Unsupervised Label Learning on Manifolds by Spatially Regularized Geometric Assignment

Artjom Zern[1], Matthias Zisler[1], Freddie Åström[1],
Stefania Petra[2], and Christoph Schnörr[1]

[1] Image and Pattern Analysis Group, Heidelberg University, Germany
[2] Mathematical Imaging Group, Heidelberg University, Germany

Abstract. Manifold models of image features abound in computer vision. We present a novel approach that combines unsupervised computation of representative manifold-valued features, called labels, and the spatially regularized assignment of these labels to given manifold-valued data. Both processes evolve dynamically through two Riemannian gradient flows that are coupled. The representation of labels and assignment variables are kept separate, to enable the flexible application to various manifold data models. As a case study, we apply our approach to the unsupervised learning of covariance descriptors on the positive definite matrix manifold, through spatially regularized geometric assignment.

1 Introduction

Manifold-based methods define an active research area in computer vision [19]. Covariance descriptors, in particular, play a prominent role [5]. Covariance descriptors are typically applied to the detection and classification of entire images (e.g. faces, texture) or videos (e.g. action recognition). An important task in this context is to compute a *codebook of covariance descriptors* that can be used solving a task at hand by nearest-neighbor search [6].

The recent work [10] defines a *geometric state-of-the-art method* for computing such codebooks. Embedding descriptors into a Hilbert space (see (4) below) enables to approximate given data by kernel expansion [11] and to determine a sparse subset by ℓ_1-regularization of the expansion coefficients. This method works *entirely in feature space* and ignores the *spatial* structure of codebook assignments to data. Figure 1 illustrates that when covariance descriptors are used as 'labels' for representing *local* image structure, rather than encoding *global* second-order statistics of entire images or videos, then the *spatial structure* of label assignments should also drive the evolution of labels in feature space for unsupervised label learning.

The classical approach for the unsupervised learning of feature prototypes ('labels') is the mean-shift iteration [8, 7], which iteratively seeks modes (local peaks) of the feature density distribution through the averaging of features within local neighborhoods. This has been generalized to *manifold-valued* features by [17] by replacing ordinary mean-shifts by Riemannian (Fréchet,

Karcher) means [13]. The common way to take into account the *spatial structure* of label assignments is to *augment* the feature space by *spatial coordinates*, e.g. turn a color feature (r, g, b) into the feature vector (x, y, r, g, b). This *merge* of feature space and spatial domain has a conceptual drawback, however: The *same* color vector $(\overline{r}, \overline{g}, \overline{b})$ observed at two *different* locations $(x_1, y_1, \overline{r}, \overline{g}, \overline{b})$, $(x_2, y_2, \overline{r}, \overline{g}, \overline{b})$ defines two *different* feature vectors. Furthermore, clustering spatial coordinates into *centroids* by mean-shifts (together with the features) *differs* from *unbiased spatial regularization* as performed by variational approaches or graphical models, that do *not* depend on the location of centroids and the corresponding shape of local density modes. This work, therefore, studies the problem of representing a given manifold-valued input image with few prototypes, which are learned in an unsupervised way, while performing unbiased spatial regularization in the image domain.

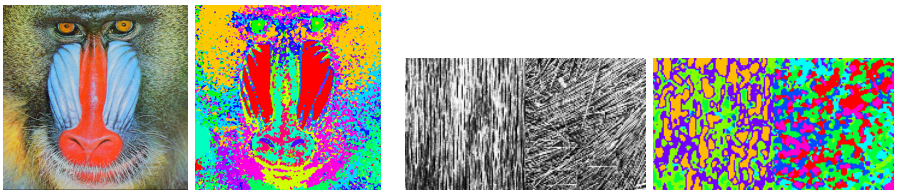

Fig. 1. Local assignments of covariance descriptors $\{G_j\}_{j\in J} \subset \mathcal{P}_d$ from a codebook J to data $\{F_i\}_{i\in I}$ are *noisy*. Different colors represent different codebook entry assignments but do not have any other specific meaning. By definition, *unsupervised* learning rules out the possibility of feature parameters tuning. Rather, covariance descriptors should *evolve* along a flow $G(t)$ driven by *spatially regularized* assignments $W(t)$ that are *not* biased towards spatial centroids and still enable *global* communication on the feature manifold $\mathcal{P}_d$. This is accomplished by our approach (1).

Contribution. We introduce a novel approach with the following properties:

(i) The approach incorporates and performs *unsupervised learning of manifold-valued features*, henceforth called *labels*. We work with covariance descriptors as a case study. But the approach applies to any feature manifold [17] for which the corresponding Riemannian feature means are well-defined and computationally feasible.

(ii) The evolution of labels (unsupervised learning) is driven by the *spatial regularization of assignments* that is *not* biased towards spatial centroids. This is accomplished by applying the *geometric* approach to image labeling by assignment, recently introduced by [2].

(iii) The smooth settings of both (i), (ii) enables to define a *smooth coupled flow*

$$(\dot{G}, \dot{W}) = \mathcal{V}(G, W) \tag{1}$$

where the evolution of labels G and the evolution of spatially regularized assignments (of labels to data) W *interact*. This interaction keeps

both domains (i) and (ii) *separate*, which enables to use alternative feature manifolds with the *same* regularized assignment mechanism.

Organization. Section 2 sketches basic material required to understand the approaches (i) and (ii), which are described in Section 3. Our approach is presented in Section 4. The concrete iterative scheme corresponding to (1) and geometric numerical integration is given by system of equations (30). We present and discuss experimental results in Section 5.

2 Preliminaries

To make this paper self-contained, we briefly sketch three methods that are relevant to our approach: (1) Geometry of the domain of covariance descriptors, S-divergence and Hilbert space embedding; (2) Soft-k-means clustering in Euclidean spaces that will be generalized to *geometric* soft-k-means on manifolds in Section 4.1; (3) Metric clustering with performance guarantee and linear complexity for label initialization with non-sparse codebooks.

Geometry of Covariance Descriptors, S-Divergence and Hilbert Space Embedding. The open cone

$$\mathcal{P}_d = \{S \in \mathbb{R}^{d\times d} \colon S = S^\top,\ S \succ 0\} \tag{2}$$

of symmetric positive definite matrices endowed with the Riemannian *metric* $\langle S_1, S_2\rangle_S = \langle S^{-1}S_1S^{-1}, S_2\rangle = \operatorname{tr}(S^{-1}S_1S^{-1}S_2)$ forms a Riemannian manifold [3]. Since evaluating the Riemannian *distance* involves a numerically expensive generalized eigenvalue problem, *divergence* functions are used instead as a compromise between respecting the geometry of (2) and computational efficiency (cf. e.g. [6]). We focus on the *symmetric S-(matrix-)divergence* [16]

$$D_S(F,G) = \log\det\Big(\frac{F+G}{2}\Big) - \frac{1}{2}\log\det(FG), \qquad F,G \in \mathcal{P}_d \tag{3}$$

that emerges as special case of a parametric family of matrix divergence functions [4] and compares favorably to the more common log-Euclidean divergence [1]. Moreover, the S-divergence generates a valid kernel function

$$k_S(F,G) = \exp\big(-\beta D_S(F,G)\big) \quad \text{with} \quad \beta \in \{\tfrac{1}{2}, \tfrac{2}{2}, \ldots, \tfrac{d-1}{2}\} \cup [\tfrac{d-1}{2}, \infty) \tag{4}$$

for the embedding $\mathcal{P}_d \to \mathcal{H}$ of covariance descriptors into a reproducing kernel Hilbert space $\mathcal{H}$ [11]. This has been explored recently by [10], see also [5].

Euclidean Soft-k-Means Clustering. The content of this paragraph can be found in numerous papers and textbooks. We merely refer to the survey [18] and to the bibliography therein. It will be generalized to the manifold $\mathcal{P}_d$ (2) of covariance descriptors in Section 4.

Given data vectors $x^1, \ldots, x^{|I|}$, we consider the task of determining prototype vectors $M = \{m^1, \ldots, m^{|J|}\}$ by minimizing the k-means criterion $J(M) =$

$\sum_{i\in I} \min_{j\in J} \|x^i - m^j\|^2$. *Soft-k*-means is based on the *smoothed* objective

$$J_\varepsilon(M) = -\varepsilon \sum_{i\in I} \log\Big(\sum_{j\in J} \exp\big(-\frac{\|x^i - m^j\|^2}{\varepsilon}\big)\Big), \tag{5}$$

which results from approximating the inner minimization problem of $J(M)$ using the log-exponential function [14, p. 27]. Similar to the basic k-means algorithm, *soft-k*-means clustering solves the stationarity condition $\nabla_{m^j} J_\varepsilon(M) = 0,\ j \in J$ by fixed point iteration that iterates the update steps

$$p^i_{\varepsilon,j}(M) = \frac{\exp\big(-d^i_j(M)/\varepsilon\big)}{\sum_{l\in J} \exp\big(-d^i_l(M)/\varepsilon\big)}, \quad d^i_j(M) = \|x^i - m^j\|^2 \tag{6}$$

for every $i \in I, j \in J$ and

$$m^j = \sum_{i\in I} q^j_i(M) x^i, \qquad q^j_i(M) = \frac{p^i_{\varepsilon,j}(M)}{\sum_{k\in I} p^k_{\varepsilon,j}(M)}. \tag{7}$$

The distribution $p^i_\varepsilon(M) \in \Delta_k$ represents the *soft-assignment* $p^i_{\varepsilon,j}(M)$ of each data point x^i to each prototype m^j, and the distribution $q^j(M)$ determines the convex combination of data points that determines the prototypes m^j by the *mean-shift* (7).

Greedy Clustering in Metric Spaces. We adopt a simple algorithm from [9] as a preprocessing step for data reduction, due to the following properties: It works in *any metric space* (X, d_X), it has *linear* complexity $\mathcal{O}(kN)$ with respect to the problem size N, and it comes along with a *performance guarantee*.

Given data points $X_N = \{x^1, \dots, x^N\} \subset X$, the objective is to determine a k-subset $M = \{m^1, \dots, m^k\}$ that solves the combinatorially hard optimization problem

$$J^*_\infty = \min_{M\subset X_N, |M|=k} \max_{x\in X_N} d_X(x, M). \tag{8}$$

Starting from a first initial point m^1, e.g. chosen randomly, selecting the remaining $k-1$ points $m^2, \dots, m^k$ by greedy iteration yields a set M that is a 2-approximation $J_\infty(M) \le 2J^*_\infty$ of the optimum (8) [9, Thm. 4.3]. As a consequence, the subset of k points of M are uniformly distributed in X_N according to the metric d_X. Figure 2 provides an illustration.

3 Label Learning, Label Assignment

We sketch the recent work of [10] and [2] which motivated our approach, that is presented in Section 4.

Sparse Coding of Covariance Descriptors [10]. Given observations $\{F_i\}_{i\in I} \subset \mathcal{P}_d$ and the embedding $\phi\colon \mathcal{P}_d \to \mathcal{H}$ into a Hilbert space induced by the S-divergence (3) and the kernel function (4), the objective function for learning a

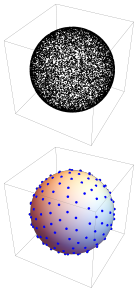

Fig. 2. Approximation of the metric clustering objective (8). LEFT: 10.000 points on the sphere regarded as manifold equipped with the cosine distance. RIGHT: 200 prototypes determined with linear runtime complexity by metric clustering are *almost uniformly located* in the data set, which qualifies them for unbiased initialization of more complex nonlinear prototype evolutions (Section 4). This works in any metric space and is applied in this paper to covariance descriptors on the positive definite matrix manifold (2), to determine non-sparse codebooks as initialization $G(t)|_{t=0}$ of the flow (1).

sparse codebook $G = \{G_j\}_{j\in J} \subset \mathcal{P}_d$ of covariance descriptors reads

$$J(G,y) = \sum_{i\in I} l_\phi(y, F_i, G), \quad l_\phi(y, F_i, G) = \Big\|\phi(F_i) - \sum_{j\in J} y_j \phi(G_j)\Big\|^2 + \alpha\|y\|_1, \tag{9}$$

where sparsity is enforced through ℓ_1-penalization of the coefficients y. The approach iterates i) a *sparse coding step* solving $y_i = \operatorname{argmin}_y l_\phi(y, F_i, G)$, $i \in [m]$ while keeping G fixed, and ii) a *dictionary learning step* evaluating the optimality condition $\nabla_{G_j} J(G) = 0$, $\forall j$. In the particular case of the S-divergence (3), (4), this condition takes the form of an *algebraic Riccati equation* which can be solved numerically with the fixed-point iteration

$$G_j - G_j \mathcal{G}_j(G) G_j = 0, \qquad G_j^{(k+1)} = \big(\mathcal{G}_j(G^{(k)})\big)^{-1}, \quad j \in J. \tag{10}$$

The map $\mathcal{G}_j(G)$ is given by

$$\mathcal{G}_j(G) = \frac{\sum_i y_{ij}\Big(k_S(F_i, G_j)\big(\frac{F_i+G_j}{2}\big)^{-1} - \sum_r y_{ir} k_S(G_j, G_r)\big(\frac{G_j+G_r}{2}\big)^{-1}\Big)}{\sum_i y_{ij}\Big(k_S(F_i, G_j) - \sum_r y_{ir} k_S(G_j, G_r)\Big)}. \tag{11}$$

As discussed in Section 1, we point out again that this approach *entirely* works on the feature manifold $\mathcal{P}_d$ with auxiliary variables y, *independent* of the spatial image structure corresponding to the data $\{F_i\}_{i\in I}$.

Regularized Image Labeling on the Assignment Manifold [2]. Given data $\{F_i\}_{i\in I}$ and labels $\{G_j\}_{j\in J}$, assignments $F_i \leftrightarrow G_j$ are represented by the components $W_{i,j}$ of *assignment vectors* $W_i \in \mathcal{S}$, $i \in I$, where $\mathcal{S}$ denotes the open probability simplex equipped with the Fisher-Rao metric. Gathering all vectors into the assignment matrix $W \subset \mathcal{W} \in \mathbb{R}^{|I|\times|J|}$ on the product manifold $\mathcal{W} := \prod_{i\in I} \mathcal{S}$, spatially regularized assignments W are determined as follows.

A distance matrix

$$D \in \mathbb{R}^{|I|\times|J|}, \quad D_i = \big(d(F_i, G_1), \ldots, d(F_i, G_{|J|})\big)^\top, \qquad i \in I \tag{12}$$

with row vectors D_i, describes the similarity of labels G_j and data F_i based on any distance function $d(\cdot,\cdot)$. Using the mapping (with componentwise multiplication of stricly positive vectors in the numerator)

$$L_p(u) = \frac{p e^u}{\langle p, e^u\rangle}, \qquad p \in \mathcal{S}, \tag{13}$$

which serves as a first-order approximation of the exponential mapping induced by the Fisher-Rao geometry [2, Prop. 3], the distance matrix D is turned into the likelihood matrix

$$L(W) \in \mathbb{R}^{|I|\times|J|}, \quad L_i(W) = L_{W_i}(-D_i/\rho) = \frac{W_i e^{-D_i/\rho}}{\langle W_i, e^{-D_i/\rho}\rangle},\ \rho > 0,\ i \in I. \quad (14)$$

These local assignments are spatially regularized through geometric averaging, resulting in the similarity matrix

$$S(W) \in \mathbb{R}^{|I|\times|J|}, \qquad S_i = S_i(W) = \frac{\mathrm{mean}_g\{L_j(W)\}_{j\in\mathcal{N}(i)}}{\big\langle \mathbb{1}, \mathrm{mean}_g\{L_j(W)\}_{j\in\mathcal{N}(i)}\big\rangle}, \quad i \in I \quad (15)$$

with $\mathrm{mean}_g\{L_j(W)\}_{j\in\mathcal{N}(i)} = \big(\prod_{j\in\mathcal{N}(i)} L_j(W)\big)^{\frac{1}{|\mathcal{N}(i)|}}$ and spatial neighborhoods $\mathcal{N}(i)$ around each pixel i. Finally, W is determined by maximizing the objective function $J : \mathcal{W} \to \mathbb{R}$, $J(W) := \langle W, S(W)\rangle$. This leads to the Riemannian gradient ascent flow

$$\dot{W}(t) = \nabla_{\mathcal{W}} J(W(t)), \qquad W(0) = \frac{1}{|J|}\mathbb{1}_{|I|}\mathbb{1}_{|J|}^{\top} := C \quad (16)$$

initialized at the barycenter C of the assignment manifold corresponding to uniform unbiased assignments. Using the approximation discussed in [2],

$$\nabla J(W(t)) \approx S(W), \quad (17)$$

where ∇ denotes the Euclidean gradient, the Riemannian gradient flow (16) explicitly reads for each vector

$$\dot{W}_i = W_i\big(S_i(W) - \langle W_i, S_i(W)\rangle \mathbb{1}\big), \qquad i \in I. \quad (18)$$

Numerical integration of (18) in order to solve for $W(t)$ can be conveniently done on the tangent space $T^I = T \times \cdots \times T := T_C\mathcal{W}$, $T := T_{\frac{1}{|J|}\mathbb{1}}\mathcal{S}$, using the framework suggested by [15]. The pullback of the flow (16) using the map L_C (14) evaluated at the barycenter C takes the form

$$\dot{V}_i(t) = \Pi_T \nabla_i J(W(t)), \qquad V_i(0) = 0, \qquad W_i(t) = L_{C_i}(V_i(t)), \qquad i \in I, \quad (19)$$

where Π_T denotes the orthogonal projection onto the tangent space T. Discretization with the simplest numerical integration method, i.e. explicit Euler-steps with stepsize $h > 0$, and taking into account approximation (17), yields the iterative scheme

$$V_i^{(k+1)} = V_i^{(k)} + h\Pi_T S_i(W^{(k)}), \quad V_i^{(0)} = 0, \quad h > 0, \quad (20a)$$

$$W_i^{(k+1)} = L_{C_i}(V_i^{(k+1)}), \quad i \in I. \quad (20b)$$

This update step for supervised smooth geometric image labeling has high potential for parallel implementations and corresponding speed-ups.

Moreover, it has been shown recently [12] how the approach can be used in order to evaluate any given discrete graphical model for image labeling.

4 Label Learning by Assignment

In this section, we detail our approach (1) in two steps concerning the G- and $W-$component, respectively.

(1) *G-component:* We extend the basic soft-k-means clustering approach (Section 2, eqns. (6), (7)) to an arbitrary feature manifold. In the particular case of the S-divergence, it turns out that the resulting fixed point iteration takes the form of an algebraic Riccati equation, as does the approach [10] – cf. (10), with different mappings $\mathcal{G}_j$, of course.
(2) *W-component:* Since the assignment variables due to (1) turn out to be *probabilities*, we can replace them by the assignment variables of the flow (16) and thus seemlessly enforce *spatial regularization* for manifold-valued soft-k-means clustering of features. Conversely, the resulting evolution of labels $\{G_j\}_{j\in J}$ affects the assignment flow through the distance vectors (12). This defines the W-component of our approach.

We point out once more that covariance descriptors are used here as a case study. They can be replaced or augmented by any other manifold-valued features for which the corresponding operations are mathematically well-defined and computationally feasible.

4.1 Manifold-Valued Soft-k-Means Clustering

Given data and labels $\{F_i\}_{i\in I}, \{G_j\}_{j\in J} \subset \mathcal{P}_d$ as covariance descriptors, we adopt the S-divergence (3) and rewrite the soft-k-means objective (5) in the form

$$J(G) := J(G_1,\ldots,G_{|J|}) = -\varepsilon \sum_{i\in I} \log\Big(\sum_{j\in J} \exp\big(-\frac{D_S(F_i,G_j)}{\varepsilon}\big)\Big), \quad \varepsilon > 0. \tag{21}$$

The Riemannian metric of the positive definite manifold $\mathcal{P}_d$,

$$g_A(U,V) = \operatorname{tr}(A^{-1}UA^{-1}V), \qquad A \in \mathcal{P}_d, \qquad U,V \in S_{\mathrm{sym}}(d,\mathbb{R}) \tag{22}$$

is also induced by the S-divergence D_S [4, Prop. 3.8]. We regard the argument $G = (G_1,\ldots,G_n)$ of J as points on the product manifold $\prod_{j\in J}\mathcal{P}_d$. The j-th component of the Riemannian gradient of J is a symmetric matrix $(\operatorname{grad} J)_j$ satisfying

$$g_{G_j}\big((\operatorname{grad} J)_j, V\big) = d_j J(V), \qquad \forall\, V \in S_{\mathrm{sym}}(d,\mathbb{R}), \tag{23}$$

where $d_j J(V)$ denotes the differential of J with respect to G_j applied to a tangent matrix V. Thus, if $d_j J(V) = \operatorname{tr}(UV)$ for some symmetric matrix U, then $(\operatorname{grad} J)_j(G) = G_j U G_j$. We have

$$d_j J(G)(V) = \sum_{i\in I} \underbrace{\frac{\exp\big(-\frac{D_S(F_i,G_j)}{\varepsilon}\big)}{\sum_{l\in J}\exp\big(-\frac{D_S(F_i,G_l)}{\varepsilon}\big)}}_{p_{ij}(G)} d_j D_S(F_i,G_j)(V) \tag{24}$$

where $p_{ij}(G) \in \Delta_n$ are the assignment probabilities of datum F_i to each prototype G_j. Evaluating the differential $d_j D_S(F_i, G_j)$, we have $d_j D_S(F_i, G_j)(V) = \operatorname{tr}\big(((F_i + G_j)^{-1} - \frac{1}{2}G_j^{-1})V\big)$ and thus obtain from (24)

$$(\operatorname{grad} J)_j(G) = \sum_{i \in I} p_{ij}(G) G_j (F_i + G_j)^{-1} G_j - \frac{P_j(G)}{2} G_j, \quad j \in J \tag{25}$$

with $P_j(G) = \sum_{i \in I} p_{ij}(G)$. Setting the gradient to zero and rearranging yields

$$G_j - \sum_{i \in I} q_{ij}(G) G_j \Big(\frac{F_i + G_j}{2}\Big)^{-1} G_j = 0 \quad q_{ij}(G) := \frac{p_{ij}(G)}{P_j}, \quad j \in J \tag{26}$$

where the variable symbold p and q highlight the extension of their counterparts in the Euclidean case (6). Moreover, rewriting the preceding optimality condition in the form

$$G_j - G_j \mathcal{G}_j(G) G_j = 0, \qquad \mathcal{G}_j(G) = \sum_{i \in I} q_{ij}(G) \Big(\frac{F_i + G_j}{2}\Big)^{-1}, \quad j \in J \tag{27}$$

reveals a structure analogous to condition (10) resulting from the approach [10], with different mappings $\mathcal{G}_j$, however. The major difference is that (10) includes the embedding $\mathcal{P}_d \to \mathcal{H}$ of covariance descriptors into the Hilbert space generate by the kernel (4), whereas (27) is directly defined on the feature manifold $\mathcal{P}_d$.

Exploiting this analogy, we adopt the fixed point iteration (10),

$$G_j^{(k+1)} = \big(\mathcal{G}_j(G^{(k)})\big)^{-1}, \quad j \in J \tag{28}$$

with $\mathcal{G}_j$ given by (27). Since $(q_{ij})_{i \in I}$ is a probability vector, it is immediate that $\mathcal{G}_j \colon \mathcal{P}_d \to \mathcal{P}_d$ maps the feature manifold onto itself and that eq. (28) is well-posed.

4.2 Joint Label Learning and Label Assignment

We modify in this section the schemes (28) and (20) so as to obtain an interaction $\{G_j\} \leftrightarrow \{W_i\}$ of label evolution $G(t)$ and label assignments $W(t)$, in both directions. And we explain why these modifications make sense.

$\{G_j\} \to \{W_i\}$: Changing labels $G(t)$ change the distance matrix (12) to $D(t) = D\big(G(t)\big)$, which affects the matrices (14) and (15) and in turn the assignment flow (16). Since the distance function $d(\cdot, \cdot)$ defining D by (12) is the S-divergence $D_S(\cdot, \cdot)$ (3) in this paper, and since labels G_j satisfying (27) minimize $D_S(F_i, G_j)$ by minimizing the objective (21), they also maximize the likelihood vectors (14) which are *spatially regularized* to form the similarity matrix (15). The similarity matrix $S(W)$ drives the assignment flow through (20) so as to maximize the correlation $\langle W, S(W) \rangle$ between pixelwise label assignments W_i and similarity vectors $S_i(W)$, which represent assignments in the spatial context through the *non-local* geometric diffusion process (15).

$\{W_i\} \to \{G_j\}$: The *row* vectors W_i can be interpreted as posterior probabilities $W_{ij} = \Pr(G_j|F_i)$ of assigning prototypes $G_j, j \in J$ to data $F_i, i \in I$. Hence these row vectors are of primary importance for the assignment flow, as just discussed. Conversely, the *column* vectors $W^j,\ j \in J$ with components $(W^j)_i = W_{ij}$ represent *weights* that associate each data point F_i with a label G_j. Normalization $\frac{W_{ij}}{\langle \mathbb{1}, W^j \rangle}$ turns these weights into probability vectors that exactly show up as probabilities $q_{ij}(G)$ in the optimality condition (27), which defines the evolution of labels G_j. Consequently, in order to affect label evolution by *spatially regularized* assignments, we exchange these probabilities so that the mappings $\mathcal{G}_j,\ j \in J$ defining the optimality condition (27) now read

$$\mathcal{G}_j(W) = \sum_{i \in I} q_{ij}(W) \Big(\frac{F_i + G_j}{2} \Big)^{-1}, \quad q_{ij}(W) = \frac{W_{ij}}{\langle \mathbb{1}, W^j \rangle} \in \mathcal{S}_{|I|},\ j \in J. \tag{29}$$

4.3 Summary and Discussion

Summing up, the joint flow (1) of labels $\{G_j\}_{j \in J}$ and assignments $\{W_i\}_{i \in I}$ is implemented, with $h > 0$ by the discrete iterative scheme

$$V_i^{(k+1)}(G) = V_i^{(k)}(G) + h \Pi_T S_i\big(W^{(k)}(G)\big), \quad G = G^{(k)}, V_i^{(0)} = 0 \tag{30a}$$

$$W_i^{(k+1)}(G) = L_{C_i}\big(V_i^{(k+1)}(G)\big), \quad i \in I \tag{30b}$$

$$G_j^{(k+1)}(W) = \big(\mathcal{G}_j(W)\big)^{-1}, \quad W = W^{(k+1)}, \quad j \in J \tag{30c}$$

$$\mathcal{G}_j(W) = \sum_{i \in I} q_{ij}(W) \Big(\frac{F_i + G_j}{2} \Big)^{-1}, \quad q_{ij}(W) = \frac{W_{ij}}{\langle \mathbb{1}, W^j \rangle}. \tag{30d}$$

We adopt the **termination criterion** from [2]: The iteration (30) stops when the average entropy of the assignment variables drops below 10^{-3}, which signals an (almost) unique label assignment and hence also stationarity of the label evolution.

While the dependency $G = G(W)$ is explicit by (30d), the dependency $W = W(G)$ is not: it is given through the distance matrix (12) which defines the vectors S_i of (30a) through (14) and (15). As a consequence, the computation of these matrices, though not expensive, has to be repeated at every step of the iteration (30).

To conclude this section, we point out and discuss few **further aspects** that characterize our approach **and differences to established work**.

(a) Our approach is affected by *specific* properties of the feature manifold only through the divergence function D_S of (21), and through the structure of the resulting optimality condition (27). The divergence function should induce the Riemannian metric, like (22) in our present case study using covariance descriptors. The optimality condition should admit a numerically convenient iterative scheme, like the fixed point iteration (28). If these properties are satisfied for some feature manifold, our approach can be applied.

(b) The approach [10] works entirely on the feature manifold, as discussed in Section 3, whereas our approach additionally takes into account the spatial

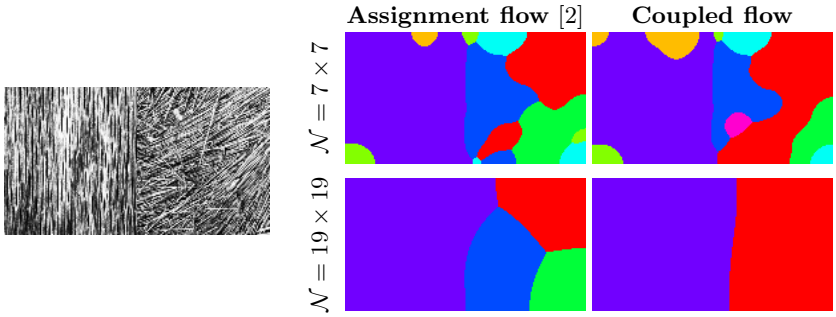

Fig. 3. LEFT: Image combining homogeneous (left) and heterogeneous texture (right). RIGHT: Supervised flow (fixed labels) vs. unsupervised flow (with label evolution) together with weak and strong spatial regularization (bottom vs. top). Our approach (Coupled flow) moves few labels in a proper position on the feature manifold and suppresses redundant ones, through the *sparsifying effect* of spatial regularization and adaptation of labels done *simultaneously*.

structure of assignments through regularization. Mean-shift approaches, on the other hand, merge both representations by augmenting feature vectors with spatial coordinates. As a consequence, spatial regularization through averaging is based on the corresponding centroids (spatial coordinates of prototypes). Our approach keeps both representations separate, and spatial regularization through the geometric diffusion process (17) does not involve any centroids.

(c) A natural idea is to replace manifold-valued soft-k-means clustering, i.e. the G-component of our approach, by the approach [10]. This is not directly practicable because the sparse coefficients y of the kernel expansion (9) are *signed*, whereas our approach works with assignment *probabilities*. We leave this problem for future work.

5 Experiments

In this section we show numerical results to illustrate the impact of geometric spatial regularization on unsupervised label evolution on the feature manifold, the empirical convergence rate and the influence of parameter values.

We compare to two state-of-the-art methods: *Assignment Flow* [2] for supervised labeling on the manifold (fixed labels), *Harandi et al.* [10] for unsupervised label learning (no spatial regularization), and the *Local Method*, which is nearest-neighbor labeling based on given initial labels.

Experiment set-up. We did not undertake any 'feature-engineering' but constructed only a basic set of covariance feature descriptors consisting of intensities, first and second order partial derivatives: $g = \big(\partial^{0,0}, \partial^{1,0}, \partial^{0,1}, \partial^{2,0}, \partial^{1,1}, \partial^{0,2}\big)$ which, in the case of color images, results in descriptors of size 18×18. For all methods we used the S-divergence as distance. Initial labels were computed by

metric clustering to limit the otherwise infeasible large label space when using the complete data set.

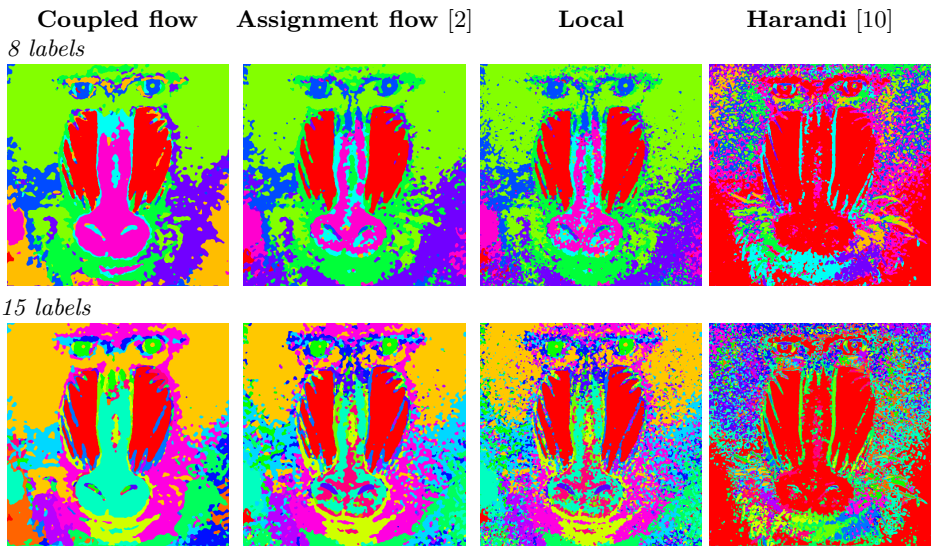

Fig. 4. Impact of spatial regularization (Assignment flow, Coupled flow) and unsupervised label evolution (Harandi, Coupled flow). The Assignment flow consistently suppresses fine details whereas Harandi's approach captures all small variations in the data including noise. The Coupled flow provides a good compromise between suppressing spurious labelings and obtaining a compact, spatially coherent representation through label evolution that preserves visual features (eye region, tip of the nose).

Fixed labels vs. evolving labels, with and without spatial regularization. Figure 3 shows an image with two textures. The texture on the left is more homogeneous than the texture on the right. We compare the *supervised* assignment flow (fixed labels) and our *unsupervised* approach (with label evolution) performing both weak and stronger spatial regularization (top vs. bottom row). Starting from an overcomplete dictionary with 8 initial labels, the supervised approach easily clusters the homogeneous left texture whereas partitions emerge on the right that reflect the heterogeneous texture structure. Our unsupervised approach manages this task with fewer labels. This demonstrates i) the effect of label evolution and ii) the sparsifying effect of spatial regularization *done simultaneously*, in order to move few labels in a proper position on the feature manifold and to get rid of the remaining ones.

Figure 4 shows a larger problem instance including the outcome of the Local method and Harandi's unsupervised clustering. We point out that, since label evolutions differ between the approaches, colors merely index label assignments but cannot be compared. Rather, the relative frequencies of label assignments are informative and enable to compare the methods.

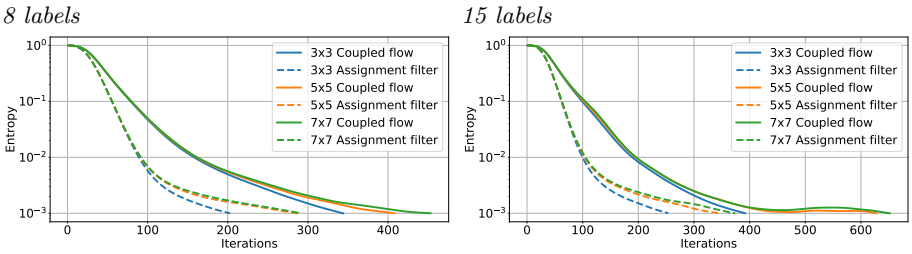

Fig. 5. Entropy of label assignments depending on the iteration number for the Assignment flow and the Coupled flow, for different strengths of spatial regularization. We observe i) that both labeling processes converge rapidly until the termination criterion (entropy $\leq 10^{-3}$) is met, ii) stronger spatial regularization requires more iterations to resolve label assignment conflicts, and iii) the Coupled flow needs more iterations than the supervised Assigment flow, due to the interaction between label evolution and regularized label assignment.

We performed two experiments with 8 and 15 initial labels (top vs. bottom row) and fixed strength of spatial regularization. Both the local method and especially Harandi's method are susceptible to small details and noise. The noise suppressing effect of spatial regularization is clearly visible for the supervised assignment flow. Our approach (Coupled flow) additionally combines regularization with label evolution and yields a more coherent result without suppressing details, e.g. at the eyes and the tip of the nose.

Empirical convergence rate. Figure 5 displays the entropy measure of label assignments used as termination criterion for the Assignment flow and the Coupled flow, for different strenghts of spatial regularization. Due to the interaction between label evolution and regularized label assignments, the Coupled flow needs more iterations to converge. The total number of iterations is quite small, however, and the approach has high potential for parallel implementation on modern hardware.

Euclidean color clustering (special case: spatially regularized mean-shift; Figure 6). We demonstrate the flexible applicability of our approach: replacing the positive definite manifold (covariance descriptors) by the Euclidean RGB-space results in spatially regularized mean-shift clustering as a special case. The S-divergence and geometric averaging were replaced by the squared Euclidean distance and arithmetic averaging. We used 200 initial labels which suffice to represent the image structure (compare Input vs. Local). Comparing the Assignment flow and the Coupled flow demonstrates the sparsifying effect of combining label evolution with spatial regularization. In particular, regions around the eyes and the nose are encoded by a smaller number of prototypes.

Real Images and Texture. We applied the assignment of covariance descriptors to representative urban scenes, that comprise both low-rank image structure (edges of walls, windows, etc.) and texture (trees, roofs, etc.).

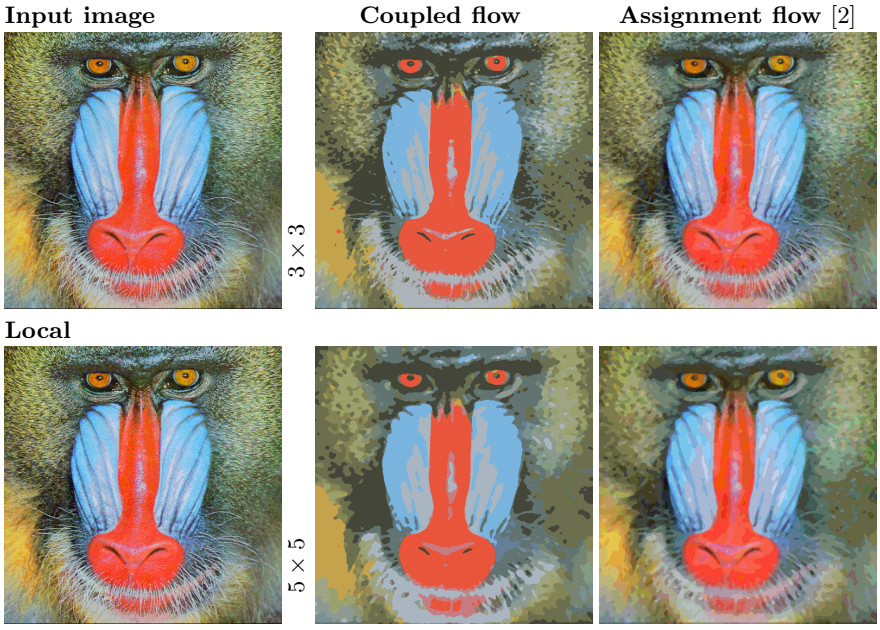

Fig. 6. Replacing the covariance descriptor manifold $\mathcal{P}_d$ by the $\mathbb{R}^3$ color space results in *spatially regularized mean-shift* as a special case. Comparing the Assignment flow (fixed labels) with the Coupled flow (evolving labels), for 200 initial labels and different strenghts of spatial regularization (top vs. bottom), demonstrates the sparsifying effect of the *coupled* process, resulting in a more compact representation of the image. In particular, regions around the eyes and the nose are encoded by a smaller number of prototypes.

Figure 7 shows the results for 15 initial labels whereas **Figure 8** shows the results for 100 initial labels. In both scenarios, the unsupervised method of Harandi consistently returns non-compact oversegmentations, despite performing label evolution. This is significantly different for the Coupled flow, where label evolution *and* spatial regularization adapt labels on the feature manifold and thus enable more compact representions of textured regions corresponding to the tree and the roof in Fig. 8.

6 Conclusion

We introduced a novel approach in terms of two coupled Riemannian gradient flows that perform simultaneously label evolution on a feature manifold and spatially regularized label assignment. This unsupervised method returns compact representations of local image structure in high-dimensional feature spaces that are statistically significant and hence useful for subsequent image interpretation. The modular design enables flexible applications to various feature manifolds. The algorithm has high potential for fine-grained parallel implementation.

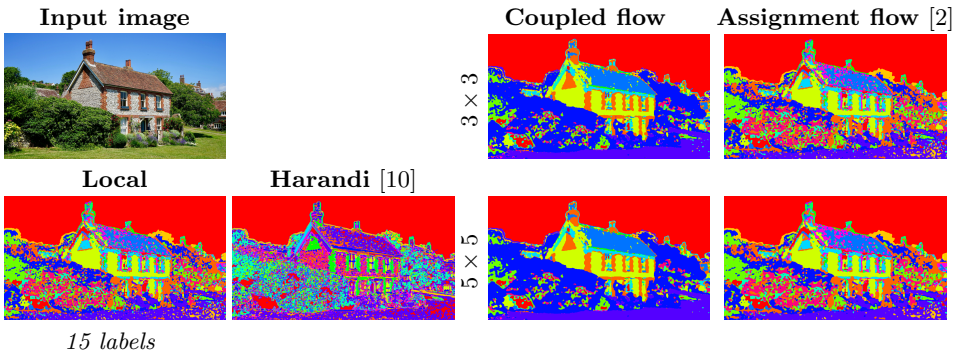

Fig. 7. The Coupled flow yields more homogeneous regions, which can be observed in particular in regions corresponding to the roof, windows and grass. Without prototype adaptation the Assignment flow returns large disconnected regions that are less useful for subsequent image interpretation. The global clustering method of Harandi produces visually pleasing results, but fails to return spatially coherent regions, especially in textured regions of the scene.

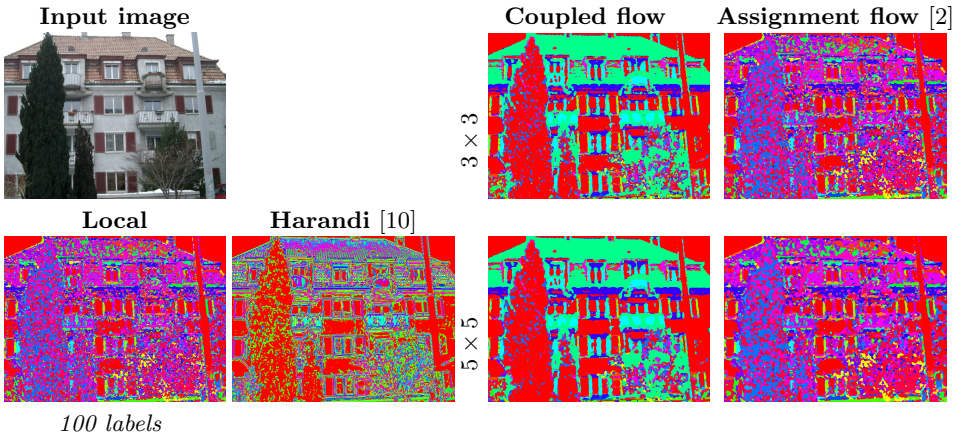

Fig. 8. Same set-up as Fig. 7 using 100 initial labels. The observations stated in the caption of Figure 7 hold here as well. We only point out that with sufficiently strong spatial regularization (bottom row), the Coupled flow managed to adapt labels on the feature manifold, and thus represents more compactly textured regions (tree, roof) in a spatially coherent way.

Acknowledgements. This work was supported by the German Research Foundation (DFG), grant GRK 1653.

References

1. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices. SIAM J. Matrix Anal. Appl. **29**(1), 328–347 (2006)
2. Åström, F., Petra, S., Schmitzer, B., Schnörr, C.: Image Labeling by Assignment. Journal of Mathematical Imaging and Vision **58**(2), 211–238 (2017)
3. Bhatia, R.: Positive Definite Matrices. Princeton Univ. Press (2006)
4. Chebbi, Z., Moakher, M.: Means of Hermitian Positive-Definite Matrices Based on the Log-Determinant α-Divergence Function. Linear Algebra and its Applications **436**(7), 1872–1889 (2012)
5. Cherian, A., Sra, S.: Positive Definite Matrices: Data Representation and Applications to Computer Vision. In: Minh, H., Murino, V. (eds.) Algorithmic Advances in Riemannian Geometry and Applications, pp. 93–114. Springer (2016)
6. Cherian, A., Sra, S., Banerjee, A., Papanikolopoulos, N.: Jensen-Bregman LogDet Divergence with Application to Efficient Similarity Search for Covariance Matrices. IEEE PAMI **35**(9), 2161–2174 (2013)
7. Comaniciu, D., Meer, P.: Mean Shift: a Robust Approach Toward Feature Space Analysis. IEEE Trans. Patt. Anal. Mach. Intell. **24**(5), 603–619 (2002)
8. Fukunaga, K., Hostetler, L.: The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. IEEE Trans. Inform. Theory **21**(1), 32–40 (1975)
9. Har-Peled, S.: Geometric Approximation Algorithms. AMS (2011)
10. Harandi, M., Hartley, R., Lovell, B., Sanderson, C.: Sparse Coding on Symmetric Positive Definite Manifolds Using Bregman Divergences. IEEE Transactions on Neural Networks and Learning Systems **27**(6), 1294–1306 (2016)
11. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel Methods in Machine Learning. Ann. Statistics **36**(3), 1171–1220 (2008)
12. Hühnerbein, R., Savarino, F., Åström, F., Schnörr, C.: Image Labeling Based on Graphical Models Using Wasserstein Messages and Geometric Assignment. SIAM J. Imaging Science **11**(2), 1317–1362 (2018)
13. Karcher, H.: Riemannian Center of Mass and Mollifier Smoothing. Comm. Pure Appl. Math. **30**, 509–541 (1977)
14. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, 3rd edn. (2009)
15. Savarino, F., Hühnerbein, R., Aström, F., Recknagel, J., Schnörr, C.: Numerical Integration of Riemannian Gradient Flows for Image Labeling. In: Proc. SSVM. LNCS, vol. 10302. Springer (2017)
16. Sra, S.: Positive Definite Matrices and the Symmetric Stein Divergence. CoRR abs/1110.1773 (2013)
17. Subbarao, R., Meer, P.: Nonlinear Mean Shift over Riemannian Manifolds. Int. J. Comp. Vision **84**(1), 1–20 (2009)
18. Teboulle, M.: A Unified Continuous Optimization Framework for Center-Based Clustering Methods. J. Mach. Learning Res. **8**, 65–102 (2007)
19. Turaga, P., Srivastava, A. (eds.): Riemannian Computing in Computer Vision. Springer (2016)