Empirical Risk Minimization for Metric Learning Using Privileged Information

Xun Yang,[†] **Meng Wang,**[†] **Luming Zhang,**[†] and **Dacheng Tao**[‡]
[†]School of Computer and Information, Hefei University of Technology, China
[‡]Centre for Quantum Computation & Intelligent Systems, FEIT, University of Technology Sydney, Australia
{hfutyangxun, eric.mengwang, zglumg}@gmail.com;
dacheng.tao@uts.edu.au;

Abstract

Traditional metric learning methods usually make decisions based on a fixed threshold, which may result in a suboptimal metric when the inter-class and inner-class variations are complex. To address this issue, in this paper we propose an effective metric learning method by exploiting privileged information to relax the fixed threshold under the empirical risk minimization framework. Privileged information describes useful high-level semantic information that is only available during training. Our goal is to improve the performance by incorporating privileged information to design a locally adaptive decision function. We jointly learn two distance metrics by minimizing the empirical loss penalizing the difference between the distance in the original space and that in the privileged space. The distance in the privileged space functions as a locally adaptive decision threshold, which can guide the decision making like a *teacher*. We optimize the objective function using the Accelerated Proximal Gradient approach to obtain a global optimum solution. Experiment results show that by leveraging privileged information, our proposed method can achieve satisfactory performance.

1 Introduction

Learning a suitable distance metric from the given training instances plays an important role in many machine learning and computer vision tasks. Over the past decades, several distance metric learning (DML) methods have been proposed, e.g., information-theoretic metric leaning (ITML) [Davis *et al.*, 2007], logistic discriminant metric learning (LDML) [Guillaumin *et al.*, 2009], and discriminative deep metric learning [Hu *et al.*, 2014]. Existing algorithms for metric learning have been shown to perform well empirically on a variety of applications, e.g., classification and clustering. While, most of them restrict the distance between a pair of similar/dissimilar instances to be lower/higher than a fixed threshold. Such fixed threshold based constraints may suffer from suboptimal performance when coping with some real world tasks with complex inter-class and intra-class variations (See Figure 1).

A natural solution to alleviate the limitations of fixed threshold based DML method is to design a locally adaptive decision rule. [Li *et al.*, 2013] proposed to learn a second-order local decision function in the original feature space to replace the fixed threshold. [Wang *et al.*, 2014] introduced an adaptive shrinkage-expansion rule to shrink/expand the Euclidean distance as an adaptive threshold. These two earlier works both leverage the information from the original feature space to guild the decision making. However, the guidance from the original feature space might be relatively weak, since original feature is usually noisy and less discriminative. Another way is to incorporate additional knowledge beyond the original space.

It has been shown in [Vapnik and Izmailov, 2015] that a more reliable and effective model can be learned if some high-level additional knowledge is exploited during the training stage. Such high-level knowledge is called privileged information and is only available in training stage. It typically describes some important semantic properties of the training instance, such as the attributes, tags, textual descriptions or other high-level knowledge. This idea of privileged information is inspired by the human teaching-learning in which the students will learn better if a teacher can provide some explanations, comments, comparison or other supervision. It was first incorporated into SVM in the form of SVM+ [Vapnik and Vashist, 2009] and has been utilized in object localization [Feyereisl *et al.*, 2014] and image categorization [Li *et al.*, 2014] in recent years.

Motivated by [Vapnik and Izmailov, 2015], in this paper we develop a new DML method using privileged information under the generic empirical risk minimization (ERM) framework, termed as ERMML+. First, we represent each training instance with two forms of feature representations: one is original feature representation and the other is privileged information representation. Thereby an emerging problem is how to exploit privileged information for metric learning. Generally, there exists a semantic gap between the original space and the privileged space, since the original feature is at the low-level space while privileged information is a high-level semantic knowledge. To bridge the semantic gap, we jointly learn two Mahalanobis distance metrics with positive semi-definite (PSD) constraints by minimizing the empirical loss penalizing the difference between the distance in the original space and the distance in the privileged information

space. The distance in the privileged space functions as a local decision threshold to guide the metric learning like a *teacher*, which can effectively relax the fixed threshold. We optimize the objective function using the Accelerated Proximal Gradient (APG) approach to search a global optimum solution with a fast convergence rate. We have evaluated the proposed method on two real world problems: person re-identification and face verification. Experiment results show that by leveraging the privileged information, our method outperforms not only the classical metric learning algorithms, but also the state-of-the-art methods in the computer vision community.

2 Related Work

During the past decade, many algorithms have been developed to learn a Mahalanobis distance metric, e.g., [Weinberger and Saul, 2009; Davis *et al.*, 2007; Guillaumin *et al.*, 2009; Bian and Tao, 2012]. Here we only briefly review several most relevant works.

[Guillaumin *et al.*, 2009] proposed a logistic discriminant metric learning (LDML) method which is related to our work, but LDML doesn't use any regularization term including the PSD constraint, which easily suffers from overfitting problem. The PSD constraint can provide a useful regularization to smooth the solution of the metric. [Bian and Tao, 2012] developed a loss minimization framework for metric learning, which is quite rigid since it relies on a strong assumption that the learned metric is bounded. Besides, [Guillaumin *et al.*, 2009; Bian and Tao, 2012] both adopt the fixed threshold based constraints. Compared with them, we learn a PSD metric by exploiting the privileged information to construct a local decision function, which is more robust and shows better performance.

Recently, [Fouad *et al.*, 2013] proposed a two stage strategy to exploit privileged information for metric learning based on ITML. They first learn a metric using the privileged information to remove some outlier pairs and then use the remaining pairs to learn a metric based on the original feature. Following [Fouad *et al.*, 2013], [Xu *et al.*, 2015] proposed a ITML+ method, in which privileged information is used to design a slack function to replace the slack variables in ITML. Compared with two ITML based methods [Fouad *et al.*, 2013; Xu *et al.*, 2015], we provide a new scheme to leverage privileged knowledge for distance metric learning under the generic ERM framework, which has good statistical property and can be extended easily by incorporating different loss functions and regularization terms. Moreover, we apply low rank selection for the learned metric in each iteration, which allows us to work directly with higher dimensional input data. While ITML based methods aim to learn a full matrix for the target distance metric that is in the square of the dimensionality, making it computationally unattractive for high dimensional data and prone to overfitting [Mignon and Jurie, 2012]. It is shown in the experiments that our method ERMML+ performs better than the ITML+ method.

Figure 1: Two examples from the VIPeR dataset where $\mathbf{M}$ is dominated by $\mathbf{M}^+$. Note that $\mathbf{M}$ is the learned metric based on a fixed decision threshold (only original feature is used), while $\mathbf{M}^+$ is learned by incorporating privileged information to design a local decision function. Here, privileged information denotes the pedestrian attributes such as gender, age, short/long hair, cloth color, etc. (a, b) and (d, e) are two similar pairs but with large inner-class variations. (b, c) and (e, f) are two dissimilar pairs while sharing large inter-class similarities.

3 ERM for DML

Assume we have a pairwise constrained training instances set $\mathcal{Z} = \{(\mathbf{x}_i, \mathbf{z}_i, \ell_i) \,|i = 1, 2, \ldots, n\}$, where $\mathbf{x}_i \in \mathbb{R}^d, \mathbf{z}_i \in \mathbb{R}^d$ are both defined on the same space, i is the index of the i-th pair of training instance, and ℓ_i is the label of the pair $(\mathbf{x}_i, \mathbf{z}_i)$ defined by

$$\ell_i = \begin{cases} 1, & \text{if}(\mathbf{x}_i, \mathbf{z}_i) \in \mathcal{S} \\ -1, & \text{if}(\mathbf{x}_i, \mathbf{z}_i) \in \mathcal{D}, \end{cases} \tag{1}$$

where $\mathcal{S}$ denotes the set of similar pairs and $\mathcal{D}$ denotes the set of dissimilar pairs. The goal of DML is to learn a Mahalanobis distance metric defined by

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{z}_i) = \sqrt{(\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{M} (\mathbf{x}_i - \mathbf{z}_i)}, \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{d \times d}$ is the learned PSD metric. The learned Mahalanobis distance $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{z}_i)$ is expected to be small if $\mathbf{x}_i$ and $\mathbf{z}_i$ are similar, or large if they are dissimilar.

Given a metric $\mathbf{M}$, how to determine whether two instances are similar or dissimilar? A common way is to compare their distance with a fixed decision threshold σ [Mignon and Jurie, 2012], then the decision function f can be defined by

$$f(\mathbf{x}_i, \mathbf{z}_i; \mathbf{M}) = \sigma - (\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{M} (\mathbf{x}_i - \mathbf{z}_i). \tag{3}$$

If they are similar, the decision function $f > 0$, otherwise $f < 0$.

The problem of DML can be cast in the generic ERM framework by minimizing the empirical risk $E(\mathbf{M})$

$$\min_{\mathbf{M} \succeq 0} E(\mathbf{M}) = \frac{1}{n} \sum_{i=1}^{n} L\big(\ell_i f(\mathbf{x}_i, \mathbf{z}_i; \mathbf{M})\big), \tag{4}$$

where $L(\cdot)$ is a convex loss function (decrease progressively), e.g., log loss and smooth hinge loss. Previous DML methods [Guillaumin *et al.*, 2009] and [Mignon and Jurie, 2012] are both under this framework.

4 ERM for DML using Privileged information

4.1 Problem Formulation

Traditional pairwise constrained DML methods usually adopt the fixed threshold based decision function f, which is too rough to obtain a reasonable metric. In this paper, we aim to design a locally adaptive decision rule to alleviate the limitations of fixed threshold based methods.

Motivated by the [Vapnik and Izmailov, 2015], we exploit privileged information to design an adaptive decision function in the training stage. First, each training instance is represented with two forms of feature representations: one is $\mathbf{x}_i \in \mathbb{R}^d$ from the original feature space; the other is $\mathbf{x}_i^* \in \mathbb{R}^{d^*}$ from the privileged space. The training set is reformulated as $\mathcal{Z} = \{(\mathbf{x}_i, \mathbf{x}_i^*, \mathbf{z}_i, \mathbf{z}_i^*, \ell_i) \,|i = 1, 2, \ldots, n\}$. Then, we replace the fixed threshold σ in (3) using the squared distance $d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*)$, where $\mathbf{P} \in \mathbb{R}^{d^* \times d^*}$ is the distance metric matrix in the privileged information space. $d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*)$ functions as the adaptive decision threshold. Our locally adaptive decision function is formulated by

$$\begin{aligned} &f(\mathbf{x}_i, \mathbf{z}_i; \mathbf{x}_i^*, \mathbf{z}_i^*; \mathbf{M}, \mathbf{P}) \\ &= d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*) - d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{z}_i) \\ &= (\mathbf{x}_i^* - \mathbf{z}_i^*)^T \mathbf{P} (\mathbf{x}_i^* - \mathbf{z}_i^*) - (\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{M} (\mathbf{x}_i - \mathbf{z}_i) \end{aligned} . \tag{5}$$

To be simplified, we rewrite the decision function as

$$f(\mathbf{A}_i, \mathbf{S}) = \mathrm{tr}(\mathbf{S}^T \mathbf{A}_i) = \langle \mathbf{A}_i, \mathbf{S} \rangle, \tag{6}$$

by introducing two block-diagonal matrices $\mathbf{A}_i$ and $\mathbf{S}$. $\langle \cdot, \cdot \rangle$ denotes the matrix inner product. $\mathbf{A}_i \in \mathbb{R}^{(d+d^*)\times(d+d^*)}$ and $\mathbf{S} \in \mathbb{R}^{(d+d^*)\times(d+d^*)}$ are defined by

$$\mathbf{A}_i = \mathrm{diag}\Big((\mathbf{x}_i^* - \mathbf{z}_i^*)(\mathbf{x}_i^* - \mathbf{z}_i^*)^T, -(\mathbf{x}_i - \mathbf{z}_i)(\mathbf{x}_i - \mathbf{z}_i)^T \Big), \tag{7}$$

$$\mathbf{S} = \mathrm{diag}(\mathbf{P}, \mathbf{M}) = \begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{M} \end{pmatrix}. \tag{8}$$

We also take the severe imbalance of similar/dissimilar pairs into consideration by weighting each pair in our empirical risk function. Combining the new decision function f and the pair weight w, our problem can be formulated as

$$\hat{\mathbf{S}} = \arg\min_{\mathbf{S} \in \mathcal{Q}} E(\mathbf{S}) = \sum_{i=1}^{n} w_i L\big(\ell_i f(\mathbf{A}_i, \mathbf{S})\big), \tag{9}$$

where $\mathcal{Q} = \{\mathbf{M} \succeq 0; \mathbf{P} \succeq 0\}$, w_i is the weight of the similar/dissimilar pairs. If $\ell_i = 1$ (-1), w_i is defined by the reciprocal of the number of similar (dissimilar) pairs. Compared with ITML+, our formulation has a general form that follows the ERM framework.

In our model, $d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*)$ can effectively guide the decision making in the original feature space like a *teacher* by controlling *student*'s concept of similarity between training instances. See Figure 1 for two representative examples from the VIPeR dataset. In Figure 1, (a, b) and (d, e) are two similar pairs but with large inner-class variations; (b, c) and (e, f) are two dissimilar pairs while sharing large inter-class similarities. We can see that the learned metric $\mathbf{M}$ results in a suboptimal metric when only uses original feature for training, while the metric $\mathbf{M}^+$ performs better by exploiting the privileged information in the training stage. Note that we don't leverage privileged information in the testing stage.

4.2 Optimization using APG

In this paper, we use the well known log loss function as an example,[1] which is defined by $L(u) = \ln(1 + \exp(-u))$. Hence, the objective function in (9) is rewritten as

$$\min_{\mathbf{S} \in \mathcal{Q}} E(\mathbf{S}) = \sum_{i=1}^{n} w_i \ln\big(1 + e^{-\ell_i f(\mathbf{A}_i, \mathbf{S})}\big). \tag{10}$$

Then, we exploit the APG method [Nesterov, 2004] to solve the optimization problem. It has been proved that APG method can achieve the optimal convergence rate at $\mathcal{O}(1/k^2)$, where k is the number of iteration steps. However, APG method requires the condition that the objective function is Lipschitz continuous with a Lipschitz constant $\mathcal{L}$. According to Property 1, we know that $E(\mathbf{S})$ in (10) has a Lipschitz-continuous gradient with a Lipschitz constant.

Property 1 *Given any direction $\Delta \in \mathbb{R}^{(d+d^*)\times(d+d^*)}$, the empirical risk $E(\mathbf{S})$ satisfies*

$$\big\langle \nabla^2 E(\mathbf{S})\Delta, \Delta \big\rangle \le \mathcal{L} \|\Delta\|_F^2, \tag{11}$$

where $\mathcal{L} = \frac{1}{4} \sum_{i=1}^{n} w_i \|\mathbf{A}_i\|_F^2$, and $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. The gradient of $E(\mathbf{S})$ is computed by

$$\nabla E(\mathbf{S}) = \sum_{i=1}^{n} \frac{-w_i \ell_i \mathbf{A}_i}{1 + e^{\ell_i \langle \mathbf{A}_i, \mathbf{S} \rangle}}. \tag{12}$$

Define function as $\phi(\varrho) = \langle \nabla E(\mathbf{S} + \varrho\Delta), \Delta \rangle$ with $\varrho > 0$, and we have

$$\begin{aligned} &\phi(\varrho) - \phi(0) = \langle \nabla E(\mathbf{S} + \varrho\Delta) - \nabla E(\mathbf{S}), \Delta \rangle \\ &= \sum_{i=1}^{n} \Big\langle \frac{-w_i \ell_i \mathbf{A}_i \left(1 - e^{\varrho \ell_i \langle \mathbf{A}_i, \Delta \rangle}\right)}{\left(1 + e^{\ell_i \langle \mathbf{A}_i, \mathbf{S} + \varrho\Delta \rangle}\right)\left(1 + e^{-\ell_i \langle \mathbf{A}_i, \mathbf{S} \rangle}\right)}, \Delta \Big\rangle. \end{aligned} \tag{13}$$

Hence,

$$\begin{aligned} &\big\langle \nabla^2 E(\mathbf{S})\Delta, \Delta \big\rangle = \phi'(0) = \lim_{\varrho \to 0} \frac{\phi(\varrho) - \phi(0)}{\varrho} \\ &= \sum_{i=1}^{n} \frac{w_i \ell_i^2 \langle \mathbf{A}_i, \Delta \rangle^2}{\left(1 + e^{\ell_i \langle \mathbf{A}_i, \mathbf{S} \rangle}\right)\left(1 + e^{-\ell_i \langle \mathbf{A}_i, \mathbf{S} \rangle}\right)} \\ &\le \sum_{i=1}^{n} \frac{w_i \ell_i^2 \langle \mathbf{A}_i, \Delta \rangle^2}{4} \le \frac{1}{4} \sum_{i=1}^{n} w_i \|\mathbf{A}_i\|_F^2 \|\Delta\|_F^2 \end{aligned} . \tag{14}$$

We can see that the log loss based $E(\mathbf{S})$ has a Lipschitz gradient with a constant $\mathcal{L} = \frac{1}{4} \sum_{i=1}^{n} w_i \|\mathbf{A}_i\|_F^2$.

This completes the proof.

Below, we detail the optimization procedure at step k.

Suppose the latest two approximate solutions $\mathbf{S}_{k-1}$ and $\mathbf{S}_{k-2}$ are known, we can construct the search point $\mathbf{Z}_k$ at current step as a linear combination of the latest two approximate solutions

$$\mathbf{Z}_k = \mathbf{S}_{k-1} + \left(\frac{\alpha_{k-1} - 1}{\alpha_k}\right)(\mathbf{S}_{k-1} - \mathbf{S}_{k-2}), \tag{15}$$

[1] Due to the space limits, we only use one example for analysis. Other convex loss functions, e.g., smooth hinge loss and squared loss, can also be used in our model.

Algorithm 1 ERMML+

Input: The training set $(\mathbf{x}_i, \mathbf{x}_i^*, \mathbf{z}_i, \mathbf{z}_i^*, \ell_i)$, $i = 1, 2, \ldots, n$.
Output: $\mathbf{S} = \mathrm{diag}(\mathbf{P}, \mathbf{M})$, where $\mathbf{M}$ is the learned metric.
Initialize: L_0, γ, ϵ, α_0, $\mathbf{M}_0 = \mathbf{M}_{-1} = \mathbf{I}^{d\times d}$, $\mathbf{P}_0 = \mathbf{P}_{-1} = \mathbf{I}^{d^*\times d^*}$.
Iterative: $k = 1, 2, \cdots$

1: Set $\alpha_k = \frac{1}{2}\left(1 + \sqrt{1 + 4\alpha_{k-1}^2}\right)$.
2: Compute $\mathbf{Z}_k$ by (15).
3: Compute $E(\mathbf{Z}_k)$ and $\nabla E(\mathbf{Z}_k)$ with a loss function.
4: Set $t_k = 1/\mathcal{L}_{k-1}$.
5: Obtain $\mathbf{S}_k$ by solving (18) using (22) and (23).
6: If the condition in (24) is not satisfied, set $\mathcal{L}_{k-1} = \gamma\mathcal{L}_{k-1}$ and return to step 4.
7: Set $\mathcal{L}_k = \mathcal{L}_{k-1}$

Until $|E(\mathbf{S}_k) - E(\mathbf{S}_{k-1})| < \epsilon$.

where $\alpha_k = \frac{1}{2}\left(1 + \sqrt{1 + 4\alpha_{k-1}^2}\right)$. The problem can be reformulated equivalently as a proximal regularization of the linearized function $E(\mathbf{S})$ at $\mathbf{Z}_k$ as

$$\begin{aligned}\mathbf{S}_k &= \arg\min_{\mathbf{S}\in\mathcal{Q}} Q_{t_k}(\mathbf{S}, \mathbf{Z}_k) \\ &= \arg\min_{\mathbf{S}\in\mathcal{Q}} E(\mathbf{Z}_k) + \langle \mathbf{S} - \mathbf{Z}_k, \nabla E(\mathbf{Z}_k)\rangle + \frac{1}{2t_k}\|\mathbf{S} - \mathbf{Z}_k\|_F^2,\end{aligned} \tag{16}$$

where t_k is the step size of the gradient method. The gradient $\nabla E(\mathbf{Z}_k)$ is block-diagonal

$$\nabla E(\mathbf{Z}_k) = \mathrm{diag}\left(\nabla E_{\mathbf{P}}(\mathbf{Z}_k), \nabla E_{\mathbf{M}}(\mathbf{Z}_k)\right). \tag{17}$$

By ignoring the constant term $E(\mathbf{Z}_k)$ and adding another constant term $\frac{t_k}{2}\|\nabla E(\mathbf{Z}_k)\|_F^2$, the problem in (16) can be expressed equivalently as

$$\begin{aligned}\mathbf{S}_k &= \arg\min_{\mathbf{S}\in\mathcal{Q}} \frac{1}{2t_k}\|\mathbf{S} - (\mathbf{Z}_k - t_k\nabla E(\mathbf{Z}_k))\|_F^2 \\ &= \arg\min_{\mathbf{S}\in\mathcal{Q}} \frac{1}{2t_k}\|\mathbf{S} - \mathbf{H}_k\|_F^2,\end{aligned} \tag{18}$$

where $\mathbf{H}_k$ is a block-diagonal matrix and it can be rewrote as

$$\mathbf{H}_k = \mathrm{diag}\left(\mathbf{H}_k^{\mathbf{P}}, \mathbf{H}_k^{\mathbf{M}}\right), \tag{19}$$

where

$$\begin{aligned}\mathbf{H}_k^{\mathbf{P}} &= \mathbf{P}_{k-1} + \frac{\alpha_{k-1}-1}{\alpha_k}(\mathbf{P}_{k-1} - \mathbf{P}_{k-2}) - t_k\nabla E_{\mathbf{P}}(\mathbf{Z}_k), \\ \mathbf{H}_k^{\mathbf{M}} &= \mathbf{M}_{k-1} + \frac{\alpha_{k-1}-1}{\alpha_k}(\mathbf{M}_{k-1} - \mathbf{M}_{k-2}) - t_k\nabla E_{\mathbf{M}}(\mathbf{Z}_k).\end{aligned} \tag{20}$$

Since the two diagonal blocks $\mathbf{M}$ and $\mathbf{P}$ in $\mathbf{S}$ are not coupled, $\mathbf{M}_k$ and $\mathbf{P}_k$ can be obtained independently. We can compute $\mathbf{M}_k$ and $\mathbf{P}_k$ by projecting $\mathbf{H}_k^{\mathbf{M}}$ and $\mathbf{H}_k^{\mathbf{P}}$ into the positive semi-definite cone, respectively. We take $\mathbf{M}_k$ as an example. To obtain $\mathbf{M}_k$, we conduct singular value decomposition (SVD) on $\mathbf{H}_k^{\mathbf{M}}$ and then we have

$$\mathbf{H}_k^{\mathbf{M}} = \mathbf{U}^{\mathbf{M}}\mathbf{\Lambda}^{\mathbf{M}}\left(\mathbf{U}^{\mathbf{M}}\right)^T, \tag{21}$$

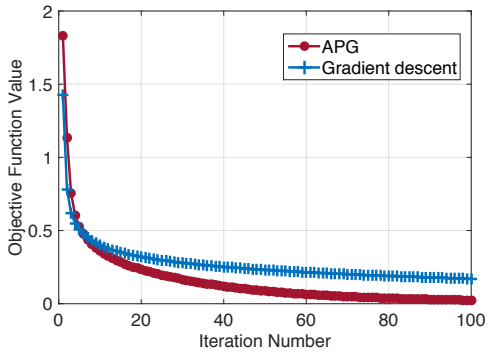

Figure 2: Convergence curve of our proposed method with two kinds of optimization algorithms. The red one is the APG method and the blue one is the standard gradient descent method.

where $\mathbf{U}^{\mathbf{M}}$ and $\mathbf{\Lambda}^{\mathbf{M}}$ are obtained from the SVD. $\mathbf{\Lambda}_{\mathbf{M}}$ is a diagonal matrix that contains all the singular values of $\mathbf{H}_k^{\mathbf{M}}$. Since $\mathbf{M} \succeq 0$, we can obtain $\mathbf{M}_k$ by

$$\mathbf{M}_k = \mathbf{U}^{\mathbf{M}}\mathbf{\Lambda}_+^{\mathbf{M}}\left(\mathbf{U}^{\mathbf{M}}\right)^T, \tag{22}$$

where $\mathbf{\Lambda}_+^{\mathbf{M}}$ is diagonal with $\left(\mathbf{\Lambda}_+^{\mathbf{M}}\right)_{ii} = \max\left\{0, \mathbf{\Lambda}_{ii}^{\mathbf{M}}\right\}$. $\mathbf{P}_k$ can be obtained in the same way by

$$\mathbf{P}_k = \mathbf{U}^{\mathbf{P}}\mathbf{\Lambda}_+^{\mathbf{P}}\left(\mathbf{U}^{\mathbf{P}}\right)^T. \tag{23}$$

Step Size Estimation

It's vital to choose an appropriate step size t_k for APG algorithm. In [Nesterov, 2004], the reciprocal $\frac{1}{\mathcal{L}}$ of Lipschitz constant functions as the step size . However, it is too conservative to set $t_k = \frac{1}{\mathcal{L}}$ for all k and it is also very time consuming to compute the Lipschitz constant directly.

Following [Beck and Teboulle, 2009], we first initialize $\mathcal{L}$ with a small value $\mathcal{L}_0$ and increase this estimate with a multiplicative factor γ ($\gamma > 1$) repeatedly until the following condition is satisfied

$$E(\mathbf{S}_k) \le Q_{t_k}(\mathbf{S}_k, \mathbf{Z}_k), \tag{24}$$

This procedure ensures that the step size is suitable. The process of the optimization is summarized in Algorithm 1.

Convergence Analysis

We show in the following theorem that by perform the APG method, the proposed method can achieve the optimal convergence rate at $\mathcal{O}(1/k^2)$.

Theorem 1 *Let $\{\mathbf{S}_k\}$ be generated by the Algorithm 1. Then for any $k > 1$ we have*

$$E(\mathbf{S}_k) - E(\hat{\mathbf{S}}) \le \frac{2\gamma\mathcal{L}\|\mathbf{S}_0 - \hat{\mathbf{S}}\|_F^2}{(k+1)^2}, \tag{25}$$

where $\hat{\mathbf{S}}$ is the optimal solution.

The proof of this theorem can follow the same strategy as in [Beck and Teboulle, 2009]. We omit the proof here for the space limits. We empirically verify the convergence rate of

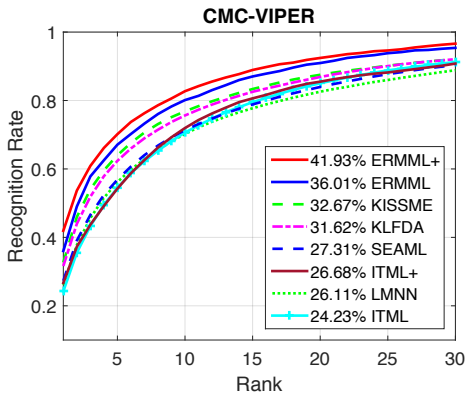

Figure 3: CMC curves of average recognition rates on VIPeR. Rank-1 recognition rate is used to rank the methods. ERMML+ and ERMML are applied with full PCA components.

our method with APG in comparison with the standard gradient descent method using the same step size estimation strategy in Figure 2. We run this experiment on the LFW dataset as an example. We can see that the APG method converges fast than the standard gradient descent method.

5 Experiment

To validate the effectiveness of our method, we conduct experiments on three real-world datasets: VIPeR [Gray and Tao, 2008], iLIDS [Zheng *et al.*, 2009], and LFW [Huang *et al.*, 2007]. VIPeR and iLIDS are two person re-identification (re-id) datasets and LFW is a face verification dataset. In each experiment, we present results by comparing the proposed ERMML+ method with various methods including LMNN, ITML, KLFDA [Xiong *et al.*, 2014], KISSME [Kostinger *et al.*, 2012], SEAML [Wang *et al.*, 2014], and ITML+ [Xu *et al.*, 2015]. Among them, KLFDA and KISSME are two state-of-the-art methods proposed recently for person re-id and face verification. The fixed threshold based method in section 3 is realized as a baseline, termed as ERMML. For a fair comparison, similar/dissimilar pairs weighting is also incorporated in ERMML. The fixed threshold is set as the mean of the squared Euclidean distances between all pairs of training instances. In the following experiments, we implement ERMML+ and ERMML using the log loss function.

5.1 Person Re-identification on VIPeR and iLIDS

VIPeR is a widely used person re-id dataset containing 632 pedestrians in which each person has a pair of images taken from widely differing views. The large viewpoint change of 90 degrees or more as well as huge lighting variations in VIPeR make it one of the most challenging person re-id datasets. iLIDS has 476 images of 119 pedestrians, which is collected at an airport and has severe occlusions.

We adopt the single-shot experiment setting in [Xiong *et al.*, 2014] for all DML methods. The datasets are randomly divided into two parts and the testing set has p individuals. We repeat the random partition 10 times to get an average performance. For easy comparison, we evaluate the re-id results by the cumulative matching characteristic (CMC) curve, which is an estimate of finding the correct match in the top n match. To obtain the privileged information for ERMML+ and ITML+, we have trained several attribute detectors to detect the pedestrian attributes as the privileged information.

Table 1: The recognition rate (%) of various metric learning algorithms on VIPeR with the first 100-D PCA components.

Method	Rank=1	Rank=10	Rank=20
ERMML+	35.32	77.94	89.15
ERMML	32.41	77.78	89.46
KLFDA	31.62	75.63	86.87
KISSME	32.67	76.83	87.52
SEAML	27.31	71.36	83.96
ITML+	26.68	71.95	85.51
LMNN	26.11	70.26	82.58
ITML	24.23	70.64	85.04

Table 2: Comparison of newly reported results (%) on VIPeR. ERMML+ and ERMML are applied with full PCA components.

Method	Rank=1	Rank=10	Rank=20	Reference
ERMML+	41.93	82.72	92.44	Ours
ERMML	36.01	80.13	90.95	\
SLKFP	36.8	83.7	91.7	CVPR 2015
QALF	30.17	62.44	73.81	CVPR 2015
XQDA	40	80.51	91.08	CVPR 2015
PRCSL	34.8	82.3	91.8	ICCV 2015
MLAPG	40.73	82.34	92.37	ICCV 2015
CVPDL	33.99	77.53	88.58	IJCAI 2015

We utilize the weighted histograms of overlapping stripes descriptor[2] for original feature representation. The descriptor is a 5138 dimensional feature vector. For all methods, PCA is first used for dimension reduction. ERMML+ and ERMML are applied with all PCA components, since they employ the low-rank projection to obtain the PSD constrained metric. Other algorithms are applied with the first 100 dimensional PCA components.

Figure 3 plots the CMC curves on VIPeR with $p = 316$ for eight DML methods . We can see from Figure 3 that ERMML+ achieves the best performances 41.93% at rank=1, which is better than all the other methods including SEAML, ITML+ , and two state-of-the-art methods KISSME and KLFDA. The baseline method ERMML obtains the second best result 36.01% at rank=1 which can be owing to the effectiveness of log loss based ERM framework. ERMML+ achieves 5.92% improvement over ERMML and ITML+ achieves 2.45% improvement over ITML, which indicates that privileged information is helpful to learn a better metric and can play a more effective role in the ERM framework than the ITML framework. ERMML+ also achieves 14.62% improvement over the adaptive DML method SEAML at rank=1. ITML and LMNN perform not very well due to the complex inter-class and inner-class variations in VIPeR. The success of ERMML+ mainly benefits from the exploring of

[2]http://www.micc.unifi.it/lisanti/source-code/whos/

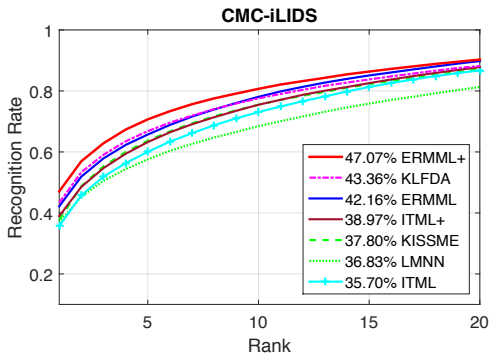

Figure 4: CMC curves of average recognition rates on iLIDS. Rank-1 recognition rate is used to rank the methods. ERMML+ and ERMML are applied with full PCA components.

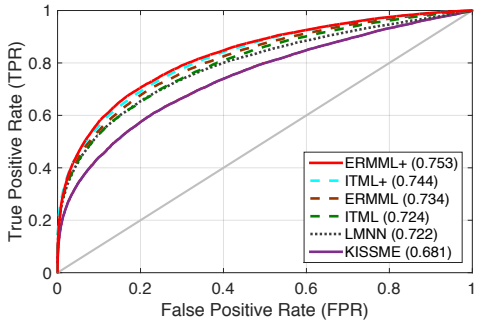

Figure 5: The ROC curves of various methods on LFW. Only one fold is used for training ($K = 1$). The average recognition rate is used to rank the methods.

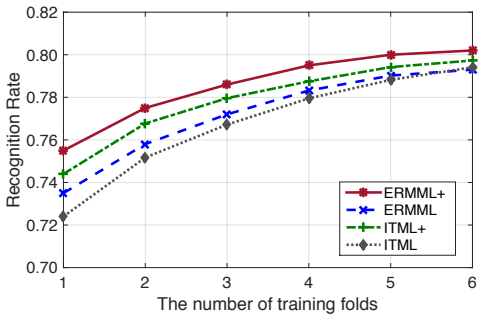

Figure 6: Performance comparison of ERMML+, ITML+, ERMML, and ITML by varying the number of training folds from $K = 1$ to $K = 6$ on LFW.

privileged information and the ERM framework.

For a fair comparison, we also show the results of all methods with the same feature at Table 1. We can note that ERMML+ still perform the best even with the first 100 dimensional PCA components. We also compare the performance of ERMML+ with six newly released results on the VIPeR dataset using the same protocol , including SLKFP [Chen *et al.*, 2015], QALF [Zheng *et al.*, 2015], XQDA[Liao *et al.*, 2015], PRCSL [Shen *et al.*, 2015], MLAPG [Liao and Li, 2015], CVPDL[Li *et al.*, 2015]. We can see that our method results in a new state-of-the-art performance on VIPeR.

Figure 4 compares ERMML+ with other DML methods by plotting the CMC curves on iLIDS with $p = 60$. It is apparent that by exploiting privileged information, our locally decision function in (5) significantly improves the performance of traditional method with a fixed threshold. Similarly as in Figure 3, ERMML+ achieves a significant improvement over ITML+ again, which shows that we can learn a better metric by incorporating the privileged information into the ERM framework.

5.2 Face Verification on LFW

LFW is a widely used face images database containing more than 13000 face images from 5749 individuals. We extract the 3456 dimensional SIFT descriptors as the original features, which are reduced to 200 dimensions using PCA for all methods. We use the face attribute[3] as the privileged information for ERMML+ and ITML+.

The dataset is divided into 10 folds, in which each fold has 300 similar image pairs and 300 dissimilar image pairs. In this experiment, we randomly choose K folds for training and the rest is used for testing. The procedure is repeated 10 times to report an average result. We only consider the pairwise constraints given by the similar/dissimilar pairs.

Figure 5 plots the ROC curves of different DML methods on LFW. It shows that privileged information based methods ERMML+ and ITML+ perform better. Our method improves ERMML by 2.1% and is slightly better than ITML+ by 1.1%. Note that the improvement brought by the privileged information is relatively small on LFW. It is mainly due to the fact that SIFT is already a strong visual descriptors, which can learn a good metric independently if sufficient training samples are provided.

To analyse the influence of privileged information in more depth, we compare the performance of ERMML+, ITML+, ERMML, and ITML by varying the number of training folds from $K = 1$ to $K = 6$ in Figure 6. Figure 6 shows that the benefit of privileged information tends to reduce with the increasing of the size of training set. That is because the metric learning algorithm may suffer from overfitting when the training samples are too limited while the incorporation of privileged information can effectively relieve the overfitting by introducing necessary correcting information. Our method would be particularly useful if there exist only a few training data or the original feature is weak.

6 Conclusion

In this paper, we propose to learn an effective metric learning method by exploiting privileged information in the generic empirical risk framework. We solve the problem efficiently by the Accelerated Proximal Gradient method. Our proposed method ERMML+ generalizes from the traditional metric learning methods using a fixed threshold by designing a locally adaptive decision rule based on privileged information. We apply the proposed method to solve two real world problems: person re-identification and face verification. Experi-

[3]http://www.cs.columbia.edu/CAVE/projects/faceverification/

ment results have shown that our method can outperform the state-of-the-art metric learning methods.

Acknowledgments

This work is partially supported by Australian Research Council (ARC) Projects DP-140102164, FT-130101457, and LE140100061, and the National 973 Program of China under grant 2014CB347600, and the National Nature Science Foundation of China under grants 61272393, 61322201, and 61432019, and the China Scholarship Council (CSC).

References

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Bian and Tao, 2012] Wei Bian and Dacheng Tao. Constrained empirical risk minimization framework for distance metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1194–1205, 2012.

[Chen *et al.*, 2015] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, pages 1565–1573, 2015.

[Davis *et al.*, 2007] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[Feyereisl *et al.*, 2014] Jan Feyereisl, Suha Kwak, Jeany Son, and Bohyung Han. Object localization based on structural svm using privileged information. In *NIPS*, pages 208–216, 2014.

[Fouad *et al.*, 2013] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider. Incorporating privileged information through metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24(7):1086–1098, 2013.

[Gray and Tao, 2008] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. 2008.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505, 2009.

[Hu *et al.*, 2014] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, pages 1875–1882, 2014.

[Huang *et al.*, 2007] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[Kostinger *et al.*, 2012] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[Li *et al.*, 2013] Zhen Li, Shiyu Chang, Feng Liang, T.S. Huang, Liangliang Cao, and J.R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, pages 3610–3617, 2013.

[Li *et al.*, 2014] Wen Li, Li Niu, and Dong Xu. Exploiting privileged information from web data for image categorization. In *ECCV*, pages 437–452. 2014.

[Li *et al.*, 2015] Sheng Li, Ming Shao, and Yun Fu. Cross-view projective dictionary learning for person re-identification. In *IJCAI*, pages 2155–2161, 2015.

[Liao and Li, 2015] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Mignon and Jurie, 2012] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012.

[Nesterov, 2004] Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

[Shen *et al.*, 2015] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *ICCV*, pages 3200–3208, 2015.

[Vapnik and Izmailov, 2015] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

[Vapnik and Vashist, 2009] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009.

[Wang *et al.*, 2014] Qilong Wang, Wangmeng Zuo, Lei Zhang, and Peihua Li. Shrinkage expansion adaptive metric learning. In *ECCV*, pages 456–471. 2014.

[Weinberger and Saul, 2009] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(8):207–244, 2009.

[Xiong *et al.*, 2014] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16. 2014.

[Xu *et al.*, 2015] Xinxing Xu, Wen Li, and Dong Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(12):3150–3162, 2015.

[Zheng *et al.*, 2009] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, 2009.

[Zheng *et al.*, 2015] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *CVPR*, pages 1741–1750, 2015.