Nonlinear Statistical Shape Modeling for Ankle Bone Segmentation Using A Novel Kernelized Robust PCA

Jingting Ma[1,2], Anqi Wang[3], Feng Lin[1], Stefan Wesarg[3], and Marius Erdt[1,2]

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2] Fraunhofer IDM@NTU, Nanyang Technological University, Singapore
[3] Visual Healthcare Technologies, Fraunhofer IGD, Darmstadt, Germany
{jma012;asflin}@ntu.edu.sg {marius.erdt}@fraunhofer.sg
{anqi.wang;stefan.wesarg}@igd.fraunhofer.sg

Abstract. Statistical shape models (SSMs) are widely employed in medical image segmentation. However, an inferior SSM will degenerate the quality of segmentations. It is challenging to derive an efficient model because: (1) often the training datasets are corrupted by noise and/or artifacts; (2) conventional SSM is not capable to capture nonlinear variabilities of a population of shape. Addressing these challenges, this work aims to create SSMs that are not only robust to abnormal training data but also satisfied with nonlinear distribution. As Robust PCA is an efficient tool to seek a clean low-rank linear subspace, a novel kernelized Robust PCA (KRPCA) is proposed to cope with nonlinear distribution for statistical shape modeling. In evaluation, the built nonlinear model is used in ankle bone segmentation where 9 bones are separately distributed. Evaluation results show that the model built with KRPCA has a significantly higher quality than other state-of-the-art methods.

Keywords: Statistical shape models · Corrupted training data · Nonlinear distribution · Kernelized Robust PCA

1 Introduction

Statistical shape models (SSMs) [1] play an important role in medical image segmentation, where significant variabilities of a class of an anatomical structure are learned by principal component analysis (PCA) to guide and constrain segmentations. However, often the existing models are derived from a set of abnormal training shapes, which leads to a bias model as PCA is sensitive to outliers. Furthermore, the conventional SSM assumes training data to follow linear distribution, namely it is not capable to cope with nonlinear subspaces. To achieve higher accuracy and more flexibility, robustness to outliers and fitness to nonlinear distribution are desired properties for an ideal model. In this work, we aim to create such a model that can be used in segmentation.

The problem of data contamination has raised intensively attention in many applications. A large amount of approaches have been investigated to improve the robustness of PCA in dimensionality reduction. Robust PCA (RPCA) [2] and Low-rank Representation (LRR) [3] are the most frequently used ones: RPCA decomposes the data matrix X into a latent low-rank matrix L and a sparse matrix E of errors, which is $\min_{L,E} \|L\|_* + \lambda \|E\|_1 , s.t. X = L + E$; the basic form of LRR is defined as $\min_{Z,E} \|Z\|_* + \lambda \|E\|_{2,1} , s.t. X = AZ + E$, where Z indicates the lowest rank representation of X with respect to a "dictionary" subset A, usually $A = X$ to avoid a loss of generality. However, neither RPCA nor LRR is capable to cope with nonlinear or multimodal subspaces. It is well known that Kernel PCA (KPCA) is a powerful technique that allows to generalize PCA to nonlinear dimensionality reduction. With the inspirations from KPCA, numerous works on kernelizing LRR have been proposed [4][5]. Unfortunately, a sufficiently large subset of training datasets is crucial to build a dictionary that can represent all the global structures. On the other hand, the topic of kernelizing RPCA is rarely covered in literature [6]. This is because minimizing l_1 norm of an implicit matrix is more mathematically challenging than $l_{2,1}$ norm.

In this work, we propose a novel kernelized RPCA algorithm to create nonlinear SSMs. The fundamental idea is to map the input data onto a feature space where RPCA can be performed. Subsequently, a model is built with the pattern knowledge learned from KRPCA. In evaluation, the built model is applied in ankle bone segmentation and compared with the competing models built with PCA, KPCA and a latest kernelized version LRR respectively.

2 KRPCA for Statistical Shape Modeling

In this section, we provide the derivation of KRPCA and the procedure of modeling via KRPCA. Our derivation is based on a notation that given a subset of n training shapes, a matrix $X \in \mathbb{R}^{m \times n} = \{x_1, \ldots, x_n\}$ is constructed with column x_i representing the i^{th} shape vector.

2.1 Kernel RPCA

Deriving a nonlinear mapping from the input space $\mathcal{I} \in \mathbb{R}^{m \times n}$ to a high dimensional feature space $\mathcal{F} \in \mathbb{R}^{d \times n}$ via a mapping function Φ, where $d \gg m$ and d is unknown. Thus each shape vector $x_i \in \mathcal{I}$ is projected onto feature space where it becomes $\phi(x_i)$. As the projected data matrix $\Phi(X) \in \mathcal{F} = \{\phi(x_1), \ldots, \phi(x_n)\}$ is implicitly represented, a kernel function $\kappa(x, y)$ is induced to present the similarity between shapes vectors in the input space. In this work we apply RBF Gaussian kernel that $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \exp(- \|x_i - x_j\|^2 / 2\sigma^2)$.

Afterwards, RPCA is performed in $\mathcal{F}$ without explicitly constructing the nonlinear mapping Φ. Analogous to RPCA in linear subspace, we decompose $\Phi(X)$ into two parts: $\Phi(X) = \Phi(L) + \Phi(E)$, where $\Phi(L)$ represents the feature space projection of a clean low-rank matrix $L \in \mathcal{I}$ and $\Phi(E)$ represents that of a sparse matrix $E \in \mathcal{I}$. However, this complex system is infeasible to be directly

solved due to the implicitness. We reformulate the problem by considering the distributive property over matrix addition:

$$\Phi(X)^T\Phi(X) = \Phi(X)^T\Phi(L) + \Phi(X)^T\Phi(E)\ , \tag{1}$$

$K = \Phi(X)^T\Phi(X)$ be the constant kernel matrix where $K_{ij} = \phi(x_i)^T\phi(x_j) = \kappa(x_i, x_j)$. We define the matrix $\Phi(X)^T\Phi(L) = K_L$ that:

$$\Phi(X)^T\Phi(L) = \begin{bmatrix} \phi(x_1)^T\phi(l_1) & \dots & \phi(x_1)^T\phi(l_n) \\ \vdots & \ddots & \vdots \\ \phi(x_n)^T\phi(l_1) & \dots & \phi(x_n)^T\phi(l_n) \end{bmatrix} = \begin{bmatrix} \kappa(x_1, l_1) & \dots & \kappa(x_1, l_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n, l_1) & \dots & \kappa(x_n, l_n) \end{bmatrix} = K_L, \tag{2}$$

where $K_L \in \mathbb{R}^{n\times n}$. Similarly, $\Phi(X)^T\Phi(E) = K_E$ is defined and the decomposition problem in feature space is rewritten as:

$$\min_{K_L, K_E} \ \|K_L\|_* + \lambda \|K_E\|_1\ , \quad s.t.\ K = K_L + K_E\ . \tag{3}$$

In this way, the low-rank model is applied on the kernel matrix K that determines the similarity of shapes in input space, in order to find the underlying clusters of similar shapes. We employ the augmented Lagrange multiplier (ALM) to solve Eq(3):

$$\mathcal{L}(K_L, K_E, Y, \mu) = \|K_L\|_* + \lambda \|K_E\|_1 + \langle Y, K - K_L - K_E\rangle + \frac{\mu}{2}\|K - K_L - K_E\|_F^2\ , \tag{4}$$

where Y decides the multiplier, μ is a positive parameter for adaptive penalization and λ is used to balance nuclear and l_1 norms. With an iteration strategy, $K_L^{(k+1)}$ and $K_E^{(k+1)}$ are obtained for the $(k+1)^{th}$ iteration.

Solving K_L. Fix the other variables, $K_L^{(k+1)}$ can be obtained by solving the subproblem:

$$K_L^{(k+1)} = \min_{K_L} \ \|K_L\|_* + \frac{\mu^{(k)}}{2}\left\| K_L - \left(K - K_E^{(k)} + \frac{1}{\mu}Y^{(k)}\right)\right\|_F^2\ , \tag{5}$$

the analytical solution to Eq(5) is given below and the proof is provided in Lemma 1.

$$K_L^{(k+1)} = \mathcal{D}_{1/\mu^{(k)}}\Big[syl\Big(\mu^{(k)}I,\ \big((K_L^{(k)})^T K_L^{(k)}\big)^{-\frac{1}{2}},\ -\mu^{(k)}(K - K_E^{(k)} + \frac{1}{\mu^{(k)}}Y^{(k)})\Big)\Big]. \tag{6}$$

Lemma 1. *Let $F(X) = \|X\|_* + \theta\|X - H\|_F^2$, where θ and H are constant. The solution X^* can be given by deriving the subgradient of $F(X)$ and seeking its stationary point as $F(X)$ is convex. To reduce dimensionality, a shrinkage operator $\mathcal{D}_\tau[X] = Udiag(\Sigma_{ii} - \tau)V^T$ [7] is leveraged to shrink the rank of X^*, where $U\Sigma V^T$ is the singular value decomposition of X.*

First we set the subgradient of $F(X)$ with respect to X zero and have:

$$\frac{\partial}{\partial X}F(X) = X(X^TX)^{-\frac{1}{2}} + 2\theta X - 2\theta H = 0 \ , \tag{7}$$

obviously it is not feasible to get X^* and $(X^TX)^{-\frac{1}{2}}$ simultaneously, we consider $(X^TX)^{-\frac{1}{2}}$ as constant in computation of X^* and iteratively obtain X^* and $(X^TX)^{-\frac{1}{2}}$. Thus the problem of Eq(7) is well known as Sylvester equation that $AX + XB + C = 0$. Here it is solved by the function syl from C++ Armadillo Library, that $\hat{X} = syl(A, B, C)$. By applying the shrinkage operator $\mathcal{D}_\tau$ to $\hat{X}$, we arrive at the solution:

$$X^* = \mathcal{D}_\tau[\ syl(\ 2\theta I, (X^TX)^{-\frac{1}{2}}, -2\theta H\ )\ ] \ . \tag{8}$$

Solving K_E. Fix the other variables, K_E^{k+1} is obtained by solving the subproblem:

$$K_E^{(k+1)} = \min_{K_E} \ \lambda \left\|K_E\right\|_1 + \frac{\mu^{(k)}}{2}\left\|K_E - \left(K - K_L^{(k+1)} + \frac{1}{\mu^{(k)}}Y^{(k)}\right)\right\|_F^2 \ , \tag{9}$$

the problem can be efficiently solved via the soft-thresholding operator $\mathcal{S}_\tau[X] = \max(X - \tau, 0) + \min(X + \tau, 0)$ in [7]. Thus $K_E^{(k+1)}$ is defined as:

$$K_E^{(k+1)} = \mathcal{S}_{\lambda/\mu^{(k)}}\left[K - K_L^{(k+1)} + \frac{1}{\mu^{(k)}}Y^{(k)}\right] \ . \tag{10}$$

By updating $Y^{(k+1)} = Y^{(k)} + \mu^{(k)}(K - K_L^{(k+1)} - K_E^{(k+1)})$ and $\mu^{(k+1)} = \min(\rho\mu^{(k)}, \mu_{max})$, a new iteration is generated. The procedure converges to a point when $\left\|K_L^{(k+1)} - K_L^{(k)}\right\|_F \to 0$ and $\left\|K - K_L^{(k+1)} - K_E^{(k+1)}\right\|_2 < \epsilon$, where the optimal solution K_L^* to the optimization is reached.

2.2 Applying KRPCA to statistical shape modeling

To reduce dimensionality in feature space, K_L^* is leveraged to compute the first k^{th} eigenvectors $\boldsymbol{v^\alpha}$ and eigenvalues λ^α by $K_L^* v_j^\alpha = n\lambda^\alpha v_j^\alpha$, $j = 1$ to n and $\alpha = 1$ to k. Therefore, a lower dimensional KPCA space is constructed. For a sample $x \in \mathcal{I}$, we extract the principal components in KPCA space β^α by projecting x onto the selected eigenvectors that

$$\beta^\alpha(x) = \sum_{i=1}^{n} v_i^\alpha \kappa(x_i, x), \quad \alpha = 1, \ldots, k \ . \tag{11}$$

To apply KPCA into SSM, Davies et al. gave a definition of probability density function (PDF) $\tilde{p}$ of KPCA model in [1]:

$$\tilde{p}(x) \propto \sum_{\alpha=1}^{k} \beta^\alpha(x)\beta^\alpha(x) = \sum_{i,j=1}^{n}\sum_{\alpha=1}^{k} v_i^\alpha v_j^\alpha \kappa(x_i, x)\kappa(x_j, x) \ , \tag{12}$$

representing the square of distance to the origin in KPCA space. As a result, the model captures nonlinear patterns by considering this "proximity to data" measure $\tilde{p}$. Up to now, the KRPCA is built and the whole procedure is summarized in Algorithm 1.

Algorithm 1 Algorithm of Statistical Shape Modeling Using KRPCA

Input: Observation matrix $X \in \mathbb{R}^{m \times n}$
Initialize: $Y = K_L = K_E = 0$, $K = \Phi(X)^T \Phi(X)$, $k = 0$
1. **while** not converge **do**
2. Estimate $K_L^{(k+1)}$ by Eq(6)
3. Estimate $K_E^{(k+1)}$ by Eq(10)
4. Update $Y^{(k+1)}$, $\mu^{(k+1)}$, $k \to k+1$
5. **end while**
7. Perform Eigen Decomposition $K_L^* v_j^\alpha = n\lambda^\alpha v_j^\alpha$
8. Compute principal components in KPCA space $\beta^\alpha(x) = \sum_{i=1}^n v_i^\alpha \kappa(x_i, x)$
9. Get the PDF $\tilde{p}$ for nonlinear statistical shape model
Output: the built KRPCA model

Usually it is trivial to back project parameter vector $\beta^\alpha(x)$ onto the input space, however, it is necessary to generate sample vectors in the input space for further evaluation. Thus we employ the reversed reconstruction approach in [8] to get the reconstructed shape $\hat{x} \in \mathcal{I}$ by:

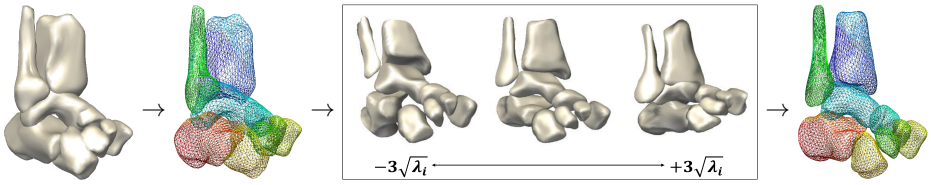
$$\hat{x} = \frac{\sum_{i=1}^n \gamma_i \kappa(\hat{x}, x_i) x_i}{\sum_{i=1}^n \gamma_i \kappa(\hat{x}, x_i)} \ , \quad \gamma_i = \sum_{\alpha=1}^k \beta^\alpha v_i^\alpha \ . \tag{13}$$

3 Evaluation

The evaluation consists of two parts: (1) we investigate the performance of KRPCA by evaluating and comparing the quality of models built with PCA, KPCA, KRLRR [5] and our proposed KRPCA, where 25 corrupted ankle shapes with ground truth randomly chosen from the total 35 datasets are used; (2) we apply the built models in ankle bone segmentation as the segmentation quality reflects the model efficiency, where the remaining 10 unseen datasets are employed.

3.1 Model evaluation

We set the same parameters $\mu_0 = 1.25$, $\lambda = 1/\sqrt{n}$, $\mu_{max} = 1e6$, $\rho = 1.6$, $\epsilon = 1e7$ and $\sigma = 5$ for all methods that are learned from [2][8]. Fig. 1-(a) shows an example of corrupted training shape where bones are abnormally overlapped. A necessary pre-processing step before modeling is establishing correspondence among all the shapes to generate the same number of landmarks (from Fig. 1-(a) to (b), the color represents the order of landmark). Here a shape of ankle contains 9 independent bones with 5148 landmarks in total.



(a) (b) (c) (d)

Fig. 1: Flow chart of modeling and back projection

Table 1: Back projection error MAD

	Mean (mm)	Min (mm)	Max (mm)
Training Datasets	7.77 ± 3.51	2.66	17.03
PCA	7.72 ± 3.45	2.72	16.87
KPCA	7.17 ± 3.35	2.41	16.78
KRLRR	9.20 ± 6.67	2.56	25.98
KRPCA (Ours)	**6.68** ±3.27	**2.39**	**16.10**

Back projection error. To evaluate the robustness to outliers of the competing models, the corrupted training shapes are projected back onto the model (from Fig. 1-(b) to (c)) to generate the corrected reconstructions (Fig. 1-(d)). We define the back projection error by computing the mean absolute distance MAD between the reconstructions $\widetilde{X} = \{\tilde{x}_i\}_{1\leq i\leq n}$ and ground truth $\widehat{X} = \{\hat{x}_i\}_{1\leq i\leq n}$ as $MAD = 1/n \sum_{i=1}^{n} \Psi(\tilde{x}_i, \hat{x}_i)$, where $\Psi(\tilde{x}_i, \hat{x}_i)$ denotes the Euclidean distance between the shape $\tilde{x}_i$ and $\hat{x}_i$. Table 1 shows the MAD results, where the row "Training Datasets" is MAD between training shapes and ground truth that indicates the degree of corruption of training shapes.

Generalization ability and Specificity. To evaluate the quality of SSMs, the most frequently used measures are Generalization ability and Specificity [1]. Fig. 2 shows the results for Generalization ability and Specificity for the first 8 modes of all competing models.

3.2 Application in ankle bone segmentation

Even though the shape of ankle has a zero mean, it is still challenging for the SSM to learn the exact variation of each bone. As a result, often the segmented bones are abnormally overlapped, which motivates us to create a nonlinear model. We apply the built models to an existing segmentation approach [9] based on conventional SSM. The accuracy of segmentation is measured by Hausdorff distance, Dice coefficient and overlap volume percentage (compare Table 2). For an intuitive view, Fig. 3 shows the comparison of segmentation results from the PCA model, KPCA model, KRLRR model and our KRPCA model.

In summary, our proposed KRPCA model has a significantly higher quality in terms of all the measurements; KRLRR and KPCA both perform better than

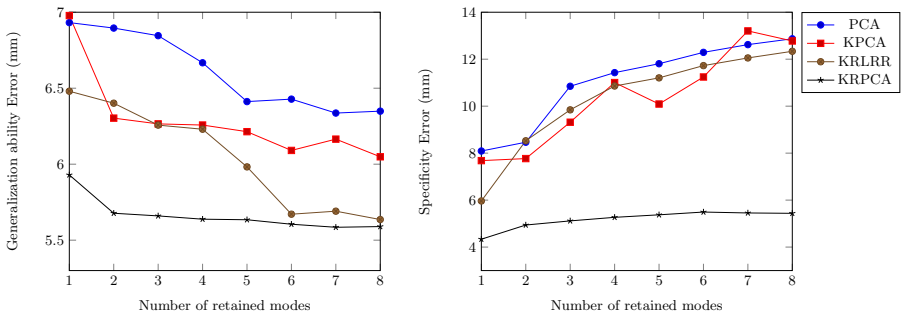


Fig. 2: Generalization ability and Specificity for the models built with PCA, KPCA, KRLRR and KRPCA. Smaller value indicates better result.

Table 2: Segmentation results for reference PCA model and competing models. Note that smaller Hausdorff distance indicates better result; for Dice coefficient and overlapping volume percentage, larger value represents better result.

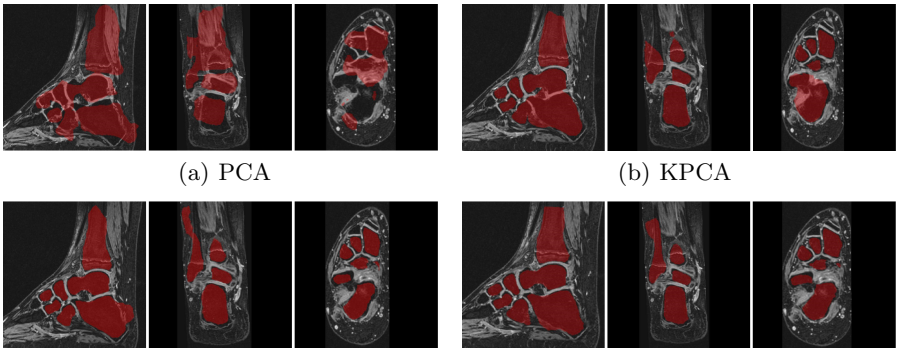
	Hausdorff Distance (mm)			Dice Coefficient			Volume Overlap (%)		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
PCA	7.39±3.30	3.22	15.00	0.86±0.10	0.60	0.93	77.94±12.90	43.02	88.62
KPCA	8.13±3.59	3.64	14.41	0.87±0.07	0.76	0.93	82.38±10.99	61.42	87.43
KRLRR	10.32±5.12	3.22	20.99	0.90±0.03	0.87	0.95	82.38 ± 4.32	76.94	90.46
KRPCA	**7.28±2.93**	**2.36**	**11.05**	**0.91±0.02**	**0.88**	**0.96**	**83.67±4.31**	**79.22**	**92.29**

PCA model, although KRLRR has a larger back projection error and Hausdorff distance than PCA. However, the large standard deviation explains the larger error, that is, KRLRR is not robust enough in conditions of a limited subset of training data. This is also the strong motivation for us to kernelize RPCA rather than LRR.

4 Discussion

The motivation of this work is to create SSMs that are robust to abnormalities in training data and satisfied with nonlinear distribution. A novel kernelized RPCA approach is proposed for modeling. Evaluation results show that the model built with KRPCA has a better quality compared with competed models in conditions that the number of training datasets is relatively limited. In future work, we will emphasize on segmentation with applying the built nonlinear model.

Acknowledgments. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.



(a) PCA (b) KPCA

(c) KRLRR (d) KRPCA

Fig. 3: Figure shows the comparison of segmentation results of competing models in Sagittal, Coronal and Axial position respectively.

References

1. R. H. Davies, C. J. Twining, and C. J. Taylor. *Statistical Models of Shape - Optimization and Evaluation*. Springer, 2008.
2. Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
3. Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010.
4. Canyi Lu, Zhouchen Lin, and Shuicheng Yan. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, 24(2):646–654, 2015.
5. Shijie Xiao, Mingkui Tan, Dong Xu, and Zhao Yang Dong. Robust kernel low-rank representation. *IEEE transactions on neural networks and learning systems*, 27(11):2268–2281, 2016.
6. Lin Wu and Yang Wang. Robust hashing for multi-view data: Jointly learning low-rank kernelized similarity consensus and hash functions. *Image and Vision Computing*, 57:58–66, 2017.
7. Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems*, pages 612–620, 2011.
8. Quan Wang. Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538*, 2012.
9. Sebastian Steger, Florian Jung, and Stefan Wesarg. Personalized articulated atlas with a dynamic adaptation strategy for bone segmentation in ct or ct/mr head and neck images. In *SPIE Medical Imaging*, pages 90341I–90341I. International Society for Optics and Photonics, 2014.

The final publication is available at Springer via
https://doi.org/10.1007/978-3-319-66182-7_16