# Robust and Accurate Appearance Models based on Joint Dictionary Learning

**Data from the Osteoarthritis Initiative**

Anirban Mukhopadhyay[1], Oscar Salvador Morillo Victoria[2], Stefan Zachow[1,2] and Hans Lamecker[1,2]

[1]Zuse Institute Berlin, [2]1000 Shapes Gmbh.

**Abstract.** Deformable model-based approaches to 3D image segmentation have been shown to be highly successful. Such methodology requires an appearance model that drives the deformation of a geometric model to the image data. Appearance models are usually either created heuristically or through supervised learning. Heuristic methods have been shown to work effectively in many applications but are hard to transfer from one application (imaging modality/anatomical structure) to another. On the contrary, supervised learning approaches can learn patterns from a collection of annotated training data. In this work, we show that the *supervised joint dictionary learning technique* is capable of overcoming the traditional drawbacks of the heuristic approaches. Our evaluation based on two different applications (liver/CT and knee/MR) reveals that our approach generates appearance models, which can be used effectively and efficiently in a deformable model-based segmentation framework.

**Keywords:** Dictionary Learning, Appearance Model, Liver CT, Knee MR, 3D segmentation

## 1 Introduction

Deformable model-based methods are widely used in medical image analysis for performing anatomical segmentation. These methods consist of two main parts, a cost function representing the appearance model and a Statistical Shape Model (SSM) based regularizer. One of the most common representation of deformable models are point clouds or (e.g. triangle) meshes. In this representation, the cost function a.k.a. 'detector' associated with each point (henceforth called *landmark point*) of the model is used to predict a new landmark location, followed by a deformation of the model towards the targeted positions. SSM based regularizer is used to ensure a smooth surface after deformation. This paper is mainly focused on the general design of cost function.

Many applications rely on heuristically learnt landmark detectors. Even though these detectors are highly successful in particular application scenarios [6, 7], they are hard to transfer and generalize [5]. Systematic learning procedures can successfully resolve the aforementioned issues. E.g. Principal Component Analysis (PCA) on the Gaussian smoothed local profiles have been introduced as

a learning-based cost function (henceforth called *PCA*) in the classical Active Shape Model (ASM) segmentation method [3]. However, this method is not very robust in challenging settings [8]. A more advanced approach is using normalized correlation with a globally constrained patch model [4] and sliding window search with a range of classifiers [2, 11]. Most recently, Lindner et al. have proposed random-forest regression voting (RFRV) as the cost function [8]. Even though its performance is considered state of the art in 2D image analysis, memory and time consumption issues currently renders RFRV impractical in 3D scenarios.

The ability to learn generic appearance model independent of modalities during training and efficient and effective sparse representation calculation during testing, make Dictionary Learning (DL) an interesting choice to encounter the 3D landmark detection problem. In this work we adopt the method of Mukhopadhyay et al. [9] to sparsely model the background and foreground classes in separate dictionaries during training, and compare the representation of new data using these dictionaries during testing. However, unlike the focus of [9] in developing a sel-sufficient 2D+t segmentation technique for CP-BOLD MR segmentation, in this work the DL framework of [9] is exploited within the cost function premise by introducing novel sampling and feature generation strategy.

The non-trivial development of a special sampling strategy and gradient orientation-based rotation invariant features, exploits the full potential of Joint Dictionary Learning (JDL) as a *general and effective landmark prediction method* applicable to deformable-model based segmentation across different anatomies and 3D imaging modalities. According to our knowledge, athough DL has been used previously as a 2D deformable model regularizer [14], *this is the first time*, when DL is employed as a 3D landmark detector.

The proposed landmark detection method is tested on 2 challenging datasets with wide inter subject variability namely High Contrast Liver CT and MR of Distal Femur. To emphasize the strength of JDL, structure of the learning framework is kept unchanged, i.e. parameter are not changed or adapted across applications, and the results are compared with that of ASM.

## 2 Method

Our proposed Joint Dictionary Learning (JDL) cost function for iterative segmentation is described here in details.

### 2.1 Active Shape Model

ASMs combine local appearance-based landmark detectors with global shape constraints for model-based segmentation. An SSM is trained by applying principal component analysis (PCA) on a number of aligned landmark points. This results in a linear model that encodes shape variation in the following way: $x_l = T_\theta(\bar{x_l} + M_l b)$, where $x_l$ is the mean position of landmark $l \in \{1...L\}$, $M_l$ is a set of modes for variation and $b$ are the SSM parameters. $T_\theta$ measures the global transformation to align the landmark points. During segmentation

of a new image, landmarks are aligned to optimize an overall quality of fit $Q = \sum_{l=1}^{L}(C_l(T_\theta(\bar{x}_l + M_l b)))$ s.t. $b^T S_b^{-1} b \leq M_t$. $C_l$ is the cost function for locally fitting the landmark point $l$. $S_b$ is the covariance matrix of the SSM parameters $b$ and $M_t$ is a threshold (98% samples of multivariate Gaussian distribution) on the Mahalanobis distance. In this work, we have shown Dictionary Learning as an effective way of systematically modeling the cost function from a set of annotated training images.

### 2.2 Joint Dictionary Learning

This section describes the way Dictionary Learning is utilized as a landmark detector. In particular, Foreground and Background dictionaries are learnt during training. During testing, a weighted sum of approximation error is utilized for representing the cost function. Details of the method is described below.

**Training:** Given a set of 3D training images and corresponding ground truth landmarks, our goal is to learn a joint appearance model representing both foreground and background. Two classes ($C$) of matrices, $Y^B$ and $Y^F$ are samples from the training images for containing the background and foreground information respectively. Information is collected from image patches: cubic patches are sampled around each landmark point of the 3D training images and 144-bin ($12 \times 12$) rotation invariant SIFT-style feature histograms (described in Section 2.3) are calculated for representing those patches.

Each column $i$ of the matrix $Y^F$ is obtained by taking the normalized vector of rotation invariant SIFT-style feature histograms at all the landmarks locations across all training images (similar features are obtained for matrix $Y^B$ from the background locations aligned along the normals of landmarks) as shown in Figure 1. JDL takes as input these two classes of training matrices, to learn two dictionary classes, $D^B$ and $D^F$. These Dictionaries are learnt using K-SVD algorithm [1]. In particular, the learning process is summarized in Algorithm 1.

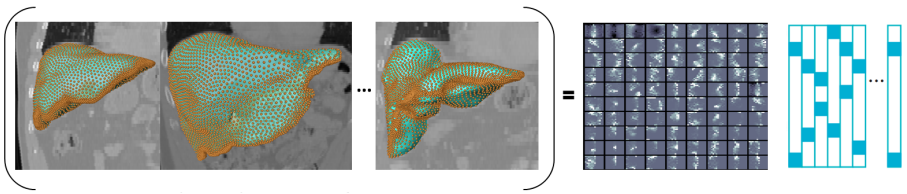

**Fig. 1.** Foreground Dictionary Learning using JDL. See text for details.

**Testing:** During segmentation of a new image, at each iteration we gather a set of test matrices $Y_l$ corresponding to each landmark $l$. $Y_l$ is obtained by sampling cubic patches along the profile and generating SIFT-like features of these patches in the similar way as training ( Section 2.3). The goal is to assign to each voxel on the profile of landmark a cost, i.e. establish if the pixel belongs to the background or the foreground as shown in Figure 2.

**Algorithm 1** Joint Dictionary Learning (JDL)

**Input:** Training patches for background and the landmarks: $Y^B$ and $Y^F$
**Output:** Dictionaries for background and the landmarks: $D^B$ and $D^F$

1: **for** C={B,F} **do**
2: Compute $Y^C$
3: Learn dictionaries with K-SVD algorithm

$$\underset{D^C, X^C}{\text{minimize}} \|Y^C - D^C X^C\|_2^2 \quad \text{s. t.} \quad \|X_i^C\|_0 \leq S$$

4: **end for**

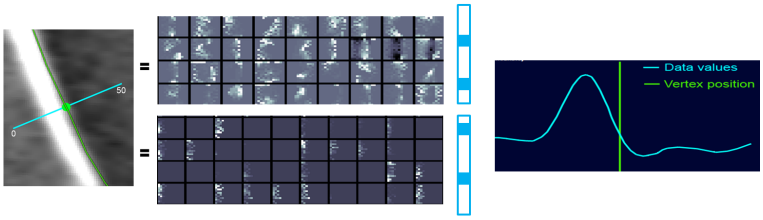


**Fig. 2.** Cost function: Weighted sum of approximation errors from representations by background and foreground dictionaries.

To perform this procedure, we use the dictionaries, $D^B$ and $D^F$, previously learnt with JDL. Orthogonal Matching Pursuit (OMP) [13] is used to compute, the sparse feature matrices $\hat{x}_{l,p}^B$ and $\hat{x}_{l,p}^F$. The cost is assigned based on the weighted sum of approximation errors. More precisely, for the cubic patch corresponding to profile voxel $p$ of landmark $l$, a cost of $\lambda(1 - R_{l,p}^B) + (1 - \lambda)R_{l,p}^F$ is assigned, as detailed in Algorithm 2. The cost is motivated by the fact that for an "ideal" location, there will be high BG approximation error and low FG approximation error. The parameter $\lambda$ balances the weight associated with approximation errors.

### 2.3 Sampling and Feature Description

The goal of sampling and rotation invariant feature description is to identify and characterize image patterns which are independent of global changes in anatomical pose and appearance. We have exploited our model-based segmentation strategy during sampling, by considering sample boxes aligned w.r.t. the surface normals. The advantages of this sampling strategy are twofold. During training, all the foreground voxel patches can encode the boundary appearance and the background voxel patches can encode the completely inside/ outside appearance. Whereas, during testing, the optimization along normal profile ensures that both foreground and background agrees on the final position. The main problem of this sampling strategy is that, the appearance of the sample strongly depends on the global rotation of the anatomy.

**Algorithm 2** Cost Function Calculation (CFC)

**Input:** Testing patches along profile of current landmark locations: $\{Y_{l,p}^T\}_{l=1}^L$, Learnt Shape Model, Dictionaries for background and the landmarks: $D^B$ and $D^F$
**Output:** Predicted Landmark location

**for** $l = 1...L$ **do**
  **for** $p =$ each location on the profile of current Landmark $l$ **do**
    **for** C={B,F} **do**
      Compute $Y_{l,p}^T$
      $R_{l,p}^C = \|y_{l,p}^T - D^C \hat{x}_{l,p}^C\|_2^2$
    **end for**
      $P_{l,p} = \lambda(1 - R_{l,p}^B) + (1 - \lambda)R_{l,p}^F$
  **end for**
**end for**

The problem of global rotation associated with sampling, is resolved during feature description. A 3D rotation invariant gradient orientation histogram derived from 3D SIFT [12] is used as a feature descriptor. In the first step, image gradient orientations of the sample are assigned to a local histogram of spherical coordinate $H$. In the next step, three primary orientations are retrieved from $H$ in the following way: $\hat{\theta_1} = argmax\{H\}$, $\hat{\theta_2}$ is the secondary orientation vector in the great circle orthogonal to $\hat{\theta_1}$ and with maximum value in $H$ and $\hat{\theta_3} = \hat{\theta_1} \times \hat{\theta_2}$. Finally, The sample patch is aligned to a reference coordinate system based on these primary orientations, and a new 144-bins ($12 \times 12$) gradient orientation histogram is generated to encode rotation invariant image features.

## 3 Results

The aim of the proposed method is to fully automatically detect the unique landmark locations from the dense annotation of 3D landmarks along the surface. In particular, we have considered 2 different anatomies acquired at 2 different modalities, to test the robustness of our proposed method: CT of livers and MR of distal femurs.

### 3.1 Data Preparation and Parameter Settings

The liver dataset consists of contrast enhanced CT data of 40 healthy livers, each with an approximate dimension of $256 \times 256 \times 50$. The corresponding surface of each liver is represented by 6977 landmark points. The distal femur MR dataset, obtained from the Osteoarthritis Initiative (OAI) database, available for public access at [10], consists of 48 subjects with severe pathological condition (Kellgren-Lawrence Osteoarthritis scale: 3). Each data has an approximate dimension of $160 \times 384 \times 384$. The corresponding distal femur surfaces are represented by 11830 landmarks each one.

For all experiments the mean shapes of respective dataset are used as initial shape. The experiment consists of a $k$-fold cross validation with $k = 10$ and 12

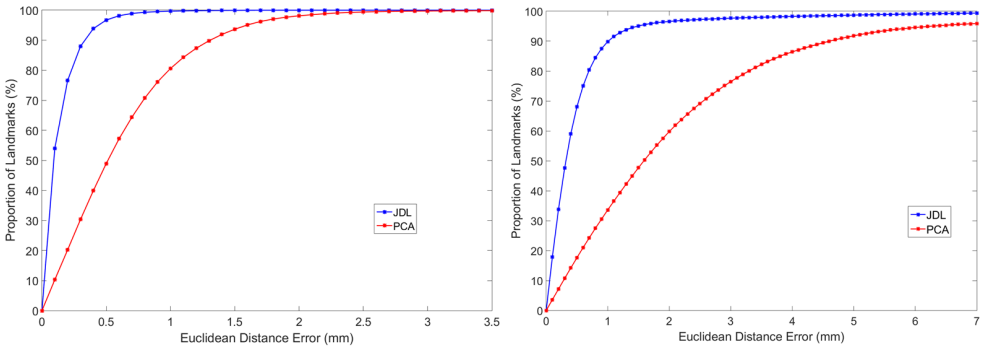


**Fig. 3.** Quantitative comparison: Local search result starting from the mean shape at the correct pose for JDL and PCA on high contrast liver CT (left) and distal femur MR (right) datasets.

for the liver and the distal femur respectively. We have set a fixed sample box size of $5 \times 5 \times 5$, dictionary of size 500 with sparsity $S = 4$ and $\lambda = 0.5$. No additional parameters are adjusted during any of the following experiments.

### 3.2 Quantitative Analysis

To compare the performance of JDL with PCA, we have performed a local search in the following way. Starting from the mean shape at the correct pose, we have computed the cost of detection for each possible landmark position along the profile. Possible positions for each landmark are considered equidistantly in 15 positions along the profile of length $\pm 7.5$ mm. As we are only interested on the performance of the landmark detector, each vertex is displaced solely based on the displacement derived from the cost of landmark detection, *without any SSM-based regularization*. The detection error for each vertex w.r.t. the ground-truth location is calculated using Euclidean Distance metric. To emphasize the superior performance of the proposed method in local search, we have compared JDL with PCA for both high contrast CT of liver as shown in Figure 3 (left) and MR of distal femur in Figure 3 (right). It is important to note that, JDL outperforms PCA in both cases. For high contrast CT of liver, 99% of the landmarks are within 1 mm of the ground-truth for JDL, compared to 80% for PCA. On the other hand, for distal femur MR, 90% of the landmarks are within 1 mm of the ground-truth for JDL, compared to only 37% for PCA.

### 3.3 Qualitative Analysis

The features learnt by JDL are discriminative enough for representing the foreground separately from the background. In particular, a set of 144-bin feature histograms (rotation invariant gradient orientation) represented by patches of size $12 \times 12$, learnt for the foreground and background are shown in Figure 4

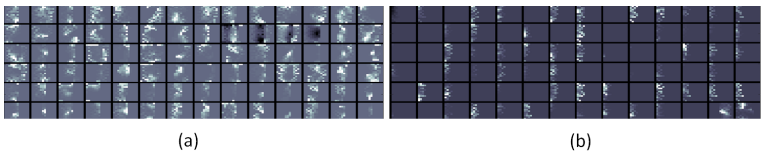


**Fig. 4.** Exemplar Foreground (a) and Background (b) dictionaries learnt from 144-bin feature vectors ($12 \times 12$) of the distal femur MR.

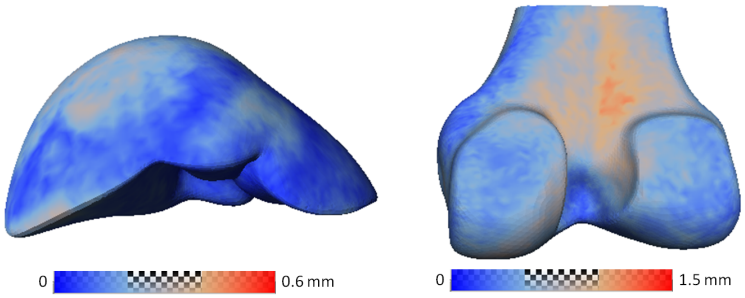


**Fig. 5.** Average accuracy for describing each landmark by JDL is superimposed as colormap on the mean liver (left) and distal femur (right) surface.

to illustrate the quality of the learnt features. It is interesting to see that, the rotation invariant gradient information is more spread out for foreground in comparison to background, as modeled by these dictionaries, resulting in an overall brighter foreground dictionary w.r.t. the background one as shown in Figure 4.

The quality of JDL is emphasized further by plotting the mean detection error for each landmark as color map on the mean liver and mean distal femur surface in Figure 5. For most of the points, low mean detection error ensures superior quality during segmentation. More importantly, Figure 5 localizes areas more prone to landmark detection failure.

## 4 Discussions and Conclusion

This study motivates us to rethink the standard cost function related assumptions of model-based segmentation for 3D images, especially regarding accommodation of several anatomies and modalities in a general framework. Deviating from heuristic learning techniques (hard to transfer and generalize across anatomies and modalities) towards systematic data-driven ones can benefit in multitude of ways: from operating with minimal manual setup to better handling of variability in image contrast, modalities and anatomies. In particular, by using JDL in 3D setting, we have shown that it is possible to address the scalability issues of Random Forest based landmark detectors in a similar situation. The performance of JDL is demonstrated in 2 challenging settings of liver CT

and distal femur MR. Furthermore, the error prone areas for landmark detection is identified, which will be addressed in future to improve the performance. JDL can be an effective tool across challenging datasets where inter-acquisition and -anatomical variability prohibits the effectiveness of heuristic learning based landmark detectors. Finally, such landmark detection tools are expected to be instrumental in advancing the utility of fully automatic segmentation techniques towards clinical translation.

**Acknowledgement:** This work is supported by Forschungscampus MODAL MedLab. The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

## References

1. Aharon et al., K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, TSP, 2006.
2. Belhumeur et al., Localizing parts of faces using a consensus of exemplars, CVPR 2011.
3. Cootes et al., Active Shape Models - Their training and application, CVIU, 1995.
4. Cristinacce et al., Automatic feature localisation with Constrained Local Models, J of Patt. Rec., 2008.
5. Heimann et al., Statistical shape models for 3D medical image segmentation: a review, MIA, 2009.
6. Kainmueller et al., Shape Constrained Automatic Segmentation of the Liver based on a Heuristic Intensity Model, MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge, 2007.
7. Kainmueller et al., An articulated statistical shape model for accurate hip joint segmentation, EMBC, 2009.
8. Lindner et al., Robust and Accurate Shape Model Matching using Random Forest Regression-Voting, PAMI, 2015.
9. Mukhopadhyay et al., Data-Driven Feature Learning for Myocardial Segmentation of CP-BOLD MRI, FIMH, 2015.
10. `http://www.oai.ucsf.edu/`
11. Saragih et al., Deformable Model Fitting by Regularized Landmark Mean-Shift, IJCV, 2011.
12. Toews et al., Efficient and Robust Model-to-Image Alignment using 3D Scale-Invariant Features, MIA, 2013.
13. Tropp et al., Signal recovery from random measurements via orthogonal matching pursuit. T Inf. Theo., 2007.
14. Zhang et al., Deformable segmentation via sparse representation and dictionary learning, MIA, 2012.