Medical Image Analysis 18 (2014) 487–499

Contents lists available at ScienceDirect

Medical Image Analysis

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

Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements

C. Chen [a,*], W. Xie [a], J. Franke [b], P.A. Grutzner [b], L.-P. Nolte [a], G. Zheng [a,*]

[a] *Institute for Surgical Technology and Biomechanics, University of Bern, Stauffacherstr. 78, CH-3014 Bern, Switzerland*
[b] *BG Trauma Centre Ludwigshafen at Heidelberg University Hospital, Ludwig-Guttmann-Str. 13, D-67071 Ludwigshafen, Germany*

ARTICLE INFO

Article history:
Received 20 May 2013
Received in revised form 20 December 2013
Accepted 10 January 2014
Available online 5 February 2014

Keywords:
Landmark detection
Shape segmentation
X-ray image
Data-driven estimation
Femur

ABSTRACT

In this paper, we propose a new method for fully-automatic landmark detection and shape segmentation in X-ray images. To detect landmarks, we estimate the displacements from some randomly sampled image patches to the (unknown) landmark positions, and then we integrate these predictions via a voting scheme. Our key contribution is a new algorithm for estimating these displacements. Different from other methods where each image patch independently predicts its displacement, we jointly estimate the displacements from all patches together in a data driven way, by considering not only the training data but also geometric constraints on the test image. The displacements estimation is formulated as a convex optimization problem that can be solved efficiently. Finally, we use the sparse shape composition model as the a priori information to regularize the landmark positions and thus generate the segmented shape contour. We validate our method on X-ray image datasets of three different anatomical structures: complete femur, proximal femur and pelvis. Experiments show that our method is accurate and robust in landmark detection, and, combined with the shape model, gives a better or comparable performance in shape segmentation compared to state-of-the art methods. Finally, a preliminary study using CT data shows the extensibility of our method to 3D data.

1. Introduction

In clinical practice, X-ray radiography is widely used for various purposes due to its convenience and low cost. Segmenting shape contours such as femur and pelvis benefits many applications, such as computer aided disease diagnosis (Chen et al., 2005; Lindner et al., 2012), image based surgery planning and intervention (Gottschling et al., 2005). In addition, 3D reconstruction of anatomical models can also be performed with the segmented 2D contours (Baka et al., 2011; Dong and Zheng, 2008; Zheng et al., 2007, 2009a). Traditionally, shape segmentation in X-ray images, despite its extreme usefulness, is seldom done in clinical practice due to its difficulty. In cases where it is ever done, it is carried out manually by doctors, which is both time-consuming and error-prone. Therefore, in this paper our attention is on fully-automatic techniques, which will immediately make this traditionally useful but difficult task widely applicable. However, automatic segmentation of X-ray images faces many challenges. The poor and non-uniform image contrast, along with the noise, makes the segmentation very difficult. Occlusions and the overlap between bones make it difficult to identify local features of bone contours. Furthermore, the existence of implants often drastically changes the visual appearance of the relevant anatomical regions.

A typical pipeline of X-ray segmentation consists of two steps: landmark detection and shape regularization (Lindner et al., 2012, 2013), as depicted in Fig. 1. In this paper we also follow this pipeline. Given an image, we first detect the positions of a set of landmarks which are defined along the shape contour. Then, the landmark detection output is regularized using a statistical shape model. In this way, the final contour is controlled by both the image cue encoded in the landmark detection output, and the shape prior information conveyed in the statistical shape model.

In the above pipeline, accurately detecting landmarks is crucial for a good segmentation performance. In this paper, we propose a new method for this task. We estimate the displacements from a set of randomly sampled local image patches to the landmark based on patch appearance, and the individual predictions are then combined in a voting scheme to produce the predicted landmark position. In previous methods, the displacement from each patch to the landmark is estimated independently using a pre-trained model. Our method is fundamentally different, as we jointly

[*] Corresponding authors. Tel.: +41 765854810 (C. Chen), tel.: +41 786793908 (G. Zheng).

E-mail addresses: cheng.chen@istb.unibe.ch (C. Chen), guoyan.zheng@istb.unibe.ch (G. Zheng).

http://dx.doi.org/10.1016/j.media.2014.01.002
1361-8415/

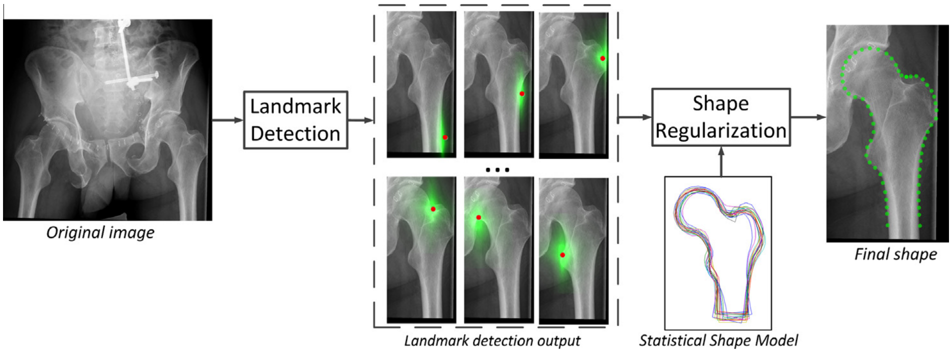

Fig. 1. The general pipeline of shape segmentation which is composed of two steps: landmark detection and shape regularization.

estimate the displacements from all patches to landmarks together in a data-driven way. This joint estimation scheme allows us to exploit the mutual interactions among the displacements that are being estimated by considering the geometric relations between the patches in the test image. Combining the information from training data and the geometry constraints, our displacement estimation method achieves better accuracy.

After landmark detection, these predicted landmark positions are regularized by a statistical shape model to get the final segmented shape contour. In this paper we exploit the sparse shape composition model (Zhang et al., 2011a), which is shown to be better than classical PCA based shape models.

We tested our method on large and challenging datasets involving three anatomic structures: complete femur, proximal femur and pelvis. These datasets contain a considerable amount of images with an implant and images with low contrast. In the experiments, we show that both the landmark detection method and the shape regularization improve the performance, and that by combining them together we get better or comparable results compared to other methods. Finally, we also performed a preliminary 3D study using CT data to show the 3D extensibility of our method.

The paper is organized as follows: We first briefly summarize the related work in Section 2. Then, in Section 3 we introduce our new landmark detection algorithm, followed by Section 4 which presents the shape regularization method using the sparse shape composition model. The experiments are presented in Section 5. We conclude the paper in Section 6.

2. Related work

In recent literature, there has been a considerable amount of work in landmark detection. Some methods utilize low-level image features such as gradients and edges (Chen et al., 2005; Cristinacce and Cootes, 2008; Smith et al., 2009). For example, Chen et al. (2005) locate candidate femoral shafts and heads by detecting parallel lines and circles. This type of methods often suffers from the large appearance variation and image noise encountered in X-ray images. To alleviate this problem, some similar methods such as (Bergtholdt et al., 2010; Donner et al., 2010; Gamage et al., 2010; Schmidt et al., 2007) incorporate the topological constraints in a model-based way, where they search for the best configuration of the model given the image cue revealed by the low-level image features.

To overcome the challenge of appearance variation, some machine learning based methods have been proposed which have shown promising performance. For example, in (Zheng et al., 2007; Dong and Zheng, 2008, 2009), a particle filter-based approach is first used to determine the morphological parameters, and then a belief propagation based approach is used to extract contours from multiple calibrated X-ray images. Zhou and Comaniciu (2007) introduce the so-called shape regression machine to segment in real time the left ventricle endocardium from an echocardiogram of an apical four chamber view. Zheng et al. (2008, 2009b) use marginal space learning for localizing the heart chambers, and then estimate the 3D shape through learning-based boundary delineation.

In recent years, random forest (RF) (Breiman, 2001) based methods are becoming more and more popular. RF (Breiman, 2001) was originally proposed for general classification or regression, and the class-specific Hough forest was presented in (Gall and Lempitsky, 2009) for object detection. Since then, RF has shown very promising results in tasks related to landmark detection or organ localization in medical data (Criminisi et al., 2010; Lindner et al., 2012, 2013). The basic idea is as follows: First, some local patches are sampled in the image. Then, the displacements from the patches to the landmark are estimated by RF regression. Finally, the landmark position is estimated by a voting scheme considering the individual estimations from all the patches. Pauly et al. (2011) localize organs in MR images using Random Ferns which has a similar idea with RF except that a fern systematically applies the same decision function for each node of the same level of the tree. There are two key components behind the success of RF-like voting based methods. The first is the strategy of positioning landmarks by estimating its relative displacements with regard to other image parts. Here the fact of medical image being highly structured is exploited to improve the localization using relational displacement prediction. The second is the discriminative power of the RF model. In this paper, we follow the framework of predicting relational displacements from image patches. However, instead of using RF, we propose a new method to improve the displacement prediction by a data-driven approach. The significant difference of our method is that we predict the displacement of test patches not only by comparing the test patch with the training patches, but also exploit the fact that the location of these test patches are known to us and can be used to enforce a geometric constraint on the displacements: the displacement from different patches to a common landmark position should be consistent with the geometric relation between the test patches. By utilizing this information as a regularization on the displacement being predicted, we improve the prediction accuracy.

Recently, Donner et al. (2013) proposed a new landmark detection method by combining the RF-based prediction with the high-level topological relation between the landmarks, and they get very good results on X-ray images and 3D CT data. First, RF classification and regression give the candidates for each landmark, and then an MRF model encoding the global configuration of landmarks is employed to get the final landmark positions. This method aims at disambiguating landmark candidates using a global model *on top of* the individual landmark predictions, which is especially suit-

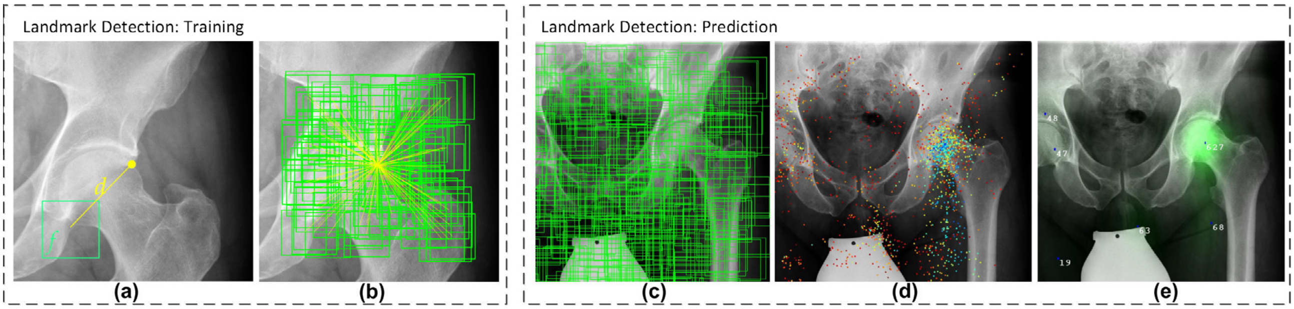

Fig. 2. Overview of our landmark detection algorithm. During training stage (a) and (b), a set of patches are sampled around the ground-truth landmark position. During testing stage, given an image, a set of patches are randomly sampled (c), each patch makes a prediction of the landmark position (d), and then the predictions are aggregated to produce a response image (e).

able for repetitive anatomical patterns. In contrast, our method aims at improving the prediction for individual landmarks by exploiting the *local* geometric relations between the sampled image patches and the landmarks. The global landmark relation is instead encoded in the sparse shape composition model used for shape regularization. On the other hand, since our method directly improves the localization of individual landmarks, it can be easily combined with other high-level regularizors (such as the MRF model in Donner et al. (2013)) to provide more accurate "basic localizations" on top of which the global reasoning can be performed.

Another related topic of this paper is the statistical shape model which is typically used to regularize the landmark positions using global topological information. Apart from the methods inspired by the popular Active Shape Models (ASM) (Behiels et al., 1999; Cootes and Taylor, 1992; Cristinacce and Cootes, 2008; Pilgram et al., 2008), different new shape models have been proposed, such as the models based on mixture of Gaussians (Cootes et al., 1997), sparse PCA (Sjostrand et al., 2007), manifold learning (Etyngier et al., 2007; Zhang et al., 2011b). Recently, Zhang et al. (2011a, 2012) proposed sparse shape composition, which is based on sparse representation techniques (Candes and Tao, 2006; Donoho, 2004). This method has the advantage of keeping the local shape information and is shown to have better performance than previous methods, and in this paper we use this model for shape regularization. Also, we comment that, from a generalized point of view, models such as MRF in Donner et al. (2010, 2013), which regularize the individual landmark positions by global topological constraints, can also be viewed as a shape model.

3. Landmark detection by jointly estimating image displacements

In this study, a shape is defined by an ordered set of landmarks along its boundary. Each shape is mathematically represented by the concatenation of the coordinates of each landmark. In this way, a 2D or 3D shape is a vector in the $\mathbb{R}^{2L}$ or $\mathbb{R}^{3L}$ space, respectively. As introduced in Section 1, similar to Lindner et al. (2012, 2013) and Zhang et al. (2012), we adopt a two-step segmentation framework where landmark detection is followed by shape regularization. In this section we present our landmark detection algorithm. The shape regularization will introduced in Section 4 and can be found in details in Zhang et al. (2011a).

3.1. Basic idea

The overview of our landmark detection algorithm is given in Fig. 2. In the training step, as shown in (a), a rectangular image patch is randomly sampled around the ground-truth landmark position, with *f* denoting the visual feature of the patch, and *d* denoting the displacement from this patch to the landmark position. In the same way, we randomly sample a number of patches around the ground-truth landmark position, as shown in (b). The features and corresponding displacements of all these training patches constitute the training data. Then, in the prediction step, given a new image, we also randomly sample a number of image patches, as in (c). Since now we do not know the landmark position, these test patches are sampled everywhere in the image.[1] The visual features of these patches will be calculated, and based on their features, the corresponding displacements with regard to the (unknown) landmark position can be estimated. In this way, each patch makes a vote on the landmark prediction as in (d), where each vote contains a position (depicted by dots) and uncertainty (color-coded). Then, from these votes, we construct the *response image* as in (e), which can be viewed as the probability of the landmark position on every image location.

The crucial part of the procedure presented above is the estimation of displacements for the test patches, which is also the key of our contribution. To better explain the idea of our method, we illustrate a simplified scenario as in Fig. 3. The training data in the left consists of the features and displacements of seven training patches $\left\{\left(\tilde{\mathbf{f}}_k, \tilde{\mathbf{d}}_k\right)\right\}_{k=1\ldots7}$. During the test step, we randomly sample two test patches; "Patch 1" centered at image location $\mathbf{c}_1$ with visual feature $\mathbf{f}_1$, and "Patch 2" centered at image location $\mathbf{c}_2$ with visual feature $\mathbf{f}_2$. Our task is now to estimate the corresponding displacements $\mathbf{d}_1$ and $\mathbf{d}_2$ of the two patches (red[2] in the figure). To this end, we consider the following two factors.

- *Exploitation of training data*
  The training data should serve as a basic guidance to estimate the displacements. Specifically, the displacement of a test patch should be similar to those of training patches with similar visual features. To this end, we establish links between test data and training data based on feature proximity. For example, in Fig. 3, test patch 1 is linked to training patches 1 and 6 because they have similar features, and similarly, test patch 2 is linked to training patches 4 and 7. Then, a criterion to estimate $\mathbf{d}_1$ and $\mathbf{d}_2$ is to minimize the discrepancy between the displacements that are linked:

$$\text{minimize } \|\mathbf{d}_1 - \tilde{\mathbf{d}}_1\|^2 + \|\mathbf{d}_1 - \tilde{\mathbf{d}}_6\|^2 + \|\mathbf{d}_2 - \tilde{\mathbf{d}}_4\|^2 + \|\mathbf{d}_2 - \tilde{\mathbf{d}}_7\|^2 \quad (1)$$

[1] Or the relevant ROI (region of interest) of the image. See Section 5.1.2 for details of our multi-resolution implementation.

[2] For interpretation of color in Fig. 3, the reader is referred to the web version of this article.

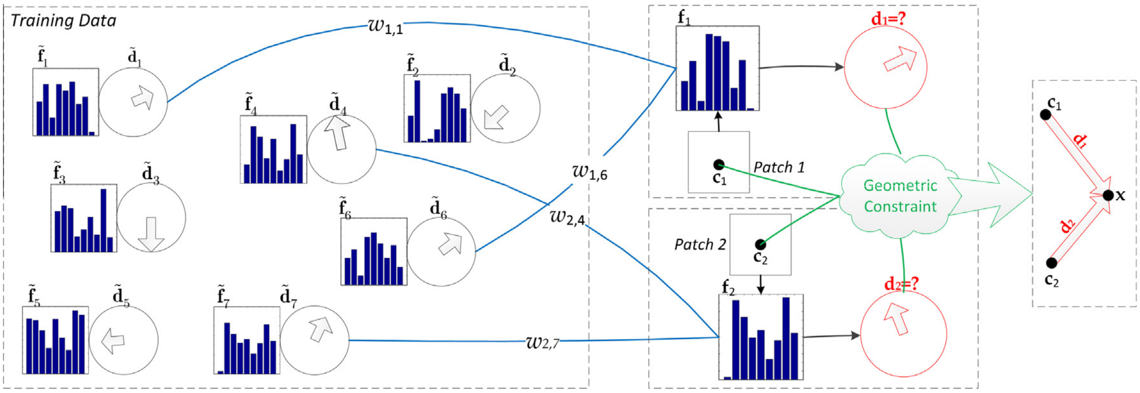

Fig. 3. A simplified situation where we try to estimate the displacements of the test patches. See text for details.

where $\|\cdot\|$ gives the length of the displacement vector. If we denote $w_{k,i}$ as the "proximity flag", where $w_{k,i} = 1$ if training patch k is linked to test patch i, and $w_{k,i} = 0$ otherwise, we can compactly write Eq. (1) as:

$$\text{minimize} \sum_k \sum_i w_{k,i} \|\tilde{\mathbf{d}}_k - \mathbf{d}_i\|^2 \tag{2}$$

- *Exploitation of the geometric constraint*
  There is a potential geometric constraint between $\mathbf{d}_1$ and $\mathbf{d}_2$, as they are displacements from two different patches to the same landmark. Although the true position of this landmark, $\mathbf{x}$, is unknown, we do know that $\mathbf{c}_1, \mathbf{c}_2$ and $\mathbf{x}$ form a triangle, and therefore we have $\mathbf{d}_1 - \mathbf{d}_2 = \mathbf{c}_2 - \mathbf{c}_1$. Therefore, we also want to minimize the discrepancy:

$$\text{minimize} \ \|(\mathbf{d}_1 - \mathbf{d}_2) - (\mathbf{c}_2 - \mathbf{c}_1)\|^2 \tag{3}$$

Using i and j to index the test patches, Eq. (3) can be written compactly as:

$$\text{minimize} \sum_i \sum_j \|(\mathbf{d}_i - \mathbf{d}_j) - (\mathbf{c}_j - \mathbf{c}_i)\|^2 \tag{4}$$

The idea of our method is thus to design an objective function with regard to the displacements to be estimated (e.g. $\mathbf{d}_1, \mathbf{d}_2$ in the above figure), by considering both the training data and the geometric constraint.

The above explanation is a simplified illustration. In the following, we formally present our method.

3.2. Problem formulation

Training step. Assume that we are interested in L landmarks whose ground-truth position is known in a set of training images. As shown in Fig. 4(a), $\tilde{\mathbf{x}}_l \in \mathbb{R}^2$ is the position of the lth landmark. We randomly sample a number of rectangular patches around all the landmarks. For the kth patch, we denote $\tilde{\mathbf{c}}_k \in \mathbb{R}^2$ as its center position, $\tilde{\mathbf{f}}_k \in \mathbb{R}^{d_f}$ as its visual feature, and $\left(\tilde{\mathbf{d}}_k^l\right)_{GT} = \tilde{\mathbf{x}}_l - \tilde{\mathbf{c}}_k \in \mathbb{R}^2$ as its ground-truth (GT) displacement to the lth landmark. In total, we sample $\widetilde{K}$ patches over all the training images, and we denote $\tilde{\mathbf{F}} = \left[\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_{\widetilde{K}}\right] \in \mathbb{R}^{d_f \times \widetilde{K}}$ as the matrix of features of all training patches, and $\left(\tilde{\mathbf{D}}\right)_{GT} \in \mathbb{R}^{2L \times \widetilde{K}}$, whose element $\left(\tilde{\mathbf{D}}_{ij}\right)_{GT} = \left(\tilde{\mathbf{d}}_j^i\right)_{GT}$, as the matrix of all training displacements.

Prediction (test) step. In the prediction step, we are given a new test image, on which we want to estimate the positions of the L landmarks, as shown in Fig. 4(b). To this end, we randomly sample K patches, where $\mathbf{c}_k \in \mathbb{R}^2$ and $\mathbf{f}_k \in \mathbb{R}^{d_f}$ are the center position and the visual feature of the kth patch. We denote $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_k] \in \mathbb{R}^{d_f \times K}$as the matrix of features of all test patches.

Strategy. To estimate the position of the L landmarks on the test image, we first want to estimate $\{\mathbf{d}_k^l\}_{k=1\ldots K, l=1\ldots L}$, which is the displacement from each patch to each landmark. Then, $\left\{\mathbf{c}_k + \mathbf{d}_k^l\right\}_{k=1\ldots K}$ will be the set of votes of the lth landmark's position from all the test patches, from which we can compute the landmark position by a voting scheme (details in Section 3.5). Therefore, if we denote $\mathbf{D} \in \mathbb{R}^{2L \times K}$, whose element $\mathbf{D}_{ij} = \mathbf{d}_j^i$, as the matrix of displacements in the test image, our goal is to estimate $\mathbf{D}$.

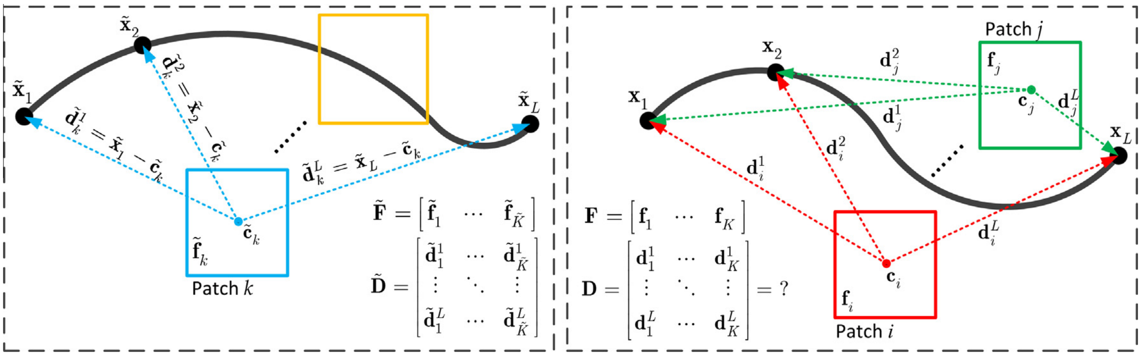
$\tilde{\mathbf{x}}_1$
$\tilde{\mathbf{x}}_2$
$\tilde{\mathbf{x}}_L$
$\tilde{\mathbf{d}}_k^1 = \tilde{\mathbf{x}}_1 - \tilde{\mathbf{c}}_k$
$\tilde{\mathbf{d}}_k^2 = \tilde{\mathbf{x}}_2 - \tilde{\mathbf{c}}_k$
$\tilde{\mathbf{d}}_k^L = \tilde{\mathbf{x}}_L - \tilde{\mathbf{c}}_k$
$\tilde{\mathbf{f}}_k$
$\tilde{\mathbf{c}}_k$
Patch k
$\tilde{\mathbf{F}} = \left[\tilde{\mathbf{f}}_1 \ \cdots \ \tilde{\mathbf{f}}_{\widetilde{K}}\right]$
$\tilde{\mathbf{D}} = \begin{bmatrix} \tilde{\mathbf{d}}_1^1 & \cdots & \tilde{\mathbf{d}}_{\widetilde{K}}^1 \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{d}}_1^L & \cdots & \tilde{\mathbf{d}}_{\widetilde{K}}^L \end{bmatrix}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_L$
Patch j
$\mathbf{f}_j$
$\mathbf{c}_j$
$\mathbf{d}_j^2$
$\mathbf{d}_j^1$
$\mathbf{d}_j^L$
$\mathbf{d}_i^1$
$\mathbf{d}_i^2$
$\mathbf{d}_i^L$
$\mathbf{c}_i$
$\mathbf{f}_i$
Patch i
$\mathbf{F} = \left[\mathbf{f}_1 \ \cdots \ \mathbf{f}_K\right]$
$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^1 & \cdots & \mathbf{d}_K^1 \\ \vdots & \ddots & \vdots \\ \mathbf{d}_1^L & \cdots & \mathbf{d}_K^L \end{bmatrix} = ?$

Fig. 4. Problem formulation of the joint estimation of image displacements. (left) Training data. (right) Prediction data.

3.3. Objective function

First, we construct a compound displacement matrix which contains jointly the training displacements and the test displacements to be estimated:

$$\widehat{\mathbf{D}} = \left[\tilde{\mathbf{D}}\ \mathbf{D}\right] = \begin{bmatrix} \tilde{\mathbf{d}}_1^1 & \cdots & \tilde{\mathbf{d}}_{\widetilde{K}}^1 & \mathbf{d}_1^1 & \cdots & \mathbf{d}_K^1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{d}}_1^L & \cdots & \tilde{\mathbf{d}}_{\widetilde{K}}^L & \mathbf{d}_1^L & \cdots & \mathbf{d}_K^L \end{bmatrix} \in \mathbb{R}^{2L\times(\widetilde{K}+K)} \tag{5}$$

The left part (the first $\widetilde{K}$ columns) of $\widehat{\mathbf{D}}$ contains the displacements in the training images, and the right part (the last K columns) is the displacements in the test image. Note that we can write $\tilde{\mathbf{D}} = \widehat{\mathbf{D}}\mathbf{P}$ and $\mathbf{D} = \widehat{\mathbf{D}}\mathbf{Q}$ by defining appropriate (0,1) matrices $\mathbf{P}$ and $\mathbf{Q}$ which select corresponding columns.

Treating $\widehat{\mathbf{D}}$ as a variable, we design an objective with regard to $\widehat{\mathbf{D}}$:

$$E(\widehat{\mathbf{D}}) = E_g(\widehat{\mathbf{D}}) + \alpha E_f(\widehat{\mathbf{D}}) + \beta E_p(\widehat{\mathbf{D}}) \tag{6}$$

Please note that although we are ultimately interested in estimating the displacements of test patches $\mathbf{D}$, our objective function is defined on $\widehat{\mathbf{D}}$, which is the combination of training and test displacements. In this way we can embed the relations between training and test data into our objective function. After we get the optimal $\widehat{\mathbf{D}}$, the optimal $\mathbf{D}$ is simply given by $\mathbf{D} = \widehat{\mathbf{D}}\mathbf{Q}$. Below we define each term in the objective function.

3.3.1. Ground-truth Discrepancy $E_g(\widehat{\mathbf{D}})$

The left part of $\widehat{\mathbf{D}}$ should be close to the ground-truth displacements in the training data, which is encoded in $\left(\tilde{\mathbf{D}}\right)_{GT}$. Therefore, we want to minimize the Ground-truth Discrepancy:

$$E_g(\widehat{\mathbf{D}}) = \frac{1}{2L\widetilde{K}}\left\|\widehat{\mathbf{D}}\mathbf{P} - (\tilde{\mathbf{D}})_{GT}\right\|_F^2 \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm.[3]

3.3.2. Feature propagation discrepancy $E_f(\widehat{\mathbf{D}})$

First, we construct a compound feature matrix $\widehat{\mathbf{F}} = \left[\tilde{\mathbf{f}}\ \ \mathbf{f}\right] \in \mathbb{R}^{d_f\times(K+K)}$. Now, each column of $\widehat{\mathbf{F}}$ is the feature of a (training or test) patch, and the corresponding column of $\widehat{\mathbf{D}}$ is the displacement vector (to all landmarks) of that patch. We denote $\|\mathrm{col}_i(\widehat{\mathbf{F}}) - \mathrm{col}_j(\widehat{\mathbf{F}})\|_{L2}$ as the $L2$ feature distance of a pair of patches (i,j), where $\mathrm{col}_i()$ denotes the ith column. From all pairwise distances, we construct a binary affinity matrix $\mathbf{S} \in \{0,1\}^{(\widetilde{K}+K)(K+K)}$, where $s_{ij} = 1$ if and only if the ith and the jth patches are mutually ρ nearest neighbors ($\rho = 10$ in this paper) in the feature space. Note that the edges in the affinity matrix might link two training patches, two test patches, or a training patch and a test patch.

For every pair of patches (i,j), if they are similar in the feature space, their displacements to landmarks should also be similar. We define the Feature Propagation Discrepancy $E_f(\widehat{\mathbf{D}})$ as the violation from this assumption:

$$E_f(\widehat{\mathbf{D}}) = \frac{1}{2L\sum_{i\neq j}}\sum_{i\neq j} s_{ij}\left\|\mathrm{col}_i(\widehat{\mathbf{D}}) - \mathrm{col}_j(\widehat{\mathbf{D}})\right\|_{L2}^2 \tag{8}$$

For each pair of patches, E_f introduces a high penalty if the two patches are similar in the feature space (i.e. $s_{ij} = 1$) but their displacements are very different (i.e. $\left\|\mathrm{col}_i(\widehat{\mathbf{D}}) - \mathrm{col}_j(\widehat{\mathbf{D}})\right\|_{L2}$ is large). If we construct $\mathbf{M}$ as the (trace normalized) Laplacian matrix (Kokiopoulou et al., 2011) of $\mathbf{S}$, E_f can be compactly written as:

$$E_f(\widehat{\mathbf{D}}) = \frac{1}{L}\mathrm{Tr}\left(\widehat{\mathbf{D}}\mathbf{M}\widehat{\mathbf{D}}^\top\right) \tag{9}$$

In short, this term favors the consistency between feature proximity and displacement proximity. In this way, the ground-truth displacements are propagated to the test data via the links between training and test patches.

3.3.3. Patch offset penalty $E_p(\widehat{\mathbf{D}})$

Each column of $\mathbf{D}$ is the displacements from a single patch in the test image to all the landmarks. The subtraction of two columns can be written as $\mathrm{col}_i(\mathbf{D}) - \mathrm{col}_j(\mathbf{D}) = \mathbf{D}(\mathbf{e}_i^K - \mathbf{e}_j^K) = \begin{bmatrix} \mathbf{d}_i^1 - \mathbf{d}_j^1 \\ \cdots \\ \mathbf{d}_i^L - \mathbf{d}_j^L \end{bmatrix}$, where $\mathbf{e}_i^K$ is a K dimensional column vector whose ith element is 1 and all other elements are 0s. From Fig. 4(b), we can see that $\mathbf{d}_i^1 - \mathbf{d}_j^1 = \ldots = \mathbf{d}_i^L - \mathbf{d}_j^L = \mathbf{c}_j - \mathbf{c}_i$, because $(\mathbf{d}_i^1, \mathbf{d}_j^1), \ldots, (\mathbf{d}_i^L, \mathbf{d}_j^L)$ form triangles with the same edge $\mathbf{c}_j - \mathbf{c}_i$. Therefore, we impose a penalty $E_p^{i-j}(\mathbf{D}) = \left\|\mathbf{D}(\mathbf{e}_i^K - \mathbf{e}_j^K) - \bar{\mathbf{c}}_{j-i}\right\|_F^2$, where $\bar{\mathbf{c}}_{j-i}$ is the L times vertical replicate of $\mathbf{c}_j - \mathbf{c}_i$. We can include a penalty for each pair (i,j) of columns. For efficiency reasons, we eliminate redundancies and use $K - 1$ pairs:

$$E_p(\widehat{\mathbf{D}}) = \frac{1}{2LK}\sum_{i=1}^{L-1} E_p^{i-(i+1)}(\mathbf{D}) = \frac{1}{2LK}\left\|\widehat{\mathbf{D}}\mathbf{Q}\mathbf{U} - \bar{\mathbf{C}}\right\|_F^2 \tag{10}$$

where $\mathbf{U} = \left[\mathbf{e}_1^K - \mathbf{e}_2^K, \ldots, \mathbf{e}_{K-1}^K - \mathbf{e}_K^K\right]$ and $\bar{\mathbf{C}} = \left[\bar{\mathbf{c}}_{2-1} \quad \ldots \quad \bar{\mathbf{c}}_{K-(K-1)}\right]$.

3.4. Optimization

Substituting Eqs. (7), (9) and (10) into Eq. (6), we get the final objective function. We can prove that Eq. (6) is convex, and therefore to find the global optimum, we need to solve the equation:

$$\partial E(\widehat{\mathbf{D}})/\partial\widehat{\mathbf{D}} = \widehat{\mathbf{D}}\mathcal{A} + \mathcal{G} = \mathbf{0} \tag{11}$$

where $\mathcal{A} = \frac{1}{L\widetilde{K}}\mathbf{P}\mathbf{P}^\top + \frac{2\alpha}{L}\mathbf{M} + \frac{\beta}{LK}\mathbf{Q}\mathbf{U}\mathbf{U}^\top\mathbf{Q}^\top$, and $\mathcal{G} = -\frac{(\tilde{\mathbf{D}})_{GT}\mathbf{P}^\top}{L\widetilde{K}} - \frac{\beta\bar{\mathbf{C}}\mathbf{U}^\top\mathbf{Q}^\top}{LK}$. The optimal solution is given by $\widehat{\mathbf{D}} = -\mathcal{G}\mathcal{A}^{-1}$.

3.5. Constructing response image

After we find the optimum $\widehat{\mathbf{D}}$, we have $\mathbf{D} = \widehat{\mathbf{D}}\mathbf{Q}$, and $\left\{\mathbf{c}_k + \mathbf{d}_k^l\right\}_{k=1\ldots K}$ will be the set of votes for the position of the lth landmark. We write $\mathbf{v}_k = \mathbf{c}_k + \mathbf{d}_k^l$ as the position vote made by the kth patch. For each vote, there is also an uncertainty $\mathbf{\Sigma}_k$, which is calculated as the (diagonal) variance of the training displacements that are linked to the kth test patch when we calculated the feature propagation discrepancy $E_f(\widehat{\mathbf{D}})$ in Section 3.3. The next step is to calculate the probability of landmark on different image locations, from the votes $\{(\mathbf{v}_k, \mathbf{\Sigma}_k)\}_{k=1\ldots K}$. We view each vote $(\mathbf{v}_k, \mathbf{\Sigma}_k)$ as a Gaussian distribution $G(\cdot|\mu, \mathbf{\Sigma})$ with mean $\mathbf{v}_k$ and variance $\mathbf{\Sigma}_k$. Then, the probability of landmark at an image coordinate (x,y) is given by accumulating the contribution of all votes on this image location:

$$I(x,y) = \sum_{k=1}^{K} G((x,y)|\mathbf{v}_k, \mathbf{\Sigma}_k) \tag{12}$$

Table 1
Computational complexity of different steps of our method.

Train	Test			
Feature	Feature	Build $\mathcal{A}$ and $\mathcal{G}$	Optimization	Response
$O(\widetilde{K})$	$O(K)$	$O((\widetilde{K}+K)^2)$	$O((\widetilde{K}+K)^3)$	$O(KL)$

[3] E_g is normalized by the number of landmarks L and the number of training patches $\widetilde{K}$. The following terms are normalized in a similar way.

$I(x, y)$ is viewed as an image function which is called *response image* of the landmark, as in Fig. 2 (e), which will be used in Section 4 for shape regularization.

3.6. Computational complexity

Table 1 shows the dominant computational complexity of our method in different steps. During the training stage, we need to calculate the visual feature and displacements (column "Feature"), which is linear to $\widetilde{K}$, the number of training patches. In the prediction stage, given a new image, we first need to calculate the visual feature of the test patches (column "Feature"), which is linear to K. Then, we need to build the matrices $\mathcal{A}$ and $\mathcal{G}$. Then, the solution of the optimization problem is cubic respect to $\widetilde{K} + K$ as it requires the inversion of matrix $\mathcal{A}$. Finally, we need to construct the response image (column "Response"), which is linear both to K, and the number of landmarks L.

3.7. Discussions

An important difference of our method from previous ones is that during the estimation of image displacements, we consider both the consistency with regard to the training data and the inter-patch relations between the test patches. In our objective function Eq. (6), $E_g(\widehat{\mathbf{D}})$ minimizes the deviation of the training displacements from the ground-truth, which serves as the root of the entire inference. $E_f(\widehat{\mathbf{D}})$ links patches with similar appearance, which can be viewed as the exploitation of the training data. Finally, $E_p(\widehat{\mathbf{D}})$ ensures that the predictions made by the test patches are consistent with regard to the inter-relation between those patches. This final term does not involve the training data, and can be viewed as a regularization of the test displacements that are being estimated. All these three terms are combined to form a single objective function, where all the displacements are *jointly* estimated. In Section 5.2.3, we will see how the performance is improved by both the exploitation of training data and the inter-patch relation.

It is also noteworthy to mention the difference from (Chen et al., 2013), which contains an additional term $E_l()$ which is the row-wise counterpart of $E_p()$. Considering two landmarks l_1 and l_2, patch i makes two predictions $\mathbf{d}_i^{l_1}$ and $\mathbf{d}_i^{l_2}$, respectively. Similarly, another patch j makes two predictions $\mathbf{d}_j^{l_1}$ and $\mathbf{d}_j^{l_2}$. Then, we should have $\mathbf{d}_i^{l_1} - \mathbf{d}_i^{l_2} = \mathbf{d}_j^{l_1} - \mathbf{d}_j^{l_2}$, because both quantities are equivalent to the (unknown) location difference of the two landmarks $\mathbf{x}_{l_2} - \mathbf{x}_{l_1}$. In this paper, $E_l()$ is removed because it is implied by $E_p()$, which makes it redundant. This can be seen more easily by writing out the equations. $E_p()$ says that $\mathbf{d}_i^{l_1} - \mathbf{d}_j^{l_1} = \mathbf{d}_i^{l_2} - \mathbf{d}_j^{l_2}$ and $E_l()$ says that $\mathbf{d}_i^{l_1} - \mathbf{d}_i^{l_2} = \mathbf{d}_j^{l_1} - \mathbf{d}_j^{l_2}$. An advantage of removing this term is that the objective function now can be globally optimized in closed form.

4. Shape regularization by sparse shape composition

Given the response image of each landmark, the shape regularization step is performed using a statistical shape model to generate the optimal segmented shape contour by considering both the image cue encoded in the response images and the shape prior information encoded in the shape model. First, we initialize the shape by the mode on the response image for each landmark. Then, we update the shape iteratively. In each iteration, we perform three actions: (1) update the shape by moving each landmark locally to a better position according to its response image, (2) regularize the shape by the shape model, and (3) update the shape pose by calculating the optimal similarity (translation + rotation + scale) transformation from the shape in the image space to the model space using Procrustes Analysis. These steps are straightforward except step (2), which regularizes the locally updated shape by the shape model. Traditionally, this can be done by the Active Shape Model (Cootes and Taylor, 1992) based on PCA (Principal Component Analysis). In this paper, we instead employ the recently proposed shape model based on sparse representation (Zhang et al., 2011a, 2012). Here we briefly explain this method.

The shape model consists of a set of pre-aligned training shapes $\{\mathbf{y}_i\}_{i=1,\ldots,N}$. For a new shape $\mathbf{y}'$ to be regularized, after a similarity transformation T (which is evaluated by Proscrustes Analysis as an optimal similarity transformation which minimizes the average landmark-wise distance between the new shape and the mean training shape), it should be approximated by a linear combination involving only a small subset of the training shapes, plus a sparse error:

$$T(\mathbf{y}') \approx \mathbf{Y}\mathbf{x} + \epsilon = [\mathbf{Y} \quad \mathbf{I}]\begin{bmatrix}\mathbf{x} \\ \epsilon\end{bmatrix} = \mathbf{Y}'\mathbf{x}' \tag{13}$$

Both the linear coefficient $\mathbf{x}$ and the error ϵ are sparse. Therefore, the composite coefficient $\mathbf{x}\prime$ is also sparse. Our goal becomes solving the following L_1-regularized least squares problem:

$$\mathbf{x}'_{\text{opt}} = \arg\min_{\mathbf{x}'}\left(\left\|T(\mathbf{y}') - \mathbf{Y}'\mathbf{x}'\right\|_2^2 + \lambda\|\mathbf{x}'\|_1\right) \tag{14}$$

where λ is a parameter controlling the importance of the sparsity constraint. There are a number of solvers for Eq. (14), and we employ the method using truncated Newton interior-point method as described in Kim et al. (2007).

The interpretation of Eq. (13) is this: the shape $\mathbf{y}'$ should be approximated (after transformation T) as a linear combination of only a small number of "bases", which can either be the training shapes, or standard basis of the $\mathbb{R}^{2L}$ space. The contribution from the training shapes represents the "true" part of shape $\mathbf{y}'$ that is consistent with the shape model, and the contribution from the standard basis accommodates large but sparse errors (noise). Therefore, after we get the optimal $\mathbf{x}'_{\text{opt}}$ by Eq. (14), we decompose $\mathbf{x}'_{\text{opt}}$ by $\mathbf{x}'_{\text{opt}} = \left[\mathbf{x}_{\text{opt}}^\top, \boldsymbol{\epsilon}_{\text{opt}}^\top\right]^\top$, discard the ϵ_{opt} which corresponds to the noises, and the regularized shape is given by back-projecting the "true" part of the shape:

$$\mathbf{y}'_{\text{regularized}} = T^{-1}(\mathbf{Y}\mathbf{x}_{\text{opt}}) \tag{15}$$

5. Experiments

In this section we present our experimental results. We first introduce the experiment setup in Section 5.1. In Sections 5.2 and 5.3 we report the evaluation on the landmark detection algorithm and the shape regularization method, respectively. Individual evaluation of these two components is important, as the shape regularization may "smooth out" the errors in landmark detection. Then we evaluate the complete segmentation method in Section 5.4. Finally, we present an experiment on 3D data in Section 5.5.

5.1. Experiment setup

5.1.1. Data

We tested our method on X-ray images from our clinical partner. These images are classified into three datasets based on the involved anatomical structures: complete femur, proximal femur and pelvis. For each dataset, we randomly select some images for training and the rest for test purposes:

(1) Complete femur: 80 training images, 109 test images.
(2) Proximal femur: 100 training images, 188 test images.
(3) Pelvis: 100 training images, 163 test images.

A considerable part of the images is post-operative X-ray radiographs after trauma or joint replacement surgery, which significantly increases the challenge due to large variation of femur/pelvis appearance and the presence of implants. As an indication, we made a manual counting, which shows that 32% of the test images contain implants.

For each image in the training dataset, we manually annotate the contour of the left part (i.e., the left femur or the left semi-pelvis). To establish landmark correspondences, we randomly choose one image as the reference and other images are floating images. We evenly sample a set of landmarks along the contour on the reference image, and the corresponding landmarks in floating images are found by an Expectation Conditional Maximization (ECM)-based deformable shape registration method (Zheng, 2013). In this way, for each image, we have both the dense contour and landmarks. We establish 59, 97 and 89 landmarks for proximal femur, pelvis and complete femur, respectively.

5.1.2. Implementation details

To improve the efficiency, we adopt a multi-scale strategy with three scale levels [0.25 0.5 1], where the test image is resized to 25%, 50% and 100% of its original size in each dimension, as shown in Fig. 5. At each level, we detect the landmarks as in Section 3, and then the shape is regularized as in Section 4. The result of each level is propagated to the next level as initialization, where the landmarks are detected by sampling patches only in a limited region around the initial position. At the first level where no initialization is available, the landmarks are detected by sampling patches all through the image. In this way, we combine the global detection at the first level and local detection at the higher levels, which achieves high accuracy without exploding the computation time.

In each scale level, we use the same parameters as follows: For landmark detection, each shape is divided into subshapes of 4 successive landmarks (i.e. L=4). For each subshape, we sample $\widetilde{K} = 2000$ training patches and $K = 500$ test patches. For the visual feature of the patches, we use multi-level HoG (Histogram of Oriented Gradient) feature (Dalal and Triggs, 2005) with block sizes 1×1 and 2×2. Each block is divided into 2×2 cells and for each cell an 18 dimensional HoG feature is extracted by histogramming the gradient direction of each pixel. Therefore, our original feature dimension is $d_f = 360$. To improve the efficiency, we adopt a feature selection algorithm as proposed in Chen et al. (2011) to reduce the feature dimension to 100. This method reduces the feature dimension with an LDA (Linear Discriminant Analysis)-like criterion. The difference is that it seeks an optimal subset of feature dimensions rather than a linear transformation of the full feature. For the objective function, we use $\alpha = 0.1$ and $\beta = 0.01$ as our default value. In each scale level, the landmark detection algorithm is performed on each subshape independently, and then the shape is regularized once for the complete shape.

5.2. Evaluation of the landmark detection algorithm

In this section we evaluate our landmark detection algorithm. To isolate the influence, we drop the statistical shape model, i.e. we evaluate the detected landmarks before the shape regularization.

5.2.1. Evaluation metric

Suppose that we are considering a landmark whose ground-truth position is $\mathbf{x}$. During prediction we sample K patches, which

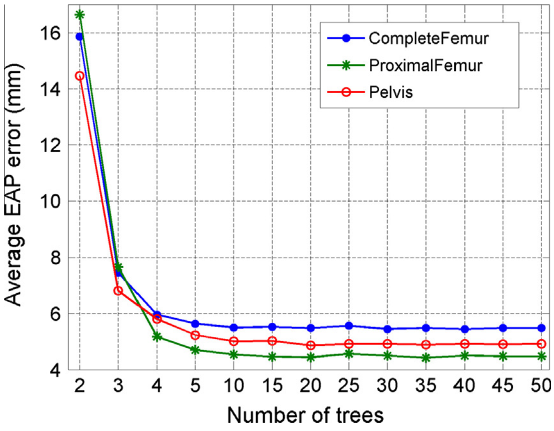

Fig. 6. Variation of EAP along with different numbers of trees in the RF ensemble.

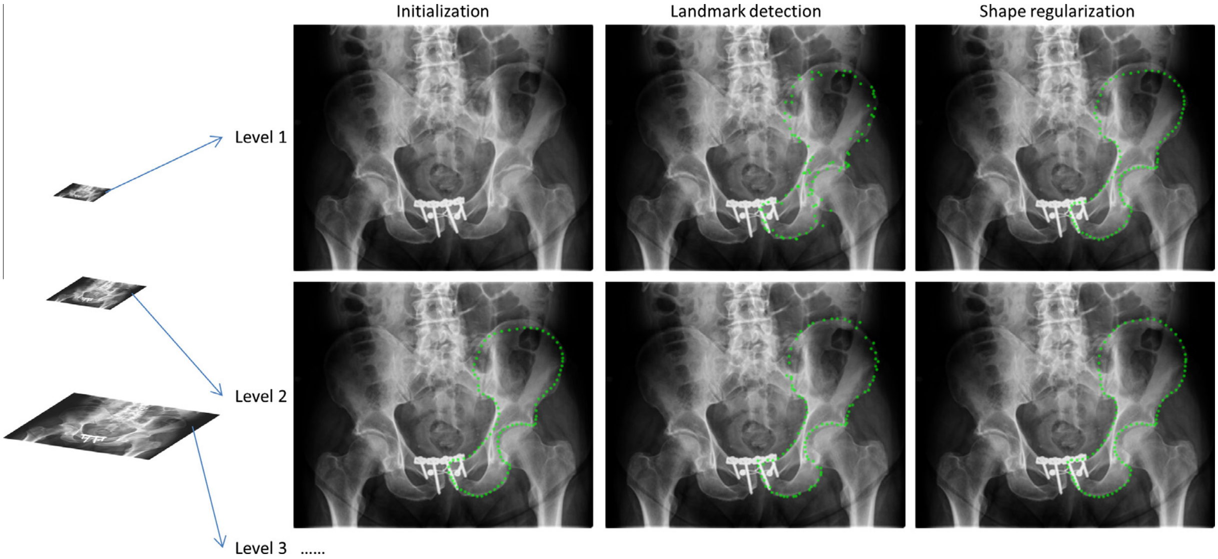

Fig. 5. Multi-scale framework. At level 1 there is no initialization, so the landmarks are detected by searching all over the image. At higher levels the landmarks are detected only in the neighborhood of the initial positions. To save space, the third level is omitted.

Table 2
Quantitative comparison of our method and RF method. The EIP and EAP measurements are in unit mm.

	Complete femur		Proximal femur		Pelvis	
	EIP	EAP	EIP	EAP	EIP	EAP
Our method (default)	8.7	4.4	8.8	4.3	8.3	4.5
RF method	14.5	5.5	14.1	4.5	15.6	5.0
p-Value	$< 1^{-30}$	$< 1^{-15}$	$< 1^{-30}$	$< 1^{-5}$	$< 1^{-30}$	$< 1^{-8}$

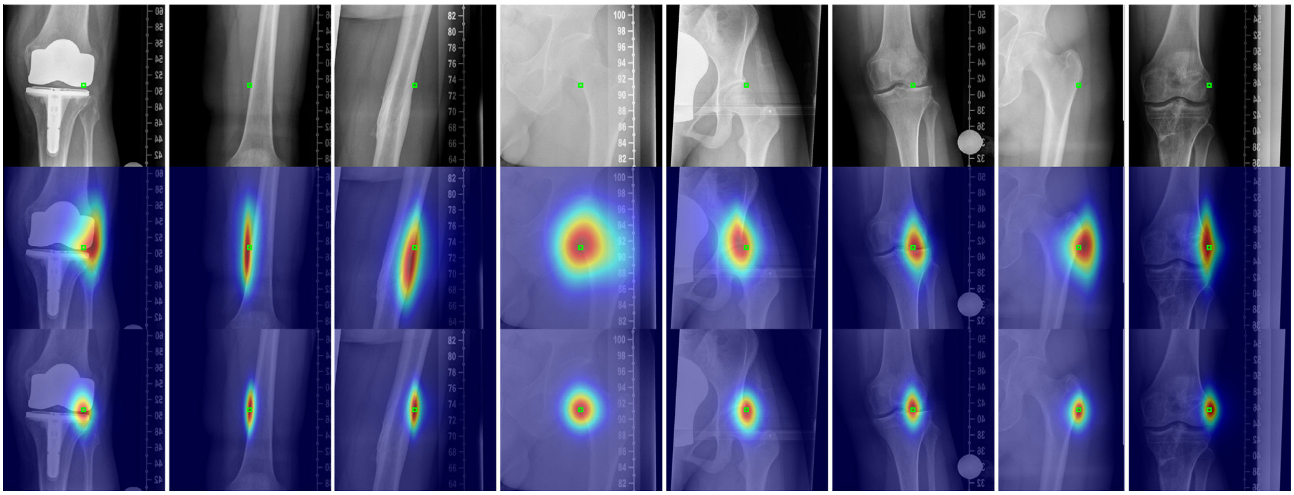

Fig. 7. Comparison of the response images. Top row: the original image. Middle row: RF method. Bottom row: our method. In each sub-image, the green square is the ground-truth landmark position. This figure is best viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

give us a set of votes $\{\mathbf{v}_k\}_{k=1\ldots K}$ as described in Section 3.5. We define two performance measurements. **EIP (Error of Individual Predictions)** and **EAP (Error of Aggregated Prediction)**:

$$EIP = \frac{1}{K}\sum_{k}\|\mathbf{x} - \mathbf{v}_k\|, \quad EAP = \|\bar{\mathbf{v}} - \mathbf{x}\| \tag{16}$$

where $\bar{\mathbf{v}}$ is the aggregated prediction, i.e. the mode of the response image. These two measurements emphasize on different aspects. EIP directly estimates the quality of each individual vote, and EAP estimates the error of the final landmark detection.

As our X-ray images are in DICOM format where the pixel spacing is known, to unify the comparability, in all our evaluations we convert the image plane distances into physical unit millimeter.

5.2.2. Comparison with Random Forest based method

We compare our method with the Random Forest based method as in Lindner et al. (2012). We use exactly the same training data $\left(\tilde{\mathbf{F}}, \tilde{\mathbf{D}}\right)$ as in our method to train the RF regressor. Since for each patch we want to estimate the displacements to L landmarks, we utilize $2L$ RF regressors for each subshape, one for each output dimension. In each tree of the forest, each node splits the training data points into either its left or right branch according to a threshold test on one dimension of the feature vector, and nodes with less than 5 data samples will stop expanding and become a leaf nodes. For RF, we use the same parameter as our method when applicable (e.g. the same multi-scale framework and visual feature). As for the ensemble size for RF, we use 10 trees for all datasets, as the performance does not change significantly with > 10 trees, as shown in Fig. 6.

Table 2 shows the quantitative comparison. We can see that in all the three datasets, our method generates better results in both EIP and EAP. Fig. 7 shows some qualitative comparisons of the response images of some landmarks using our method and RF method. In each sub-image, the green square represents the ground-truth position of the landmarks concerned. We have two observations. First, in our method the peak in the produced response image is more compact (the spread of the distribution is smaller), which contributes to a smaller EIP. This is because our method exploits the inter-relations between the individual predictions by the geometric constraints, thus the predictions are more "compatible", while in RF method the predictions of different patches are independent and less consistent. Second, the final prediction of landmark position of our method is more accurate (the mode of the distribution is more close to the ground-truth), which corresponds to smaller EAP.

We also compare the computation time needed for our method and the RF method as in Table 3, which reports the time consumed[4] to process each subshape when $L = 4$. The column "Feature" represents the sampling and feature calculation of patches. "Prediction" stands for the estimation of displacements. This includes the time to build corresponding matrices and the optimization, while in RF case this is the time to project the feature of test patches to the forest and retrieve the displacement prediction. "Response" represents the construction of response image. We note that the time spent on feature calculation is neglectable. As for the training, our method is faster as it does not involves training any regressor as RF. For the Testing stage, in both cases most time is consumed to build the response image, while our method takes a slightly longer time because we spend more time in the prediction of displacements.

[4] All the computation time in this paper is measured in our unoptimized Matlab implementation on a PC with a quad-core CPU at 3.0 GHz.

Table 3
Computational time comparison of our method and RF method per subshape with $L = 4$ landmarks. Time unit is in second.

Step	Train			Test			
	Feature	Train model	Total	Feature	Prediction	Response	Total
Time (our method)	0.2	N/A	0.2	0.1	0.5	1.9	2.5
Time (RF method)	0.2	19.1	19.3	0.1	0.2	1.9	2.2

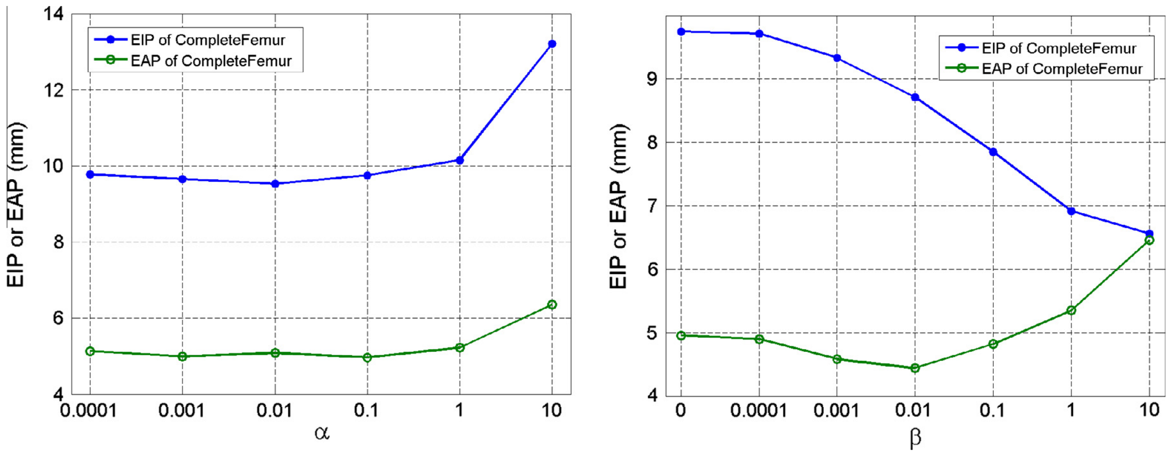

Fig. 8. Variation of EIP and EAP on the complete femur dataset, with different parameter settings of the objective function.

Table 4
EAP and computation time on complete femur dataset with different numbers of train/test patches. Bold values indicates default parameters.

$\widetilde{K} = \mathbf{2000}, K = \ldots$	50	100	200	**500**	1000
EAP (mm)	5.4	4.8	4.6	**4.4**	4.4
Time (s)	0.7	0.9	1.2	**2.5**	3.8
$K = \mathbf{500}, \widetilde{K} = \ldots$	200	500	1000	**2000**	4000
EAP (mm)	7.3	5.9	4.8	**4.4**	4.1
Time (s)	1.8	1.9	2.0	**2.5**	2.9

Table 5
EAP values (in mm) if we use different numbers of training images. “Full” means using 80/100/100 training images for the three datasets as specified in Section 5.1.1.

Num. of training images	Full	60	40	20	10	5
Proximal femur	4.3	4.3	4.3	4.4	4.9	6.4
Pelvis	4.5	4.6	4.6	4.9	5.3	7.1
Complete femur	4.4	4.4	4.5	4.6	5.0	6.6

5.2.3. Objective function parameters

There are two parameters α and β in our objective function. α (default 0.1) controls the importance of the training data, while β (default 0.01) controls the inter-patch constraint. To see the performance variance with different parameter values, we perform two experiments. First, we set $\beta = 0$ and try different α values. In this case, no inter-patch constraint is enforced, and the performance is only determined by the training data. Note that α cannot be zero. Otherwise the optimization problem will become singular and cannot be solved. The result is shown in Fig. 8 (left). Due to the space limit we report only on complete femur dataset, but on the other two datasets we have similar observations. We can see that the result is not sensitive to the α value as long as it is not too large. As a second experiment, we fix $\alpha = 0.1$ and tune β, as in Fig. 8 (right). We note that with increasing β, both EIP and EAP drop. After a certain value (around $\beta = 0.01$), continue increasing β still decreases EIP, but EAP starts to increase. This is because, with very strong emphasis on the inter-patch geometric constraint, all the predictions are squeezed together, yielding a very concentrated (but somewhat misplaced) distribution on the response image. As an extreme condition, if $\beta \to \infty$, all predictions will collapse to a single point, yielding exactly the same EIP and EAP.

5.2.4. Numbers of training/test patches

Our landmark detection algorithm works by sampling patches in the image and aggregate their predictions. Therefore, the performance is influenced by $\widetilde{K}$, number of training patches, and K, number of test patches. To better understand this, we conducted experiments on different $\widetilde{K}$ and K values on the complete femur dataset. We consider both accuracy (the EAP value) and efficiency (the required time to complete the test stage of a subshape, corresponding to the rightmost “Total” column in Table 3).

From Table 4 we see that in terms of accuracy, $\widetilde{K}$ plays a more important role than K. Increasing $\widetilde{K}$ constantly decreases the EAP error. We choose $\widetilde{K} = 2000$ as our default parameter because this seems to be a good compromise between accuracy and efficiency. As for K, increasing K also decreases EAP, but the influence seems saturated when $K \geqslant 500$. We choose $K = 500$ as our default parameter.

On the other hand, with respect to efficiency, in Table 1 we see that the optimization process of our method is cubic with regard to both $\widetilde{K}$ and K. However, in Table 4 we see that the computation time only increases moderately with increasing $\widetilde{K}$ and K. This is because, as can be seen in Table 3, the most time-consuming part of our method is the construction the response image, which is linear to K and independent of $\widetilde{K}$.

5.2.5. Selection of training images

As stated in Section 5.1.1, 80/100/100 training images are randomly selected for the three datasets. To study the sensitivity of our method with respect to the training image selection, we perform an experiment where we select a smaller number of training images, while keeping all other parameters unchanged (i.e. $\widetilde{K}$ keeps the same, which means that in the case of fewer training

Table 6
Computation time and EAP for different number of landmarks in each subshape on the Complete Femur dataset.

	$L = 1$	$L = 2$	$L = 4$	$L = 8$	$L = 16$	$L = Lmax(89)$
Time per subshape (s)	0.8	1.4	2.5	5.3	12.4	88.5
Num of subshapes	89	45	23	12	6	1
Time for all subshapes (s)	70.3	63.1	57.5	63.3	74.5	88.5
EAP (mm)	4.5	4.4	4.4	4.7	4.8	5.3

images, the patches sampled in each single image might increase), and the results are shown in Table 5. From this we can see that the performance drops when the number of training images reduces, but the deterioration is significant only when the number of training images is very small (below 20).

5.2.6. Size of subshape L

Our landmark detection algorithm works with subshapes which consists of L successive landmarks (default $L = 4$). The value of L influences both accuracy and efficiency. To study this issue, we make a study using the complete femur dataset using different sizes of subshape. The result is summarized in Table 6.

Given a fixed K which is the number of test patches, K patches are sampled around *all* landmarks *per subshape*. Therefore, for a subshape with more landmarks, the sampling region for the patches will be larger, and for each landmark, the number of nearby patches that make good predictions is smaller. Therefore, in general the accuracy drops when L increases, as suggested in Table 6.

On the other hand, the influence of L on efficiency is more complex.

(1) As shown in Table 1, the time required to process each subshape increases with L sub-linearly.
(2) On the other hand, as L increases, we need to process fewer subshapes (inverse to L).

Combining (1) and (2), it seems to suggest that the processing time in terms of the complete shape will always reduce as L increases (multiplication of a sub-linear term with an inverse-proportional term). However, from Table 6 we see that the smallest computation time (57.5 s) occurs at $L = 4$. This is related to our implementation. When constructing the response image in Section 3.5, to save time, we only calculate for locations within the area where test patches are sampled (outside this area the probability is close to zero). As L increases, this area increases, and when L is very large, we see this effect as we spend significantly more time in the construction of response image.

5.3. Evaluation of the sparse shape composition model

In this section we compare the sparse shape composition model with the traditional PCA based active shape model. To isolate the influence, landmark detection is not involved in this section.

First, we compare Davies' generality measurement (Davies, 2002; Styner et al., 2003) of the two shape models[5] using the training shapes of the three datasets, respectively. The is done by regularizing a selected test shape by a shape model constructed by all remaining shapes. The generality score is the distance from the test shape and the regularized shape (The average point-to-point distance converted to millimeters). This procedure is repeated for every shape in a leave-one-out manner, and the generality measurement on the three datasets are listed in Table 7, from which we clearly see the advantage of the sparse shape composition model.

Table 7
Comparison of generality measurement of the two statistical models. Numbers are in unit mm.

	Complete femur	Pelvis	Proximal femur
PCA based model	0.79	1.36	1.21
Sparse shape composition model	0.23	0.44	0.51

Table 8
Comparison of generality measurement in the presence of outliers on Pelvis dataset. Numbers are in unit mm.

Num. or outlier points	0	1	2	4
PCA based model	1.36	2.39	3.34	4.79
Sparse shape composition model	0.44	1.63	2.30	3.99

To evaluate the robustness of the statistical shape model with regard to outliers. We also perform the above experiment with artificially added outliers. Specifically, for the test shape, we randomly perturb the position of some landmarks by an omnidirectional noisy displacement whose magnitude is sampled from a Gaussian distribution with zero mean and 100 mm standard deviation. Due to the page limit, we only report the result on the pelvis dataset as in Table 8. From the table we can see that the sparse shape composition model is more robust against outliers.

5.4. Evaluation of the complete segmentation system

In this section, we evaluate our complete shape segmentation system as a combination of our landmark detection algorithm and the sparse shape composition model. The qualitative result is shown in Fig. 9 for proximal femur and pelvis, and Fig. 10 for complete femur. From these figures we can visually conclude that our segmentation approach generates very good result. Since we only annotated the left part of the training images, during test stage, we perform an additional pass by horizontally mirroring each image to get the segmentation of the right part. The quantitative evaluation, however, is restricted to the left part as we do not have the ground-truth for the right part on the test images.

For quantitative evaluation, we calculate the average point-to-curve distance (converted to millimeter) between the segmented shape and the manually segmented ground-truth contour. For comparison, we compare our method with the popular RF based landmark detection combined with PCA based shape model or sparse shape composition. The detailed result is shown in Table 9. Note that both methods achieve a success rate of 100%, 98.4% and 98.8% on complete femur, proximal femur and pelvis datasets, respectively,[6] and the unsuccessful cases are excluded when we calculate the errors in Table 9.

[5] Davies' other two measurements "specificity" and "compactness" do not apply here, as the sparse shape decomposition model does not permit generating new shape instances.

[6] The unsuccessful cases are determined by manual inspection, and are usually caused by extreme appearance/shape abnormality due to fracture.

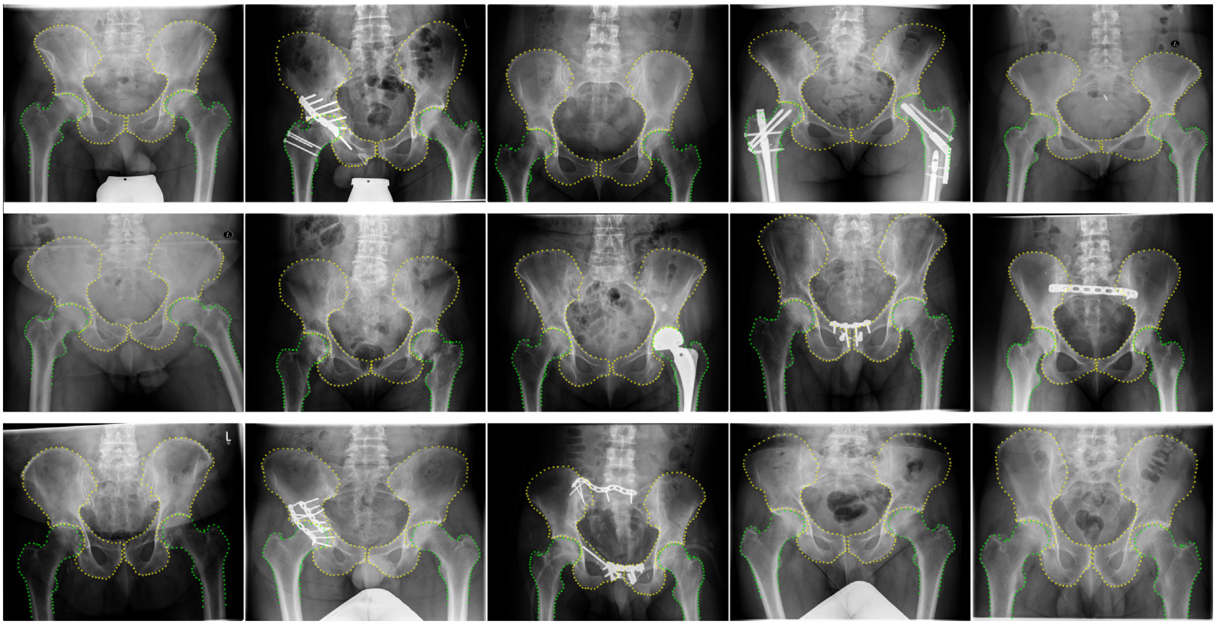

Fig. 9. Qualitative result of our complete segmentation method on proximal femur and pelvis.

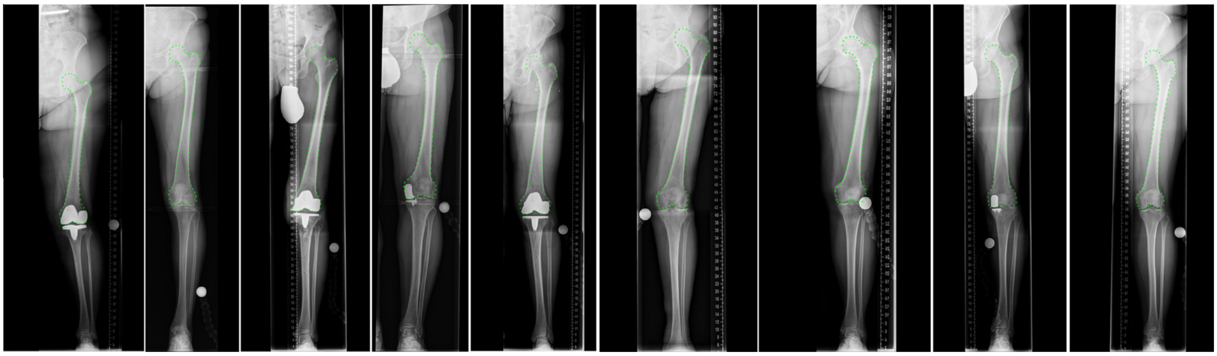

Fig. 10. Qualitative result of our complete segmentation method on complete femur.

Table 9
Quantitative comparison of the segmented shape of our method with RF method. Error values are in unit millimeter.

	Complete femur			Proximal femur			Pelvis		
	Mean	Std	*p*-Value	Mean	Std	*p*-Value	Mean	Std	*p*-Value
Our method	1.2	0.4	N/A	1.3	0.5	N/A	2.0	0.7	N/A
RF + PCA	1.7	0.5	<1e−5	1.3	0.5	0.25	2.4	0.7	<1e−4
RF + Sparse	1.6	0.5	<1e−5	1.3	0.5	0.1	2.3	0.7	<1e−3

5.5. Extension to 3D data

To test the 3D extensibility, we also implemented an extension of our method in 3D cases, where the 2D elements in the method are changed to their 3D counterparts. For example, the 2D image patches become 3D volumes, and the image displacements are now 3D vectors. As a preliminary study, we perform an experiment using a dataset containing CT data of proximal femurs from 7 patients, with the volumetric data and the corresponding ground-truth 3D femoral shape represented by 8789 vertices. The pixel size of the volumetric data is about 200 × 400 × 150, and the pixel spacing is 1 mm in each dimension. All the ground truth 3D femoral shapes are semi-automatically segmented from the corresponding volume data with a Amira software (FEI Visualization Sciences Group, France) and the vertex correspondences across the different femoral shapes are established with the diffeomorphic demons algorithm (Vercauteren et al., 2009). It is well-known that the separation of the femoral head and the acetabulum is one of the main difficulties in automatic segmentation

Table 10
Leave-one-out result on the CT data. In each column one patient is used for test and the remaining 6 are used for training.

	#1	#2	#3	#4	#5	#6	#7	Average
Segmentation error (mm)	1.6	1.8	1.7	1.3	1.5	1.2	1.5	1.5

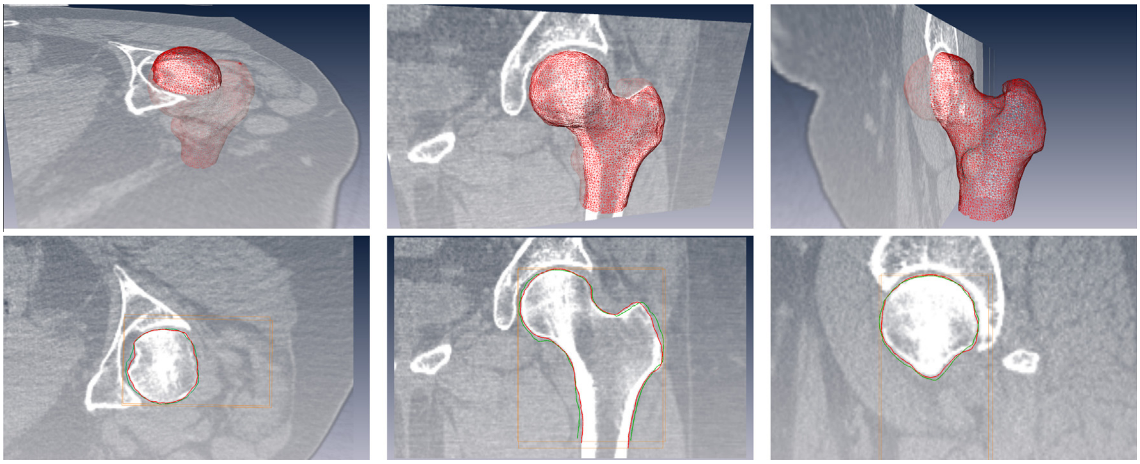

Fig. 11. Qualitative result of our segmentation method on CT data. In the bottom row the red contours are the ground-truth segmentation, and the green contours are the segmentation results with our method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of such a dataset due to the extreme narrowness of the joint space and the intensity similarity between the two neighboring structures (Yokota et al., 2009).

As for the algorithm parameters, we keep $\widetilde{K} = 2000$, $K = 500, \alpha = 0.1$ and $\beta = 0.01$. We use $L = 1$, because defining neighboring vertices on a 3D shape is more complex than 2D and is beyond the scope of this paper. For the visual feature of 3D image volumes, we devide each volume into $5 \times 5 \times 5$ blocks, and the mean intensity of each block is calculated and concatenated to form the 125 dimensional feature vector.

Since the 8789 vertices represent the 3D shape very densely, we do not perform landmark detection on every vertex. Instead, we randomly choose a subset of "key points" for landmark detection. The other points are determined in the shape regularization process. Specifically, in Eqs. (13) and (14) we use only the key points to calculate the reconstruction coefficient $\mathbf{x}$, and then, in Eq. (15), all the 8789 points are used to reconstruct the regularized shape. Similar to the 2D case, we also adopt a multi-scale strategy of two scales [0.5 1]. To further improve the efficiency, we use different numbers of key points in different levels. For the first level whose purpose is to coarsely localize the shape, we use 50 key points. For the second level we use 500 key points.

The experiment is performed in a leave-one-out manner, where in each round one data is selected for test and the remaining 6 data are used for training. The segmentation error is calculated by the average point-to-surface distance between the segmented shape and the ground-truth shape (all 8789 points are involved in this calculation). The result is shown in Table 10, where we can see that we achieve an average error of 1.5 mm. Fig. 11 shows some qualitative results, where in the top row we show the 3D view where the reconstructed shapes are superimposed with three orthogonal slices of CT data, and in the bottom we show the 2D view which is the contour intersection of the slice with the ground-truth shape (red) and our segmented shape (green).

As for the efficiency, please note that when the other parameters are fixed, the computational time of our landmark detection algorithm is linear with regard to the number of subshapes (or the number of key points in this since $L = 1$). In our unoptimized Matlab implementation, it takes about 10 min to segment a 3D shape.

Please note that this is a preliminary study, and that we use a very simple volume feature and a small dataset. Several improvements may immediately improve the performance, e.g. better feature for volumes, larger dataset with more expressive shape model, better strategy to select key points, and so on. We leave these for the further study.

6. Conclusions

In this paper, we proposed a new method for landmark detection and shape segmentation in X-ray images. Our method works by jointly estimating the image displacements of test patches using the training data and also the geometric information on the test image itself. The key contribution is the exploitation of the inter-patch relations to impose the geometric regularizations on the image displacements that are being estimated. We formulate our problem as a convex objective function which can be solved efficiently. The landmark detection output is then exploited together with the sparse shape composition model to generate the segmented shape. Our method is evaluated on three datasets concerning Complete Femur, Proximal Femur and Pelvis. The experiments show that our method indeed improves the estimation of image displacements from image patches to landmark positions. The improved predictions give us a more accurate landmark detection result, and, combined with the shape model, show an improved or comparable performance compared with other methods. We also extend our method to a 3D case involving a small CT proximal femur dataset which shows that our method also generates promising result.

In the future, we would like to study in details our method on 3D segmentation problem. We are also interested in combining our objective function of image displacements with the objective of shape regularization by statistical shape model and finally generate a unified optimization problem which deals with the image cue (landmark detection) and shape prior cue (shape model) simultaneously.

Acknowledgement

Dr. G. Zheng is supported by the Swiss KTI Project No. 13523.1 PFFLR-LS. Dr. C. Chen is supported by the Swiss National Science Foundation Project No. 51NF40-144610. Mr. W. Xie is supported by the Swiss National Science Foundation Project No. 205321-138009.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.media.2014.01.002.

References

Baka, N., Kaptein, B., de Bruijne, M., van Walsum, T., Giphart, J., Niessen, W., Lelieveldt, B., 2011. 2d–3d Shape reconstruction of the distal femur from stereo X-ray imaging using statistical shape models. Med. Image Anal. 15, 840–850.

Behiels, G., Vandermeulen, D., Maes, F., Suetens, P., Dewaele, P., 1999. Active shape model-based segmentation of digital X-ray images. In: MICCAI, pp. 128–137.

Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C., 2010. A study of parts-based object class detection using complete graphs. Int. J. Comput. Vis. 87, 93–117. http://dx.doi.org/10.1007/s11263-009-0209-1.

Breiman, L., 2001. Random forests. In: Machine Learning, pp. 5–32.

Candes, E.J., Tao, T., 2006. Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inform. Theory 52, 5406–5425. http://dx.doi.org/10.1109/tit.2006.885507.

Chen, Y., Ee, X., Leow, W.K., Howe, T.S., 2005. Automatic extraction of femur contours from hip X-ray images. In: Proceedings of the ICCV Workshop on Computer Vision for Biomedical Image Applications, pp. 200–209.

Chen, C., Yang, Y., Nie, F., Odobez, J.M., 2011. 3d human pose recovery from image by efficient visual feature selection. Comput. Vis. Image Understand. 115, 290–299.

Chen, C., Xie, W., Franke, J., Grutzner, P.A., Nolte, L.P., Zheng, G., 2013. Fully automatic X-ray image segmentation via joint estimation of image displacements. In: MICCAI.

Cootes, T., Taylor, C.J., 1992. Active shape models – smart snakes. In: In British Machine Vision Conference. Springer-Verlag, pp. 266–275.

Cootes, Taylor, Cootes, T.F., Taylor, C.J., 1997. A mixture model for representing shape variation. In: Image and Vision Computing. BMVA Press, pp. 110–119.

Criminisi, A., Shotton, J., Robertson, D.P., Konukoglu, E., 2010. Regression forests for efficient anatomy detection and localization in ct studies. In: MCV, pp. 106–117.

Cristinacce, D., Cootes, T.F., 2008. Automatic feature localisation with constrained local models. Pattern Recog. 41, 3054–3067.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: In CVPR, pp. 886–893.

Davies, R.H., 2002. Learning Shape: Optimal Models for Analysing Natural Variability. University of Manchester.

Dong, X., Zheng, G., 2008. Automatic extraction of proximal femur contours from calibrated X-ray images using 3d statistical models. In: MIAR, pp. 421–429.

Dong, X., Zheng, G., 2009. Automatic extraction of proximal femur contours from calibrated X-ray images using 3d statistical models: an in vitro study. Int. J. Med. Robot. Comput. Assis. Surg. 5, 213–222. http://dx.doi.org/10.1002/rcs.253.

Donner, R., Langs, G., Micusik, B., Bischof, H., 2010. Generalized sparse MRF appearance models. Image Vision Comput. 28, 1031–1038. http://dx.doi.org/10.1016/j.imavis.2009.07.010.

Donner, R., Menze, B.H., Bischof, H., Langs, G., 2013. Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Med. Image Anal., doi:http://dx.doi.org/10.1016/j.media.2013.02.004.

Donoho, D.L., 2004. For Most Large Underdetermined Systems of Equations, The Minimal l1-norm Near-Solution Approximates the Sparsest Near-Solution. Technical Report. Comm. Pure Appl. Math.

Etyngier, P., Ségonne, F., Keriven, R., 2007. Shape priors using manifold learning techniques. In: 11th IEEE International Conference on Computer Vision. Rio de Janeiro.

Gall, J., Lempitsky, V., 2009. Class-specific Hough forests for object detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition: CVPR 2009. IEEE, Miami, Florida, pp. 1022–1029, doi:http://dx.doi.org/10.1109/CVPR.2009.5206740.

Gamage, P., Xie, S.Q., Delmas, P., Xu, W.L., 2010. Segmentation of radiographic images under topological constraints: application to the femur. Int. J. Comput. Assis. Radiol. Surg. 5, 425–435.

Gottschling, H., Roth, M., Schweikard, A., R, R.B., 2005. Intraoperative, fluoroscopy-based planning for complex osteotomies of the proximal femur. Int. J. Med. Robot. Comput. Assis. Surg. 1, 33–38.

Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D., 2007. An interior-point method for large-scale l1-regularized logistic regression. J. Mach. Learn. Res., 2007.

Kokiopoulou, E., Chen, J., Saad, Y., 2011. Trace optimization and eigenproblems in dimension reduction methods. Numer. Lin. Algeb. Appl. 18, 565–602.

Lindner, C., Thiagarajah, S., Wilkinson, J.M., Wallis, G.A., Cootes, T.F., 2012. Accurate fully automatic femur segmentation in pelvic radiographs using regression voting. In: MICCAI, vol. 3, pp. 353–360.

Lindner, C., Thiagarajah, S., Wilkinson, J.M., Wallis, G.A., Cootes, T.F., 2013. Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE Trans. Med. Imag. 32, 1462–1472.

Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Möller, A.M., Nekolla, S., Navab, N., 2011. Fast multiple organ detection and localization in whole-body mr dixon sequences, In: Proceedings of the 14th International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. Part III. Springer-Verlag, Berlin, Heidelberg, pp. 239–247.

Pilgram, R., Walch, C., Kuhn, V., Schubert, R., Staudinger, R., 2008. Proximal femur segmentation in conventional pelvic X-ray. Med. Phys. 35, 2463–2472.

Schmidt, S., Kappes, J., Bergtholdt, M., Pekar, V., Dries, S., Bystrov, D., Schnörr, C., 2007. Spine detection and labeling using a parts-based graphical model. In: Karssemeijer, N., Lelieveldt, B. (Eds.), Information Processing in Medical Imaging, Lecture Notes in Computer Science, vol. 4584. Springer, Berlin, Heidelberg, pp. 122–133.

Sjostrand, K., Rostrup, E., Ryberg, C., Larsen, R., Studholme, C., Baezner, H., Ferro, J., Fazekas, F., Pantoni, L., Inzitari, D., Waldemar, G., 2007. Sparse decomposition and modeling of anatomical shape variation. IEEE Trans. Med. Imag. 26, 1625–1635.

Smith, R., Najarian, K., Ward, K., 2009. A hierarchical method based on active shape models and directed hough transform for segmentation of noisy biomedical images; application in segmentation of pelvic X-ray images. BMC Med. Inform. Dec. Mak. 9, 1–11.

Styner, M., Rajamani, K., Nolte, L.P., Zsemlye, G., Székely, G., Taylor, C., Davies, R., 2003. Evaluation of 3d correspondence methods for model building. In: Taylor, C., Noble, J. (Eds.), Information Processing in Medical Imaging, Lecture Notes in Computer Science, vol. 2732. Springer, Berlin Heidelberg, pp. 63–75.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 45, S61–S72.

Yokota, F., Okada, T., Takao, M., Sugano, N., Tada, Y., Sato, Y., 2009. Automated segmentation of the femur and pelvis from 3d ct data of diseased hip using hierarchical statistical shape model of joint structure. In: MICCAI, vol. 1. pp. 811–818.

Zhang, S., Zhan, Y., Dewan, M., Huang, J., Metaxas, D.N., Zhou, X.S., 2011a. Sparse shape composition: a new framework for shape prior modeling. In: CVPR, pp. 1025–1032.

Zhang, W., Yan, P., Li, X., 2011b. Estimating patient-specific shape prior for medical image segmentation. In: ISBI, pp. 1451–1454.

Zhang, S., Zhan, Y., Metaxas, D.N., 2012. Deformable segmentation via sparse representation and dictionary learning. Med. Image Anal. 16, 1385–1396.

Zheng, G., 2013. Expectation conditional maximization-based deformable shape registration. In: CAIP, vol. 1, pp. 548–555.

Zheng, G., Dong, X., Gonzalez Ballester, M., 2007. Unsupervised reconstruction of a patient-specific surface model of a proximal femur from calibrated fluoroscopic images. In: Ayache, N., Ourselin, S., Maeder, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007, Lecture Notes in Computer Science, vol. 4791. Springer, Berlin Heidelberg, pp. 834–841.

Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D., 2008. Four-chamber heart modeling and automatic segmentation for 3-d cardiac ct volumes using marginal space learning and steerable features. IEEE Trans. Med. Imag. 27, 1668–1681.

Zheng, G., Gollmer, S., Schumann, S., Dong, X., Feilkas, T., Ballester, M.A.G., 2009a. A 2d/3d correspondence building method for reconstruction of a patient-specific 3d bone surface model using point distribution models and calibrated X-ray images. Med. Image Anal. 13, 883–899.

Zheng, Y., Georgescu, B., Comaniciu, D., 2009b. Marginal space learning for efficient detection of 2d/3d anatomical structures in medical images. In: IPMI, pp. 411–422.

Zhou, S.K., Comaniciu, D., 2007. Shape regression machine. In: IPMI, pp. 13–25.