Spatial and Hierarchical Riemannian Dimensionality Reduction and Dictionary Learning for Segmenting Multichannel Images

Faezeh Fallah
Institute of Signal Processing and System Theory
University of Stuttgart
Stuttgart, Germany

Karim Armanious
Institute of Signal Processing and System Theory
University of Stuttgart
Stuttgart, Germany

Bin Yang
Institute of Signal Processing and System Theory
University of Stuttgart
Stuttgart, Germany

Fabian Bamberg
Center for Diagnostic and Therapeutic Radiology
University of Freiburg
Freiburg, Germany

***Abstract*—In this paper, we proposed an automated method for segmenting objects of weak boundaries and similar intensities on volumetric multichannel images. This method relied on a multiresolution classifier that tackled class overlaps by using the Riemannian geometry of the RCDs of the multiscale patches of every multichannel image and reducing the dimensionality of these RCDs through a novel method that incorporated the intra- and inter-class neighborhoods of the RCDs in the Riemannian space and the spatial and hierarchical relationships between their corresponding patches. The reduced dimensional RCDs were then used to learn resolution-specific dictionaries for coding and classifications. To speed up the optimizations and to avoid convergence to local extrema, the dictionaries and the codes got initialized by a novel scheme that used the Riemannian geometry of the RCDs. This method was evaluated on the challenging task of segmenting cardiac adipose tissues on fat-water MR images.**

***Index Terms*—Riemannian Manifolds, Nonlinear Dimensionality Reduction, Dictionary Learning, Locality Constrained Coding, Segmenting Multichannel Images**

I. Introduction

Hierarchical classifiers have enabled low complex localization and segmentation of objects on multichannel images [1], [2]. However, they have mostly relied on flat Euclidean geometry of the image descriptors. This limited their performance in challenging segmentation tasks. Natural data, including region covariance descriptors (RCDs), lie on (curved) Riemannian manifolds that obey non-Euclidean geometry [3]. RCDs, computed from real-valued n-dimensional features, are real-valued $n \times n$ symmetric positive definite (SPD) matrices that represent points in the interior of a convex cone. This Riemannian manifold is denoted by S^n_{++} and lies in an $n(n+1)/2$-dimensional Euclidean space [3].

RCDs by computing average and second-order statistics, provide a natural way to fuse various types of features. They also reduce impacts of noise or artifacts, provide scale and rotation invariance, and can be efficiently computed through integral images [4]. To benefit from high-dimensional RCDs in sophisticated clustering/classification tasks, several methods have been introduced to reduce their dimensionality in favor of increasing the discriminative power of the clusterer/classifier [3], [5]–[7]. Goh et al. extended Euclidean-based nonlinear dimensionality reduction methods to their Riemannian counterparts to cluster data on separated sub-manifolds of a Riemannian manifold of known dimension [5]. This hindered its application to cases where the sub-manifolds, representing different classes, belong to different Riemannian manifolds of different dimensionalities and these dimensionalities are unknown a priori. Additionally, this clustering method relied on the k-nearest neighbors algorithm which required sub-manifolds to be k-disconnected from each other and k-connected internally. This also hindered its application to many real life scenarios where different classes (sub-manifolds) have overlaps in the Riemannian feature space. It also allowed classification only in a transductive way, as no parametric map to the lower-dimensional space was provided. Harandi et al. improved the discriminative power of the clusterer/classifier by reducing the dimensionality of the Riemannian manifold of the RCDs with regard to their inter- and intra-class neighborhoods [3]. However, it neglected the proximity of the corresponding patches in the spatial domain which was an important factor for image segmentation. Moreover, it determined the lower dimensionality and the size of the inter- and intra-class neighborhoods via cross-validations. These increased its computations and made the optimized hyperparameters specific to the application and data quality.

In this paper, we proposed an automated method for segmenting objects of weak boundaries and similar intensities on volumetric multichannel images. This method relied on a classifier that tackled class overlaps by using a novel scheme based on the Riemannian geometry of the RCDs of the multiscale patches to reduce their dimensionality and to learn their resolution-specific dictionaries for coding and classifications. This method was evaluated on the challenging task of segmenting cardiac adipose tissues on fat-water MR images. In the following, a fat-water image/patch refers to a *volumetric* fat-water image/patch.

II. Materials and Methods

A. Framework of Automatic Segmentation

The proposed method used a hierarchical decision tree classifier that encoded a multiresolution image pyramid to

Magnetic resonance images of this work were acquired under the grant BA 4233/4-1 from German Research Foundation (DFG).

Corresponding author email: int87517@stud.uni-stuttgart.de

978-9-0827-9703-9/19/$31.00

process samples from coarse to fine. Samples of the coarsest resolution were fed to the root node of the tree to be processed while being decomposed into finer resolutions. The finest samples reached the leaf nodes. Every decision node of this classifier: 1) reduced the dimensionality of the RCDs of its received samples by the *dimensionality reduction maps* learnt in the previous (coarser) *and* the current layer. 2) encoded the reduced dimensional RCDs by a dictionary learnt based on a kernelized locality constrained coding (kLCC) [8]. 3) classified every sample based on the code of its RCD. 4) decomposed samples into a finer resolution and assigned them to the nodes of the next (finer) layer according to their estimated labels.

B. The Multiscale RCDs

The fat-water images were divided into a training and a test set and were processed by a multiresolution pyramid to form multiresolution training and test samples. The training samples involved multiscale fat-water patches of the training images, their RCDs and their reference labels. These samples were used to optimize the parameters of the classifier. The test samples involved multiscale fat-water patches of the test images and their RCDs. They were used to evaluate the classifier performance. The pyramid was built from coarse to fine and involved L resolution layers with $l = L$ and $l = 1$ denoting the coarsest and the finest layer, respectively. In the l^{th} layer, cubic patches of $(2l+1)^3$ voxels were extracted from every fat or water image. These patches covered the entire image without having overlaps with each other. Then, from the 26-connected neighborhood of every nonborder voxel of each fat/water patch, 45 features were extracted. These features included median of intensities, average gradient magnitude, histogram of oriented gradients quantized by 20 vectors of a regular icosahedron, and $14 \times 3 = 42$ isotropic (angle-invariant) features of angular mean, range, and standard deviation of 14 Haralick features of a 3D gray-level co-occurrence matrix of 1 voxel displacement in 13 directions [9]. Thus, from every fat-water (2-channel) patch, $45 \times 2 = 90$ intra-channel features were extracted. Also, 6 inter-channel features, including fat fraction ratio, fat-water ratio, absolute differences in the median of intensities and average gradient magnitudes, and l_1 norm of differences in histogram of oriented gradients and isotropic Haralick features, were extracted from it. These gave an 96-dimensional feature vector for every fat-water patch at every resolution layer. Using these features, the integral image of the patch was computed and yielded its $n \times n$ RCD with $n = 96$ [4]. For the j^{th} fat-water patch at the l^{th} layer, this RCD was denoted by $\mathbf{C}_{l,j} \in S^n_{++} \subset \mathbb{R}^{n \times n}$.

C. Spatial and Hierarchical Dimensionality Reduction

In every layer l of the classifier, prior to the classifications, the dimensionality of the $n \times n$ RCDs of the received samples were reduced by the dimensionality reduction maps learnt in the previous (coarser) layers and the current layer. The reduced dimensional RCDs were then used to classify the samples. Then these samples were decomposed into a finer resolution to be processed by the next (finer) layer of the classifier.

The dimensionality reduction in every layer l was done by a map $f_{\mathbf{W}_l} : S^{n_l}_{++} \to S^{m_l}_{++}$ with $f_{\mathbf{W}_l}(\mathbf{X}_{l,j}) = \mathbf{W}_l^T \mathbf{X}_{l,j} \mathbf{W}_l$, $m_l = \frac{n_l}{r_l}$, $n_l = \frac{n}{r_L \times r_{L-1} \times \cdots \times r_{l+1}}$, and $\mathbf{W}_l \in G(m_l, n_l) \subset \mathbb{R}^{n_l \times m_l}$ where $G(m_l, n_l)$ denoted the Grassmannian manifold.

Because this map was to be used by the next (finer) layers as well, during the training, RCDs of the training samples of the current layer, $\{\mathbf{C}_{l,j} \in S^n_{++}\}_{j=1}^{N_l}$, and the RCDs of the samples resulted from their hierarchical decomposition, $\{\mathbf{C}_{l-1,k} \in S^n_{++}\}_{k=1}^{N_{l-1}}$, were considered.

To this end, first the dimensionality of $\{\mathbf{C}_{l,j} \in S^n_{++}\}_{j=1}^{N_l}$ and $\{\mathbf{C}_{l-1,k} \in S^n_{++}\}_{k=1}^{N_{l-1}}$ were reduced by the dimensionality reduction maps learnt in the previous (coarser) layers. These formed $\{\mathbf{X}_{l,j} \in S^{n_l}_{++} \subset \mathbb{R}^{n_l \times n_l}\}_{j=1}^{N_l}$ and $\{\mathbf{X}_{l-1,k} \in S^{n_l}_{++} \subset \mathbb{R}^{n_l \times n_l}\}_{k=1}^{N_{l-1}}$ with $\mathbf{X}_{l,j} = \mathbf{W}_{l+1}^T \cdots \mathbf{W}_{L-1}^T \mathbf{W}_L^T \mathbf{C}_{l,j} \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_{l+1}$ and $\mathbf{X}_{l-1,k} = \mathbf{W}_{l+1}^T \cdots \mathbf{W}_{L-1}^T \mathbf{W}_L^T \mathbf{C}_{l-1,k} \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_{l+1}$. Thus, $\mathbf{C}_{l,j} \in S^n_{++}$ and $\mathbf{X}_{l,j} \in S^{n_l}_{++}$ were from the same sample at the l^{th} layer. $\mathbf{C}_{l-1,k} \in S^n_{++}$ and $\mathbf{X}_{l-1,k} \in S^{n_l}_{++}$ were from a child of this sample at the $(l-1)^{\text{th}}$ layer.

The map $f_{\mathbf{W}_l} : S^{n_l}_{++} \to S^{m_l}_{++}$ aimed at reducing the complexity of dictionary learning for classification while increasing the discriminative power of the classifier by minimizing the intra-class distances and maximizing the inter-class distances between RCDs of the samples. Additionally, and in particular for image segmentation, we aimed to maintain sharp edges between different objects and a smooth segmentation within one object region. To achieve these, we proposed to weigh the above (Riemannian) distances by a spatial affinity function that measured neighborhood related label-correspondence between RCDs of the same layer. Also, to make the dimensionality reduction maps, learnt by each layer, applicable to the next (finer) layer, we introduced a hierarchical affinity function that measured hierarchically related label-correspondence between RCDs of samples at adjacent layers.

The spatial $a_s : S^{n_l}_{++} \times S^{n_l}_{++} \to \mathbb{R}$ and the hierarchical $a_h : S^{n_l}_{++} \times S^{n_l}_{++} \to \mathbb{R}$ affinity functions were defined as

$$a_s(\mathbf{X}_{l,j}, \mathbf{X}_{l,k}) = \begin{cases} +2, \text{ if } c_{l,j} = c_{l,k} \text{ and } \mathbf{X}_{l,j} \leftrightarrow \mathbf{X}_{l,k} \\ +1, \text{ if } c_{l,j} = c_{l,k} \text{ and } \mathbf{X}_{l,j} \nleftrightarrow \mathbf{X}_{l,k} \\ -1, \text{ if } c_{l,j} \neq c_{l,k} \text{ and } \mathbf{X}_{l,j} \nleftrightarrow \mathbf{X}_{l,k} \\ -2, \text{ if } c_{l,j} \neq c_{l,k} \text{ and } \mathbf{X}_{l,j} \leftrightarrow \mathbf{X}_{l,k} \end{cases}, \quad (1)$$

$$a_h(\mathbf{X}_{l,j}, \mathbf{X}_{l-1,k}) = \begin{cases} +2, \text{ if } c_{l,j} = c_{l-1,k} \text{ and } \mathbf{X}_{l,j} \updownarrow \mathbf{X}_{l-1,k} \\ +1, \text{ if } c_{l,j} = c_{l-1,k} \text{ and } \mathbf{X}_{l,j} \downarrow \mathbf{X}_{l-1,k} \\ -1, \text{ if } c_{l,j} \neq c_{l-1,k} \text{ and } \mathbf{X}_{l,j} \downarrow \mathbf{X}_{l-1,k} \\ -2, \text{ if } c_{l,j} \neq c_{l-1,k} \text{ and } \mathbf{X}_{l,j} \updownarrow \mathbf{X}_{l-1,k} \end{cases}, \quad (2)$$

where $\mathbf{X}_{l,j}, \mathbf{X}_{l,k}, \mathbf{X}_{l-1,k} \in S^{n_l}_{++}$ were the reduced dimensional RCDs; $c_{l,j}, c_{l,k}$ and $c_{l-1,k}$ were the reference labels of the training samples of $\mathbf{X}_{l,j}, \mathbf{X}_{l,k}$ and $\mathbf{X}_{l-1,k}$, respectively; $\leftrightarrow$ $(\nleftrightarrow)$ denoted patches of the RCDs being (not being) 26-connected neighborhood of each other; $\updownarrow$ $(\downarrow)$ denoted patches of the RCDs having (not having) a hierarchical (parent-child) relationship with each other. Patch of $\mathbf{X}_{l,j} \in S^{n_l}_{++}$ was the parent of the patch of $\mathbf{X}_{l-1,k} \in S^{n_l}_{++}$ if patch of $\mathbf{X}_{l-1,k} \in S^{n_l}_{++}$ was part of the patch of $\mathbf{X}_{l,j} \in S^{n_l}_{++}$.

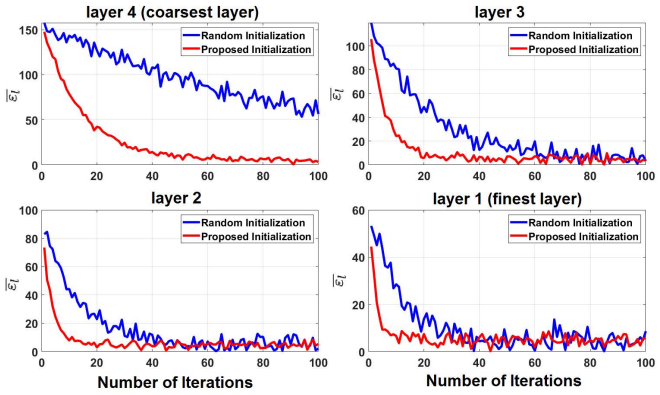

Fig. 1. Profile of the averaged reconstruction errors of dictionary learning and coding over iterations of every layer of the hierarchical classifier.

Then the optimum parameters $\mathbf{W}_l \in G(m_l, n_l) \subset \mathbb{R}^{n_l \times m_l}$ for the map $f_{\mathbf{W}_l} : S_{++}^{n_l} \to S_{++}^{m_l}$ at the l^{th} layer were determined as

$$\begin{aligned} \mathbf{W}_l = \arg\min_{\hat{\mathbf{W}}_l} \Bigg(& \sum_{\substack{j,k=1 \\ j \neq k}}^{N_l} a_s(\mathbf{X}_{l,j}, \mathbf{X}_{l,k}) \cdot \tau^2(\hat{\mathbf{W}}_l^T \mathbf{X}_{l,j} \hat{\mathbf{W}}_l, \hat{\mathbf{W}}_l^T \mathbf{X}_{l,k} \hat{\mathbf{W}}_l) \\ & + \sum_{j=1}^{N_l} \sum_{k=1}^{N_{l-1}} a_h(\mathbf{X}_{l,j}, \mathbf{X}_{l-1,k}) \cdot \tau^2(\hat{\mathbf{W}}_l^T \mathbf{X}_{l,j} \hat{\mathbf{W}}_l, \hat{\mathbf{W}}_l^T \mathbf{X}_{l-1,k} \hat{\mathbf{W}}_l) \Bigg), \\ \text{s.t.} \quad & \mathbf{W}_l^T \mathbf{W}_l = \mathbf{I}_{m_l}, \end{aligned} \tag{3}$$

where $\tau^2(\cdot)$ was the Riemannian distance between RCDs and $\mathbf{I}_{m_l}$ was the $m_l \times m_l$ identity matrix. The unitary constraint ensured that $\forall j,k, \mathbf{W}_l^T \mathbf{X}_{l,j} \mathbf{W}_l$ and $\mathbf{W}_l^T \mathbf{X}_{l-1,k} \mathbf{W}_l$ were valid SPD matrices, i.e. $\mathbf{W}_l \in \mathbb{R}^{n_l \times m_l}$ was of full (m_l) rank [3].

The minimization in (3) was done by an iterative Riemannian conjugate gradient method on the Grassmannian manifold $G(m_l, n_l) \ni \mathbf{W}_l$ [3]. To speed up this minimization, in the initial iterations $\tau^2(\cdot) = \tau_{le}^2(\cdot)$, where $\tau_{le}^2(\mathbf{X}, \mathbf{Y}) = \|\log(\mathbf{X}) - \log(\mathbf{Y})\|_F^2$ was the log-Euclidean metric based on the Frobenius norm $\|\cdot\|_F^2$ and the principal matrix logarithm $\log(\cdot)$. In the final iterations, $\tau^2(\cdot) = \tau_S^2(\cdot)$, with $\tau_S^2(\mathbf{X}, \mathbf{Y}) = \log|\frac{\mathbf{X}+\mathbf{Y}}{2}| - \frac{1}{2}\log|\mathbf{X}\mathbf{Y}|$ being the Stein divergence [3]. This way, the reduced dimensional RCDs of the training samples at the l^{th} layer were $\{\mathbf{W}_l^T \mathbf{X}_{l,j} \mathbf{W}_l = \mathbf{Y}_{l,j} \in S_{++}^{m_l}\}_{j=1}^{N_l}$. These were used in the following steps to learn dictionaries for classifying test samples.

D. Riemannian Dictionary Learning

The proposed hierarchical classifier relied on a Riemannian kernelized locality constrained coding (kLCC) to classify the test samples [8]. To this end, during the training, in every layer l, a dictionary $\mathbb{D}_l = \{\mathbf{D}_{l,i} \in S_{++}^{m_l}\}_{i=1}^{A_l}$ was learnt from the reduced dimensional RCDs $\mathbb{Y}_l = \{\mathbf{Y}_{l,j} \in S_{++}^{m_l}\}_{j=1}^{N_l}$ of the training samples. During the test, the reduced dimensional RCD of every test sample was encoded by the learnt dictionary and was then classified based on the resulting code.

In an approximated kLCC, the prior on the codes was eliminated by representing each RCD with a number of nearest dictionary atoms [8]. In this case, the closed-form solution of the code was unique if this number was less than or equal to the dimension of the RCD. Based on this and without loss of generality, we assumed that in every layer l of the classifier, the $m_l \times m_l$ RCDs of the training samples of a specific class $1 \leq c \leq C$ could be represented by m_l class-specific dictionary atoms. In this regard, the RCDs $\mathbb{Y}_l = \{\mathbf{Y}_{l,j} \in S_{++}^{m_l}\}_{j=1}^{N_l}$ of the training samples and the dictionary atoms $\mathbb{D}_l = \{\mathbf{D}_{l,i} \in S_{++}^{m_l}\}_{i=1}^{A_l}$ were partitioned according to their class labels. These formed $\mathbb{Y}_l = \bigcup_{c=1}^{C} \mathbb{Y}_{l,c}$ and $\mathbb{D}_l = \bigcup_{c=1}^{C} \mathbb{D}_{l,c}$ with $\mathbb{Y}_{l,c} = \{\mathbf{Y}_{l,c,j} \in S_{++}^{m_l}\}_{j=1}^{N_{l,c}}$, $\mathbb{D}_{l,c} = \{\mathbf{D}_{l,c,i} \in S_{++}^{m_l}\}_{i=1}^{A_{l,c}=m_l}$, $N_l = \sum_{c=1}^{C} N_{l,c}$, and $A_l = A_{l,c} \cdot C$. Having $A_{l,c} = m_l$, not only guaranteed uniqueness of the resulting codes but also implied that by moving towards finer layers and reducing m_l the dictionary atoms become more discriminative.

In the Riemannian kLCC [6], the nonlinear geometries of $S_{++}^{m_l}$ was avoided by embedding it into a linear Hilbert space H_l through a map $\phi_l : S_{++}^{m_l} \to H_l$ that used a reproducing positive definite kernel $k_l(\mathbf{X}, \mathbf{Y}) = \exp(-\beta_l \cdot \tau_S^2(\mathbf{X}, \mathbf{Y}))$ based on the Stein divergence $\tau_S^2(\cdot)$. The kernel $k_l : S_{++}^{m_l} \times S_{++}^{m_l} \to \mathbb{R}$ fulfilled: $k_l(\mathbf{X}, \mathbf{Y}) = (\phi_l(\mathbf{X}))^T \cdot \phi_l(\mathbf{Y})$. Accordingly,

$$\forall j, \phi_l(\mathbf{Y}_{l,j}) \approx \sum_{i=1}^{A_l} \alpha_{l,j,i} \cdot \phi_l(\mathbf{D}_{l,i}), \quad \text{s.t.} \sum_{i=1}^{A_l} \alpha_{l,j,i} = 1, \tag{4}$$

where $\boldsymbol{\alpha}_{l,j} = (\alpha_{l,j,1}, \cdots, \alpha_{l,j,A_l})^T$ was the coding vector of $\phi_l(\mathbf{Y}_{l,j})$ with respect to $\phi_l(\mathbb{D}_l) = \{\phi_l(\mathbf{D}_{l,i})\}_{i=1}^{A_l}$. In [6], [7], the dictionaries and the codes were optimized iteratively and alternately. That is, in every trial, one unknown (dictionary or code) was fixed and the other unknown was optimized. However, these methods provided no way for the initialization other than using randomized dictionaries and codes. Regarding the nonconvexity of the cost functions involved in such a dictionary learning and coding [6], [7], this random initialization could lead to a convergence to local extrema and an increased processing time. Additionally, in case of using the embedding $\phi_l : S_{++}^{m_l} \to H_l$, the parameter β_l could significantly impact the resulting dictionary and thus needed to be optimized as well. This hindered the initialization of the dictionaries in the Hilbert space by a linear combination of $\{\phi_l(\mathbf{Y}_{l,j})\}_{j=1}^{N_l}$ as β_l was unknown initially. Thus, to apply [6] in the training (optimization), we proposed a systematic way to initialize the dictionary and the codes in every layer of the classifier.

E. Layer-specific Initialization of Dictionary and Codes

In our approach, for every layer l, the dictionary $\mathbb{D}_l = \{\mathbf{D}_{l,i} \in S_{++}^{m_l}\}_{i=1}^{A_l} = \bigcup_{c=1}^{C} \mathbb{D}_{l,c}$ with $\mathbb{D}_{l,c} = \{\mathbf{D}_{l,c,i} \in S_{++}^{m_l}\}_{i=1}^{A_{l,c}=m_l}$ was initialized by exploiting the Riemannian geometry of $\mathbb{Y}_l = \{\mathbf{Y}_{l,j} \in S_{++}^{m_l}\}_{j=1}^{N_l} = \bigcup_{c=1}^{C} \mathbb{Y}_{l,c}$ with $\mathbb{Y}_{l,c} = \{\mathbf{Y}_{l,c,j} \in S_{++}^{m_l}\}_{j=1}^{N_{l,c}}$. First the Karcher mean $\overline{\mathbf{Y}}_{l,c} = \arg\min_{\mathbf{Y} \in S_{++}^{m_l}} \sum_{j=1}^{N_{l,c}} \|\log_{\mathbf{Y}}(\mathbf{Y}_{l,c,j})\|_{\mathbf{Y}}^2$ of every $\mathbb{Y}_{l,c}$, $1 \leq c \leq C$, was computed iteratively [5]. Here, $\log_{\mathbf{Y}} : S_{++}^{m_l} \to T_{\mathbf{Y}} S_{++}^{m_l}$ was a map from the Riemannian manifold $S_{++}^{m_l}$ to its m_l-dimensional tangent space at $\mathbf{Y} \in S_{++}^{m_l}$ denoted by $T_{\mathbf{Y}} S_{++}^{m_l}$; $\|\log_{\mathbf{Y}}(\mathbf{Y}_{l,c,j})\|_{\mathbf{Y}}^2$ was the norm of the vector $\log_{\mathbf{Y}}(\mathbf{Y}_{l,c,j})$ in $T_{\mathbf{Y}} S_{++}^{m_l}$.

Then a principal geodesic analysis was conducted by mapping $\mathbb{Y}_{l,c}$ into $T_{\overline{\mathbf{Y}}_{l,c}} S_{++}^{m_l}$ and finding all m_l principal directions (eigen vectors) of the covariance matrix $\Sigma_{\overline{\mathbf{Y}}_{l,c}} = \frac{1}{N_{l,c}} \sum_{j=1}^{N_{l,c}} [(\log_{\overline{\mathbf{Y}}_{l,c}}(\mathbf{Y}_{l,c,j})) \cdot (\log_{\overline{\mathbf{Y}}_{l,c}}(\mathbf{Y}_{l,c,j}))^T]$. These vectors, denoted by $\{\mathbf{y}_{l,c,i}\}_{i=1}^{A_{l,c}=m_l}$, formed an orthonormal basis for

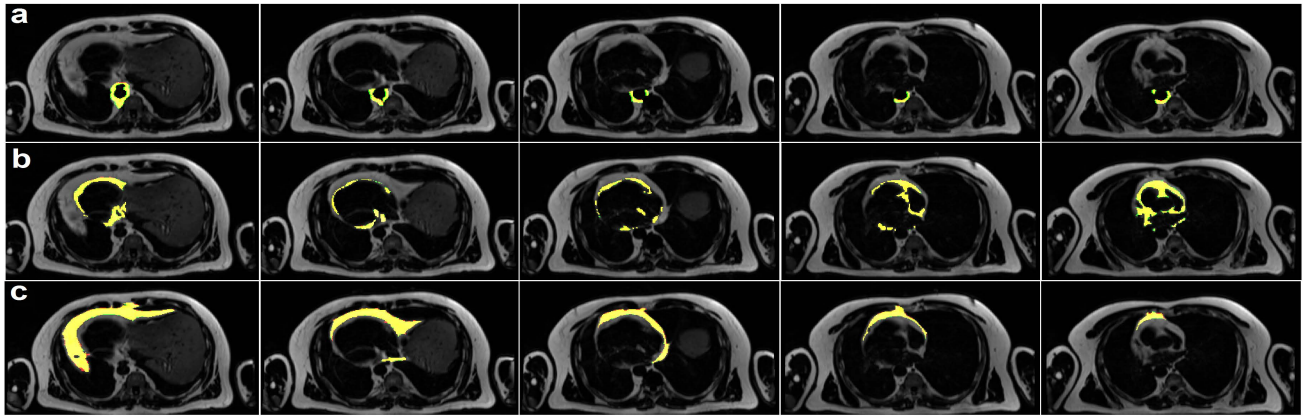

Fig. 2. The automatically segmented masks (red), the reference masks (green), and their overlaps (yellow) for PvAT (a), EpAT (b), and PeAT (c) shown on axial slices of a test fat image.

$T_{\overline{\mathbf{Y}}_{l,c}} S^{m_l}_{++}$. Accordingly, *vector-counterparts* of the initial dictionary atoms for representing samples of the c^{th} class were defined as: $\{\mathbf{d}^{(0)}_{l,c,i} = \sigma_{l,c,i} \cdot \mathbf{y}_{l,c,i}\}_{i=1}^{A_{l,c}=m_l}$ where $\sigma_{l,c,i}$ was the standard deviation of the samples along $\mathbf{y}_{l,c,i}$ or $\sigma_{l,c,i} = \sqrt{\lambda_{l,c,i}}$ with $\lambda_{l,c,i}$ being the eigenvalue of $\mathbf{y}_{l,c,i}$. Finally, by using the map $\log^{-1}_{\overline{\mathbf{Y}}_{l,c}} = \exp_{\overline{\mathbf{Y}}_{l,c}} : T_{\overline{\mathbf{Y}}_{l,c}} S^{m_l}_{++} \to S^{m_l}_{++}$, these vectors were converted to points (matrices) on $S^{m_l}_{++}$ to form the *initial* dictionary atoms for representing samples of the c^{th} class. That is, the initial dictionary $\mathbb{D}^{(0)}_l = \{\mathbf{D}^{(0)}_{l,i}\}_{i=1}^{A_l} = \bigcup_{c=1}^{C} \mathbb{D}^{(0)}_{l,c}$ was made from $\mathbb{D}^{(0)}_{l,c} = \{\mathbf{D}^{(0)}_{l,c,i} = \exp_{\overline{\mathbf{Y}}_{l,c}}(\mathbf{d}^{(0)}_{l,c,i})\}_{i=1}^{A_{l,c}=m_l}$.

Then for every $\mathbf{Y}_{l,j} \in \mathbb{Y}_{l,c}$, that had the class label c, the *initial* codes were defined as

$$\alpha^{(0)}_{l,j,i} = \begin{cases} 1/m_l, & \text{if } \mathbf{D}^{(0)}_{l,i} \in \mathbb{D}^{(0)}_{l,c} \\ 0, & \text{otherwise} \end{cases}, \quad 1 \le i \le A_l. \tag{5}$$

These formed $\boldsymbol{\alpha}^{(0)}_{l,j} = (\alpha^{(0)}_{l,j,1}, \cdots, \alpha^{(0)}_{l,j,A_l})^T$ with $\sum_{i=1}^{A_l} \alpha^{(0)}_{l,j,i} = 1$.

Having the dictionaries and the codes initialized, the iterative and alternating approach of [6] was followed to optimize the dictionary $\mathbb{D}^{(J)}_l = \{\mathbf{D}^{(J)}_{l,i} \in S^{m_l}_{++}\}_{i=1}^{A_l} = \bigcup_{c=1}^{C} \mathbb{D}^{(J)}_{l,c}$ with $\mathbb{D}^{(J)}_{l,c} = \{\mathbf{D}^{(J)}_{l,c,i} \in S^{m_l}_{++}\}_{i=1}^{A_{l,c}=m_l}$, the codes $\{\boldsymbol{\alpha}^{(J)}_{l,j} \in \mathbb{R}^{A_l}\}_{j=1}^{N_l}$, and the $\beta^{(J)}_l \in \mathbb{R}$ over J iterations. The $\beta^{(J)}_l$ defined $\phi^{(J)}_l : S^{m_l}_{++} \to H_l$ and $k^{(J)}_l(\mathbf{X}, \mathbf{Y}) = (\phi^{(J)}_l(\mathbf{X}))^T \cdot \phi^{(J)}_l(\mathbf{Y})$.

III. Evaluation

A. Image Data Sets and Reference Labeling

Thirty nine fat-water MR images were acquired on a clinical 3 T MR scanner and were then divided into 21 images for training and 18 images for testing. On every fat-water image, reference masks of pericardial (PeAT), epicardial (EpAT), and cardiac perivascular adipose tissues (PvAT) were manually segmented by an experienced radiologist using the interactive tools of Medical Imaging Interaction Toolkit (MITK) [10]. These masks determined the voxel-wise and the patch-wise reference labels for $C = 4$ classes including background.

B. Segmentation of the Test Images

To automatically segment cardiac adipose tissues on a test fat-water image, an image pyramid of $L = 4$ layers was built. The resulting test samples were processed from coarse to fine. In every layer $1 \le l \le 4$ of the classifier, first the dimensionality of the RCD of every test sample j was reduced by a factor of 2 and then it was mapped into a linear Hilbert space yielding $\phi^{(J)}_l(\mathbf{Y}_{l,j})$. This was then encoded by the mapped dictionary $\phi^{(J)}_l(\mathbb{D}_l) = \{\phi^{(J)}_l(\mathbf{D}_{l,i})\}_{i=1}^{A_l}$ to get the classification code $\boldsymbol{\alpha}^{(J)}_{l,j} = (\alpha^{(J)}_{l,j,1}, \cdots, \alpha^{(J)}_{l,j,A_l})^T$ of the test sample.

The dictionary atoms were class-specific, i.e. every atom $\mathbf{D}_{l,i} \in \mathbb{D}_l$ had a class label $b_{l,i} = c$ if $\mathbf{D}_{l,i} \in \mathbb{D}_{l,c} \subset \mathbb{D}_l$. Accordingly, for every $\phi^{(J)}_l(\mathbf{Y}_{l,j})$, a class-specific code $\boldsymbol{\alpha}_{l,c,j} = (\alpha^{(J)}_{l,j,1} \cdot \delta(b_{l,1} - c), \cdots, \alpha^{(J)}_{l,j,A_l} \cdot \delta(b_{l,A_l} - c))^T$ could be obtained [6], [7]. This code determined error of reconstructing $\phi^{(J)}_l(\mathbf{Y}_{l,j})$ with $\phi^{(J)}_l(\mathbb{D}_l)$ for the c^{th} class as

$$\begin{aligned} \varepsilon_{l,c}(\mathbf{Y}_{l,j}) &= \left\| \phi^{(J)}_l(\mathbf{Y}_{l,j}) - \sum_{i=1}^{A_l} \alpha^{(J)}_{l,j,i} \cdot \delta(b_{l,i} - c) \cdot \phi^{(J)}_l(\mathbf{D}_{l,i}) \right\|^2 \\ &= -2\boldsymbol{\alpha}^T_{l,c,j} \cdot \mathbf{K}(\mathbf{Y}_{l,j}, \mathbb{D}_l) + \boldsymbol{\alpha}^T_{l,c,j} \cdot \mathbb{K}(\mathbb{D}_l, \mathbb{D}_l) \cdot \boldsymbol{\alpha}^T_{l,c,j} \end{aligned}, \tag{6}$$

where $\mathbf{K}(\mathbf{Y}_{l,j}, \mathbb{D}_l) = [k^{(J)}_l(\mathbf{Y}_{l,j}, \mathbf{D}_{l,i})]_{i=1,\cdots,A_l}$ and $\mathbb{K}(\mathbb{D}_l, \mathbb{D}_l) = [k^{(J)}_l(\mathbf{D}_{l,j}, \mathbf{D}_{l,i})]_{i,j=1,\cdots,A_l}$. Accordingly, the class label of every $\mathbf{Y}_{l,j}$ was estimated as $\hat{c}_{l,j} = \arg\min_{1 \le c \le C} \varepsilon_{l,c}(\mathbf{Y}_{l,j})$. By assigning these labels to the corresponding fat-water patches, masks of the automatically segmented adipose tissues were obtained.

C. Quantitative Metrics

The automatically segmented masks of the cardiac adipose tissues were compared against the reference masks using the quantitative metrics of dice coefficient (Dice), mean symmetric surface distance (MSSD), and Hausdorff distance (HSD) [11].

IV. Results

A. Convergence and Processing Times

The kLCC-based dictionary was iteratively learnt once with the proposed initialization and once with a random initialization. To measure the convergence of these dictionaries in every layer l of the classifier, in each iteration, the updated dictionaries and β_l were used to reconstruct the training samples in that layer using (6). The resulting class-specific reconstruction

errors were averaged yielding: $\overline{\varepsilon}_l = \frac{1}{C}\sum_{c=1}^{C}\sum_{\forall j}\varepsilon_{l,c}(\mathbf{Y}_{l,j})$. Fig. 1 shows the profile of these errors over iterations.

In each layer of the proposed classifier, the computational complexity of the Riemannian dimensionality reduction and dictionary learning were like the previous methods [3], [5]–[7]. The computational complexity of the proposed initialization was in the order of a principal geodesic analysis [12].

On a PC with 32 GB RAM and a quad-core CPU of 3.10 GHz frequency, the multiscale RCDs of 21 training fat-water images got computed in 2.4 hours and the layer-specific dictionaries with the proposed/random initialization were learned in 2.2/2.7 hours. On the same PC, segmentation of a test fat-water image took 18±5.4 minutes.

B. Objective and Subjective Results

On 18 test fat-water images, the proposed method achieved Dice = 87.9±2.1, MSSD = 1.2±0.22, and HSD = 4.12±1.08 in the automatically segmented EpAT. Dice = 89.7±1.1, MSSD = 0.8±0.11, and HSD = 3.86±0.92 in the automatically segmented PeAT. Dice = 82.5±1.9, MSSD = 1.23±0.42, and HSD = 4.88±1.12 in the automatically segmented PvAT.

Fig. 2 shows the automatically segmented and the reference masks of these adipose tissues on a test fat image.

V. Discussion and Conclusion

The present method expanded on [3], by incorporating the spatial (neighborhood) and the hierarchical (parent-child) relationships between patches into their intra- and inter-class Riemannian distances. This, not only enhanced the Riemannian dimensionality reduction for the segmentations, but also eliminated the need to considering intra- and inter-class neighborhood sizes as hyperparameters. This improved the generalizability of the proposed method to any segmentation/classification task. Also, over the hierarchies of the proposed classifier, dimensions of the Riemannian manifolds of the RCDs were reduced in proportion to the reduced number of available samples. This could enhance the discriminative power of any hierarchical classifier [1].

The present method introduced a novel way to initialize the dictionaries and the codes of a kLCC [6], [8]. This was done by exploiting the Riemannian geometry of the reduced dimensional RCDs of the training samples. Despite of involving additional computations, this initialization led to a faster convergence than a random initialization. It could also enhance any kernelized dictionary learning by allowing a systematic initialization and optimization of the kernel size. Moreover, we used dictionary learning not only to represent training samples in a compact way, but also to generate discriminative codes for the classification of the test samples. To this end, the kLCC [6], [8] was preferred to a sparse coding since it provided closed-form unique solutions to the codes and showed a higher classification performance [6]. Also, this coding avoided degenerations caused by the sensitivity of the Lasso regularization to numerical inconsistencies and did not need to normalize the second norm of its updated dictionary at the end of each iteration. In the approximated kLCC [8], the number of nearest dictionary atoms was a hyperparameter. To avoid its optimization, we set it to the maximum value that guaranteed uniqueness of the codes' solution [6]. Moreover, in contrast to [5], our method made no assumption about the number or dimensionality of the involved manifolds. It also did not demand a certain intra-class connectivity or a certain inter-class disconnectivity between samples as the classifier boundaries were derived supervisedly.

In a previous approach [13], Riemannian dimensionality reduction and dictionary learning were jointly done by considering their interactions. In contrast, we performed those steps independently. Use of the aforementioned interactions could reduce the computational costs of the present method and would be a subject of a future work. The present method was evaluated on the challenging task of segmenting cardiac adipose tissues on fat-water MR images. To the best of our knowledge, no previous method has addressed this automatic segmentation. Thus a comparison with the state-of-the-art was not possible. Future work will be a comparative evaluation of the present method on other segmentation/classification tasks. Finally, the proposed hierarchical method could extract features at different resolutions and abstraction levels. This reduced the need to Gabor, local binary pattern, and other morphological descriptors. However, the impact of these additional features can be evaluated in an extended work.

References

[1] F. Fallah, B. Yang, S. S. Walter, and F. Bamberg, "A hierarchical ensemble classifier for multilabel segmentation of fat-water MR images," in *Proc Eur Signal Process Conf*, Sep. 2018.

[2] ——, "Hierarchical feature-learning graph-based segmentation of fat-water MR images," in *Proc IEEE Conf Signal Process Algorithms Archit Arrange Appl*, 2018, pp. 37–42.

[3] M. Harandi, M. Salzmann, and R. Hartley, "Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods," vol. 40, no. 1, pp. 48–62, 2018.

[4] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," vol. 30, no. 10, pp. 1713–1727, 2008.

[5] A. Goh and R. Vidal, "Clustering and dimensionality reduction on Riemannian manifolds," in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2008, pp. 1–7.

[6] M. Harandi and M. Salzmann, "Riemannian coding and dictionary learning: Kernels to the rescue," in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2015, pp. 3926–3935.

[7] M. T. Harandi, R. Hartley, B. Lovell, and C. Sanderson, "Sparse coding on symmetric positive definite manifolds using Bregman divergences," *IEEE Trans Neural Netw Learn Syst*, vol. 27, no. 6, pp. 1294–1306, 2016.

[8] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2010, pp. 3360–3367.

[9] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," no. 6, pp. 610–621, 1973.

[10] I. Wolf, M. Vetter, I. Wegner, T. Böttger, M. Nolden, M. Schöbinger, M. Hastenteufel, T. Kunert, and H. P. Meinzer, "The medical imaging interaction toolkit," *Med Image Anal*, vol. 9, no. 6, pp. 594–604, 2005.

[11] F. Fallah, S. S. Walter, F. Bamberg, and B. Yang, "Simultaneous volumetric segmentation of vertebral bodies and intervertebral discs on fat-water MR images," *IEEE J Biomed Health Inform*, pp. 1–10, 2018.

[12] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans Med Imaging*, vol. 23, no. 8, pp. 995–1005, 2004.

[13] H. Kasai and B. Mishra, "Riemannian joint dimensionality reduction and dictionary learning on symmetric positive definite manifolds," in *Proc Eur Signal Process Conf*, 2018, pp. 2010–2014.