Pattern Recognition 75 (2018) 188–198

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Efficient multi-modal geometric mean metric learning

Jianqing Liang[a], Qinghua Hu[a,*], Pengfei Zhu[a], Wenwu Wang[b]

[a] School of Computer Science and Technology, Tianjin University, Tianjin, China
[b] Department of Electrical and Electronic Engineering, University of Surrey, United Kingdom

ARTICLE INFO

Article history:
Received 16 November 2016
Revised 5 January 2017
Accepted 27 February 2017
Available online 2 March 2017

Keywords:
Metric learning
Multi-modality
Efficiency
Geometric mean

ABSTRACT

With the fast development of information acquisition, there is a rapid growth of multi-modality data, e.g., text, audio, image and video, in health care, multimedia retrieval and many other applications. Confronted with the challenges of clustering, classification or regression with multi-modality information, it is essential to effectively measure the distance or similarity between objects described with heterogeneous features. Metric learning, aimed at finding a task-oriented distance function, is a hot topic in machine learning. However, most existing algorithms lack efficiency for high-dimensional multi-modality tasks. In this work, we develop an effective and efficient metric learning algorithm for multi-modality data, i.e., Efficient Multi-modal Geometric Mean Metric Learning (EMGMML). The proposed algorithm learns a distinctive distance metric for each view by minimizing the distance between similar pairs while maximizing the distance between dissimilar pairs. To avoid overfitting, the optimization objective is regularized by symmetrized LogDet divergence. EMGMML is very efficient in that there is a closed-form solution for each distance metric. Experimental results show that the proposed algorithm outperforms the state-of-the-art metric learning methods in terms of both accuracy and efficiency.

1. Introduction

Multi-modality data are booming with the ubiquitous usage of digital devices and social network. In multi-media retrieval, there exists a large variety of data, e.g., text, audio, image, and video on the website. In biometric recognition, a person can be identified by retina, face, iris, signature, fingerprint, or palmprint [1–5]. For face recognition, a face image may be captured by cell phones, near-infrared cameras or depth cameras [6–13]. An object is usually described by different modalities with complementary information in many computer vision and pattern recognition tasks.

Learning a task-driven metric from massive multi-modality data automatically is meaningful to diverse applications such as computer vision, bioinformatics and information retrieval. Metric learning, which aims to train an appropriate measure from data, has stimulated wide interests over the past decade. A large number of approaches have been proposed, most of which intend to learn a Mahalanobis-like metric. Generally, according to the optimality of the solution, metric learning can be categorized into global methods and local methods. Global methods can be regarded as learning a linear geometric transformation over the input space [14–17]. While the simplicity promotes their wide application, the global metrics still suffer from the curse of dimensionality. Compared with global metrics, local metrics have been shown to be able to flexibly capture geometric variations across different feature spaces [18–20]. However, a major drawback of local metric learning is that it may lead to overfitting [21]. In addition, they are generally confronted with high computational cost.

Despite the large amount of work on single modality, learning metrics for multiple modalities still remains largely unexplored [22]. Since single metrics ignore consensus & complementarity properties between different modalities, they may fail in multi-modal learning. Under such circumstance, multiple kernel techniques, which map the data to high-dimensional feature spaces with a set of nonlinear kernel matrices, have been introduced to address these issues [7,23–25]. To our best knowledge, McFee and Lanckriet [23] first utilized multiple kernel learning techniques to integrate heterogeneous modalities into a single, unified similarity space. In their work, an optimal ensemble of kernel transformations is learned. Unfortunately, it is not applicable to large-scale tasks due to the high computational costs. Lu et al. [24] proposed a weighted kernel embedding technique for metric learning, which is shown to be effective in combining multiple features. Recently, Lu et al. [7] exploited statistical information to represent image sets and developed a localized multi-kernel metric learning (LMKML) method. Liang et al. [25] developed a semi-supervised online multi-kernel similarity learning framework, which is a multi-stage

* Corresponding author.
E-mail addresses: liangjianqing@tju.edu.cn (J. Liang), huqinghua@tju.edu.cn (Q. Hu), zhupengfei@tju.edu.cn (P. Zhu), w.wang@surrey.ac.uk (W. Wang).

http://dx.doi.org/10.1016/j.patcog.2017.02.032
0031-3203/

algorithm consisting of feature selection, selective ensemble learning, active sample selection and triplet generation. While state-of-the-art performance has been achieved, it remains an open problem to develop an efficient strategy to improve the speed. Generally speaking, although multiple kernel learning may capture the complex data structure and avoid the curse of dimensionality, the time-consuming process in terms of parameter adjustment limits its scalability in large-scale tasks.

A distance metric learning algorithm is evaluated in terms of both accuracy and efficiency. Although these aforementioned methods outperform the state-of-the-arts, the high time complexity limits their scalability in practical applications, especially in handling multi-modality data. As time cost as well as the memory requirement dramatically increases when dealing with large-scale data represented with high-dimensional multiple modalities, how to develop an effective and efficient metric learning method has become a hot topic. To solve the problem, online learning techniques have been considered [26,27]. In [26], a novel online multiple kernel similarity (OMKS) learning framework is proposed to learn a flexible proximity function with multiple kernels. In [27], an online multi-modal distance metric learning (OMDML) scheme is presented, which aims at learning distinctive metrics in individual modality space and the weights for combining different modalities via a joint formulation. While online approaches are more scalable compared with the batch processing techniques, they are more likely to suffer from high computational cost in projections in that the iteration process used often involves a gradient descent method.

As the iterative gradient descent or eigenvalue decomposition is used in solving the optimization problem, most of these metric learning algorithms are computationally expensive. Remarkably, Zadeh et al. [28] developed geometric mean metric learning (GMML), which formulates metric learning as an unconstrained smooth and strictly convex optimization problem. GMML is very efficient for large-scale tasks in that it admits a closed form solution. Additionally, for multi-modal learning, the commonness and individuality should be made good use of to improve the discrimination ability of the learned metrics.

In this paper, we develop a novel efficient multi-modal geometric mean metric learning (EMGMML) framework to handle data with multiple modalities, which is here referred specifically to multiple visual features extracted from media objects. EMGMML learns the metrics for multiple features in a joint optimization problem by pulling similar pairs close whereas pushing dissimilar pairs away. To exploit the complementarities among different modalities, the learned metrics for different modalities are required to be close to a common prior metric by symmetrized LogDet divergence. Meanwhile, to highlight the difference of multi-modalities, we assign a weight to each modality. Specifically, the metric associated with each modality can be addressed in a closed form solution. Then, the metric learning problem can be converted into a quadratic programming in terms of weights. Compared with existing metric learning approaches, EMGMML is highly scalable and efficient since the commonly used kernel mapping and the optimization of a semi-definite programming problem are no longer required. Empirical results on benchmark datasets with hundreds of dimensions verify that multiple weighted metrics obtained by our algorithm give prominent performance boost in terms of visual search.

The remainder of this paper is organized as follows. Section 2 briefly reviews the GMML algorithm in [28]. In Section 3, we introduce our proposed EMGMML algorithm for high-dimensional multi-modal data. Section 4 analyzes experimental results on both qualitative and quantitative point of view. Section 5 concludes our study and gives an outlook for our future work.

2. Geometric mean metric learning model

In this section, we review geometric mean metric learning (GMML) [28] algorithm.

2.1. Formulation

We aim to learn a Mahalanobis distance

$$d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}-\boldsymbol{x}')^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}'), \tag{1}$$

where $\boldsymbol{x},\boldsymbol{x}' \in \mathbb{R}^d$ are data vectors and $\boldsymbol{A}$ is a $d \times d$ real and symmetric positive definite (SPD) matrix to be solved. Constraints are provided in the form of positive / negative pairs

$$\mathcal{S} := \{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are in the same class}\}$$

$$\mathcal{D} := \{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are in different classes}\}.$$

The objective is to minimize the sum of distances between similar points with a matrix $\boldsymbol{A}$ and distances between dissimilar points with $\boldsymbol{A}^{-1}$

$$\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{D}} d_{\boldsymbol{A}^{-1}}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{2}$$

The idea is that increasing the distance $d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{y})$ between dissimilar pairs is equivalent to decreasing $d_{\boldsymbol{A}^{-1}}(\boldsymbol{x},\boldsymbol{y})$. The gradients of $d_{\boldsymbol{A}}$ and $d_{\boldsymbol{A}^{-1}}$ are in opposite directions, which can confirm the rationality of the idea.

Substituting the distance with traces, we get

$$\min_{\boldsymbol{A}\succ 0} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} \operatorname{tr}(\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T) + \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{D}} \operatorname{tr}(\boldsymbol{A}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T) \tag{3}$$

We denote two crucial matrices

$$\begin{aligned}\boldsymbol{S} &:= \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} (\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T,\\ \boldsymbol{D} &:= \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{D}} (\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T,\end{aligned} \tag{4}$$

that are the similarity and dissimilarity matrices, respectively. Utilizing (4), we rewrite (3) as

$$\min_{\boldsymbol{A}\succ 0} h(\boldsymbol{A}) := \operatorname{tr}(\boldsymbol{A}\boldsymbol{S}) + \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{D}). \tag{5}$$

$h(\boldsymbol{A})$ has some key properties such as geodesic convexity. Here are several concepts of geodesically convex functions.

Geodesic convexity is a generalization of linear convexity for sets and functions to nonlinear Riemannian manifolds [29]. The geodesic curve locally minimizes the Riemannian distances between two points. The connection between $\boldsymbol{A}$ and $\boldsymbol{B}$ on the SPD manifold is defined as

$$\boldsymbol{A}\sharp_t\boldsymbol{B} = \boldsymbol{A}^{1/2}(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2})^t\boldsymbol{A}^{1/2}, \quad t\in[0,1].$$

On the entire set of SPD, the definition of geodesically convex functions is given as follows [30]

Definition 1. A function f on a geodesically convex subset of a Riemannian manifold is *geodesically convex*, if for all points $\boldsymbol{A}$ and $\boldsymbol{B}$ in this set, it satisfies

$$f(\boldsymbol{A}\sharp_t\boldsymbol{B}) \le t f(\boldsymbol{A}) + (1-t) f(\boldsymbol{B}), \quad t\in[0,1].$$

If for $t \in (0, 1)$ the above inequality is strict, the function is called strictly geodesically convex.

Key properties of $h(\boldsymbol{A})$ is summarized as follows [28]

Theorem 1. *The cost function h in (5) is both strictly convex and strictly geodesically convex on the SPD manifold.*

2.2. Solution

According to the convexity of the objective function, we can obtain its global minimum by setting the gradient as zero

$$\nabla h(\boldsymbol{A}) = \boldsymbol{S} - \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{A}^{-1} = 0$$

Thus

$$\boldsymbol{A}\boldsymbol{S}\boldsymbol{A} = \boldsymbol{D}. \tag{6}$$

Actually, the sole solution of (6) is the midpoint on the geodesic connecting $\boldsymbol{S}^{-1}$ and $\boldsymbol{D}$ [31], namely

$$\boldsymbol{A} = \boldsymbol{S}^{-1}\sharp_{1/2}\boldsymbol{D} = \boldsymbol{S}^{-1/2}(\boldsymbol{S}^{1/2}\boldsymbol{D}\boldsymbol{S}^{1/2})^{1/2}\boldsymbol{S}^{-1/2}.$$

Following the above definition, we know that $\boldsymbol{A}$ is SPD.

While GMML obtains a closed-form solution, owing to the inverse matrix calculation, it is still computationally expensive in handling high-dimensional multi-modal tasks. Furthermore, the performance may suffer due to the ignorance of the correlation between different modalities. To solve this problem, we propose the framework of EMGMML as follows.

3. Efficient multi-modal geometric mean metric learning model

In this section, we describe how to learn a geometric mean metric on multi-modality data.

3.1. Formulation

Given a set of samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, each sample $\boldsymbol{x}_i$ is represented with m modalities $\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \ldots, \boldsymbol{x}_i^m$, we aim at learning such a weighted Mahalanobis distance

$$d_{\{\boldsymbol{A}_p, w_p\}_{p=1}^m}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{p=1}^{m} w_p \left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)^T \boldsymbol{A}_p \left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right), \tag{7}$$

where $\boldsymbol{x}_i^p, \boldsymbol{x}_j^p \in \mathbb{R}^{d_p}$ are the ith and jth point on the pth modality, respectively. w_p is a weight that determines the importance of the pth modality in distance metric learning. $\boldsymbol{A}_p$ is a $d_p \times d_p$ real and symmetric positive definite matrix to be learned for the pth modality. Similarly, supervision information is given by sets of pairs in terms of each modality

$$\mathcal{S}_p := \left\{\left(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p\right) \mid \boldsymbol{x}_i^p \text{ and } \boldsymbol{x}_j^p \text{ are in the same class}\right\}$$
$$\mathcal{D}_p := \left\{\left(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p\right) \mid \boldsymbol{x}_i^p \text{ and } \boldsymbol{x}_j^p \text{ are in different classes}\right\}.$$

Referring to the GMML algorithm, the objective can be

$$\sum_{p=1}^{m} w_p \Big(\sum_{(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \in \mathcal{S}_p} d_{\boldsymbol{A}_p}\left(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p\right) + \sum_{(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \in \mathcal{D}_p} d_{\boldsymbol{A}_p^{-1}}\left(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p\right) \Big) \tag{8}$$

Rewriting the objective with traces, we turn (8) into

$$\begin{aligned} \min_{\{\boldsymbol{A}_p\}_{p=1}^m \succ 0} \sum_{p=1}^{m} w_p \Big(& \sum_{\left(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p\right) \in \mathcal{S}_p} \operatorname{tr}\left(\boldsymbol{A}_p \left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)\left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)^T\right) \\ & + \sum_{(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \in \mathcal{D}_p} \operatorname{tr}\left(\boldsymbol{A}_p^{-1} (\boldsymbol{x}_i^p - \boldsymbol{x}_j^p)\left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)^T\right) \Big) \end{aligned} \tag{9}$$

We now define the following two matrices $\boldsymbol{S}_p$ and $\boldsymbol{D}_p$ to represent similarity and dissimilarity matrices for the pth modality

$$\begin{aligned} \boldsymbol{S}_p &:= \sum_{\left(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p\right) \in \mathcal{S}_p} \left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)\left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)^T, \\ \boldsymbol{D}_p &:= \sum_{(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \in \mathcal{D}_p} \left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)\left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p\right)^T, \end{aligned} \tag{10}$$

Therefore, we can get the basic formulation of EMGMML

$$\min_{\{\boldsymbol{A}_p\}_{p=1}^m \succ 0} h(\{\boldsymbol{A}_p\}_{p=1}^m) := \sum_{p=1}^{m} w_p(\operatorname{tr}(\boldsymbol{A}_p\boldsymbol{S}_p) + \operatorname{tr}(\boldsymbol{A}_p^{-1}\boldsymbol{D}_p)). \tag{11}$$

As the matrix $\boldsymbol{S}_p$ may be near-singular or non-invertible, we add a regularizer to the objective [28]

$$\min_{\{\boldsymbol{A}_p\}_{p=1}^m \succ 0} \sum_{p=1}^{m} w_p(\operatorname{tr}(\boldsymbol{A}_p S_p) + \operatorname{tr}(\boldsymbol{A}_p^{-1}\boldsymbol{D}_p)) + \lambda \sum_{p=1}^{m} w_p D_{sld}(\boldsymbol{A}_p, \boldsymbol{A}_0), \tag{12}$$

where $\boldsymbol{A}_0$ is a prior metric (set as discussed in Section 4.4) and D_{sld} is the symmetrized LogDet divergence

$$D_{sld}(\boldsymbol{A}_p, \boldsymbol{A}_0) := \operatorname{tr}(\boldsymbol{A}_p\boldsymbol{A}_0^{-1}) + \operatorname{tr}(\boldsymbol{A}_p^{-1}\boldsymbol{A}_0) - 2d, \tag{13}$$

It is noteworthy that another variable is w_p. To ensure the distance is positive, we require w_p to be non-negative. However, as the distance and divergence are both non-negative, the objective obtains the minimum when each w_p equals 0. Since we hope each modality can make its own contribution, most w_p should be positive. Thus, we let the sum of w_p be a constant. At this point, the objective becomes a linear programming, which, as a result, may lead most of the weights to nearly zero. To avoid overfitting, we introduce a regularizer term of w_p. Ultimately, the regularized version of EMGMML is

$$\begin{aligned} \min_{\{\boldsymbol{A}_p, w_p\}_{p=1}^m} & \sum_{p=1}^{m} w_p(\operatorname{tr}(\boldsymbol{A}_p\boldsymbol{S}_p) + \operatorname{tr}(\boldsymbol{A}_p^{-1}\boldsymbol{D}_p)) \\ & + \lambda \sum_{p=1}^{m} w_p D_{sld}(\boldsymbol{A}_p, \boldsymbol{A}_0) + \gamma \sum_{p=1}^{m} w_p^2, \\ & s.t.\ \boldsymbol{A}_p \succ 0, \quad p = 1, 2, \ldots, m \\ & w_p \geq 0, \quad p = 1, 2, \ldots, m \\ & \sum_{p=1}^{m} w_p = 1 \end{aligned} \tag{14}$$

Let $\boldsymbol{w} = [w_1, w_2, \ldots, w_m]$ be an m-dimensional vector, then $\sum_{p=1}^m w_p^2$ equals $\|\boldsymbol{w}\|_2^2$.

3.2. Solution

In the following, we develop an efficient optimization approach to solve (14). An alternating strategy is introduced in the solving procedure. Observing that the only constraint of $\boldsymbol{A}_p$ is the positive definiteness, we consider to solve $\boldsymbol{A}_p$ at first. For simplicity, we denote the function

$$\begin{aligned} \mathcal{L}(\{\boldsymbol{A}_p\}_{p=1}^m) = & \sum_{p=1}^{m} w_p(\operatorname{tr}(\boldsymbol{A}_p\boldsymbol{S}_p) + \operatorname{tr}(\boldsymbol{A}_p^{-1}\boldsymbol{D}_p)) \\ & + \lambda \sum_{p=1}^{m} w_p D_{sld}(\boldsymbol{A}_p, \boldsymbol{A}_0) \end{aligned} \tag{15}$$

The derivative of $\mathcal{L}$ with respect to $\boldsymbol{A}_p$ is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}_p} = w_p(\boldsymbol{S}_p - \boldsymbol{A}_p^{-1}\boldsymbol{D}_p\boldsymbol{A}_p^{-1}) + \lambda w_p(\boldsymbol{A}_0^{-1} - \boldsymbol{A}_p^{-1}\boldsymbol{A}_0\boldsymbol{A}_p^{-1})$$

Setting it to zero leads to

$$w_p = 0, \text{ or } \boldsymbol{S}_p - \boldsymbol{A}_p^{-1}\boldsymbol{D}_p\boldsymbol{A}_p^{-1} + \lambda(\boldsymbol{A}_0^{-1} - \boldsymbol{A}_p^{-1}\boldsymbol{A}_0\boldsymbol{A}_p^{-1}) = 0$$

However, if $w_p = 0$ holds for all $p = 1, 2, \ldots, m$, then we can not satisfy the constraint $\sum_{p=1}^{m} w_p = 1$. Therefore

$$\boldsymbol{S}_p - \boldsymbol{A}_p^{-1}\boldsymbol{D}_p\boldsymbol{A}_p^{-1} + \lambda(\boldsymbol{A}_0^{-1} - \boldsymbol{A}_p^{-1}\boldsymbol{A}_0\boldsymbol{A}_p^{-1}) = 0. \tag{16}$$

We can obtain the solution

$$\boldsymbol{A}_p = (\boldsymbol{S}_p + \lambda \boldsymbol{A}_0^{-1})^{-1} \sharp_{1/2} (\boldsymbol{D}_p + \lambda \boldsymbol{A}_0), \tag{17}$$

From the form of geometric mean, we may conclude that $\boldsymbol{A}_p$ is SPD. Once the $\boldsymbol{A}_p$ is determined, the problem (14) is transformed to a quadratic programming on w_p.

3.3. Weighted version

To generalize the scope of the solution, we propose the weighted EMGMML objective with the optimal w_p [28,31]

$$\min_{\{\boldsymbol{A}_p\}_{p=1}^m \succ 0} h_t(\{\boldsymbol{A}_p\}_{p=1}^m) := (1-t) \sum_{p=1}^m w_p \delta_R^2(\boldsymbol{A}_p, \boldsymbol{S}_p^{-1}) + t \sum_{p=1}^m w_p \delta_R^2(\boldsymbol{A}_p, \boldsymbol{D}_p), \tag{18}$$

where δ_R is the Riemannian distance on SPD matrices

$$\delta_R(\boldsymbol{X}, \boldsymbol{Y}) := \| \log(\boldsymbol{Y}^{-1/2} \boldsymbol{X} \boldsymbol{Y}^{-1/2}) \|_F \quad for\ \boldsymbol{X}, \boldsymbol{Y} \succ 0,$$

As the w_p is fixed and positive, $\boldsymbol{S}_p$ and $\boldsymbol{D}_p$ are known, the problem (18) is equivalent to the following m tasks:

$$\min_{\boldsymbol{A}_p \succ 0} \quad h_t(\boldsymbol{A}_p) = (1-t)\delta_R^2(\boldsymbol{A}_p, \boldsymbol{S}_p^{-1}) + t\delta_R^2(\boldsymbol{A}_p, \boldsymbol{D}_p), \tag{19}$$

The unique solution is the weighted geometric mean

$$\boldsymbol{A}_p = \boldsymbol{S}_p^{-1} \sharp_t \boldsymbol{D}_p, \tag{20}$$

Therefore, the regularized form of the solution is

$$\boldsymbol{A}_p = (\boldsymbol{S}_p + \lambda \boldsymbol{A}_0^{-1})^{-1} \sharp_t (\boldsymbol{D}_p + \lambda \boldsymbol{A}_0), \quad t \in [0, 1]$$

The algorithm is summarized in Algorithm 1.

3.4. Discussion

Let the dimension of the pth modality be d_p and $d_{\max} = \max_{p \in [1,m]} d_p$. The total dimension is $d = \sum_{p=1}^m d_p$. The number of the pairs is denoted as T. The time cost of GMML mainly lies in two parts: the computation of matrices $\boldsymbol{S}$, $\boldsymbol{D}$ and distance matrix $\boldsymbol{A}$. The time cost of the first part is $O(Td^2)$. The second part involves the matrix power and multiplication, which costs both $O(d^3)$. Therefore, the total time cost for GMML should be $O(Td^2 + d^3)$. As for EMGMML, the first part costs $O(mTd_{\max}^2)$ while the second one is $O(md_{\max}^3)$. The extra term induced by the quadratic programming is $O(m^2)$. As m is much smaller than $d_{\max}$, the time complexity of EMGMML is $O(mTd_{\max}^2 + md_{\max}^3)$. From the above analysis, we know that as an extended version, EMGMML inherits the advantages of GMML in scalability. It is even more efficient in dealing with multi-modality high-dimensional data.

Overall, our proposed EMGMML framework projects multiple modalities onto distinctive feature subspaces, and then exploits a weighted combination to integrate corresponding metrics. An alternating strategy is used for solving the joint objective of metrics as well as weights, which is shown to be both effective and efficient by empirical results in Section 4.

Algorithm 1 The optimization of EMGMML.

Input:

Constraint sets in terms of positive pairs $\{\mathcal{S}_p\}_{p=1}^m$ and negative pairs $\{\mathcal{D}_p\}_{p=1}^m$,

Step length of geodesic t, Regularization parameters λ, γ, Prior knowledge $\boldsymbol{A}_0$

1: **for** $p = 1$ to m **do**

2: Compute the similarity and dissimilarity matrices

$$\boldsymbol{S}_p = \sum_{(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \in \mathcal{S}_p} (\boldsymbol{x}_i^p - \boldsymbol{x}_j^p)(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p)^T,$$

$$\boldsymbol{D}_p = \sum_{(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \in \mathcal{D}_p} (\boldsymbol{x}_i^p - \boldsymbol{x}_j^p)(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p)^T$$

3: Return the transformation matrix

$$\boldsymbol{A}_p = (\boldsymbol{S}_p + \lambda \boldsymbol{A}_0^{-1})^{-1} \sharp_t (\boldsymbol{D}_p + \lambda \boldsymbol{A}_0)$$

4: **end for**

5: Take $\boldsymbol{A}_p$ into (14) and solve the quadratic programming

Output:

Transformation matrices $\{\boldsymbol{A}_p\}_{p=1}^m$ and combination weights $\{w_p\}_{p=1}^m$

Table 1
Basic descriptions of datasets.

Datasets	# Classes	# Dimensions	# Samples
Corel 800	10	2835	800
ImageCLEF	10	2323	800
Indoor	10	2835	600
Caltech 10	10	2835	800
Birds	6	2835	600
Corel 5k	50	2835	5000

4. Experiments

In this section, we empirically analyze the performance of EMGMML. We first describe the datasets and descriptors as well as the evaluation criterion. Then we elaborate the compared methods and parameter setting and tuning. Finally, we compare EMGMML with state-of-the-arts in terms of effectiveness and efficiency on retrieval.

4.1. Datasets and environment

We carry out the experiments on image datasets including Corel [15], ImageCLEF[1], Indoor[2], Caltech256[3] and Birds [32]. Some images are shown in Fig. 1. For each dataset, several types of visual descriptors are exploited. Global features contain color histogram (256 dimensions for gray images and 768 dimensions for color images), GLCM coefficients (16 dimensions), LBP (59 dimensions) and GIST features (512 dimensions). Local features include the SIFT, dense-SIFT, SURF, Geometric Blur and PHOG (680 dimensions) descriptors. All of these local descriptors are represented by Bag-of-Words (BOW) with vocabulary size as 200 except the last one. The basic information of these datasets is listed in Table 1. For image retrieval, we split the dataset into several parts: 50% for training (5% labeled and 45% unlabelled), 10% for validation, 10% for query, and the remaining 30% as pooling set. The experiment is performed on a machine with 3.40 GHz Intel processor and 8 GB memory, and the Matlab software.

Referring to early literatures [33], we generate similar pairs by selecting two samples from the same category and dissimilar pairs by picking up two samples from distinct classes. The only difference is that we exploit all the samples from the training set instead of performing random selection.

4.2. Evaluation criterion

In this paper, we use mean average precision (MAP) to evaluate the performance of image retrieval. MAP is defined on the retrieved ranking list of queries. It is such a measurement of how

[1] http://imageclef.org/.
[2] http://web.mit.edu/torralba/www/indoor.html.
[3] http://www.vision.caltech.edu/Image_DataSets/Caltech256/.

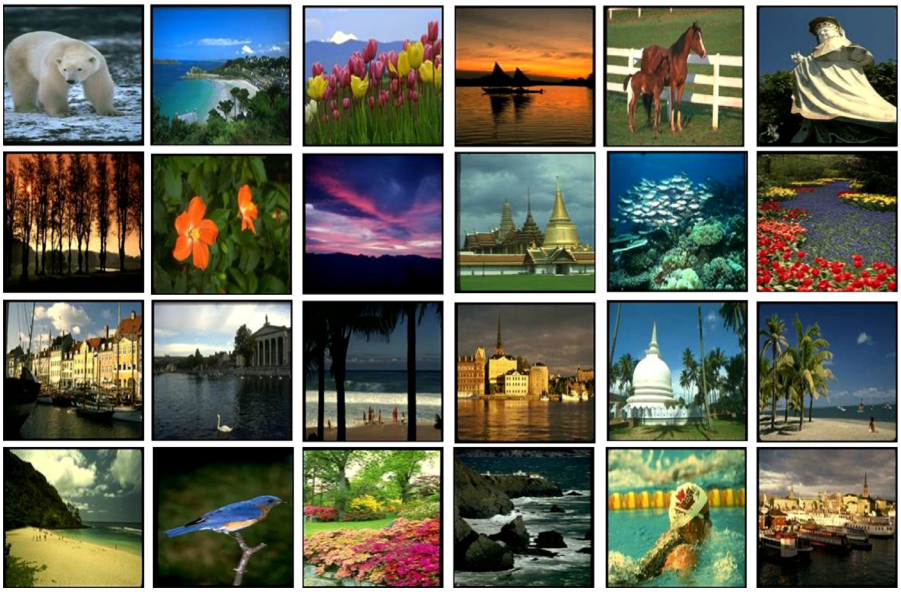


Fig. 1. Several image examples in our experiments.

the retrieved samples relate to the query. Given a query and its R retrieved images, the Average Precision is defined as [34]

$$AP = \frac{1}{L}\sum_{r=1}^{R} prec(r)\delta(r), \tag{21}$$

where L is the number of relevant samples in the retrieved set, $prec(r)$ is the precision at the rth position. $\delta(r)$ represents whether the rth retrieved image is relevant to the query or not. $\delta(r) = 1$ when they are relevant and 0 otherwise. The MAP is computed as the average AP of all the queries. We set R as the number of each class in the pooling set for small datasets, while we set R as 10 for large datasets like Corel 5k.

4.3. Comparison methods

We compare the proposed algorithm with eight baseline methods.

- **DCA**. An efficient metric learning scheme which exploits both positive and negative constraints [15].
- **LRML**. A novel metric learning technique that integrates both labeled and unlabelled data into an effective graph regularization framework [16].
- **OASIS**. A supervised online dual approach that learns a bilinear similarity measure [35].
- **EMR**. A scalable graph-based manifold ranking algorithm [36].
- **DML-eig**. An efficient eigenvalue optimization framework for metric learning [37].
- **OMKS**. An efficient online metric learning algorithm which learns a flexible nonlinear proximity function with multiple kernels for improving visual search [26].
- **SERAPH**. An information-theoretic semi-supervised metric learning approach that does not rely on the manifold assumption [17].
- **GMML**. A supervised metric learning method that is based on geometric intuition and has a closed form solution [28].

To observe the effect of weights on performance, we add another method called **UGMML**, which learns an optimal metric with GMML for each modality, and then uniformly combines all these metrics. All of the distance metric learning approaches, except EMGMML, UGMML as well as OMKS, are performed on the concatenated feature vectors from different modalities.

4.4. Parameter setting and tuning

As for parameters, we only tune several key parameters on validation datasets for the best results and set all the others to default values. For GMML, we set the parameter $\lambda = 0.1$. The prior matrix $\boldsymbol{A}_0$ is set as an identity matrix [28]. The step length t is adjusted in [0, 1] with a step size 0.1. For EMGMML, the parameter γ is tuned with the grid-search" strategy from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$. The parameter settings of λ, $\boldsymbol{A}_0$ and t are the same with GMML. Fig. 2 gives the influence of λ on EMGMML. In fact, λ controls the importance of the regularization term with respect to each learned metric $\boldsymbol{A}_p$. It is clear that the performance on ImageCLEF is sensitive to the choice of the parameter λ, while for other datasets the performance remains relatively stable. For DML-eig, we tune the parameter k in kNN from

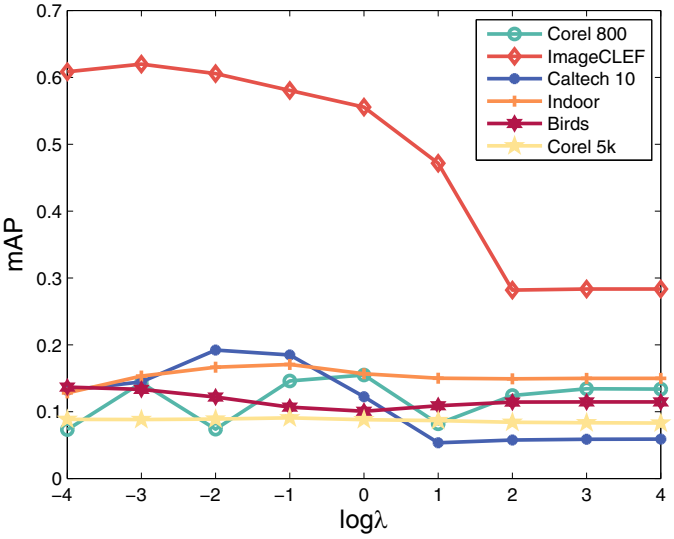


Fig. 2. Retrieval performance versus logλ on EMGMML. Other parameters are tuned to the best on the validation set.

Table 2
MAP of nine competing metrics for image retrieval. The best and the second best results are shown in bold and underlined, respectively.

Datasets						
Algorithm	Corel 800	ImageCLEF	Indoor	Caltech 10	Birds	Corel 5k
EMR	0.0419	0.0562	0.0434	0.0436	0.0498	0.0102
PCA+EMR	0.0419	0.0560	0.0434	0.0436	0.0498	0.0102
LRML	0.0649	0.1047	0.0649	0.0721	0.0607	0.0253
PCA+LRML	0.0705	0.1213	0.0673	0.0762	0.0677	0.0398
SERAPH	0.1205	0.1408	0.1319	0.1376	0.0965	0.0393
PCA+SERAPH	0.0621	0.0867	0.0512	0.0536	0.0577	0.0236
OASIS	0.0528	0.0500	0.0753	0.0324	0.0637	0.0156
PCA+OASIS	0.0274	0.0361	0.0316	0.0368	0.0473	0.0065
DML-eig	0.0429	0.0628	0.0551	0.0501	0.0538	0.0209
PCA+DML-eig	0.0513	0.0546	0.0528	0.0576	0.0558	0.0241
DCA	0.1281	0.3665	0.1183	0.0915	0.0626	0.0738
PCA+DCA	0.1363	0.3840	0.1006	0.1034	0.0661	0.0747
OMKS	**0.1373**	0.4372	0.1583	**0.2170**	0.1280	0.0504
GMML	0.1183	0.4288	0.1231	0.1611	0.0964	0.0628
PCA+GMML	0.1215	0.4399	0.1237	0.1620	0.0970	0.0680
UGMML	0.1088	0.4775	0.1198	0.1716	0.0960	0.0596
EMGMML	0.1337	**0.5492**	**0.1846**	0.2012	**0.1395**	**0.0916**

1 to the number of the labeled training images per class minus one [37]. As for LRML, we set the regularization parameters γ_s, γ_d as 1 and vary the parameter k of k-NN in 5–20 [16]. We set the number of the landmarks picked p in EMR as 50. In OMKS, there are three parameters to be tuned, that is, the Gaussian kernel parameter γ, discount weight β as well as the trade-off parameter C [26]. γ is tuned from 0.01 to 0.1 with 0.01 interval. β is adjusted in the range of 0–1 with 0.01 interval and C is tuned in [0.001, 0.01] with 0.001 interval.

4.5. Performance comparisons

We report the MAP values for all the competing methods in Table 2. Methods with 'PCA' as their prefixes indicate that we use PCA to reduce the dimension of original feature vectors to 200, and then perform retrieval with the corresponding metric. It can be seen that EMGMML consistently outperforms GMML in retrieval tasks. From top to down are unsupervised, semi-supervised and supervised methods. EMGMML improves the most on the Indoor dataset with an increasement about 49.96%. On Corel 800 dataset, the performance of OMKS is equivalent to that of EMGMML, perhaps owing to the capability of non-linear metrics for capturing subtle differences. While UGMML is sometimes inferior to GMML, for instance, on Corel 800, Indoor and Corel 5k datasets, our EMGMML achieves a great improvement due to the learned appropriate weights. In our method, multiple metrics and weights are jointly performed to achieve the optimality, thus yielding much better performance.

Fig. 3 presents the top-n (n = 1,2,...,5) precision results on two datasets. It is clear that OMKS and EMGMML show comparative performance on Caltech 10. However, EMGMML significantly outperforms all the other state-of-the-art metric learning algorithms on the Indoor dataset.

Fig. 4 shows the performance with respect to parameter t and γ on the validation set. In general, when t is relatively smaller and γ is comparatively larger, we obtain better retrieval performance. Actually, when t gets closer to 0, the learned metric for each modality $\boldsymbol{A}_p$ approaches $\boldsymbol{S}_p + \lambda \boldsymbol{A}_0^{-1}$. When γ is large, the regularization term works and each modality can contribute fully to the learning tasks. Among these datasets, ImageCLEF is more sensitive to these parameters, which is partly due to the fact that it is the only gray image dataset and thus much simpler.

Our EMGMML method learns weights for each modality, which represents its importance in learning metrics. Intuitively, the modality that has good performance should be assigned a large weight. To observe the correlation, we run the GMML method with each modality feature and then compare the results with its weight value. Fig. 5 shows the learned weights versus the mAP values of each modality with GMML on four datasets. From the plot, we observe that these two variables reveal positive correlation in general. We also utilize the correlation coefficient to examine the relations. The coefficient is 0.1721 on Corel 800, 0.2718 on Caltech 10, 0.4861 on ImageCLEF and 0.1027 on Indoor. In statistics, two variables are viewed as real correlated if their coefficient is between $\pm$ 0.3–$\pm$ 0.5, significantly related with coefficient in the range of $\pm$ 0.5–$\pm$ 0.8. According to the criterion, most of the learned weights for each modality can be regarded as positively related to its retrieval performance.

Table 3 lists the running time of each metric learning algorithm on datasets. The time is computed for fixed values of parameters tuned. It is clear that our EMGMML runs faster than GMML. Compared with GMML, the speed of EMGMML upgrades about 5 times on small datasets, i.e. Corel 800, Birds, Caltech 10 and so on. In fact, for color images d=2835, $d_{\max}$ = 768 and m = 9. Take these parameters into the time complexity expressions, we get the ratio 5.589, which is consistent with our experimental results. By comparison with UGMML, the quadratic programming only takes a few seconds. OASIS is substantially time-consuming and it takes about 5 hours. Considering the fact that its complexity grows rapidly with dimension, we conclude that it is not applicable to deal with high-dimensional data. Although the unsupervised metrics such as EMR reveal their superiority in efficiency due to the lack of training process, it is much more inferior with respect to effectiveness. In addition, it is noteworthy that EMGMML is more scalable in handling the large datasets, i.e. Corel 5k. However, the multiple kernel method OMKS takes a long time, almost 13 hours to converge in an iteration.

The experiments above are performed with a set of fixed labeled training data which accounts for 10% in the training set. In the following section, we discuss the influence of different labeling rates for EMGMML as well as GMML, UGMML and BGMML which outputs the best results of multiple modalities with GMML. Fig. 6 presents the retrieval results of different GMML-like metrics with various labeling rates on two datasets. It is clear our EMGMML significantly outperforms all the other metric learning methods under varied labeling rates, and BGMML follows next. From the results, we notice that as the labeling rate increases, different approaches reveal different trends. Specifically, EMGMML and BGMML achieve better performance with larger labeling rate, while GMML as well as UGMML does not. This may be partly due to the ignorance of

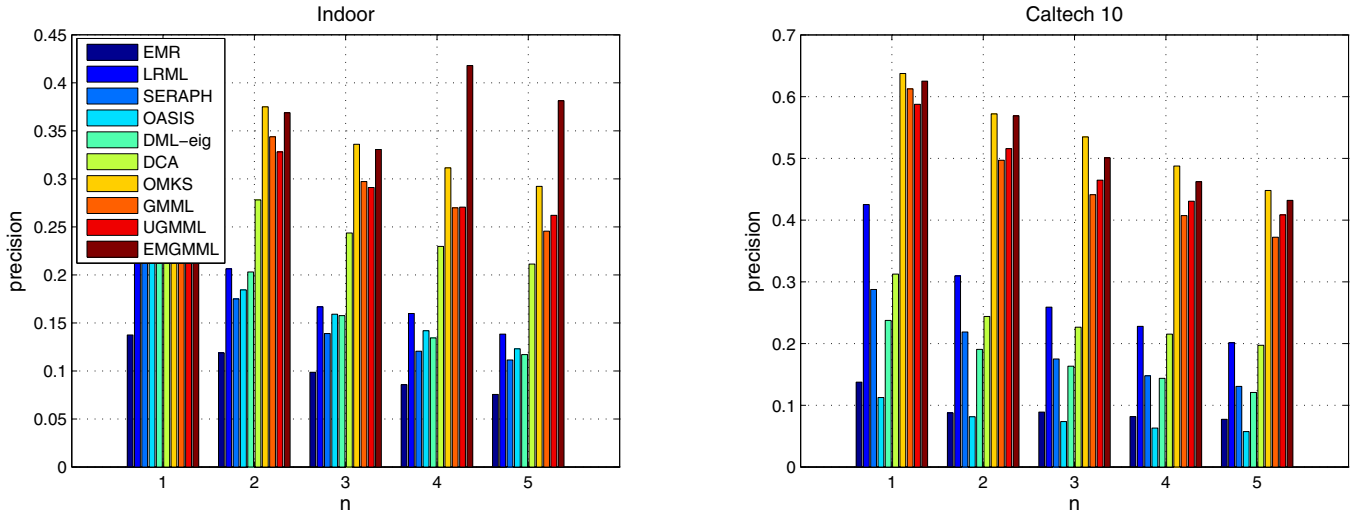
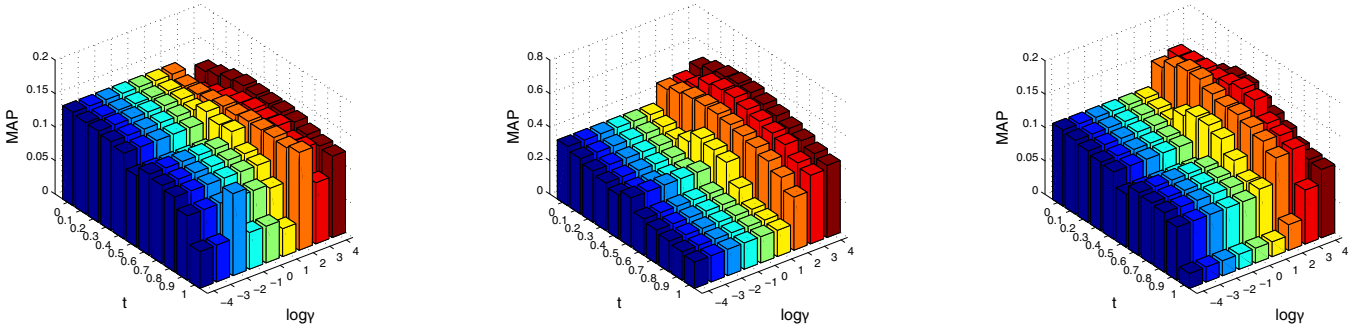
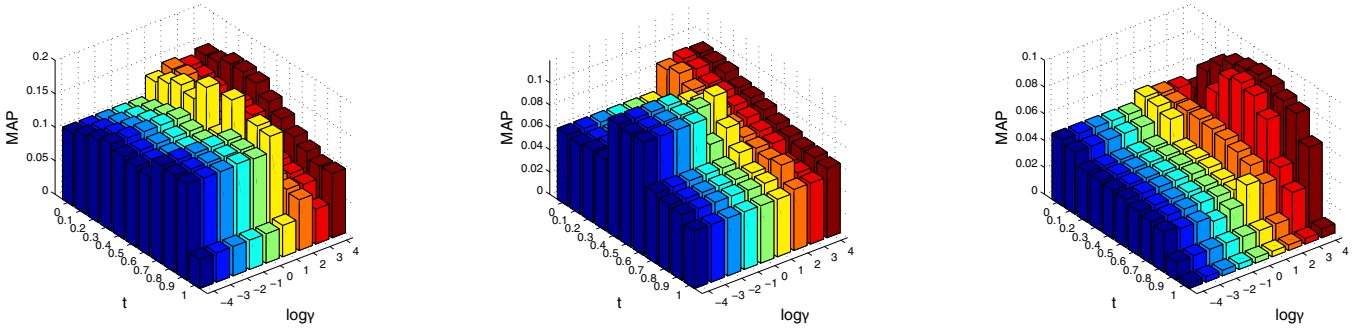


Fig. 3. Top-n precision results on Indoor and Caltech 10 datasets.



(a) Corel 800 (b) ImageCLEF (c) Indoor



(d) Caltech (e) Birds (f) Corel 5k

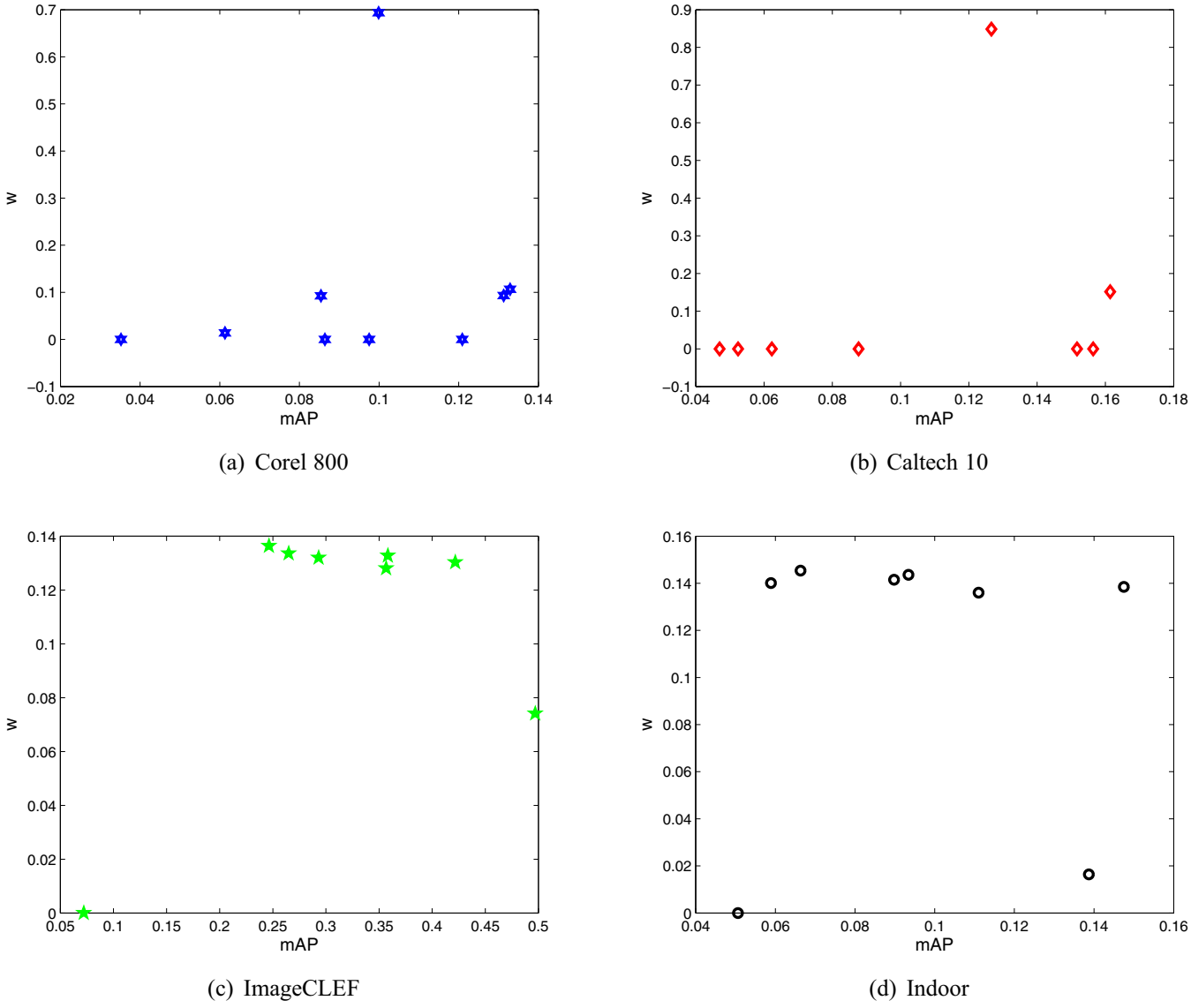
Fig. 4. Retrieval performance along with EMGMML in terms of t and γ.

complementarity between different modalities, as UGMML treats all of the modalities equally. As for GMML, although it learns metrics in a supervised manner, it handles multiple modalities as a single modality in a high-dimensional feature space, more labeled training data can not guarantee the performance improvement.

In the end, we randomly sample several query images and compare the top 5 ranked images retrieved with different metrics. Fig. 7 shows the qualitative comparisons of six different queries obtained by GMML and EMGMML. Generally, EMGMML retrieves more relevant images compared with GMML. For instance, for query 4, EMGMML obtained all of the 5 images, while GMML only obtained 2. This visual result clearly shows that EMGMML is much more effective than GMML in learning metrics for multiple modalities.

5. Conclusion and future work

We have introduced a general framework of multi-modal metric learning based on geometric mean metric learning to learn a metric for high-dimensional multi-modal data. Traditional metric learning approaches aim to learn a global linear metric, which is not appropriate for handing multiple modalities. In this study,



(a) Corel 800 (b) Caltech 10

(c) ImageCLEF (d) Indoor

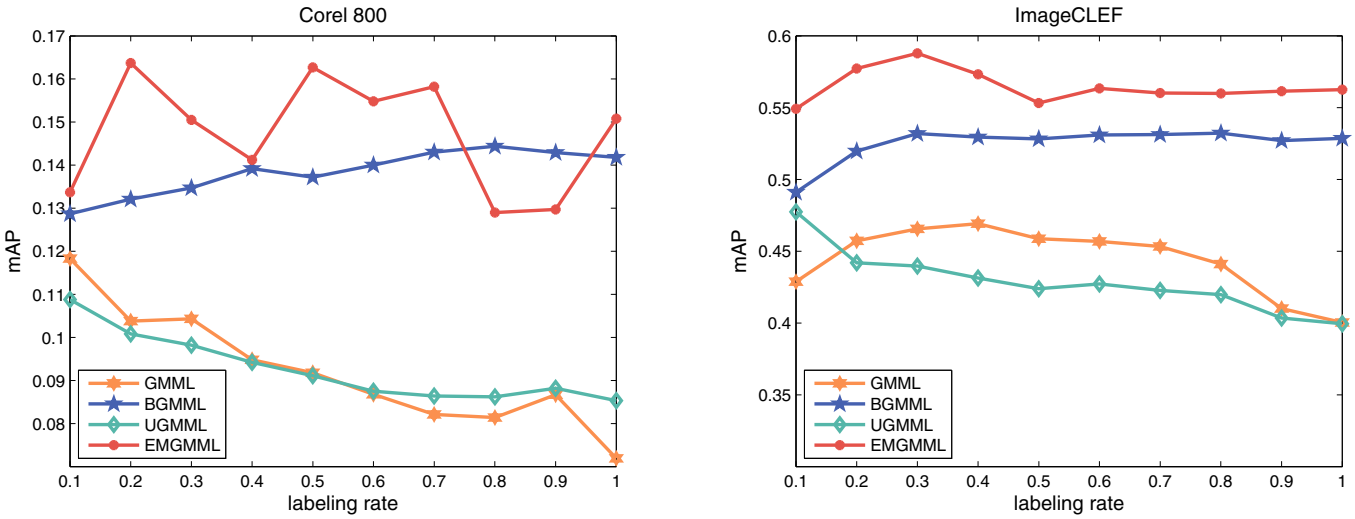
Fig. 5. Scatter plots of weights versus mAP values on Corel 800, Caltech 10, ImageCLEF and Indoor datasets.

Table 3
Time cost (seconds) of nine competing metrics for image retrieval. The best and the second best results are shown in bold and underlined, respectively.

Datasets						
Algorithm	Corel 800	ImageCLEF	Indoor	Caltech 10	Birds	Corel 5k
EMR	<u>3.10</u>	<u>2.02</u>	<u>2.48</u>	<u>2.57</u>	**2.07**	68.83
LRML	**1.87**	**1.37**	**1.88**	**1.85**	<u>1.94</u>	**3.97**
SERAPH	172.65	54.60	181.87	91.96	121.27	152.52
OASIS	18018.31	10249.82	17082.07	15078.70	20788.12	16877.20
DML-eig	78.90	42.45	31.48	57.73	40.64	43.32
DCA	7.82	5.10	9.16	8.22	5.58	99.44
OMKS	70.04	51.61	58.36	68.59	50.11	47567.19
GMML	25.17	14.30	28.32	27.97	25.82	25.81
UGMML	4.54	3.53	4.22	4.26	4.40	<u>9.16</u>
EMGMML	5.00	3.90	4.87	5.36	5.07	9.78

we have studied the potential of exploiting the consensus & complementarity properties among different modalities. The proposed method has the following advantages over most of existing methods: 1) the learned metric achieves excellent performance compared with the state-of-the-arts; 2) its time complexity is only related to the maximum dimension of the modalities rather than the entire dimension nor the sample size. Extensive experiments on image data for visual search demonstrate the excellent performance of our method.

In practical applications, only a small amount of data are labeled while the majority remain unlabelled. Therefore, how to make full use of these massive unlabelled data remains an open



(a) Corel 800 (b) ImageCLEF

Fig. 6. Evaluation of labeling rate on Corel 800 and ImageCLEF datasets.

Fig. 7. Examples of image retrieval on Corel 800, Caltech 10 and Birds from top to bottom by GMML (first row) and EMGMML (second row). "√" represents the images of the same class with the queries, and "× " represents the images from different classes.

problem. Moreover, the kernel technique which has shown advantages in mining complex patterns, has great potential for metric learning. In future work, we would like to consider geometric mean metric for semi-supervised and multiple kernel learning scenarios [38].

Acknowledgments

This work is partly supported by National Program on Key Basic Research Project under Grant 2013CB329304, National Natural Science Foundation of China under Grants 61432011, U1435212 and 61502332.

References

[1] K. Chang, K.W. Bowyer, S. Sarkar, B. Victor, Comparison and combination of ear and face images in appearance-based biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1160–1165.

[2] S. Sarkar, Z. Liu, Evaluation of Gait Recognition, in: Encyclopedia of Biometrics, Springer, 2009, pp. 281–289.

[3] Q. Zheng, A. Kumar, G. Pan, Suspecting less and doing better: new insights on palmprint identification for faster and more accurate matching, IEEE Trans. Inf. Forensics Secur. 11 (3) (2016) 633–641.

[4] A. Morales, A. Kumar, M.A. Ferrer, Interdigital palm region for biometric identification, Comput. Vision Image Understanding 142 (2016) 125–133.

[5] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, J. Zhou, Neighborhood repulsed metric learning for kinship verification, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 331–345.

[6] J. Lu, Y.-P. Tan, G. Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 39–51.

[7] J. Lu, G. Wang, P. Moulin, Localized multifeature metric learning for image-set-based face recognition, IEEE Trans. Circ. Syst. Video Technol. 26 (3) (2016) 529–540.

[8] B.Y. Li, M. Xue, A. Mian, W. Liu, A. Krishna, Robust RGB-D face recognition using kinect sensor, Neurocomputing 214 (2016) 93–108.

[9] I.A. Kakadiaris, G. Passalis, G. Toderici, M.N. Murtuza, T. Theoharis, 3d face recognition., in: BMVC, 2006, pp. 869–878.

[10] R. Wang, S. Shan, X. Chen, Q. Dai, W. Gao, Manifold-manifold distance and its application to face recognition with image sets, IEEE Trans. Image Process. 21 (10) (2012) 4466–4479.
[11] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[12] Y. Chen, J. Su, Sparse embedded dictionary learning on face recognition, Pattern Recognit. 64 (2017) 51–59.
[13] C. Hu, X. Lu, M. Ye, W. Zeng, Singular value decomposition and local near neighbors for face recognition under varying illumination, Pattern Recognit. 64 (2017) 60–83.
[14] F. Wang, B. Zhao, C. Zhang, Unsupervised large margin discriminative projection, IEEE Trans. Neural Netw. 22 (2011) 1446–1456.
[15] S.C. Hoi, W. Liu, M.R. Lyu, W.-Y. Ma, Learning distance metrics with contextual constraints for image retrieval, in: Proceedings of the Nineteenth IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2072–2078.
[16] S.C. Hoi, W. Liu, S.-F. Chang, Semi-supervised distance metric learning for collaborative image retrieval, in: Proceedings of the Twenty First IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.
[17] G. Niu, B. Dai, M. Yamada, M. Sugiyama, Information-theoretic semi-supervised metric learning via entropy regularization, Neural Comput. 26 (8) (2014) 1717–1762.
[18] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[19] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[20] Y. Mu, W. Ding, D. Tao, Local discriminative distance metrics ensemble learning, Pattern Recognit. 46 (8) (2013) 2337–2349.
[21] A. Bellet, A. Habrard, M. Sebban, A survey on metric learning for feature vectors and structured data, Technical Report, arXiv:1306.6709 (2013).
[22] J. Hu, J. Lu, J. Yuan, Y.-P. Tan, Large margin multi-metric learning for face and kinship verification in the wild, in: Asian Conference on Computer Vision, Springer, 2014, pp. 252–267.
[23] B. McFee, G. Lanckriet, Learning multi-modal similarity, J. Mach. Learn. Res. 12 (2) (2011) 491–523.
[24] X. Lu, Y. Wang, X. Zhou, Z. Ling, A method for metric learning with multiple-kernel embedding, Neural Process. Lett. 43 (3) (2016) 905–921.
[25] J. Liang, Q. Hu, W. Wang, Y. Han, Semi-supervised online multi-kernel similarity learning for image retrieval, IEEE Trans. Multimedia (2016).
[26] H. Xia, S.C. Hoi, R. Jin, P. Zhao, Online multiple kernel similarity learning for visual search, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 536–549.
[27] P. Wu, S.C. Hoi, P. Zhao, C. Miao, Z.-Y. Liu, Online multi-modal distance metric learning with application to image retrieval, IEEE Trans. Knowl. Data Eng. 28 (2) (2016) 454–467.
[28] P.H. Zadeh, R. Hosseini, S. Sra, Geometric mean metric mearning, in: Proceedings of the Thirty Third International Conference on Machine Learning, 2016.
[29] A. Papadopoulos, Metric spaces, convexity and nonpositive curvature, Eur. Math. Soc., 2005.
[30] S. Sra, R. Hosseini, Conic geometric optimization on the manifold of positive definite matrices, SIAM J. Optim. 25 (1) (2015) 713–739.
[31] R. Bhatia, Positive Definite Matrices, Princeton university press, 2009.
[32] S. Lazebnik, C. Schmid, J. Ponce, A maximum entropy framework for part-based texture and object recognition, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, 1, IEEE, 2005, pp. 832–838.
[33] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell, Distance metric learning with application to clustering with side-information, Adv. Neural Inf. Process Syst. 15 (2003) 505–512.
[34] Y. Zhen, D.-Y. Yeung, A probabilistic model for multimodal hash function learning, in: Proceedings of the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 940–948.
[35] G. Chechik, V. Sharma, U. Shalit, S. Bengio, Large scale online learning of image similarity through ranking, J. Mach. Learn. Res. 11 (3) (2010) 1109–1135.
[36] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, J. Luo, Efficient manifold ranking for image retrieval, in: Proceedings of the Thirty Fourth International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2011, pp. 525–534.
[37] Y. Ying, P. Li, Distance metric learning with eigenvalue optimization, J. Mach. Learn. Res. 13 (1) (2012) 1–26.
[38] J. Liang, Y. Han, Q. Hu, Semi-supervised image clustering with multi-modal information, Multimedia Syst. 22 (2) (2016) 149–160.

Jianqing Liang received the B.E. degree from the School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi, China, in 2013. She is currently pursuing the Ph.D. degree with the College of Computer Science and Technology in Tianjin University. Her current research interests include metric learning, semi-supervised learning and machine learning.

Qinghua Hu received B. E., M. E. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China in 1999, 2002 and 2008, respectively. He once worked with Harbin Institute of Technology as assistant professor and associate professor from 2006 to 2011 and a postdoctoral fellow with the Hong Kong Polytechnic University. He is now a full professor with Tianjin University. His research interests are focused on intelligent modeling, data mining, knowledge discovery for classification and regression. He is the PC co-chair of RSCTC 2010, CRSSC 2012, and ICMLC 2014 and severs as referee for a great number of journals and conferences. He has published more than 100 journal and conference papers in the areas of pattern recognition, machine learning and data mining.

Pengfei Zhu received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin, China, in 2009 and 2011, respectively, and the Ph.D. degree from Hong Kong Polytechnic University, Hong Kong, in 2014. He is currently an associate professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. His current research interests include machine learning and computer vision.

Wenwu Wang received the B.Sc. degree in automatic control in 1997, the M.E. degree in control science and control engineering in 2000, and the Ph.D. degree in navigation guidance and control in 2002, all from Harbin Engineering University, Harbin, China. He then joined Kings College, London, U.K., in May 2002, as a postdoctoral research associate and transferred to Cardiff University, Cardiff, U.K., in January 2004, where he worked in the area of blind signal processing. In May 2005, he joined the Tao Group Ltd. (now Antix Labs Ltd.), Reading, U.K., as a DSP engineer working on algorithm design and implementation for real-time and embedded audio and visual systems. In September 2006, he joined Creative Labs, Ltd., Egham, U.K., as an engineer, working on 3D spatial audio for mobile devices. Since May 2007, he has been with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., where he is currently a Reader in Signal Processing, and a co-director of the Machine Audition Lab. He is a member of the Ministry of Defence (MoD) University Defence Research Collaboration (UDRC) in Signal Processing (since 2009), a member of the BBC Audio Research Partnership (since 2011), an associate member of Surrey Centre for Cyber Security (since 2014), and a member of the MRC/EPSRC Microphone Network (since 2015). During spring 2008, he has been a visiting scholar at the Perception and Neurodynamics Lab and the Center for Cognitive Science, The Ohio State University. His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 150 publications in these areas, including two books Machine Audition: Principles, Algorithms and Systems (IGI Global, 2010) and Blind Source Separation: Advances in Theory, Algorithms and Applications (Springer, 2014). He is currently an associate editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is also publication co-chair of ICASSP 2019 (to be held in Brighton, UK). He was a tutorial speaker on ICASSP 2013, UDRC Summer School 2014, 2015 and 2016, and SpaRTan/MacSeNet Spring School 2016.