A Novel Earth Mover's Distance Methodology for Image Matching with Gaussian Mixture Models

Peihua Li[1], Qilong Wang[2], Lei Zhang[3]
[1]Dalian University of Technology, [2]Heilongjiang University, [3]The Hong Kong Polytechnic University
peihuali@dlut.edu.cn, wangqilong.415@163.com, cslzhang@comp.polyu.edu.hk

Abstract

The similarity or distance measure between Gaussian mixture models (GMMs) plays a crucial role in content-based image matching. Though the Earth Mover's Distance (EMD) has shown its advantages in matching histogram features, its potentials in matching GMMs remain unclear and are not fully explored. To address this problem, we propose a novel EMD methodology for GMM matching. We first present a sparse representation based EMD called SR-EMD by exploiting the sparse property of the underlying problem. SR-EMD is more efficient and robust than the conventional EMD. Second, we present two novel ground distances between component Gaussians based on the information geometry. The perspective from the Riemannian geometry distinguishes the proposed ground distances from the classical entropy- or divergence-based ones. Furthermore, motivated by the success of distance metric learning of vector data, we make the first attempt to learn the EMD distance metrics between GMMs by using a simple yet effective supervised pair-wise based method. It can adapt the distance metrics between GMMs to specific classification tasks. The proposed method is evaluated on both simulated data and benchmark real databases and achieves very promising performance.

1. Introduction

Image similarity (or distance) measures play a crucial role in content-based image retrieval, classification, segmentation, and tracking, etc. The histogram has been commonly used for image modeling, while the similarity measures between them have been studied for decades. These similarity functions mainly include information theoretic-based ones such as Kullback-Leibler (K-L) or Jesson-Shannon divergence, statistic-based ones such as χ^2-distance, and ℓ_p-norm based ones. The Earth Mover's Distance (EMD) [24] can compare cross bins of histograms (or signatures), which has proven its advantages over the conventional bin-to-bin based measures. Recently, many methods have been proposed to improve the efficiency and accuracy of EMD for histogram matching [16, 25, 22, 21, 30].

An alternative and widely used image modeling method is the parametric Gaussian Mixture Model (GMM) [29, 7, 11, 3]. The K-L divergence between GMMs is often adopted for distribution matching. As there is no closed-form solution, approximating K-L divergence via Monte-Carlo procedure is a natural choice but this is unfortunately computationally intensive. In [29], Vasconcelos *et al*. presented an approximation method for K-L divergence between GMMs based on Mahalanobis distance. Goldberger *et al*. [7] also focused on the approximation of K-L divergence and studied two strategies, which are based on the analytical-form solution between component Gaussians and on unscented transform, respectively. In [11] a variational approximation based on K-L divergence was presented for GMM matching. Beecks *et al*. [3] proposed the Gaussian Quadratic Form Distance (GQFD) as similarity measure, which has a closed form for GMMs of diagonal covariance matrices and is well suited for high-dimensional image descriptors.

Adopting EMD as a similarity measure between GMMs was first used in [18] for music classification and later in vision fields of texture classification [32] and visual tracking [15]. As opposed to the great advances of EMD in comparing histograms [16, 25, 22, 21], the potentials of EMD in measuring similarity between GMMs remain unclear and appear to be rarely explored. In this paper, we propose a novel EMD methodology for GMM based image matching. Fig. 1 illustrates this methodology for applications in image retrieval or classification. Our contributions are three-fold. (1) By exploiting the sparsity of the underlying problem, we develop a sparse representation-based EMD, namely SR-EMD. Compared with the conventional EMD, SR-EMD is more robust to image noise and is more efficient, particularly when the problem size is large. (2) An appropriate ground distance plays a fundamental role for the effectiveness of EMD. From the viewpoint of information geometry, by embedding the space of Gaussians into

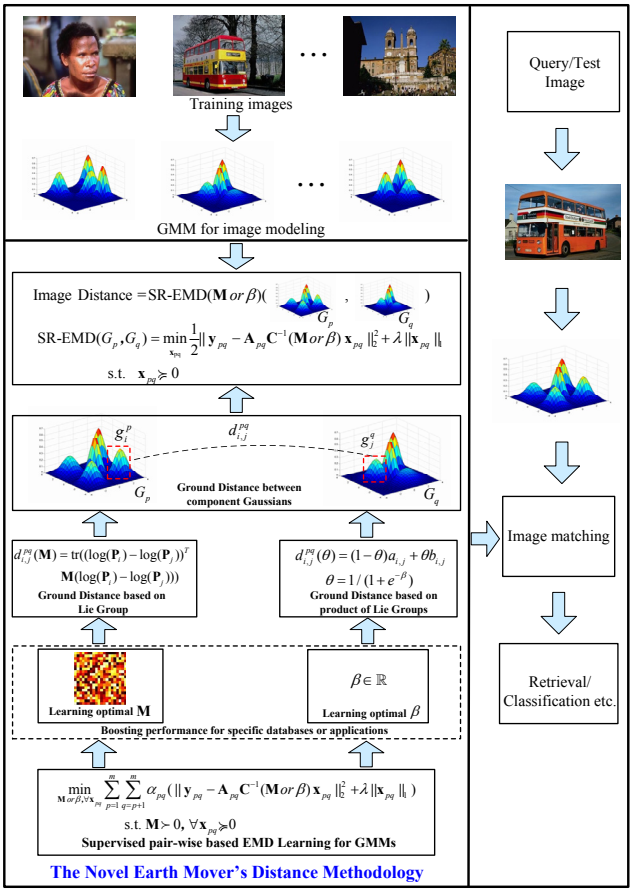


Figure 1. Overview of the proposed EMD methodology for image retrieval or classification. Every image is represented by a GMM, and image matching is accomplished by comparing GMMs of the corresponding images via sparse representation-based EMD (SR-EMD). Two novel ground distances are presented based on theory of information geometry to boost the effectiveness of EMD. We also made the first attempt to learn the distance metrics between GMMs, aiming to adapt the metrics to specific vision tasks. Please refer to Section 2 for details.

a Lie group or regarding it as the product of Lie groups, we present two novel ground distances between component Gaussians. Unlike the traditional ones, the new ground distances can characterize the intrinsic distance of the underlying Riemannian manifold of the space of Gaussians. (3) We propose a simple yet effective supervised EMD learning method for GMMs, in order to adapt the distance metrics between GMMs to specific applications. Though metric learning methods for vector data have been extensively studied, little work has been done on the metric learning for GMMs. Note that a recent paper [30] studies the EMD learning for histogram, which is however not applicable to our problem.

2. SR-EMD for GMMs and Supervised SR-EMD Learning

In this paper, we use GMM for image modeling. Given an input image, we first extract local image descriptors at dense image grids. Estimation of GMM is accomplished by the Expectation-Maximization (EM) algorithm. The number of component Gaussians that best fit the data can be estimated based on minimum description length (MDL) criterion. The estimated GMM which characterizes the probability distribution of the image descriptor $\mathbf{f}$ can be written as

$$G(\mathbf{f}) = \sum_{i=1}^{n} w_i \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where $\mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ (abbreviated as g_i below) is a component Gaussian with prior probability w_i, mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$.

2.1. Sparse Representation-based EMD (SR-EMD)

Suppose that we have two GMMs $G_p = \{(g_i^p, w_i^p)\}_{i=1,\ldots,n_p}$ and $G_q = \{(g_j^q, w_j^q)\}_{j=1,\ldots,n_q}$. Let $d_{i,j}^{pq}$ denote the ground distance between g_i^p and g_j^q. The EMD is the minimal cost moving the "goods" from GMM G_p to GMM G_q with unit transportation cost of d_{ij}^{pq}. It is modeled as a classical transportation problem, which can be written in the standard matrix form

$$\begin{aligned} \mathrm{EMD}(G_p, G_q) &= \min_{\mathbf{z}_{pq}} \mathbf{c}_{pq}^T \mathbf{z}_{pq}, \\ \text{s.t. } \mathbf{A}_{pq}\mathbf{z}_{pq} &= \mathbf{y}_{pq},\ \mathbf{z}_{pq} \succcurlyeq 0, \end{aligned} \tag{2}$$

where $\mathbf{c}_{pq}$ is an $n_p n_q$-dimensional vector which is the vectorization of ground distance matrix $\{d_{ij}^{pq}\}_{n_p \times n_q}$, $\mathbf{A}_{pq}$ is a $(n_p+n_q)\times(n_p n_q)$ matrix with 0 or 1 entries, and the weight vector $\mathbf{y}_{pq} = [w_1^p, \ldots, w_{n_p}^p, w_1^q, \ldots, w_{n_q}^q]$. Eq. (2) can be solved by the simplex algorithm or interior-point algorithm. As $\sum_i^{n_p} w_i = \sum_j^{n_q} w_j = 1$, EMD (2) is guaranteed to converge to the optimal solution $\mathbf{z}_{pq}$ in which the non-zero entry number is less than $(n_p + n_q)/(n_p n_q)$ [6, Chap. 6]. Therefore $\mathbf{z}_{pq}$ is sparse, for example, there are only less than 20% non-zero entries if $n_p = n_q = 10$.

Let us consider a more general case where the data is contaminated by noise, that is, $\mathbf{A}_{pq}\mathbf{z}_{pq} = \mathbf{y}_{pq} + \mathbf{v}_{pq}$, where $\mathbf{v}_{pq}$ is Gaussian noise of zero mean. Recalling the sparse property of (2), we let $\mathbf{C}_{pq}\mathbf{z}_{pq} = \mathbf{x}_{pq}$, where $\mathbf{C}_{pq}$ is a diagonal matrix whose diagonal entries are the elements of $\mathbf{c}_{pq}$. After some manipulations, we re-write (2) as

$$\begin{aligned} &\mathrm{SR\text{–}EMD}(G_p, G_q) = \\ &\min_{\mathbf{x}_{pq}} \frac{1}{2}\|\mathbf{y}_{pq} - \mathbf{A}_{pq}\mathbf{C}_{pq}^{-1}\mathbf{x}_{pq}\|_2^2 + \lambda\|\mathbf{x}_{pq}\|_1, \text{s.t. } \mathbf{x}_{pq} \succcurlyeq 0 \end{aligned} \tag{3}$$

where $\lambda > 0$ is a regularizing constant, $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 and ℓ_1-norm, respectively. As the ground distance may be zero, we add a very small real number, say,

1e-15, to the diagonal entries of $\mathbf{C}_{pq}$ so that every element is larger than zero. Here the distance is the sum of two terms: the first term measures the error of the system of linear constraint equations while the second one measures the transportation cost. As the EMD (2) always has an optimal solution, the value of the first term is actually very small and negligible compared with that of the second. One additional advantage of (3) is that the constraint is assimilated and the objective function is differential with respect to the parameters involved, which greatly facilities the minimization procedures to optimize these parameters (see Section 2.3).

Formulating the conventional EMD (2) as a sparse representation problem in (3) brings two benefits: robustness to noise and efficiency. It is known that, from the Bayesian viewpoint, the objective function (3) leads to the optimal solution if $\mathbf{v}_{pq}$ is Gaussian distributed and the prior distribution of $\mathbf{x}_{pq}$ is Laplacian. Hence, the sparse representation based EMD (SR-EMD) is naturally more tolerable to noise than the classical EMD. Similar interpretation is made in [4].

The LARS/Homotopy algorithms [9, Sec. 3.4] are well suited for solving SR-EMD. As the algorithm converges in about no more than $n_p + n_q$ iterations, and in each iteration the complexity is comparable to that of the least squares [9, Sec. 3.4], the computational complexity of SR-EMD is $O((n_p + n_q)^3 n_p n_q)$. In the conventional EMD, as each iteration involves operation of Gaussian elimination of the constraint system, the computational complexity is $O(n_l(n_p + n_q)^2 n_p n_q)$, where n_l denotes the number of iterations involved. Generally we have $n_l \gg (n_p + n_q)$, and n_l increases multinomially or even exponentially with $(n_p + n_q)$ [6]. Hence, the efficiency of SR-EMD is significantly higher than the conventional EMD, particularly for large size problems.

2.2. Ground Distances Between Component Gaussians

We propose two novel ground distances between component Gaussians based on information geometry [13], which conerns the study of the probability and statistics from the Riemannian geometrical point of view. By embedding the space of Gaussians into a Lie group or regarding it as a product of Lie groups, we can measure the intrinsic distance between Gaussians in the underlying Riemannian manifold. The proposed ground distances make the metric learning for GMMs possible.

Ground distance based on Lie Group The space of multivariate Gaussians is a Riemannian manifold and can be embedded into the space of SPD matrices [19]. Let $\mathcal{N}(\mathbf{0}, \mathbf{I})$ denote a k-dimensional Gaussian distribution with the mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$ (identity matrix). We denote by $|\cdot|$ the matrix determinant. It is known that if the random vector $\mathbf{x}$ follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then its affine transformation $\mathbf{Q}\mathbf{x} + \boldsymbol{\mu}$ follows $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has a decomposition $\boldsymbol{\Sigma} = \mathbf{Q}^T\mathbf{Q}, |\mathbf{Q}| > 0$, and vice versa. As such the Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be characterized by the affine transformation $(\boldsymbol{\mu}, \mathbf{Q})$. Let τ_1 be the mapping from the affine group $\mathcal{AFF}_k^+ = \{(\boldsymbol{\mu}, \mathbf{Q}) | \boldsymbol{\mu} \in \mathbb{R}^k, \mathbf{Q} \in \mathbb{R}^{k\times k}, |\mathbf{Q}| > 0\}$ to the special general linear group $\mathcal{SL}_{k+1} = \{\mathbf{A} | \mathbf{A} \in \mathbb{R}^{(k+1)\times(k+1)}, |\mathbf{A}| > 0\}$, and τ_2 be the mapping from $\mathcal{SL}_{k+1}$ to the space of SPD matrices $\mathcal{SPD}_{k+1}^+ = \{\mathbf{P} | \mathbf{P} \in \mathbb{R}^{(k+1)\times(k+1)}, \mathbf{P} = \mathbf{P}^T, |\mathbf{P}| > 0\}$, i.e.,

$$\tau_1 : \mathcal{AFF}_k^+ \mapsto \mathcal{SL}_{k+1} \qquad \tau_2 : \mathcal{SL}_{k+1} \mapsto \mathcal{SPD}_{k+1}^+$$
$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto C_{\mathbf{Q}} \begin{bmatrix} \mathbf{Q} & \boldsymbol{\mu} \\ \mathbf{0}^T & 1 \end{bmatrix}, \qquad \mathbf{S} \mapsto \mathbf{S}\mathbf{S}^T$$

where $C_{\mathbf{Q}} = |\mathbf{Q}|^{-1/(k+1)}$. Through these two mappings, a k-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be embedded into $\mathcal{SPD}_{k+1}^+$ and thus is uniquely represented by a $(k+1) \times (k+1)$ SPD matrix $\mathbf{P}$; that is,

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathbf{P} = |\boldsymbol{\Sigma}|^{-\frac{1}{k+1}} \begin{bmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix} \tag{4}$$

One may refer to [19] for detailed theory on the embedding process.

The space of SPD matrices is a Lie group that forms a Riemannian manifold, and in this paper we use the Log-Euclidean metric [2] to measure the intrinsic distance in this space. In the Log-Euclidean framework, the logarithmic multiplication $\odot$ and the scalar logarithmic multiplication $\otimes$ are defined such that $\mathcal{SPD}_n^+$ has a linear space structure. The Lie algebra $\mathcal{S}_n$ of $\mathcal{SPD}_n^+$ is a linear space with regular matrix addition $+$ and scalar matrix multiplication $\cdot$. The matrix exponential map $\exp : \mathcal{S}_n \mapsto \mathcal{SPD}_n^+$ is one to one and onto, smooth and isometry. This enables us, through matrix logarithm, to regard operations on $\mathcal{S}_n$ as being equivalent to those in $\mathcal{SPD}_n^+$. The geodesic distance between two SPD matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ is $d(\mathbf{P}_1, \mathbf{P}_2) = \|\log(\mathbf{P}_1) - \log(\mathbf{P}_2)\|_F$, where $\log$ denotes the matrix logarithm and $\|\cdot\|_F$ denotes the matrix Frobenius norm.

Let $\mathbf{P}_i$ and $\mathbf{P}_j$ be the embedding SPD matrices corresponding to two Gaussians g_i and g_j, respectively. The ground distance between them is defined as

$$d_{g_i,g_j}(\mathbf{M}) = \mathrm{tr}((\log(\mathbf{P}_i) - \log(\mathbf{P}_j))^T \mathbf{M} (\log(\mathbf{P}_i) - \log(\mathbf{P}_j))) \tag{5}$$

where $\mathbf{M}$ is an SPD matrix. If $\mathbf{M}$ is the identity matrix, (5) reduces to the geodesic distance between $\mathbf{P}_i$ and $\mathbf{P}_j$. As $d(\mathbf{P}_i, \mathbf{P}_j)$ is a metric, $d_{g_i,g_j}(\mathbf{M})$ is also a metric (matrix mahalanobis norm). Let $\mathbf{M} = \mathbf{A}^T\mathbf{A}$. By re-writing (5) as

$$d_{g_i,g_j}(\mathbf{A}) = \|\mathbf{A}(\log(\mathbf{P}_i) - \log(\mathbf{P}_j))\|_F, \tag{6}$$

we see that the ground distance (5) can be seen as a linear transformation of the matrix in the logarithmic domain. The

underlying reason that we can define the ground distance (5) or (6) is that in the Log-Euclidean framework, $\mathcal{SPD}_n^+$ is endowed with a linear space structure. This is similar to what has been commonly done for vector data in the Euclidean space [33]. It can be interpreted as that while retaining the geodesic distance, we attempt to seek a linear transformation so that the distance is more discriminative. The matrix $\mathbf{M}$ can be learned to adapt to particular applications or databases.

Ground distance based on product of Lie groups An n-dimensional Gaussian is determined by its mean vector, $\boldsymbol{\mu} \in \mathbb{R}^n$, and the covariance matrix, $\boldsymbol{\Sigma} \in \mathcal{SPD}_n^+$. It is known that $\mathbb{R}^n$ is a Lie group under vector addition and $\mathcal{SPD}_n^+$ is a Lie group under logarithmic multiplication $\odot$: $\boldsymbol{\Sigma}_1 \odot \boldsymbol{\Sigma}_2 = \exp(\log(\boldsymbol{\Sigma}_1) + \log(\boldsymbol{\Sigma}_2))$ for $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{SPD}_n^+$. Consider the product group $\mathbb{R}^n \times \mathcal{SPD}_n^+$ and define the operation $\boxdot$:

$$(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \boxdot (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 \odot \boldsymbol{\Sigma}_2) \quad (7)$$

It can be shown that this product group is also a Lie group and its Lie algebra is $\mathbb{R}^n \times \mathcal{S}_n$.

Now let us consider the distance between two Gaussians. It is noteworthy that the distance in the Lie group $\mathbb{R}^n$ between two mean vectors should be appropriately weighted by the associative covariance matrices, and it is also reasonable to balance their distance in $\mathbb{R}^n$ and that in $\mathcal{SPD}_n^+$. With the above considerations, we propose the following ground distance:

$$d_{g_i,g_j}(\theta) = (1-\theta)a_{g_i,g_j} + \theta b_{g_i,g_j}, \quad (8)$$

where

$$a_{g_i,g_j} = \big((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\big)^{1/2},$$
$$b_{g_i,g_j} = \big\| \log(\boldsymbol{\Sigma}_i) - \log(\boldsymbol{\Sigma}_j) \big\|_F.$$

In Eq. (8), $\theta \in [0,1]$ is a constant balancing the two terms, a_{g_i,g_j} measures the difference between mean vectors, which becomes the Mahalanobis distance (up to a scaling factor) if $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j$; b_{g_i,g_j} measures the Log-Euclidean distance between two covariance matrices, which is the geodesic distance in the Riemannian space. $d_{g_i,g_j}(\theta)$ is a metric satisfying all the metric axioms, which can be easily shown by noting that a_{g_i,g_j} and b_{g_i,g_j} are both metrics. For simplicity, hereafter d_{g_i,g_j}, a_{g_i,g_j} and b_{g_i,g_j} are abbreviated as d_{ij}, a_{ij} and b_{ij}, respectively.

Computational complexity of the ground distances In the proposed ground distances two component Gaussians are "decoupled" in the sense that the logarithm of the ebmedding SPD matrix (5), or the inverse and logarithm of the covariance matrix (8) of each component Gaussian can be computed offline. Hence, the computational complexity of (5) is $O((n+1)^3)$ while that of (8) is $O(n^3)$, where n denotes the feature dimension. The logarithm of an SPD matrix $\mathbf{P} \in \mathcal{SPD}_n^+$ can be computed via the eigen-decomposition algorithm [2], in which the computational complexity is approximately $O(10n^3)$ (tridigonalization followed by QR algorithm to compute the eigen-decomposition of the SPD matrix [12] and further the matrix multiplications to compute the logarithm).

2.3. Supervised SR-EMD Learning

Metric learning for vector data has been studied for years [33] and has shown a great success. The EMD learning for vector data was first studied in [30], which, however, is not applicable to GMMs. Metric learning for GMMs is very different from that for vector data. To the best of our knowledge, no work has been reported. Here we propose a simple supervised pair-wise based metric learning method for GMMs. Intuitively, we hope that training pairs of intra-class samples are similar while those of inter-class ones are dissimilar. Let $\mathcal{C} = \{G_1, \ldots, G_m\}$ be a collection of GMMs, where m denotes the sample number. Denote by $\mathcal{S} = \{(G_p, G_q) | G_p \text{ and } G_q \text{ belong to the same class}\}$ and by $\mathcal{D} = \{(G_p, G_q) | G_p \text{ and } G_q \text{ belong to different class}\}$ the sets of equivalence constraints and inequivalence constrains, respectively.

We first consider metric learning for the ground distance based on Lie group. The learning problem may be formulated as follows:

$$\min_{\mathbf{M}, \forall \mathbf{x}_{pq}} \sum\nolimits_{p=1}^{m} \sum\nolimits_{q=p+1}^{m} \alpha_{pq} \big(\|\mathbf{y}_{pq} - \mathbf{A}_{pq}\mathbf{C}_{pq}^{-1}(\mathbf{M})\mathbf{x}_{pq}\|_2^2 + \lambda \|\mathbf{x}_{pq}\|_1 \big), \quad \text{s.t. } \mathbf{M} \succ 0, \ \forall\, \mathbf{x}_{pq} \succcurlyeq 0, \quad (9)$$

where α_{pq} equals to $1/|\mathcal{S}|$ if $(G_p, G_q) \in \mathcal{S}$ and equals to $-1/|\mathcal{D}|$ if $(G_p, G_q) \in \mathcal{D}$. Here $|\cdot|$ denotes the cardinality of the set. The problem (9) involves the joint optimization of ground distance matrix $\mathbf{C}_{pq}(\mathbf{M})$ and vector $\mathbf{x}_{pq}$, which, similar to the problem of dictionary learning [1], is non-linear and non-convex. It becomes tractable when we minimize one while keeping another fixed. When $\mathbf{C}_{pq}(\mathbf{M})$ is fixed, seeking $\mathbf{x}_{pq}$ reduces to a typical sparse representation problem; when $\mathbf{x}_{pq}$ is fixed, we may optimize $\mathbf{C}_{pq}(\mathbf{M})$ with gradient descent algorithm. Hence, the learning algorithm can be accomplished by iterating two steps. In the 1^{st} step, fixing $\mathbf{C}_{pq}(\mathbf{M})$, for any sample pair (G_p, G_q), we compute SR-EMD(G_p, G_q) defined in (3) for $\forall p, q$; in the 2^{nd} step, we fix $\mathbf{x}_{pq}$ for $\forall p, q$ and update $\mathbf{C}_{pq}(\mathbf{M})$ by a gradient descent algorithm. Let f be the objective function in (9). Differentiating f w.r.t $\mathbf{M}$, we have

$$\frac{\partial f}{\partial \mathbf{M}} = -\sum\nolimits_{p=1}^{m} \sum\nolimits_{q=p+1}^{m} \alpha_{pq} \Big(\sum\nolimits_{i=1}^{n_p+n_q} \mathbf{r}(i) \times \sum\nolimits_{j=1}^{n_p n_q} \mathbf{A}_{pq}(i,j)\mathbf{C}_{pq}^{-3}(j,j)\mathbf{x}_{pq}(j)\mathbf{Q}_{pq}^{(j)} \Big), \quad (10)$$

In the above equation, $\mathbf{r}(i)$ denotes entry i of the vector $\mathbf{r} = \mathbf{y}_{pq} - \mathbf{A}_{pq}\mathbf{C}_{pq}^{-1}(\mathbf{M})\mathbf{x}_{pq}$, $\mathbf{A}_{pq}(i,j)$ denotes the entry at row i, column j of the matrix $\mathbf{A}_{pq}$, and $\mathbf{Q}_{pq}^{(j)} =$

$(\log(\mathbf{R}_k^p) - \log(\mathbf{R}_l^q))(\log(\mathbf{R}_k^p) - \log(\mathbf{R}_l^q))^T$, where $k = (j+1) \bmod (n_p+n_q), l = j/(n_p+n_q)+1$. The constraint that $\mathbf{M}$ should be positive definite is naturally satisfied by updating it along the geodesics in the Riemannian manifold [23]

$$\mathbf{M} = \exp(\log(\mathbf{M}) - \eta \partial f / \partial \mathbf{M}) \tag{11}$$

where η is the step length in the gradient descent.

Let us then consider the metric learning for the ground distance based on product of Lie groups. Note that the parameter θ is constrained on $[0, 1]$. By introducing the sigmod function $\theta = 1/(1 + e^{-\beta}), \beta \in (-\infty, +\infty)$, the constraint on θ can be removed. In this case, the learning problem is formulated as

$$\min_{\beta, \forall \mathbf{x}_{pq}} \sum_{p=1}^{m} \sum_{q=p+1}^{m} \alpha_{pq} (\|\mathbf{y}_{pq} - \mathbf{A}_{pq} \mathbf{C}_{pq}^{-1}(\beta) \mathbf{x}_{pq}\|_2^2 + \lambda \|\mathbf{x}_{pq}\|_1), \quad \text{s.t. } \forall\, \mathbf{x}_{pq} \succcurlyeq 0, \tag{12}$$

As (9), (12) can also be minimized by a two-stage iterative method: first fix β and solve $\forall \mathbf{x}_{pq}$; then, update β by fixing $\forall \mathbf{x}_{pq}$. The partial derivative of the objective function f w.r.t β is

$$\frac{\partial f}{\partial \beta} = -2 \sum_{p=1}^{m} \sum_{q=p+1}^{m} \alpha_{pq} \mathrm{tr}\Big(\mathbf{C}_{pq}^{-1} \mathbf{x}_{pq} (\mathbf{y}_{pq} - \mathbf{A}_{pq} \mathbf{C}_{pq}^{-1} \mathbf{x}_{pq})^T \mathbf{A}_{pq} \mathbf{C}_{pq}^{-1} \hat{\mathbf{C}}_{pq} \Big), \tag{13}$$

where $\hat{\mathbf{C}}_{pq} = \mathrm{diag}\{e^{-\beta}(-a_{ij}^{pq} + b_{ij}^{pq})/(1 + e^{-\beta})^2\}$. This metric learning problem is simpler and more efficient than the previous one since it involves only one parameter β.

3. Experiments

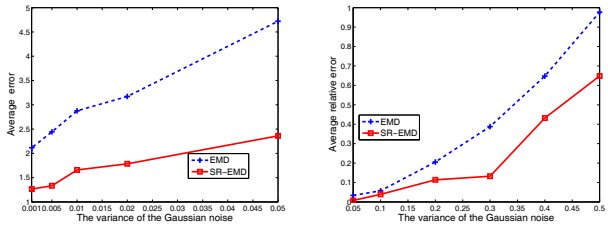
In this section, we first perform experiments on noisy images to verify the robustness and efficiency of SR-EMD. Then, we evaluate, on benchmark image retrieval databases and texture databases, respectively, the SR-EMD with Lie group-based ground distance (SR-EMD-M) and SR-EMD with product of Lie groups-based one (SR-EMD-θ), as well as SR-EMD-M and SR-EMD-θ with metric learning.

We use the CLUSTER software available at https://engineering.purdue.edu/∼bouman/software/cluster/ to estimate GMM. SR-EMD is solved with the Homotopy/LARS algorithm ($\lambda = 10^{-3}$) [9, Chap. 3.4] and EMD is solved by the method in [24]. Both algorithms are written in C++ and complied to mex files for use in Matlab. In the default cases without metric learning, we set $\theta = 0.5$ (or $\beta = 0$), $\mathbf{M} = \mathbf{I}$. The programs run on a workstation equipped with 12 2.3 GHz CPU and 12G RAM.

3.1. SR-EMD vs. EMD

This section compares the robustness and efficiency of the SR-EMD and conventional EMD. We randomly select two classes from the Corel Wang database [31]. Then we randomly select thirty images from each class. Every image is polluted by additive Gaussian noises of zero mean and increasing levels of variances. From each image a GMM with 20 component Gaussians using three-dimensional color features is estimated.

We first compare SR-EMD with the conventional EMD on "similar" GMMs. We compute the distance of two GMMs estimated from a clean image and its polluted counterpart using SR-EMD and the conventional EMD (the ground truth is 0). The errors for each noise level are averaged over sixty pairs. Fig. 2(a) shows the curves of the absolute errors (the values of SR-EMD are scaled by $1/\lambda$, c.f. (3)) vs. noise level. It can be seen that with the increase of noise level, the errors of both methods get larger; but apparently, the errors of SR-EMD are much smaller that those of the conventional EMD. Next, we compare the two methods on "dissimilar" GMMs. For each noise level, we compute the distance between a clean image and a polluted image from different classes; the relative errors (the ground truth can be calculated from the two clean images involved) are computed and averaged over all such pairs at that level. As shown in Fig. 2(b), the SR-EMD again demonstrates better robustness to noise than the conventional EMD.



(a) Similar "GMMs" (b) Dissimilar "GMMs"

Figure 2. The robustness of SR-EMD and conventional EMD to noise.

To empirically assess the efficiency of SR-EMD, we build GMM G_p with varying number (n_p=10∼150) of component Gaussians, and let $G_q = G_p$. We compare the running time of SR-EMD and EMD in computing one distance (averaged over 1000 times). Note that the computation of ground distances is the same for SR-EMD and EMD and this part is not counted since we aim to compare the optimization algorithms. Table 1 presents the running time of the two methods vs. n_p. When n_p<10, the two algorithms are comparable; however, as n_p increases, the speedups of SR-EMD over EMD becomes more and more significant.

3.2. Image Retrieval

We use two benchmark databases, Corel Wang [31] and Coil100 [20], for performance evaluation in image retrieval. The Corel Wang database consists of 1,000 images from 10 different classes. As this database contains large intra-class

Table 1. Running time comparison between SR-EMD and conventional EMD (unit: s).

n_p	10	20	30	40	50	60
EMD	6.3e-4	1.0e-2	5.5e-2	2.0e-1	5.6e-1	1.28
SR-EMD	5.9e-4	4.7e-3	1.8e-2	4.7e-2	1.2e-1	0.23
n_p	70	80	100	120		
EMD	2.80	5.24	14.7	23.7		
SR-EMD	0.40	0.78	1.79	4.02		

variations, it is considered as a very difficult benchmark database. The Coil100 database consists of 7,200 images from 100 classes. Each class contains images of the same object which is photographed from 72 different viewing angles.

We produce a GMM of 10 component Gaussians for each single image. Three kinds of local feature descriptors are extracted at dense grids (stride of 2$\sim$3 pixels) for estimating GMM: (1)128-dimensional SIFT descriptors computed via the code released in [27], the dimension of which is reduced to 15 by PCA; (2) the covariance descriptors [26], computed by taking the feature vector as R, G, B color values, and the absolute value of the first- and second-order partial derivatives of the illuminance image I, i.e., $[R, G, B, |I_x|, |I_y|, |I_{xx}|, |I_{yy}|]$, which are subject to matrix logarithm and then vectorized to 28-dimensional vectors; and (3) the 45-dimensional RGB histogram via the code released in [27].

We compare our method with Match-KL [7], EMD-KL [18], GQFD [3], and Bag-of-visual-words (BoW) methods. In GQFD the diagonal covariance matrices are used and all features involved are free of dimensionality reduction; all the other methods use full covariance matrices in estimating GMMs. The results of GQFD and BoW listed here are the best ones reported in [3]. Note that the results of Match-KL obtained in [3] were not satisfactory and we conjecture that this metric may not be suitable for high-dimensional GMMs of diagonal covariance matrices. In this paper we implement it by ourselves for comparison of GMMs with full covariance matrices. The Mean Average Precision (MAP) values of complete database rankings w.r.t 100 different random queries are used for quantitative comparison. Table 2 lists the MAP values of different methods on the two databases.

Table 2. Comparison of MAP values (%) in image retrieval

	Corel Wang [31]			Coil100 [20]		
Method	SIFT	Cov	Hist	SIFT	Cov	Hist
Match-KL [7]	40.0	36.7	36.9	28.3	36.1	41.1
GQFD [3]	45.7	46.7	36.8	45.0	47.8	60.0
BoW [3]	46.3	–	–	44.2	–	–
EMD-KL [18]	45.1	38.2	37.1	28.2	39.3	59.7
SR-EMD-M	48.7	48.3	46.9	48.4	61.7	81.8
SR-EMD-θ	49.7	51.6	45.5	52.2	64.2	82.0
SR-EMD-M (with learning)	50.1	51.0	48.7	51.4	64.5	84.9
SR-EMD-θ (with learning)	**52.6**	**53.0**	**48.9**	**55.4**	**69.8**	**85.5**

On the Corel Wang database, SR-EMD-M and SR-EMD-θ (without metric learning) are both better than the other methods while SR-EMD-θ performs the best. The MAP values of the proposed method are 3.4%, 4.9% and 9.8% higher than the current best results for SIFT, Covariance and RGBHist features, respectively. With metric learning, the MAP values of SR-EMD-M and SR-EMD-θ increase about 1.4%$\sim$4.4%. Note that the proposed methods have clear advantages over EMD-KL. As they are close relatives, we attribute the performance gains to the novel sparse representation-based formulation and the ground distances.

On the Coil100 database, while SR-EMD-M and SR-EMD-θ are both superior to other methods, the latter has obviously higher MAP values than the former. Compared with the current best results obtained by GQFD, the MAP values of SR-EMD-θ are over 7%, 16%, and 22% higher for SIFT, Covariance and RGBHist features, respectively. The improvement of MAP values of SR-EMD-M and SR-EMD-θ with metric learning is about 2.2%$\sim$5.6%. Finally, it is noteworthy to mention that the performance increase of SR-EMD-θ (with learning) is overall striking, which outperforms by over 10%, 22% and 25% the current best results of SIFT, Covariance and RGBHist features, respectively.

We compare the running time of the competing methods. The algorithms of Match-KL and GQFD are implemented in Matlab. In EMD-like methods, while the computations of the ground distances are implemented with Matlab, the algorithm for solving EMD is written with C++. Table 3 presents the the average time for one distance computation in image retrieval averaged over 1000 trials. SR-EMD-M and SR-EMD-θ are both faster than its close relatives, namely, EMD-KL, while SR-EMD-θ is more computationally efficient. It is relatively unfair to compare EMD-like algorithms with Match-KL and GQFD because of their different implementation. However, please note that in EMD-like algorithms, ground distance computation takes a large portion of the total running time. If all algorithms are implemented with C++, the EMD-like algorithms may be still comparable to Match-KL and GQFD.

Table 3. Running time of different methods (ms)

Method	Match-KL	GQFD	EMD-KL	SR-EMD-θ	SR-EMD-M
Time	6.99	8.70	10.22	5.55	7.63

3.3. Texture Classification

We further use three benchmark texture databases to evaluate the performance of the proposed method: KTH-TIPS [10], CUReT [5], and UMD http://www.cfar.umd.edu/$\sim$fer/website-texture/texture.htm. Table 4 makes a summary of them in terms of whether they contain rotation, changes of scale, illumination and viewing angles, as well as the number of classes and images (the

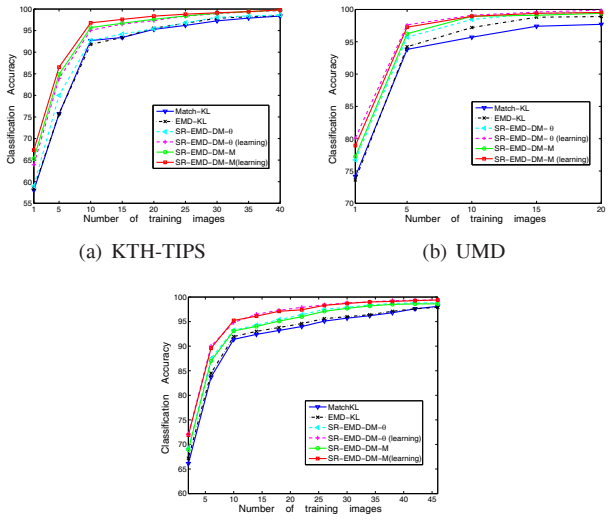
last two columns). Among SIFT, Covariance and RGBHist descriptors, the Covariance descriptors are more compact and effective and thus it is used here. We use 5-dimensional raw features, i.e., $[I, |I_x|, |I_y|, |I_{xx}|, I_{yy}|]$, and the final covariance descriptor is 15-dimensional. For a texture image, we uniformly divide it into four sub-images, on each of which we estimate a GMM with five component Gaussians. Hence, each image is represented by four GMMs. In the testing stage, each one of the four GMMs is compared to all GMMs of the training set and its label is determined by the KNN algorithm (k=3). The vote of four GMMs associated with the test image determines the class label it belongs to. We randomly select n=1$\sim$20 (or n=1$\sim$40) images as the training set and the remaining ones as the testing set. For each n, the result is averaged over 20 trials.

Table 4. Summary of benchmark texture databases

Database	Rot.	Scal.	Illu.	Ang.	Size	Class	Image
KTH-TIPS	no	yes	yes	no	200x200	10	810
UMD	yes	yes	no	yes	320x240	25	1000
CUReT	yes	no	yes	no	200x200	61	5612

We first compare the proposed methods with Match-KL[7] and EMD-KL[18]. Note that we also implemented GQFD with 128-dimensional SIFT, but the results are not satisfactory and we do not report here to save space. The results are depicted in Figures 3(a), 3(b) and 3(c) for the KTH-TIPS, UMD and CUReT databases, respectively. On KTH-TIPS, Match-KL and EMD-KL have very similar performance, both of which are obviously outperformed by the proposed methods. Between SR-EMD-M and SR-EMD-θ, the former has clear advantage and its performance is even comparable to SR-EMD-θ with metric learning. Both SR-EMD-θ and SR-EMD-M's performance can be improved by 2%$\sim$3% with metric learning. On UMD, EMD-KL has clear advantage over Match-KL when the number of training images are greater than five. SR-EMD-θ and SR-EMD-M perform similarly on this database. After metric learning, both have performance increase of 2%$\sim$3%, while SR-EMD-θ is a little better than SR-EMD-M. On CUReT, SR-EMD-θ and SR-EMD-M have very similar performance before and after metric learning, and they have clear superiority to the other two methods.

Finally, we compare SR-EMD-θ and SR-EMD-M (with metric learning) with six state-of-the-art methods on texture classification. Table 5 presents a summary of the classification rates ("-" means that the result is unavailable). On KTH-TIPS, the proposed method has over 10%, 6% and 6% improvement over the current best results when 1, 5, or 10 training images per class are used, respectively. On UMD, we have over 6% gain with only one training image, while the performance growth is obvious as the number of training samples grows. On CUReT, compared with the state-of-the-art results, the improvement of classification rates is nearly



(a) KTH-TIPS (b) UMD

(c) CUReT

Figure 3. Classification rate vs. the number of training examples on three benchmark texture databases.

4% when the number of training images is small (2 or 10), while our results are comparable to state-of-the-arts when the number of training images is large (26 or 46).

4. Conclusion

This paper presented a novel methodology of EMD for matching GMMs. Our contributions lie in the sparse representation formulation of EMD, and the novel ground distances between component Gaussians based on information geometry. In addition, we presented a simple yet effective supervised metric learning method to adapt the distance metrics between GMMs to the given data. Compared with the conventional EMD, it is more efficient and more robust to noise while enjoying significant performance gains. Though metric learning has been extensively studied for vector data, to the best of our knowledge, the problem of metric learning for GMMs is not explored yet and we made the first attempt on this problem in this paper. It would be interesting to study more advanced metric learning methods for GMMs, which may improve further the efficacy of content-based image classification applications.

Acknowledgments: The work was supported by NSFC 60973080, 61170149, Program for New Century Excellent Talents in University (NCET-10-0151), the Fundamental Research Funds for the Central Universities (DUT13RC(3)02), Key Project by Chinese Ministry of Education (210063).

Table 5. Comparison of texture classification rates (%) with state-of-the-art

	KTH-TIPS				UMD				CUReT			
Method	1	5	10	40	1	5	10	20	2	10	26	46
Zhang *et al*. [34]	55.1	80.1	90.0	96.1	–	–	–	–	53.6	80.0	91.1	95.3
Hayman *et al*. [10]	50.2	78.3	85.3	94.8	–	–	–	–	60.2	91.0	97.6	98.5
VZ-joint [28]	50.5	72.9	80.5	92.1	–	–	–	–	54.4	83.4	93.1	97.4
WMFS [14]	–	–	–	96.5	–	–	–	98.7	–	–	–	–
Liu *et al*. [17]	56.5	80.5	87.8	99.3	73.9	95.0	97.5	99.3	68.2	91.5	98.3	99.4
CLBP [8]	49.0	76.1	85.5	96.8	73.6	92.4	96.0	98.0	60.2	83.6	92.9	95.9
SR-EMD-M	**67.3**	**86.5**	**96.8**	**99.8**	78.9	97.3	98.4	99.5	71.8	95.2	98.3	99.5
SR-EMD-θ	63.9	84.0	95.1	99.6	**80.1**	**97.6**	**99.1**	**99.9**	**72.1**	**95.4**	**98.7**	**99.5**

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *TSP*, 54(11):4311–4322, 2006.

[2] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. on Matrix Analysis and Applications*, 2006.

[3] C. Beecks, A. M. Ivanescu, S. Kirchhoff, and T. Seidl. Modeling image similarity by gaussian mixture models and the signature quadratic form distance. In *ICCV*, 2011.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[5] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real-world surfaces. *ACM TOG*, 18(1):1–34, 1999.

[6] L. David G. and Y. Ye. *Linear and Nonlinear programming*. Springer, 2006.

[7] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures. In *ICCV*, 2003.

[8] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *TIP*, 19(6):1657–1663, 2010.

[9] T. Hastie, R. Tibshirani, and jerome Friedman. *The Elements of Statistial Learning*. Springer, 2009.

[10] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh. On the significance of real-world conditions for material classification. In *ECCV*, 2004.

[11] J. Hershey and P. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *ICASSP*, volume 4, pages IV–317, 2007.

[12] G. H.Golub and C. F. Loan. *Matrix Computations*. Johns Hopkins Press, 1996.

[13] S. ichi Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[14] H. Ji, X. Yang, H. Ling, and Y. Xu. Wavelet domain multifractal analysis for static and dynamic texture classification. *TIP*, 22(1):286–299, 2013.

[15] V. Karavasilis, C. Nikou, and A. Likas. Visual tracking using the Earth Mover's Distance between gaussian mixtures and kalman filtering. *Image Vision Comput.*, 29(5), Apr. 2011.

[16] H. Ling and K. Okada. An efficient Earth Mover's Distance algorithm for robust histogram comparison. *PAMI*, 29(5):840–853, 2007.

[17] L. Liu, P. Fieguth, G. Kuang, and H. Zha. Sorted random projections for robust texture classification. In *ICCV*, 2011.

[18] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *ICME*, 2001.

[19] M. Lovric, M. Min-Oo, and E. A. Ruh. Multivariate normal distributions parametrized as a Riemannian symmetric space. *JMVA*, 74(1):36–48, 2000.

[20] Nayar and H. Murase. Columbia object image library: Coil-100. Technical report, Department of Computer Science, Columbia University, 1996.

[21] O. Pele and M. Werman. A linear time histogram metric for improved SIFT matching. In *ECCV*, pages 495–508, 2008.

[22] O. Pele and M. Werman. Fast and robustEarth Mover's Distances. In *ICCV*, pages 460–467, 2009.

[23] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *IJCV*, pages 41–66, 2006.

[24] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *IJCV*, 2000.

[25] S. Shirdhonkar and D. Jacobs. Approximate Earth Mover's Distance in linear time. In *CVPR*, pages 1–8, 2008.

[26] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

[27] K. van de Sande, T. Gevers, , and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010.

[28] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *CVPR*, 2003.

[29] N. Vasconcelos. On the complexity of probabilistic image retrieval. In *ICCV*, 2001.

[30] F. Wang and L. J. Guibas. Supervised earth mover's distance learning and its computer vision applications. In *ECCV*, 2012.

[31] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *PAMI*, 23:947–963, 2001.

[32] Y. Wu, K. Chan, and H. Wang. Texture classification based on finite gaussian mixture model. In *IEEE workshop on Texture Analysis and Synthesis*, 2003.

[33] X. Yang and R. Jin. Distance metric learning: A comprehensive survey. Technical report, Michigan State University, 2007.

[34] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007.