), SCOTO and SVR for feature representation, feature selection and regression, respectively, then use the rigid registration to align the patient-specific training images to the estimated prostate-likelihood map, and finally average all the obtained aligned results and utilize the voxel-wise majority voting to get the final segmentation result (by using original and coupled features, respectively). In the implementations of the OTSU's method [33] and the level-set method [22], we first generate the slice-wise segmentation result for each individual slice using the corresponding slice-wise prostate-likelihood map, and then combine the slice-wise segmentation results for final prostate segmentation (by using original and coupled features, respectively). It is noteworthy that all the three strategies use the same prostate-likelihood map (e.g., the prostate-likelihood estimation step is same), for original and coupled features, respectively. Fig. 12 gives mean Dice ratios of all 24 patients among three strategies, which validates that the multi-atlases based label fusion outperforms the two related methods, for both the original (see Fig. 12a) and coupled features (see Fig. 12b), respectively. Also, the multi-atlases based label fusion result is statistically better than the OTSU's method [33] and the level-set method [22] with p-values of $6.56e$-9 and $1.96e$-7 (using original features), and $3.48e$-13 and $3.80e$-8 (using coupled features), respectively.

In Fig. 13, we illustrate two typical examples to explain why the multi-atlases based label fusion can generally achieve better segmentation results compared with the OTSU's method and the level-set method. The overlaid red curves indicate the radiation oncologist's manual delineations which are used as the ground-truth (see Figs. 13a and 13e), while the overlaid yellow curves indicate the respective segmentation results of three strategies with the corresponding prostate-likelihood maps (see Figs. 13b, 13c, 13d, 13f, 13g, and 13h). Although the prostate part can be roughly determined according to the prostate-likelihood map, the prostate boundaries are still hard to distinguish since the large prostate motion and appearance changes greatly deteriorate the regression results in some cases, especially for the pixels around the prostate boundaries. Both the OTSU's method and the level-set method directly work on the estimated prostate-likelihood map without fully using the prostate shape information. Our method, multi-atlases based label fusion, takes the advantages of the patient-specific prostate shape information to guide better segmentation.

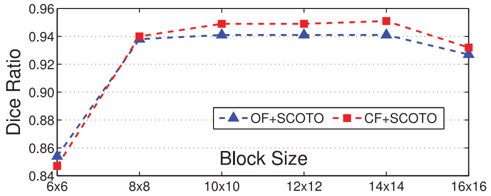
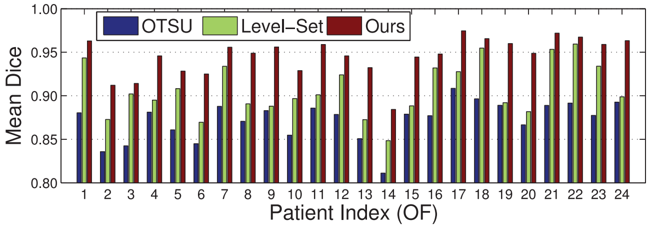
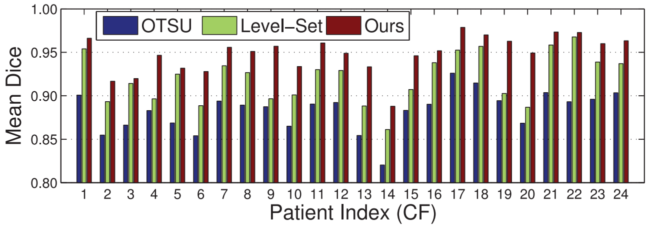


Fig. 11. The segmentation performance with different block sizes.

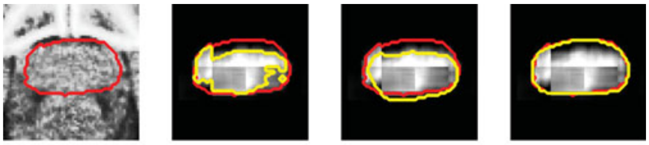


(a) Using the original features (OF)

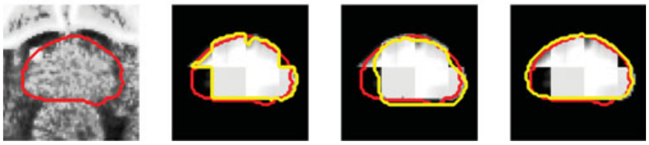


(b) Using the coupled features (CF)

Fig. 12. The mean Dice ratio of all 24 patients among three strategies: OTSU's method, level-set method, multi-atlases based label fusion.



(a) Original image (b) OTSU [33] (c) Level-Set [22] (d) Ours

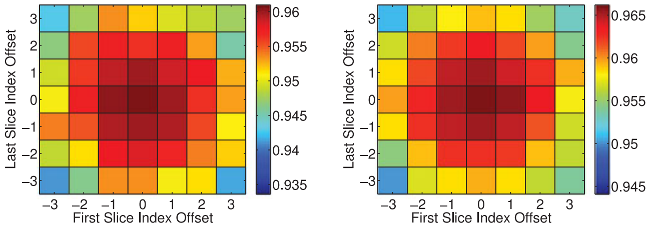


(e) Original image (f) OTSU [33] (g) Level-Set [22] (h) Ours

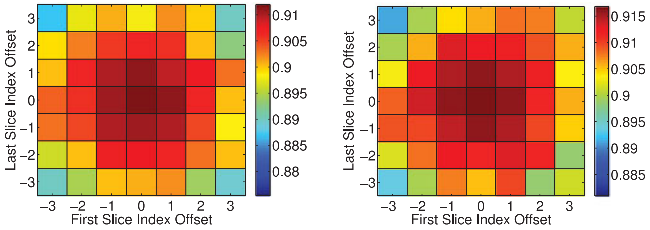
Fig. 13. The two sets of typical segmentation results by different three strategies: OTSU, level-set, and multi-atlases based label fusion, for the cases that the prostate-likelihood estimation results are not good. Red curves indicate the manual delineations, and the yellow curves indicate the respective segmentation results of three different strategies.

7.6 Influence of Inaccurate Manual Specification

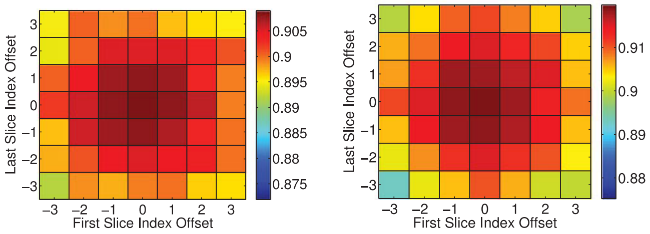
Our method belongs to a semi-automatic method, which requires that the manual specification from the radiation oncologist for the first and last slice indices of the prostate region in the treatment image. The manual specification, which needs only a few seconds, can guide better prostate segmentation (which will be validated in the following parts). However, inaccurate manual specification sometimes happens due to various reasons (e.g., fatigue), especially for the inexperienced radiation oncologists. In this section, we will experimentally investigate the sensitivity of our method with respect to inaccurate manual specification. Specifically, the first and the last slice index offsets will vary from -3 to 3 according to the ground truths (with 0 indicating the correct manual specification) along the superior-inferior direction, respectively. Thus, totally 49 (7×7) combinations of different inaccurate manual specifications have been evaluated. We perform the experiments on the 12, 16 and 17 CT images of patient 1, patient 2, and patient 3, respectively. The respective results are illustrated in Figs. 14a, 14c, and 14e for OF+SCOTO (using the original features), and Figs. 14b, 14d, and 14f for CF+SCOTO (using the coupled features), respectively.



(a) PA1 (OF+SCOTO) (b) PA1 (CF+SCOTO)



(c) PA2 (OF+SCOTO) (d) PA2 (CF+SCOTO)



(e) PA3 (OF+SCOTO) (f) PA3 (CF+SCOTO)

Fig. 14. The mean Dice ratio of different first and last slice index offsets (patient 1, patient 2, and patient 3, for OF+SCOTO and CF+SCOTO, respectively).

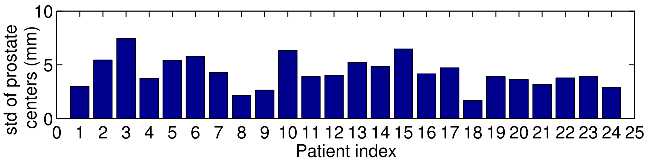


Fig. 15. The standard deviation of prostate centers for each patient.

Fig. 14 discloses that our method can generally obtain the relative satisfactory segmentation results (Dice ratio > 0.9), even in the case that both the indices of the first and the last slice are up to 9 mm away from the truth indices. Thus, in practice, the influence of small inaccurate manual specification to the final segmentation is not significant, which also demonstrates the robustness of our method.

7.7 Patients with Large Prostate Motion

We now validate if the radiation oncologist's simple manual specification can benefit the final segmentation performance for the patients with large irregular prostate motion. In all the 24 patients, it is found that patients 3, 10 and 15 have large prostate motions according to the calculated standard deviation of prostate centers in the planning and treatment images, which can be found by referring to Fig. 15.

By applying the proposed method to patients 3, 10 and 15, the obtained mean Dice ratio are 0.909, 0.928, 0.946 by OF+SCOTO, and 0.920, 0.934, 0.946 by CF+SCOTO, respectively, which are better than the corresponding results reported in [23], [36], [27], [16] (except the result of OF +SCOTO on patient 3 is slightly worse than [16]). The comparisons are also listed in Fig. 16. These results show the effectiveness of the proposed method in incorporating the manual specification from radiation oncologist when large irregular motion occurs in the prostate regions.

7.8 Comparison with Previous Methods

We now further compare the proposed method with several state-of-the-art methods for prostate segmentation in CT images developed in recent years, which include deformable-model based methods [8], [13], registration-based

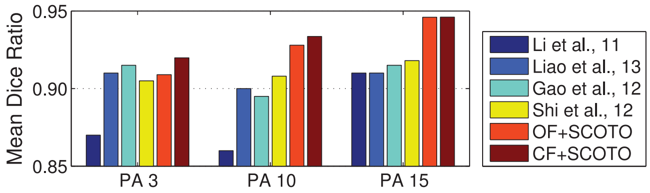


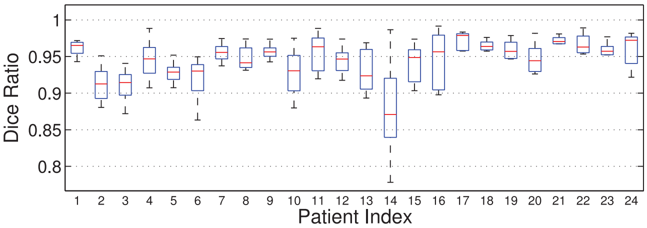
Fig. 16. Comparison on prostate segmentation results of five different methods for three patients with large prostate motions.

TABLE 4
Comparison of Different Performance Measurements (Mean Dice Ratio, Median TPF, and Mean ASD) with Other Related Methods, with the Best Performance Marked by Bold Font

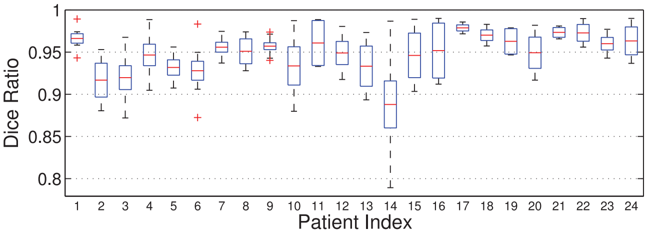
Methods	Mean Dice	Median TPF	ASD (mm)
Davis et al. [11]	0.820	N/A	N/A
Chen et al. [8]	N/A	0.840	1.10± N/A
CT dataset 1: 24 patients, 330 CT images			
Feng et al. [13]	0.893	N/A	2.08 ± 0.79
Liao and Shen [26]	0.899	N/A	1.08± N/A
Gao et al. [16]	0.913	N/A	1.24 ± 0.77
Liao et al. [27]	0.909	N/A	**0.97**± N/A
Shi et al. [36]	0.920	0.901	1.26 ± 0.78
Wu et al. [43]	0.914	N/A	1.30 ± **0.49**
OF+SCOTO	0.941	0.932	1.09 ± 0.74
CF+SCOTO	**0.949**	**0.934**	1.05 ± 0.63
CT dataset 2: 11 patients, 164 CT images			
Li et al. [23]	0.908	0.900	1.40± N/A
Shi et al. [36]	0.923	0.911	1.27 ± 0.80
OF+SCOTO	0.940	0.923	1.13 ± 0.77
CF+SCOTO	**0.947**	**0.929**	**1.06** ± **0.69**
CT dataset 3: 10 patients, 176 CT images			
Liao and Shen [26]	0.896	N/A	N/A
Shi et al. [36]	0.922	0.907	1.28 ± 0.80
OF+SCOTO	0.936	0.923	1.13 ± **0.77**
CF+SCOTO	**0.941**	**0.925**	**1.11** ± 0.78

methods [11], [26], [27], and learning-based method [23], [16], [36], [43]. The best results reported in the corresponding literatures are adopted for comparison in this section. The comparisons among different methods are listed in Table 4. The results obtained not on the same CT dataset are listed separately. Evaluated metrics include mean Dice ratio, mean ASD, and median TPF (Note that, in [8], [23], median TPF are evaluated, instead of mean TPF).

Because different CT datasets are used for experiments in Davis et al.'s work [11] and Chen et al.'s work [8], the results are here listed separately for experimental reference. For [13], [26], [36], [16], [27], [43], all the 24 patients are evaluated, which is the same with ours, so we call the 24 patients CT dataset as *CT dataset 1*. Also, two different subsets of the 24 patients are selected in [23] and [26], which are called as *CT dataset 2* and *CT dataset 3*, respectively. From the results listed in Table 4, we can notice that for OF+SCOTO in *CT dataset 1* (24 patients), the mean Dice ratio is 0.941, median TPF is 0.932, and mean ASD is 1.09 mm; in *CT dataset 2* (11 patients), the mean Dice ratio is 0.940, median TPF is 0.923, and mean ASD is 1.13 mm; in *CT dataset 3* (10 patients), the mean Dice ratio is 0.936, median TPF is 0.923, and mean ASD is 1.13 mm. Also for CF+SCOTO, in *CT dataset 1*, the mean Dice ratio is 0.949, median TPF is 0.934, and mean ASD is 1.05 mm; in *CT dataset 2*, the mean Dice ratio is 0.947, median TPF is 0.929, and mean ASD is 1.06 mm; in *CT dataset 3*, the mean Dice ratio is 0.941, median TPF is 0.925, and mean ASD is 1.11 mm. These results demonstrated that the proposed methods outperform the related methods in terms of higher mean Dice ratio, median TPF, and mean ASD (expect for [27], which is better than ours). Also, we can observe that learning-based segmentation methods are generally better than non-learning-based ones, since voxel-level regression/classification might be more useful to discover the local detail information, e.g., the slight boundary change in different treatment days.



(a) Using original features (OF+SCOTO)



(b) Using coupled features (CF+SCOTO)

Fig. 17. The box-and-whisker diagrams (OF+SCOTO and CF+SCOTO) of Dice ratio for 24 patients.

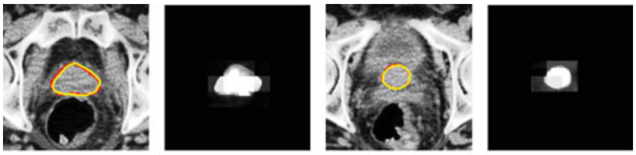
7.9 More Results by Our Method

We also report the box-and-whisker diagram of Dice ratio in Fig. 17 for each individual patient. For statistical perspective, quartile-representation is adopted, in which five horizontal lines (ascending order in values) mean the min, 25 percent percentile, median, 75 percent percentile, and the max value, respectively. Fig. 17a shows the result of OF+SCOTO, and Fig. 17b shows the result of CF+SCOTO.

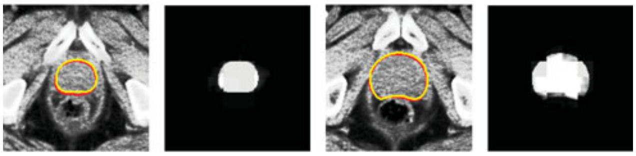
Finally, we also illustrate in Fig. 18 several typical segmented examples as well as prostate-likelihood map for the image 14 of PA 3 (with mean Dice ratio is 0.898 by OF+SCOTO, and 0.906 by CF+SCOTO), the image 10 of PA 11 (with mean Dice ratio is 0.929 by OF+SCOTO, and 0.933 by CF+SCOTO), the image 5 of PA 16 (with mean Dice ratio is 0.897 by OF+SCOTO, and 0.915 by CF+SCOTO), the image 6 of PA 21 (with mean Dice ratio is 0.978 by OF+SCOTO, and 0.979 by CF+SCOTO), and the image 8 of PA 24 (with mean Dice ratio is 0.924 by OF+SCOTO, and 0.937 by CF+SCOTO), respectively. In Fig. 18, we use the red curves to denote the manual segmentation results by the radiation oncologist, and the yellow curves to denote the segmentation results by the proposed methods. It can be observed that the predicted prostate boundaries are very close to the boundaries delineated by the radiation oncologist. Also the proposed method can accurately separate the prostate regions and background even in the base and apex slices as shown in Figs. 18a and 18d, which are usually considered very difficult to segment.

8 DISCUSSION

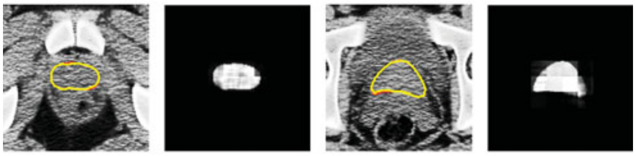
Our method can be extended to many similar segmentation tasks as detailed below. (1) For feature representation using coupled features, it can be directly extended to the similar segmentation applications, of which features are usually not



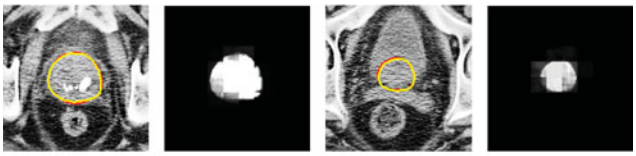
(a) Typical results of the 14^{th} image of patient 3, with Dice Ratio of 0.906.



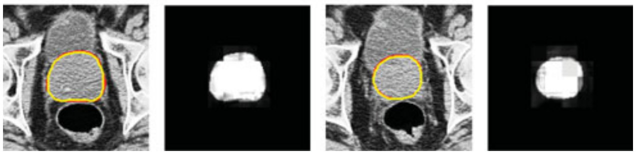
(b) Typical results of the 10^{th} image of patient 11, with Dice Ratio of 0.933.



(c) Typical results of the 5^{th} image of patient 16, with Dice Ratio of 0.915.



(d) Typical results of the 6^{th} image of patient 21, with Dice Ratio of 0.979.



(e) Typical results of the 8^{th} image of patient 24, with Dice Ratio of 0.937.

Fig. 18. Typical segmentation results and estimated prostate-likelihood maps for Patients 3,11,16,21,24. Red curves indicate manual segmentation results by radiation oncologist and the yellow curves indicate the segmentation results by our proposed method.

i.i.d., but related with others. (2) For SCOTO, as a feature selection method, it can be used in many applications, especially for medical image segmentation, of which the local regions are usually spatially connected. (3) For multi-atlases based label fusion, it can be also utilized in many related segmentation tasks, of which the previous scanned data with segmentation results can be borrowed as references for segmentation of future images.

As for the relations between SCOTO and Total Variation (TV) methods, it is known that many TV methods have already been successfully used in many tasks [19], [28]. However, the major difference between SCOTO and TV methods is that, most of TV methods currently focus on the 1D task (sequential problem), while our problem focuses on the 2D task (spatial problem). Also, according to very recent study [45], it shows that 2D TV can obtain superior results but always with very high complexity.

On the other hand, as for the relations between SCOTO and overlapping group sparsity methods [6], [7], [46], it is worth noting that, for prostate segmentation, using structure information based overlapping group strategy to partition blocks (i.e., local regions) into several overlapped groups is not a trivial task. The major challenge is that, due to large appearance changes and irregular prostate motion, it is very difficult and even infeasible to design specific grouping methods (which naturally has certain clinical meanings for neighboring blocks) for each individual slice. In addition, our method can be treated as an overlapping group sparsity methods [6], [7], [46], with very simple and general assumption (i.e., neighboring blocks prefer choosing similar features), and can be adopted to other similar tasks.

Finally, it is very challenging to develop automatic methods for localizing the start and ending slices of prostate region according to our empirical analysis and experimental evaluation, as explained below. (1) Large prostate motion. Basically, for the same patient, the start and ending slices of the prostate could change up to 10 voxel (3 cm) across different treatment days even after rigid registration with pelvic bone structures. (2) Non-prostate and prostate regions are extremely similar in the start (ending) slices. Many state-of-the-art methods [23], [13] cannot perform well in those regions. (3) Lack of training images. The treatment days usually last for 20-40 days, thus we should limit the radiation oncologist's effort for manual delineation and providing the training images (which generally include one planning and two-four treatment images for training). Thus, it is hard to predict the start and ending slices with only three-five training images, especially for the case that CT images have poor tissue contrast. Therefore, automatic localization of the start and ending slices is commonly considered as the most difficult part in prostate segmentation. In this article, our goal is mainly to incorporate the radiation oncologist's guidance to determine the most difficult part, and then utilize our method for the subsequent segmentation.

9 CONCLUSION

In this article, we have presented a novel semi-automatic learning method for prostate segmentation in CT images during the image-guided radiotherapy. Previous methods directly employ the low-level features without considering the relations of these features. Also, previous methods usually ignore the image appearance changes in different local regions of CT images during the feature selection step. Moreover, previous automatic segmentation methods seldom leverage the radiation oncologist's simple specification for performance improvement.

In our method, we first ask the radiation oncologist to spend a few seconds for the simple specification of ending slices on the current treatment image. The coupled feature representation is then employed by using both the intra- and inter-coupling information from low-level original features. Our proposed SCOTO is further used to simultaneously select the discriminative features for different local regions, for helping estimate the prostate-likelihood map. Finally, the multi-atlases based label fusion method is used to combine the segmentation results of the planning and previous treatment images for final segmentation, which allows to make the full use of the patient-specific shape information.

We extensively evaluate our method on a CT prostate dataset, which consists of 24 patients with totally 330 images, all along with manual delineation results by the experienced radiation oncologist (used as the ground truth). Experimental results show that our proposed method not only obtains superior segmentation performance (i.e., higher Dice Ratio and TPF, and lower ASD and CD) compared with the state-of-the-art methods, but also demonstrates its clinical feasibility since the segmentation performance can be improved by taking only a few seconds of radiation oncologist for simple specification of ending slices, especially for the cases with large irregular prostate motions. Moreover, several experiments also separately validate the advantages of using (1) coupled feature representation, (2) SCOTO and (3) the multi-atlases based label fusion, as well as the robustness of our method to inaccurate manual specification. In our future work, more advanced features and more intelligent way (i.e., by borrowing population information) will be investigated to further improve the current work.

ACKNOWLEDGMENTS

The work was supported by NIH (CA140413), NSFC (61432008, 61305068, 61321491, 61473190, 61422204, 61473149), Jiangsu Nature Science Foundation (NSF) (BK20130581), and Jiangsu NSF for Distinguished Young Scholar (BK20130034). Preliminary version of this work has been published in our CVPR 2013 paper [37]. Yang Gao and Dinggang Shen are the corresponding authors.

REFERENCES

[1] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[2] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learning Res.*, vol. 7, pp. 2399–2434, 2006.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[4] T. Chan, S. Esedoglu, and M. Nikolov, "Algorithms for finding global minimizers of image segmentation and denoising models," *SIAM J. Appl. Math.*, vol. 66, pp. 1632–1648, 2006.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, 2011.

[6] C. Chen, Z. Peng, and J. Huang, "O(1) algorithms for overlapping group sparsity," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 1645–1650.

[7] C. Chen, J. Huang, L. He, and H. Li, "Preconditioning for accelerated iteratively reweighted least squares in structured sparsity reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2713–2720.

[8] S. Chen, D. Lovelock, and R. Radke, "Segmenting the prostate and rectum in CT imagery using anatomical constraints," *Med. Image Anal.*, vol. 15, pp. 1–11, 2011.

[9] T. Chen, S. Kim, J. Zhou, D. Metaxas, G. Rajagopal, and N. Yue, "3D meshless prostate segmentation and registration in image guided radiotherapy," in *Proc. 12th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2009, pp. 43–50.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.

[11] B. Davis, M. Foskey, J. Rosenman, L. Goyal, S. Chang, and S. Joshi, "Automatic segmentation of intra-treatment CT images for adaptive radiation therapy of the prostate," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2005, pp. 442–450.

[12] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani, "Least angle regression," *Ann. Statistics*, vol. 32, pp. 407–499, 2003.

[13] Q. Feng, M. Foskey, W. Chen, and D. Shen, "Segmenting CT prostate images using population and patient-specific statistics for radiotherapy," *Med. Phys.*, vol. 37, pp. 4121–4132, 2010.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. 2nd Eur. Conf. Comput. Learning Theory*, 1995, pp. 23–37.

[15] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability. Philadelphia, PA, USA: SIAM, 2007.

[16] Y. Gao, S. Liao, and D. Shen, "Prostate segmentation by sparse representation based classification," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2012, pp. 452–458.

[17] Y. Gao, R. Sandhu, G. Fichtinger, and A. Tannenbaum, "A coupled global registration and segmentation framework with application to magnetic resonance prostate imagery," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1781–1794, Oct. 2010.

[18] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," presented at the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2005.

[19] J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Med. Image Anal.*, vol. 15, no. 5, pp. 670–679, 2011.

[20] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, pp. 825–841, 2002.

[21] T. Langerak, U. van der Heide, A. Kotte, M. Vigergever, M. van Vulpen, and J. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Trans. Med. Imag.*, vol. 29, no. 12, pp. 2000–2008, Dec. 2010.

[22] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3243–3254, Dec. 2010.

[23] W. Li, S. Liao, Q. Feng, W. Chen, and D. Shen, "Learning image context for segmentation of prostate in CT-guided radiotherapy," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2011, pp. 570–578.

[24] W. Li, S. Liao, Q. Feng, W. Chen, and D. Shen, "Learning image context for segmentation of the prostate in CT-guided radiotherapy," *Phys. Med. Biol*, vol. 57, no. 5, p. 1283, 2012.

[25] D. C. Li and C. W. Liu, "Extending attribute information for small data set classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 452–464, Mar. 2012.

[26] S. Liao and D. Shen, "A feature based learning framework for accurate prostate localization in CT images," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3546–3559, Aug. 2012.

[27] S. Liao, Y. Gao, J. Lian, and D. Shen, "Sparse patch-based label propagation for accurate prostate localization in CT images," *IEEE Trans. Med. Imaging*, vol. 32, no. 2, pp. 419–434, Feb. 2013.

[28] J. Liu, L. Yuan, and J. Ye, "An efficient algorithm for a class of fused lasso problems," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 323–332.

[29] C. Lu, Y. Zheng, N. Birkbeck, J. Zhang, T. Kohlberger, C. Tietjen, T. Boettger, J. Duncan, and S. Zhou, "Precise segmentation of multiple organs in ct volumes using learning-based approach and information theory," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2012, pp. 462–469.

[30] Z. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *J. Optimization Theory Appl.*, vol. 72, no. 1, pp. 7–35, 1992.

[31] G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[32] G. Mallat, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[33] M. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[35] D. Shen, Z. Lao, J. Zeng, W. Zhang, I. A. Sesterhenn, L. Sun, J. W. Moul, E. H. Herskovits, G. Fichtinger, and C. Davatzikos, "Optimized prostate biopsy via a statistical atlas of cancer spatial distribution," *Med. Image Anal.*, vol. 8, no. 2, pp. 139–150, 2004.

[36] Y. Shi, S. Liao, Y. Gao, D. Zhang, Y. Gao, and D. Shen, "Transductive prostate segmentation for CT image guided radiotherapy," in *Proc. MICCAI Workshop: Mach. Learning Med. Imag.*, 2012, pp. 1–9.

[37] Y. Shi, S. Liao, Y. Gao, D. Zhang, Y. Gao, and D. Shen, "Prostate segmentation in CT images via spatial-constrained transductive lasso," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2227–2234.

[38] Y. Shi, H.-I. Suk, Y. Gao, and D. Shen, "Joint coupled-feature representation and coupled boosting for AD diagnosis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2721–2728.

[39] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc.: Series B*, vol. 58, pp. 267–288, 1996.

[40] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Roy. Statistical Soc.: Series B*, vol. 67, pp. 91–108, 2005.

[41] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, Oct. 2010.

[42] C. Wang, Z. She, and L. Cao, "Coupled attribute analysis on numerical data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1736–1742.

[43] Y. Wu, G. Liu, M. Huang, J. Jiang, W. Yang, W. Chen, and Q. Feng, "Prostate segmentation based on variant scale patch and local independent projection," *IEEE Trans. Med. Imaging*, vol. 33, no. 6, pp. 1290–1303, Jun. 2014.

[44] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

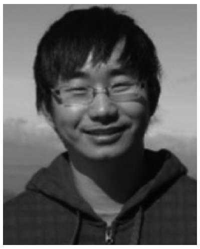
[45] B. Xin, Y. Kawaharay, Y. Wang, and W. Gao, "Efficient generalized fused Lasso and its application to the diagnosis of Alzheimers disease," in *Proc. AAAI*, 2014, pp. 2163–2169.

[46] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group Lasso," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2104–2116, Sep. 2013.

[47] Y. Zhan and D. Shen, "Deformable segmentation of 3-D ultrasound prostate images using statistical texture matching method," *IEEE Trans. Med. Imaging*, vol. 25, no. 3, pp. 256–272, Mar. 2006.

[48] Y. Zhan, D. Shen, J. Zeng, L. Sun, G. Fichtinger, J. Moul, and C. Davatzikos, "Targeted prostate biopsy using statistical image analysis," *IEEE Trans. Med. Imaging*, vol. 26, no. 6, pp. 779–788, Jun. 2007.

Yinghuan Shi received the BSc and PhD degrees from the Department of Computer Science, Nanjing University in 2007 and 2013, respectively. He is currently an assistant researcher in the Department of Computer Science and Technology, Nanjing University, China. He was a visiting scholar in the University of Technology, Sydney, from July 2014 to August 2014, a visiting student in University of North Carolina at Chapel Hill from September 2011 to September 2012, and Massey University from November 2009 to February 2010, respectively. His research interests include computer vision and medical image analysis. He has published more than 20 research papers in related journals and conferences such as the *IEEE TBME*, *IEEE TNNLS*, CVPR, and IPMI. He serves as a program committee member for several conferences, and also as a referee for several journals.



Yaozong Gao received the BE degree from the Department of Software Engineering, Zhejiang University, China, in 2008, and the MSc degree from the Department of Computer Science, Zhejiang University, China, in 2011. He is currently working toward the PhD degree in the Department of Computer Science, University of North Carolina at Chapel Hill (UNC-CH), North Carolina. Since fall 2011, he has been with the Department of Computer Science, the Department of Radiology, and Biomedical Research Imaging Center (BRIC), UNC-CH. His research interests include machine learning, computer vision, and medical image analysis.

Shu Liao received the BEng degree (first class honors and Academic Achievement Awards) in computer engineering and the MPhil and PhD degrees in computer science and engineering from The Hong Kong University of Science and Technology in 2005, 2007, and 2010, respectively. He was a postdoctoral research fellow in the Department of Radiology, Biomedical Research Imaging Center, University of North Carolina at Chapel Hill from 2010 to 2012. His research interests include medical image analysis, texture analysis, and face recognition. He is currently a senior scientist in Siemens Medical Solution, US.

Daoqiang Zhang received the BSc and PhD degrees in computer science from Nanjing University of Aeronautics and Astronautics, China, in 1999 and 2004, respectively. His research interests include machine learning, pattern recognition, and biomedical image analysis. In these areas, he has published more than100 technical papers in the refereed international journals and conference proceedings. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, received the Best Paper Award at the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), and the Best Paper Award Honorable Mention of Pattern Recognition Journal 2007. He serves as an editorial board member for the *Computational Intelligence and Neuroscience Journal*, and also serves as a program committee member for several international conferences. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation (CCF).

Yang Gao received the PhD degree from the Department of Computer Science and Technology, Nanjing University, China, in 2000. Currently, he is a professor, and also the deputy director in the Department of Computer Science and Technology, Nanjing University. He is currently directing the Reasoning and Learning Research Group in Nanjing University. He has published more than 100 papers in top-tired conferences and journals, e.g., *IEEE TNNLS*, *IEEE TBME*, *IEEE TCYB*, *CVIU*, PR, CVPR, and AAMAS. His current research interests include artificial intelligence and machine learning. He also serves as the program chair and the area chair for many international conferences.

Dinggang Shen is a professor in Radiology, Biomedical Research Imaging Center (BRIC), Computer Science, and Biomedical Engineering in the University of North Carolina at Chapel Hill (UNC-CH). He is currently directing the Center for Image Informatics, the Image Display, Enhancement, and Analysis (IDEA) Lab in the Department of Radiology, and also the medical image analysis core in the BRIC. He was a tenure-track assistant professor in the University of Pennsylvanian (UPenn), and a faculty member in the Johns Hopkins University. His research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 500 papers in the international journals and conference proceedings. He serves as an editorial board member for six international journals. He also serves in the Board of Directors, The Medical Image Computing and Computer Assisted Intervention (MICCAI) Society.