Distance Metric Learning Using Privileged Information for Face Verification and Person Re-Identification

Xinxing Xu, Wen Li, *Member, IEEE*, and Dong Xu, *Senior Member, IEEE*

***Abstract*— In this paper, we propose a new approach to improve face verification and person re-identification in the RGB images by leveraging a set of RGB-D data, in which we have additional depth images in the training data captured using depth cameras such as Kinect. In particular, we extract visual features and depth features from the RGB images and depth images, respectively. As the depth features are available only in the training data, we treat the depth features as privileged information, and we formulate this task as a distance metric learning with privileged information problem. Unlike the traditional face verification and person re-identification tasks that only use visual features, we further employ the extra depth features in the training data to improve the learning of distance metric in the training process. Based on the information-theoretic metric learning (ITML) method, we propose a new formulation called ITML with privileged information (ITML+) for this task. We also present an efficient algorithm based on the cyclic projection method for solving the proposed ITML+ formulation. Extensive experiments on the challenging faces data sets EUROCOM and CurtinFaces for face verification as well as the BIWI RGBD-ID data set for person re-identification demonstrate the effectiveness of our proposed approach.**

***Index Terms*— Distance metric learning, face verification, learning using privileged information (LUPI), person re-identification.**

I. Introduction

FACE verification and person re-identification are two important problems in computer vision, which have attracted increasing attentions from many researchers in the last two decades [1]–[4]. The face verification task is to verify whether two face images are from the same subject or not, while the person re-identification task aims to identify the subject in the probe image by comparing this probe image with a set of gallery images. Although the two applications are different, in both tasks, the training data set usually consists of a number of pairs of training images (i.e., face images or the images containing the whole head and body areas) together with side information (i.e., we only know whether each pair of images is from the same or different subjects instead of the names of those subjects in the images). Therefore, we propose to use the same learning approach to solve the two tasks in this paper.

Given only side information, a common way is to learn a Mahalanobis distance metric for face verification or person re-identification. After that, the distance between a pair of testing images is used to decide whether they are from the same subject or different subjects [4], [5]. However, most of those existing works for face verification and person re-identification are based on the RGB images only. On the other hand, with the advancement of new depth cameras, such as Kinect, one can easily capture depth information together with RGB images when collecting training data for computer vision tasks [6]. A few labeled RGB-D data sets were recently released to the public [7]–[9]. Compared with RGB images, depth information is more robust to illumination changes, complex background, and so forth, and thus it can provide useful information for many vision tasks, such as face recognition [8], gender classification [9], and object recognition [7]. Moreover, for the face verification task, the location of interested foreground regions, such as nose, mouth and eyes in the face image, can be well encoded in the depth images. However, those works require depth information and RGB information in both the training and the test stages, so those methods cannot be used in a broader range of applications, where the testing images do not contain depth information, such as the images captured by the conventional surveillance cameras.

In this paper, we propose a new scheme for recognizing RGB images by learning from a set of RGB-D training data with side information, and our method can be used for face verification and person re-identification. In this paper, the training data consist of a few pairs of RGB images and the corresponding depth images together with side information, and our goal is to decide whether a pair of RGB testing images comes from the same subject or not. In the training process, we first extract the visual features and the depth features from the RGB images and the depth images, respectively. Then, we learn a robust Mahalanobis distance metric in the visual feature space using both the visual and the depth features. In the testing process, we use the learned Mahalanobis distance metric to determine whether a pair of RGB images is from the same subject or not by using only their visual features.

To learn the Mahalanobis distance metric under the new learning scheme, we propose a novel distance-metric learning method called information-theoretic metric learning with privileged information (ITML+) by formulating a new

Manuscript received June 21, 2014; revised October 26, 2014 and January 10, 2015; accepted January 24, 2015. Date of publication March 12, 2015; date of current version November 16, 2015.

X. Xu and W. Li are with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: xuxi0006@ntu.edu.sg; wli1@ntu.edu.sg).

D. Xu was with the School of Computer Engineering, Nanyang Technological University, Singapore 639798. He is now with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia. (e-mail: dongxudongxu@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2405574

objective function based on the existing work ITML [10]. This paper is inspired by the recent work on learning using privileged information (LUPI) [11], in which a binary classification method called Support Vector Machine using Privileged Information (SVM+) was proposed to utilize privileged information in the training data. To effectively utilize the additional depth features in the training data, we model the loss term for each pair of visual training samples (i.e., the training samples with visual features) using the corresponding pair of depth training samples (i.e., the training samples with depth features). In this way, the distance between two visual training samples can be affected by their corresponding depth training samples. An efficient cyclic projection method with analytical solution is also proposed to solve the new optimization problem. Considering that some training samples may not be associated with depth information in the real-world applications, we further extend our ITML+ method to handle the scenario where only a part of training data contains depth information, and we refer to our method as partial ITML+ in this case. Our partial ITML+ method can be optimized in a similar way as in ITML+. We conduct extensive experiments on the real-world EUROCOM and CurtinFaces data sets as well as the BIWI RGBD-ID data set. The results clearly demonstrate the effectiveness of our proposed ITML+ algorithm for improving the face verification and person re-identification performances in the RGB images by utilizing the additional depth images.

This paper is organized as follows. In Section II, we briefly review the related works. The proposed ITML+ algorithm is presented in Section III and its solution is provided in Section IV. In Section V, we report the experimental results. Finally, the conclusion is drawn in Section VI.

II. Related Works

This paper is related to the distance-metric learning methods and the recent works on LUPI, as well as the existing works on face verification and person re-identification.

A. Distance Metric Learning

This paper is related to the distance-metric learning works [4], [10], [12]–[17]. The early work for the Mahalanobis distance metric learning in [12] formulates the distance-metric learning problem as a convex optimization problem that maximizes the sum of distances between dissimilar pairs while minimizing the sum of distances between similar pairs. A projected gradient descent method was proposed to solve the proposed objective function, but the SVD operation on the distance metric $\mathbf{M}$ makes the algorithm only applicable to the small-scale problems. Following [12], a large number of methods were proposed in the literature (see [16] and [17] for the comprehensive reviews of different metric learning methods). The two representative works for distance metric learning are: 1) the large margin nearest neighbors (LMNNs) method [13] and 2) the ITML [10] method.

The LMNN [13] method was proposed for the nearest neighbor classifier by constraining the data in a local way, i.e., the k-nearest neighbors of any training instance from the same class should be closer to each other, while the instances from other classes should be kept away by a margin. The constraints are thus given in a triplet form that requires two samples from the same class and one additional sample from the other class. Thus, the explicit class label information is usually required for each sample in the training set to obtain such constraints. The ITML method [10] is based on the pairwise constraints, which assumes that the positive pairs are from the same class and the negative pairs are from different classes without knowing the class label for each sample in the training set. Moreover, instead of learning a global distance metric, some works [14], [15] were proposed to learn local distance metrics for the nearest neighbor search. The unsupervised metric learning method [18] was also developed in which supervised information is not employed.

Different from the existing distance-metric learning methods [4], [10], [12]–[15], our proposed ITML+ method for distance metric learning aims to learn a robust distance metric by further exploiting additional privileged information (i.e., the depth features) in the training data. There are also several multimodal distance-metric learning methods [19]–[21], where multiple types of features are assumed to be available for both training and testing data. In these methods, the final decision is made based on all types of features. Therefore, their setting is still different from the learning setting in this paper.

B. Learning Using Privileged Information

The recently proposed LUPI method [11], [22] used privileged information to improve SVM for the supervised binary classification tasks. In SVM+ [11], privileged information is used to construct the correcting function to control the losses in the objective function. Given a set of n training data $\{\mathbf{x}_i\}|_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^h$, where h is the feature dimension of each sample. The additional privileged feature $\{\mathbf{z}_i\}|_{i=1}^n$ with $\mathbf{z}_i \in \mathbb{R}^g$ is only available for the training set, but it is not available for the test set. Note that the LUPI problem is different from the traditional multiview learning problem, where multiple types of features are available for both the training and the test data [23].

In LUPI [11], the task is to utilize the training data $\{\mathbf{x}_i, \mathbf{z}_i\}|_{i=1}^n$ as well as their labels $\{y_i\}|_{i=1}^n$ to train a classifier for classifying the test data $\{\mathbf{x}_i\}|_{i=n+1}^{n+m}$ under the SVM framework for the supervised binary classification problem. In particular, the linear target classifier $f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ is learned in order to classify the test data. At the same time, another function $\xi = \mathbf{v}'\mathbf{z} + \rho$ is learned by exploiting privileged information in the loss function. The objective function of SVM+ is proposed as follows:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{v},b,\rho} \ & \frac{1}{2}(||\mathbf{w}||^2 + \lambda||\mathbf{v}||^2) + C\sum_{i=1}^{n}(\mathbf{v}'\mathbf{z}_i + \rho) \\ \text{s.t. } & y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - (\mathbf{v}'\mathbf{z}_i + \rho) \quad \forall i = 1, \ldots, n \\ & \mathbf{v}'\mathbf{z}_i + \rho \geq 0 \quad \forall i = 1, \ldots, n. \end{aligned}$$

The above formulation can be reformulated in the dual form as a standard quadratic programming (QP) problem, which can be solved efficiently using the existing QP solvers. Following the LUPI method [11], the recent work in [24] extended SVM+

for the weakly supervised learning and domain adaptation. Another SVM based method for object recognition in RGB images by learning from RGB-D data was also proposed in [25]. Nevertheless, those works were proposed for the classification problem.

Recently, Fouad *et al.* [26] proposed a two-stage method to utilize privileged information for distance metric learning. In particular, their work first learns a distance metric using the ITML algorithm based on privileged information. Then, they remove some outlier pairs, whose distances are larger (resp., smaller) than a threshold if they are similar (resp., dissimilar) pairs. In the second stage, they use the remaining training pairs to train another distance metric using the ITML method based on the main feature. However, the two-stage method proposed in [26] can achieve only slightly better or even worse results than ITML in our experiments.

In contrast, in this paper, we design a slack function to incorporate privileged information for metric learning, which is motivated by SVM+. Using the slack function to replace the slack variables in ITML, we arrive at a unified convex objective function that can be readily solved using the cyclic projection method as in ITML. In contrast to the work in [26], which explicitly removes the outlier pairs based on the depth features, and learns the two metrics in two steps separately, in our ITML+, we jointly learn two metrics in a unified objective function. In our experiments, we show that our newly proposed ITML+ method is consistently better than ITML for different tasks, which demonstrates it is effective to utilize the slack function for modeling privileged information (see Section V for the details).

C. Face Verification and Person Re-Identification

This paper is related to the face verification works. In general, the existing face verification methods can be categorized into feature-based methods and learning-based methods. The feature-based methods [1], [27], [28] developed better face descriptors. For example, in [1], an unsupervised learning approach is proposed to encode the microstructures of a face image. In [28], the outputs of the attributes and simile classifiers are used as the midlevel features to represent a face image for the face verification task. In contrast, the learning-based works [4], [5] developed new learning methods such as the metric learning methods for the face verification task. In particular, two face images from the same person are regarded as a similar pair, while two face images from different persons are regarded as a dissimilar pair. Based on the extracted low-level visual features (i.e., SIFT [29], HOG [30], and LBP [2]) for each face image, the Mahalanobis distance metric is learned using these low-level visual features on the training samples, and the learned distance metric is applied to a pair of test samples with the same type of low-level visual features. The distance-metric learning methods have been successfully applied to the face verification task on the benchmark data sets, such as labeled faces in the wild [31]. The ITML method [10] was proposed for distance metric learning by considering the pairwise constraints as side information, while the work in [4] proposed a discriminant metric learning method that takes advantages of all pairs of samples in the data set.

Person re-identification is another related task using the images containing the whole head and body areas. Recently, many benchmark data sets have been released for the person re-identification task, such as CAVIAR4REID [32]. Many methods for person re-identification have been proposed, which include feature-based methods [33]–[36] as well as learning-based methods [37]–[39]. The feature-based methods aim to develop better descriptors for the human body areas using spatial temporal appearances [36], salience learning [35], and so on. The learning-based methods aim to develop more effective learning algorithms for the person re-identification task, such as probabilistic relative distance comparison [37], rank SVM [38], and KISSME [39].

III. Distance Metric Learning With Privileged Information

In this section, we first introduce the problem setting of our face verification and person re-identification tasks. Then, we review the objective function of ITML. After that, we propose the objective function of our new method ITML+. We also introduce a variant of our ITML+ called partial ITML+ for the case that only a part of training data was associated with privileged information.

A. Problem Statement

In our task, the training data are a few pairs of RGB-D images together with side information describing whether each pair belongs to the same subject or not. In the training process, we extract the visual features and depth features from the RGB images and depth images, respectively. Formally, let us denote the visual features as $\{\mathbf{x}_i\}|_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^h$ is the visual feature vector extracted from the RGB image of the ith training sample, and n is the number of training samples. Similarly, we denote the depth features as $\{\mathbf{z}_i\}|_{i=1}^n$, where $\mathbf{z}_i \in \mathbb{R}^g$ is the depth feature vector extracted from the depth image of the ith sample. We also use $(\mathbf{x}_i, \mathbf{z}_i)$ to denote the ith training sample.

We also have side information for the training data, namely, we have a set of similar pairs $\mathcal{S}$ and a set of dissimilar pairs $\mathcal{D}$. For each similar pair $(i, j) \in \mathcal{S}$ (resp., dissimilar pair $(i, j) \in \mathcal{D}$), the two corresponding training samples $(\mathbf{x}_i, \mathbf{z}_i)$ and $(\mathbf{x}_j, \mathbf{z}_j)$ are from the same subject (resp., different subjects). Our goal is to learn a distance metric $\mathbf{M} \in \mathbb{R}^{h\times h}$ that can be used to classify a pair of test data that only contain the RGB images. In other words, based on the RGB-D training images $\{(\mathbf{x}_i, \mathbf{z}_i)\}|_{i=1}^n$ together with side information, we aim to learn a Mahalanobis distance $d_{\mathbf{M}}(\cdot, \cdot)$ defined as

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \tag{1}$$

where we use the squared distance for the ease of representation in this paper. Intuitively, we expect the learned Mahalanobis distance $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$ can output a large value if $(i, j) \in \mathcal{D}$, and a small value if $(i, j) \in \mathcal{S}$. In the testing process, we use the learned metric to calculate the Mahalanobis distance for each pair of test samples, and determine whether the two corresponding RGB images are from the same subject or different subjects based on their Mahalanobis distance.

B. Information-Theoretic Metric Learning

The key idea of ITML is to learn the distance metric $\mathbf{M}$ by enforcing that the learned distance $d_{\mathbf{M}}$ is large for the dissimilar pairs of samples and small for the similar pairs of samples. In particular, they expect $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq u$ for a relatively small value u if $(i, j) \in \mathcal{S}$, and $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq l$ for a sufficiently large l if $(i, j) \in \mathcal{D}$. However, for the real-world applications, a feasible solution may not exist after using those strict constraints. Thus, a slack variable ξ_{ij} is introduced for each constraint. Let us define $\boldsymbol{\xi} \in \mathbb{R}^{|\mathcal{D}|+|\mathcal{S}|}$ as the vector of slack variables, where each entry ξ_{ij} corresponds to one training pair (i, j). Then, the objective function of ITML [10] is formulated as follows:

$$\begin{aligned} \min_{\mathbf{M}\succeq 0, \xi_{ij}} \; & D_{\mathrm{ld}}(\mathbf{M}, \mathbf{M}^0) + \gamma L(\boldsymbol{\xi}, \boldsymbol{\xi}^0) \\ \text{s.t. } & d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq \xi_{ij}, \quad (i, j) \in \mathcal{S} \\ & d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq \xi_{ij}, \quad (i, j) \in \mathcal{D} \end{aligned} \tag{2}$$

where $\boldsymbol{\xi}^0 \in \mathbb{R}^{|\mathcal{D}|+|\mathcal{S}|}$ is the ideal distance vector with each entry as,

$$\xi^0_{ij} = \begin{cases} u, & (i, j) \in \mathcal{S} \\ l, & (i, j) \in \mathcal{D} \end{cases}$$

$L(\boldsymbol{\xi}, \boldsymbol{\xi}^0)$ is the loss term that measures the difference between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^0$, $\mathbf{M}^0 \in \mathbb{R}^{h\times h}$ is a predefined matrix, and $D_{\mathrm{ld}}(\mathbf{M}, \mathbf{M}^0)$ is a regularizer based on LogDet divergence to avoid the trivial solution.

Following [10] and [40], given any strictly convex differentiable function $\varphi(.)$ over a convex set, the Bregman divergence over two matrices $\mathbf{M}$ and $\mathbf{M}_0$ is defined as

$$D_\phi(\mathbf{M}, \mathbf{M}_0) = \phi(\mathbf{M}) - \phi(\mathbf{M}_0) - \mathrm{tr}(\nabla\phi(\mathbf{M})'(\mathbf{M} - \mathbf{M}_0)).$$

By using the Burg entropy function $\phi(\mathbf{M}) = -\log\det(\mathbf{M})$, the LogDet divergence (or the Burg matrix divergence) can be defined as:

$$D_{\mathrm{ld}}(\mathbf{M}, \mathbf{M}^0) = \mathrm{tr}(\mathbf{M}(\mathbf{M}^0)^{-1}) - \log\det(\mathbf{M}(\mathbf{M}^0)^{-1}) - h \tag{3}$$

where h is the dimension of $\mathbf{M}$ and $\mathbf{M}^0 \in \mathbb{R}^{h\times h}$ is a predefined matrix that is often set to be the identity matrix $\mathbf{I}$. Moreover, the loss term $L(\boldsymbol{\xi}, \boldsymbol{\xi}^0)$ can be defined as $L(\boldsymbol{\xi}, \boldsymbol{\xi}^0) = D_{\mathrm{ld}}(\mathrm{diag}(\boldsymbol{\xi}), \mathrm{diag}(\boldsymbol{\xi}^0))$, which is the LogDet divergence between two diagonal matrices. Thus, ITML aims to minimize the difference between the slack variable vector $\boldsymbol{\xi}$ and the ideal distance vector $\boldsymbol{\xi}^0$ as well as enforce the learned Mahalanobis metric $\mathbf{M}$ close to the identity matrix to avoid the trivial solution.

C. Information-Theoretic Metric Learning With Privileged Information

Recall in our task, we additionally have the depth features in the training data. As ITML only considers one type of features when learning the Mahalanobis distance metric, we thus propose a new distance-metric learning method called ITML+ to learn a more robust Mahalanobis distance metric in the visual feature space by further utilizing the additional depth features in the training data.

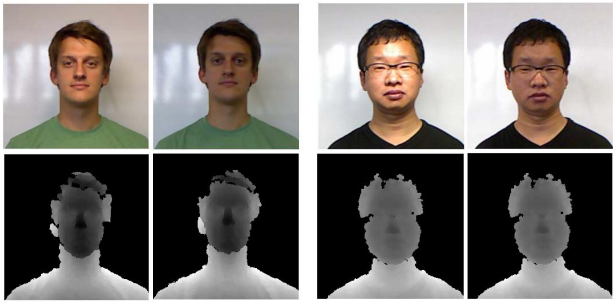


Fig. 1. Two similar pairs of training images in the EUROCOM data set. First row: RGB images captured under different lighting conditions. Second row: corresponding depth images.

Inspired by SVM+ [11], we use the additional depth features to correct the loss of each pair of training samples in the visual feature space. In particular, we replace the slack variable ξ_{ij} in (2) using a slack function in the depth feature space, i.e., $\xi_{ij} = d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j) = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}(\mathbf{z}_i - \mathbf{z}_j)$, where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the depth features of training samples from the pair (i, j), and $\mathbf{P} \in \mathbb{R}^{g\times g}$ is a Mahalanobis distance metric in the depth feature space. In this way, the distance between the training samples from the pair (i, j) in the depth feature space can serve as the correcting guidance for the distance calculated using the visual features. Accordingly, the objective function for our ITML+ is formulated as follows:

$$\begin{aligned} \min_{\mathbf{M}\succeq 0, \mathbf{P}\succeq 0} \; & \Omega(\mathbf{M}, \mathbf{P}) + \gamma \sum_{(i,j)\in\mathcal{S}\cup\mathcal{D}} \ell(d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j), \xi^0_{ij}) \\ \text{s.t. } & d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j), \quad (i, j) \in \mathcal{S} \\ & d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j), \quad (i, j) \in \mathcal{D} \end{aligned} \tag{4}$$

where $\Omega(\mathbf{M}, \mathbf{P}) = D_{\mathrm{ld}}(\mathbf{M}, \mathbf{M}^0) + \lambda D_{\mathrm{ld}}(\mathbf{P}, \mathbf{P}^0)$ is the regularization term by summing the LogDet divergence-based regularization terms related to $\mathbf{M}$ and $\mathbf{P}$, γ and λ are two tradeoff parameters, $\mathbf{M}^0$ and $\mathbf{P}^0$ are two predefined matrices (we use the identity matrices), and $\ell(d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j), \xi^0_{ij}) = D_{\mathrm{ld}}(d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j), \xi^0_{ij})$ is the LogDet divergence between $d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j)$ and ξ^0_{ij}.

Compared with the objective function of ITML in (2), the objective function of our ITML+ in (4) additionally learns a Mahalanobis distance metric $\mathbf{P}$ in the depth feature space. We also replace the original slack variable ξ_{ij} in (2) with $d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j)$ for each pair (i, j). Accordingly, the constraints become $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j), \forall(i, j) \in \mathcal{S}$, and $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j)$ otherwise.

We give some examples in Fig. 1 to explain how our ITML+ can benefit from depth information. As shown in Fig. 1, the RGB images from the same subject may have different visual appearances when they are captured under different lighting conditions. However, their depth images still look almost the same. In other words, given a training pair(i, j), their visual features $\mathbf{x}_i$ and $\mathbf{x}_j$ may be different due to some noises (e.g., illumination changes), whereas their depth features $\mathbf{z}_i$ and $\mathbf{z}_j$ are relatively robust to these noises. In this case, the distance in the visual feature space $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$ may not be good (i.e., the distance may be large if $(i, j) \in \mathcal{S}$ or small if $(i, j) \in \mathcal{D}$). However, the distance in the depth feature space $d^2_{\mathbf{P}}(\mathbf{z}_i, \mathbf{z}_j)$ can be more accurate (i.e., the distance

is small if $(i, j) \in \mathcal{S}$ or large if $(i, j) \in \mathcal{D}$). Using the constraints in (4), the learned Mahalanobis distance metric $\mathbf{M}$ in the visual feature space can be corrected using the distance metric $\mathbf{P}$ in the depth feature space. Therefore, our ITML+ can enforce similar (resp., dissimilar) pairs become more similar (resp., dissimilar) using the distances in the depth feature space as the correcting guidance. The detailed analyses of the learned distances using both ITML and ITML+ are given in Fig. 4(a) and (b) in our experiments (Section V-D).

D. Partial ITML+

In real-world applications, some training samples may not be always associated with depth information. To handle the situation where only a part of training data contains depth information, we further formulate a variant of our ITML+ method called partial ITML+. In particular, let us denote the training set as the similar pair set $\mathcal{S}_p$ and dissimilar pair set $\mathcal{D}_p$ which only contain RGB information. Then, we can formulate our partial ITML+ as follows:

$$\begin{aligned}\min_{\mathbf{M}\succeq 0,\mathbf{P}\succeq 0,\xi_{ij}} \quad & \Omega(\mathbf{M},\mathbf{P}) + \gamma\, L(\boldsymbol{\xi},\boldsymbol{\xi}^0)\\ \text{s.t. } & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \le d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j), \quad (i,j)\in\mathcal{S}-\mathcal{S}_p\\ & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \ge d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j), \quad (i,j)\in\mathcal{D}-\mathcal{D}_p\\ & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \le \xi_{ij}, \quad (i,j)\in\mathcal{S}_p\\ & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \ge \xi_{ij}, \quad (i,j)\in\mathcal{D}_p \end{aligned} \tag{5}$$

where $L(\boldsymbol{\xi},\boldsymbol{\xi}^0) = \sum_{(i,j)\in(\mathcal{S}-\mathcal{S}_p)\cup(\mathcal{D}-\mathcal{D}_p)} \ell(d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j),\xi^0_{ij}) + \sum_{(i,j)\in\mathcal{S}_p\cup\mathcal{D}_p} \ell(\xi_{ij},\xi^0_{ij})$ is the loss term with $\ell(d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j),\xi^0_{ij})$ (resp., $\ell(\xi_{ij},\xi^0_{ij})$) being the LogDet divergence between $d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j)$ (resp., ξ_{ij}) and ξ^0_{ij}, and $\Omega(\mathbf{M},\mathbf{P}) = D_{\mathrm{ld}}(\mathbf{M},\mathbf{M}^0) + \lambda D_{\mathrm{ld}}(\mathbf{P},\mathbf{P}^0)$ is defined similarly as in (4), and γ and λ are two tradeoff parameters.

In other words, we use the constraints from ITML+ for the pairs of training samples with privileged information, while we still utilize the constraints from ITML for the pairs of training samples that do not have privileged information. We observe that the formulation in (5) reduces to the ITML+ formulation in (4) if $\mathcal{S}_p = \emptyset, \mathcal{D}_p = \emptyset$, while the formulation in (5) reduces to the ITML formulation in (2) if $\mathcal{S}_p = \mathcal{S}, \mathcal{D}_p = \mathcal{D}$. In this way, the proposed partial ITML+ in (5) can naturally bridge ITML and ITML+ by varying the number of pairs of training samples with privileged information.

Moreover, our partial ITML+ method can be readily extended to handle the scenario that different samples are associated with different types of privileged information. In particular, suppose there are K types of privileged information, we can correspondingly define K distance metrics $\mathbf{P}_1,\ldots,\mathbf{P}_K$. If a training pair (i, j) is associated with the kth type of privileged information, we model the slack variable for this training pair as $\xi_{ij} = d^2_{\mathbf{P}_k}(\mathbf{z}_i,\mathbf{z}_j)$. The regularizer $D_{\mathrm{ld}}(\mathbf{P},\mathbf{P}^0)$ is accordingly replaced by $\sum_{k=1}^{K} D_{\mathrm{ld}}(\mathbf{P}_k,\mathbf{P}^0_k)$, where $\mathbf{P}^0_k$ can be an identity matrix in the implementation.

IV. Solution to ITML+

In this section, we develop a new optimization algorithm to solve our ITML+ problem in (4) using the cyclic projection method [41].

A. ITML+ With Explicit Correcting Function

The cyclic projection method cannot be directly applied to solve the new objective function in (4) for ITML+, because we have two variables $\mathbf{M}$ and $\mathbf{P}$ in the constraints. Let us introduce an intermediate variable ξ_{ij} for each constraint related to one pair (i, j), we then rewrite our ITML+ formulation in (4) as an equivalent form as follows:

$$\begin{aligned}\min_{\mathbf{M}\succeq 0,\mathbf{P}\succeq 0,\boldsymbol{\xi}} \quad & D_{\mathrm{ld}}(\mathbf{M},\mathbf{M}^0) + \lambda D_{\mathrm{ld}}(\mathbf{P},\mathbf{P}^0) + \gamma\, L(\boldsymbol{\xi},\boldsymbol{\xi}^0)\\ \text{s.t. } & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \le \xi_{ij}, \quad (i,j)\in\mathcal{S}\\ & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \ge \xi_{ij}, \quad (i,j)\in\mathcal{D}\\ & \xi_{ij} = d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j), \quad (i,j)\in\mathcal{S}\cup\mathcal{D} \end{aligned} \tag{6}$$

where $L(\boldsymbol{\xi},\boldsymbol{\xi}^0) = D_{\mathrm{ld}}(\mathrm{diag}(\boldsymbol{\xi}),\mathrm{diag}(\boldsymbol{\xi}^0))$ is the LogDet divergence between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^0$ defined similarly as in (2). The equivalence between (6) and (4) can be easily verified by substituting the correcting function $\xi_{ij} = d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j)$ back into the objective function in (6).

Now, we apply the cyclic projection method similarly as in [10]. For the ease of presentation, we further unify the two inequality constraints in (6), and write the new objective function as follows:

$$\begin{aligned}\min_{\mathbf{M}\succeq 0,\mathbf{P}\succeq 0,\boldsymbol{\xi}} \quad & D_{\mathrm{ld}}(\mathbf{M},\mathbf{M}^0) + \lambda D_{\mathrm{ld}}(\mathbf{P},\mathbf{P}^0) + \gamma\, L(\boldsymbol{\xi},\boldsymbol{\xi}^0)\\ \text{s.t. } & y_{ij} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \le y_{ij}\xi_{ij}, \quad (i,j)\in\mathcal{S}\cup\mathcal{D}\\ & \xi_{ij} = d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j), \quad (i,j)\in\mathcal{S}\cup\mathcal{D} \end{aligned} \tag{7}$$

where

$$y_{ij} = \begin{cases} 1, & (i,j)\in\mathcal{S}\\ -1, & (i,j)\in\mathcal{D}\end{cases}$$

and other terms are the same as in (6).

It can be observed that the objective function in (7) is convex. Following the cyclic projection method [10], [41], we first initialize the solution to (7) as $(\mathbf{P}_0,\mathbf{M}_0)$. Then, we iteratively pickup a pair of training samples (i, j), and update the current solution with Bregman projection such that the objective is minimized and the constraints with respect to this pair are also satisfied. The above process is repeated until all constraints are satisfied. We will give the details on Bregman projection in Section IV-B.

B. Bregman Projection

Let us denote the solution at the tth iteration as $(\mathbf{M}^t,\mathbf{P}^t)$. At the $(t + 1)$th iteration, we pickup a pair of training samples (i, j); then the new solution $(\mathbf{M}^{t+1},\mathbf{P}^{t+1})$ can be obtained with Bregman projection by optimizing the following subproblem:

$$\min_{\mathbf{M}\succeq 0,\mathbf{P}\succeq 0,\xi_{ij}} \quad D_{\mathrm{ld}}(\mathbf{M},\mathbf{M}^t) + \gamma\,\ell(\xi_{ij},\xi^t_{ij}) + \lambda D_{\mathrm{ld}}(\mathbf{P},\mathbf{P}^t) \tag{8}$$

$$\text{s.t. } y_{ij} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \le y_{ij}\xi_{ij} \tag{9}$$

$$\xi_{ij} = d^2_{\mathbf{P}}(\mathbf{z}_i,\mathbf{z}_j). \tag{10}$$

As shown in the following proposition, the above problem has analytical solutions for $\mathbf{M}$, $\mathbf{P}$, and ξ_{ij}.

Proposition 1: The optimal solution ($\mathbf{M}, \mathbf{P}$, and ξ_{ij}) to the problem in (8) can be obtained in closed form as follows:

$$\mathbf{M}^{t+1} = \mathbf{M}^t - \frac{y_{ij}\alpha_{ij}\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t}{1 + y_{ij}\alpha_{ij}r} \tag{11}$$

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \frac{\beta_{ij}\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t}{\lambda - \beta_{ij}s} \tag{12}$$

$$\xi_{ij}^{t+1} = \frac{\lambda s}{\lambda - s\beta_{ij}} \tag{13}$$

where $r = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)$, $s = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)$, and α_{ij} and β_{ij} are the dual variables that can be obtained analytically in Lemma 2.

Proof: By introducing the Lagrangian multipliers $\alpha_{ij} \geq 0$ and β_{ij} for the constraints in (9) and (10), respectively, we obtain the Lagrangian of (8) as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{M}, \mathbf{P}, \xi_{ij}) = {} & D_{\text{ld}}(\mathbf{M}, \mathbf{M}^t) + \gamma\,\ell\big(\xi_{ij}, \xi_{ij}^t\big) + \lambda D_{\text{ld}}(\mathbf{P}, \mathbf{P}^t) \\ & + \alpha_{ij}\big(y_{ij}d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) - y_{ij}\xi_{ij}\big) \\ & + \beta_{ij}\big(\xi_{ij} - d_{\mathbf{P}}^2(\mathbf{z}_i, \mathbf{z}_j)\big).\end{aligned} \tag{14}$$

By setting the derivatives of $\mathcal{L}$ with respect to $\mathbf{M}$ and $\mathbf{P}$ to zeros and denoting $\phi(\mathbf{M}) = -\log(\det(\mathbf{M}))$, we have

$$\nabla\phi(\mathbf{M}) - \nabla\phi(\mathbf{M}^t) + y_{ij}\alpha_{ij}\mathbf{A}_{ij} = 0 \tag{15}$$

$$\lambda\nabla\phi(\mathbf{P}) - \lambda\nabla\phi(\mathbf{P}^t) - \beta_{ij}\mathbf{B}_{ij} = 0 \tag{16}$$

where $\mathbf{A}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'$, and $\mathbf{B}_{ij} = (\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)'$.

Given a matrix $\mathbf{M}$, we have $\partial\det(\mathbf{M})/\partial\mathbf{M} = \det(\mathbf{M})(\mathbf{M}^{-1})'$, which gives $\nabla\phi(\mathbf{M}) = \partial\phi(\mathbf{M})/\partial\mathbf{M} = -(\mathbf{M}^{-1})'$. Thus, we derive the updating rules for the solution at the $(t+1)$th iteration from (15) and (16) as follows:

$$(\mathbf{M}^{t+1})^{-1} = (\mathbf{M}^t)^{-1} + y_{ij}\alpha_{ij}\mathbf{A}_{ij} \tag{17}$$

$$\lambda(\mathbf{P}^{t+1})^{-1} = \lambda(\mathbf{P}^t)^{-1} - \beta_{ij}\mathbf{B}_{ij}. \tag{18}$$

Next, we further simply the above equations by eliminating the matrix inverse operator. Using Sherman–Morrison inverse formula (i.e., $(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}/(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u})$ [42], we derive the equation in (17) as follows:

$$\begin{aligned}\mathbf{M}^{t+1} &= ((\mathbf{M}^t)^{-1} + y_{ij}\alpha_{ij}\mathbf{A}_{ij})^{-1} \\ &= ((\mathbf{M}^t)^{-1} + y_{ij}\alpha_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)')^{-1} \\ &= \mathbf{M}^t - \frac{y_{ij}\alpha_{ij}\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t}{1 + y_{ij}\alpha_{ij}(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)}\end{aligned} \tag{19}$$

which is exactly the solution for $\mathbf{M}^{t+1}$ as in (11) by denoting $r = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)$.

Similarly, we apply the Sherman–Morrison inverse formula to (18) and we arrive at

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \frac{\beta_{ij}\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t}{\lambda - \beta_{ij}(\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)} \tag{20}$$

which is the solution for $\mathbf{P}^{t+1}$ as in (12) by denoting $s = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)$. Note that the updating rules in (19) and (20) guarantee that the updated matrices $\mathbf{M}^{t+1}$ and $\mathbf{P}^{t+1}$ automatically satisfy the semipositive definite constraints as similarly discussed in [10].

Moreover, according to the equality constraint in (10), we have

$$\xi_{ij}^{t+1} = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^{t+1}(\mathbf{z}_i - \mathbf{z}_j). \tag{21}$$

Substituting (20) in (21), we arrive at

$$\xi_{ij}^{t+1} = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^{t+1}(\mathbf{z}_i - \mathbf{z}_j) = \frac{\lambda s}{\lambda - s\beta_{ij}} \tag{22}$$

which is exactly the solution for ξ_{ij}^{t+1} as in (13). Thus, we complete the proof. ■

C. Solutions for α_{ij} and β_{ij}

The remaining problem is to solve the two dual variables α_{ij} and β_{ij} in the updating rules in Proposition 1. Based on the KKT condition, we give the analytical solution to those two dual variables in the following.

Lemma 2: The dual variables α_{ij} and β_{ij} can be obtained in closed form as follows:

$$\alpha_{ij} = \max\left\{0, \quad \frac{\left(\frac{\gamma}{\xi_{ij}^t} + \frac{\lambda}{s} - \frac{\lambda+\gamma}{r}\right)}{y_{ij}(\lambda + \gamma + 1)}\right\} \tag{23}$$

$$\beta_{ij} = \frac{\lambda}{\lambda + \gamma}\left(\frac{\gamma}{s} - \frac{\gamma}{\xi_{ij}^t} + y_{ij}\alpha_{ij}\right) \tag{24}$$

where $r = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)$, and $s = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)$.

Proof: By setting the derivative of $\mathcal{L}$ in (14) with respect to ξ_{ij} to zero, we have

$$\gamma\,\nabla\phi(\xi_{ij}) - \gamma\,\nabla\phi\big(\xi_{ij}^t\big) - y_{ij}\alpha_{ij} + \beta_{ij} = 0. \tag{25}$$

Similar to the derivations of (17) and (18), we derive the solution of ξ_{ij}^{t+1} at the $(t+1)$th iteration as follows:

$$\gamma\left(\xi_{ij}^{t+1}\right)^{-1} = \gamma\left(\xi_{ij}^t\right)^{-1} - \alpha_{ij}y_{ij} + \beta_{ij}. \tag{26}$$

Substituting (13) in (26), we have $\gamma(\lambda - s\beta_{ij})/(\lambda s) = \gamma/\xi_{ij}^t - \alpha_{ij}y_{ij} + \beta_{ij}$, which further gives the solution for β_{ij} as shown in (24).

As α_{ij} is nonnegative, the final solution for α_{ij} is either greater than or equal to zero. In particular, according to the KKT conditions, for the inequality constraints of (9), we have

$$\alpha_{ij}: \begin{cases}\alpha_{ij} > 0 : y_{ij}[(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^{t+1}(\mathbf{x}_i - \mathbf{x}_j)] = y_{ij}\xi_{ij}^{t+1} \\ \alpha_{ij} = 0.\end{cases}$$

Thus, if $\alpha_{ij} > 0$, we must have $\xi_{ij}^{t+1} = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^{t+1}(\mathbf{x}_i - \mathbf{x}_j)$. Together with (11), we further obtain

$$\xi_{ij}^{t+1} = r - \frac{y_{ij}\alpha_{ij}r^2}{1 + y_{ij}\alpha_{ij}r} = \frac{r}{1 + ry_{ij}\alpha_{ij}}. \tag{27}$$

Combining (27) with (13), we eliminate ξ_{ij}^{t+1} and arrive at $\lambda s/(\lambda - s\beta_{ij}) = r/(1 + ry_{ij}\alpha_{ij})$, which also gives the solution $\beta_{ij} = \lambda(r - s(1 + ry_{ij}\alpha_{ij}))/(sr)$. Using (24), we further obtain the closed-form solution for α_{ij} as $\alpha_{ij} = (\gamma/\xi_{ij}^t + \lambda/s - (\lambda + \gamma)/r)/(y_{ij}(\lambda + \gamma + 1))$. As $\alpha_{ij} > 0$, we can obtain the closed form solution for α_{ij}, as shown in (23). This completes the proof. ■

Algorithm 1 Optimization Procedure for ITML+

1: Set $t = 0$, $\mathbf{M}^0 = \mathbf{I}$, $\mathbf{P}^0 = \mathbf{I}$ and initialize $\boldsymbol{\xi}^0$.
2: **repeat**
3: Pick a training pair $(i, j) \in \mathcal{S} \cup \mathcal{D}$.
4: Calculate $r = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}^t(\mathbf{x}_i - \mathbf{x}_j)$ and $s = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{P}^t(\mathbf{z}_i - \mathbf{z}_j)$, $\forall t$.
5: Obtain α_{ij} using (23) with r, s, and ξ_{ij}^t.
6: Obtain β_{ij} using (24) with s, α_{ij} and ξ_{ij}^t.
7: Update $\mathbf{M}^{t+1}$ using (11) with r, α_{ij} and $\mathbf{M}^t$.
8: Update $\mathbf{P}^{t+1}$ using (12) with s, β_{ij} and $\mathbf{P}^t$.
9: Calculate ξ_{ij}^{t+1} using (13) with s and β_{ij}.
10: Set $t \leftarrow t + 1$.
11: **until** The stop criterion is reached.

D. Overall Optimization Procedure

The detailed optimization procedure is given in Algorithm 1. We first initialize $t = 0$ and initialize the matrices $\mathbf{M}^0$ and $\mathbf{P}^0$ to $\mathbf{I}$, and also set

$$\xi_{ij}^0 = \begin{cases} u, & (i, j) \in \mathcal{S} \\ l, & (i, j) \in \mathcal{D}. \end{cases}$$

Then, we iteratively pick up a training pair (i, j) and update $\mathbf{M}^{t+1}$, $\mathbf{P}^{t+1}$, and ξ_{ij}^{t+1} according to Proposition 1. This process is repeated until the relative changes of the vector norms from the dual variables α_{ij}'s and β_{ij}'s between two successive iterations are smaller than 10^{-3} or the maximum number of iterations is reached, which is set as ten times of the number of training pairs.

Note that the semipositive definite properties for both $\mathbf{M}$ and $\mathbf{P}$ are automatically satisfied during the updating procedure at each iteration of Algorithm 1. We also observe that all the variables have closed-form solutions at each iteration. Thus, our optimization process is efficient. Moreover, the objective function in (4) is convex with linear constraints, so our optimization algorithm shares the similar convergence Property as ITML. While the convergence rate of cyclic projection method was also discussed in [43], it is still a nontrivial task to analyze the convergence rate for our optimization method, which will be studied in the future.

E. Solution to Partial ITML+

Similarly as in ITML+, we introduce the intermediate variables ξ_{ij}'s, and rewrite the objective function of partial ITML+ in (5) as follows:

$$\begin{aligned} \min_{\mathbf{M}\succeq 0, \mathbf{P}\succeq 0, \boldsymbol{\xi}} \; & D_{\mathrm{ld}}(\mathbf{M}, \mathbf{M}^0) + \gamma\, L(\boldsymbol{\xi}, \boldsymbol{\xi}^0) + \lambda D_{\mathrm{ld}}(\mathbf{P}, \mathbf{P}^0) \\ \text{s.t. } & d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) \le \xi_{ij}, \quad (i, j) \in \mathcal{S} \\ & d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) \ge \xi_{ij}, \quad (i, j) \in \mathcal{D} \\ & \xi_{ij} = d_{\mathbf{P}}^2(\mathbf{z}_i, \mathbf{z}_j), \quad (i, j) \in (\mathcal{S} - \mathcal{S}_p) \cup (\mathcal{D} - \mathcal{D}_p). \end{aligned} \tag{28}$$

Note that, for the partial ITML+ formulation in (28), part of the training pairs is associated with the correcting function based on privileged information, while the other pairs are not associated with the correcting function. When using the cyclic projection method, we update our solution by picking one training pair at each iteration. Therefore, the subproblem at each iteration can be solved in two ways. For the training pair associated with privileged information, i.e., $(i, j) \in (\mathcal{S} - \mathcal{S}_p) \cup (\mathcal{D} - \mathcal{D}_p)$, the corresponding subproblem is as the same as in (8), and we update the variables $\mathbf{M}$, $\mathbf{P}$, and ξ_{ij} according to Proposition 1. For the training pair without having privileged information, i.e., $(i, j) \in \mathcal{S}_p \cup \mathcal{D}_p$, the subproblem reduces to the same form as the subproblem in ITML [10], so we update $\mathbf{M}$ and ξ_{ij} according to the solution for the subproblem in ITML and keep $\mathbf{P}$ unchanged.

F. Computational Complexity

We now analyze the complexity of our proposed ITML+ method in Algorithm 1. In the 4th step, the time complexity for calculating r and s are $O(h^2)$ and $O(g^2)$, respectively. Only $O(1)$ time complexity is required for updating α_{ij} and β_{ij} in the fifth step and sixth step. In the seventh step, the projection of $\mathbf{M}$ for each constraint requires $O(h^2)$ time complexity using the closed-form updating solution (11), while the projection of $\mathbf{P}$ using (12) requires $O(g^2)$ time complexity in the eighth step. As we have a total number of $|\mathcal{S}|+|\mathcal{D}|$ training pairs, the time complexity for passing the whole training pairs once is $(|\mathcal{S}| + |\mathcal{D}|)O(h^2 + g^2)$. Compared with ITML, which has the time complexity of $(|\mathcal{S}| + |\mathcal{D}|)O(h^2)$ for scanning the whole training pairs once, our ITML+ is slightly more expensive, because we need to additionally optimize another distance metric $\mathbf{P}$. In practice, our ITML+ runs reasonably fast. When the feature dimensions h and g are comparable, it takes about two times of running time when compared with ITML (see Section V-D4 for the details).

V. Experiments

In this section, we compare our proposed ITML+ algorithm with several baseline algorithms for the face verification and person re-identification tasks. We use two real-world face data sets (i.e., the EUROCOM Face data set [9] and the CurtinFaces data set [8]) for the face verification task, and use the BIWI RGBD-ID data set for the person re-identification task.

A. Baseline Approaches

To the best of our knowledge, we are the first to study the face verification and person re-identification tasks in the RGB images by learning distance metric from RGB-D data. We compare our ITML+ with the following baselines.

1) $L2$ distance, we directly use the Euclidian distance in the testing stage without learning the distance metric (i.e., $\mathbf{M} = \mathbf{I}$).
2) ITML [10], the distance metric is learned based on only the visual features from the RGB images together with side information from the training pairs.
3) LMNN [13], the distance metric is learned only based on the visual features from the RGB images but together with explicit label information to construct the triplets. Note that LMNN utilizes stronger label information, because the other methods only employ side information.

4) SVM [44], it is difficult to directly apply SVM to our tasks, as the training data is given in the form of similar and dissimilar pairs. Following [28], we convert each similar (resp., dissimilar) pair as a positive (resp., negative) training sample for learning the SVM classifier. The converting function is defined as $\mathbf{z} = [(|\mathbf{x}_i - \mathbf{x}_j| \circ \mathbf{g})', (\mathbf{x}_i \circ \mathbf{x}_j \circ \mathbf{g})']'$, where $(\mathbf{x}_i, \mathbf{x}_j)$ is a training pair, $|\cdot|$ is the elementwise absolute function, $\circ$ is the element-wise product operation, and $\mathbf{g} = f(0.5(\mathbf{x}_i + \mathbf{x}_j))$ with $f(\cdot)$ being an element-wise Gaussian function with zero mean and unit variance. In this way, we obtain a $2h$-dimensional visual feature vector for each training pair $(\mathbf{x}_i, \mathbf{x}_j)$ for learning the SVM classifier.
5) SVM+ [11], similarly as in SVM, we convert each similar (resp., dissimilar) pair as a positive (resp., negative) training sample based on the visual feature or the depth feature, respectively. The training samples based on the depth features are used as privileged information for training SVM+.
6) ITML-S [26], a two-step approach to utilize privileged information for distance metric learning. Following [26], we first learn a distance metric using ITML based on the depth features, and then remove the pairs that are identified as the outliers. Finally, we train a distance metric using ITML again based on the visual features from the remaining pairs of training images.

B. Face Verification

We perform face verification on two data sets EUROCOM[1] and CurtinFaces[2], which are collected using the Microsoft Kinect. For the EUROCOM data set, the subjects are captured with different facial expressions and under different lighting and occlusion conditions. There are 14 RGB-D face images (i.e., 14 RGB images and 14 corresponding depth images) for each of the 52 subjects, including 38 males and 14 females. Therefore, a total number of 728 RGB-D images are used for the experiments. The CurtinFaces data set consists of 52 persons, and each person has 95 RGB-D face images. Thus, in total, we have 4940 RGB-D face images in the data set. These images are with the variations in facial expressions, illuminations, and poses.

1) Experimental Setup: To evaluate our proposed ITML+ algorithm for the face verification task in the RGB images, we partition the data set into a training set, a validation set, and a test set, which contains the images from 26, 13, and 13 subjects, respectively. We use the training set to learn the models, employ the validation set to select the optimal parameters for each method, and finally evaluate the performances of all methods on the test set. We assume that the training set contains both the RGB images and their corresponding depth images, while the test set and the validation set only contain the RGB images. For the EUROCOM (resp., CurtinFaces) data set, a total number of 2366 (resp., 15 000) positive/similar pairs are constructed using the samples from the same subjects in the training set, while another 7634 (resp., 15 000) negative/dissimilar pairs are randomly sampled from the pairs generated from different subjects in the training set. Therefore, the total numbers of training pairs are 10 000 and 30 000 on the EUROCOM and CurtinFaces data sets, respectively. For the test set, the same strategy is utilized to generate a total number of 5000 (resp., 30 000) pairs, including 1183 (resp., 15 000) positive and 3817 (resp., 15 000) negative pairs for performance evaluation on the EUROCOM (resp., CurtinFaces) data set. For the validation set, we also apply the same strategy to generate 5000 (resp., 30 000) pairs, including 1183 (resp., 15 000) positive and 3817 (resp., 15 000) negative pairs on the EUROCOM (resp., CurtinFaces) data set. For each method, we perform five rounds of experiments using randomly generated negative pairs. For performance evaluation, we calculate the average precision (AP) and area under curve (AUC) for each method at each round, and report the mean of AP (MAP) and the mean of AUC (MAUC) as well as the standard deviations over five rounds of experiments.

2) Feature Extraction: We extract the gradient-LBP features based on the methods in [9] and [45]. In particular, we first convert the RGB images into the grayscale images. For all the images in the data set, we crop each face into a fixed size of 120×105 pixels based on the positions of two eyes. Then, each face image is divided into 8×7 nonoverlapping subregions with the size of 15×15 pixels. We extract the gradient-LBP feature from each subregion. Finally, the gradient-LBP features from all the 56 subregions in each face image are concatenated to form a single 6888-dimensional feature vector. We also use the same strategy to extract a 6888-dimensional feature vector for each depth image. We refer to the gradient-LBP features extracted from the RGB images and the depth images as GLBP-RGB and GLBP-DEPTH, respectively. Recall that the training set contains both RGB images and depth images. Therefore, we extract both types of features, and use the GLBP-RGB features (resp., GLBP-DEPTH features) as the main features (resp., privileged information). For the test set and the validation set, we only extract the GLBP-RGB features from the RGB images as the depth images are not available. Moreover, we perform PCA for both types of features as it is computationally expensive to learn the distance metric with the original high-dimensional features. We fix the PCA dimension for both GLBP-RGB and GLBP-DEPTH features to 150 in our experiments.

3) Parameter Setting: For fair comparisons, we train the models based on the training set, and use the validation set to select the optimal parameters for each method. In particular, we set the common parameter γ for ITML, ITML-S and ITML+ in the range of $\{10^{-4}, 10^{-3.5}, 10^{-3}, \ldots, 10^{0}\}$. We also set the regularization parameter λ for ITML+ in the range of $\{10^{-2}, 10^{-1.5}, \ldots, 10^{2}\}$. Following [10], the predefined values l and u are set to be the 3rd and 97th percentages of the distances according to the $L2$ distances between all pairs of samples within the training data set. Moreover, we set the tradeoff parameter C in SVM and SVM+ as well as the tradeoff parameter γ in SVM+ in the range of $\{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$. For LMNN, the tradeoff parameter is set in the range of $\{0.1, 0.2, \ldots, 1\}$, while the parameter for KNN is set to 5.

[1]Downloaded from http://rgb-d.eurecom.fr/.
[2]Downloaded from http://impca.curtin.edu.au/downloads/datasets.cfm.

TABLE I

PERFORMANCE EVALUATION FOR DIFFERENT ALGORITHMS ON THE EUROCOM FACE DATA SET. THE MAP (PERCENTAGE) AND MAUC (PERCENTAGE), AS WELL AS THE STANDARD DEVIATIONS ARE REPORTED. THE RESULTS IN BOLDFACE ARE SIGNIFICANTLY BETTER THAN THE OTHERS, JUDGED BY THE t-TEST WITH A SIGNIFICANCE LEVEL AT 0.05

	L2 distance	SVM	ITML	LMNN	ITML-S	SVM+	ITML+
AP	58.82±0.64	66.02±1.92	84.16±0.80	84.28±0.60	83.94±0.99	66.01±0.99	**86.82±0.79**
AUC	70.37±0.19	82.02±0.86	92.76±0.21	92.96±0.25	92.57±0.48	82.98±0.57	**93.80±0.41**

TABLE II

PERFORMANCE EVALUATION FOR DIFFERENT ALGORITHMS ON THE CURTINFACES DATA SET. THE MAP (PERCENTAGE) AND MAUC (PERCENTAGE), AS WELL AS THE STANDARD DEVIATIONS ARE REPORTED. THE RESULTS IN BOLDFACE ARE SIGNIFICANTLY BETTER THAN THE OTHERS, JUDGED BY THE t-TEST WITH A SIGNIFICANCE LEVEL AT 0.05

	L2 distance	SVM	ITML	LMNN	ITML-S	SVM+	ITML+
AP	62.05±0.17	71.86±0.27	75.19±0.19	71.67±0.11	74.56±0.47	71.91±0.29	**78.73±0.39**
AUC	59.01±0.14	70.52±0.26	74.49±0.56	69.43±0.09	74.58±1.26	70.76±0.52	**79.15±0.33**

TABLE III

PERFORMANCE EVALUATION FOR DIFFERENT ALGORITHMS ON THE BIWI RGBD-ID DATA SET. THE MEAN OF RANK-1 RECOGNITION RATES (PERCENTAGE) AS WELL AS THE STANDARD DEVIATIONS ON THE TWO TEST SETS ARE REPORTED. THE RESULTS IN BOLDFACE ARE SIGNIFICANTLY BETTER THAN THE OTHERS, JUDGED BY THE t-TEST WITH A SIGNIFICANCE LEVEL AT 0.05

	L2 distance	SVM	ITML	LMNN	ITML-S	SVM+	ITML+
Walking	34.59±0.00	31.77±0.19	46.62±0.35	33.05±0.16	46.69±0.82	26.84±1.56	**48.23±0.69**
Still	85.83 ± 0.00	81.17±0.50	92.89±0.16	86.13±0.08	93.01±0.39	79.21±0.76	**95.23 ±0.31**

4) Experimental Results on the EUROCOM Data Set: The detailed experimental results are shown as in Table I. From the results, we observe that ITML and LMNN outperform the $L2$ distance method in terms of both AP and AUC, which demonstrates that it is useful to learn the distance metrics for the face verification problem. We also observe that the classification methods SVM and SVM+ achieve better results than the baseline $L2$ distance method. However, they are still worse than the distance-metric learning methods ITML and LMNN, which indicates the classification methods may not be good choices for face verification. Moreover, our ITML+ is better than ITML, which demonstrates it is beneficial to use the depth features GLBP-DEPTH as privileged information to learn a more robust distance metric for the face verification task in the RGB images.

The recently proposed ITML-S [26] method is slightly worse than ITML. A possible explanation is that the two stage approach based on the pair removal strategy is not so effective for utilizing privileged information. This also indicates that it is critical to utilize privileged information in a more effective way. In contrast, our ITML+ algorithm learns the correcting distance metric and the decision distance metric in a unified framework, and it directly models the relationship between the main feature GLBP-RGB from RGB images and the privileged feature GLBP-DEPTH from depth images, thus it is more effective than the two-step approach in [26].

5) Experimental Results on the CurtinFaces Data Set: The results of all methods on the CurtinFaces data set are reported in Table II. Again, all the distance-metric learning methods are better than the $L2$ distance method. The classification methods SVM and SVM+ are better than the $L2$ distance method, but they are still worse than ITML. We can observe from Table II that ITML+ achieves the best results and it also outperforms ITML, which again demonstrates it is beneficial to utilize extra privileged information from the training data set to improve distance metric learning for the face verification task in the RGB images. Moreover, our ITML+ again outperforms the two-step approach ITML-S in terms of both AP and AUC, which demonstrates the effectiveness of our proposed ITML+ method for utilizing privileged information in a unified framework.

C. Person Re-Identification on the BIWI RGBD-ID Data Set

In this section, we conduct the experiments on the BIWI RGBD-ID data set[3] for the person re-identification task.

The BIWI RGBD-ID data set [46] was collected using the Microsoft Kinect, and the data set consists of a training set and two testing sets (i.e., Walking and Still). The training set records 50 video sequences from 50 different subjects performing certain actions (e.g., rotation, head movements, and walking) in front of a Kinect sensor. Each video sequence corresponds to one subject. The test set is collected from 28 subjects that appear in the training set, but on a different day and with a different dress. In the Walking setting, each of the 28 subjects performs the action walking in front of the Kinect, while all subjects in the Still setting stand still in front of the Kinect with little movement. Both the RGB and the depth video sequences are recorded simultaneously.

1) Experimental Setup: In our experiments, we use the training set of the BIWI RGBD-ID data set to construct our training set and validation set, and use the two test sets for performance evaluation. For the person re-identification task,

[3]http://robotics.dei.unipd.it/reid/index.php/downloads.

we uniformly sample 20 shots of images from the video sequence of each subject. Similarly as in the face verification task, we assume our training set contains both RGB images and depth images, and the validation and test sets only contain RGB images. The 500 RGB images and 500 depth images from the first 25 subjects in the training set are used as our training set, and the 500 RGB images from the remaining 25 subjects are used as our validation set. The 560 RGB images from the Walking (resp., Still) test set are used as our first (resp., second) test set. For our training set, we construct 4750 similar pairs using the images from the same person, and randomly generate another 4750 dissimilar pairs using the images from different persons.

In the test (resp., the validation) stage, we use the first image of each subject as a probe image that leads to a set of 28 probe images for each test set (resp., 25 probe images for the validation set). The remaining 19×28 images in each test set (resp., 19×25 images in the validation set) are used as the gallery images. For each probe image, we calculate the distance between this probe image and all the gallery images using the learned distance metric, and then sort the gallery images according to their distances to this probe image in the ascending order. We use the Rank-1 recognition rate as the evaluation criterion that is the first point in the so-called cumulative matching characteristic curve. Intuitively, it measures the mean person recognition rate when finding the correct person in the top-1 match. We repeat the experiments for five rounds using different randomly sampled pairs. The mean of Rank-1 recognition rates and the standard deviation over five rounds of experiments are reported for all methods. The optimal parameters for all methods are selected according to their performances on the validation set, where the parameter ranges are the same as in the EUROCOM data set.

2) Feature Extraction: For each image, we manually crop out the person using a rectangle containing the whole head, arms, legs, and body areas of the person. We extract the RGB-D kernel descriptors (KDESs) [47] as the features, which have shown promising results for a broad range of applications using the RGB-D images [47]. Following [47], we first transform the RGB images or the depth images into the gray scale images, and resize the images to be no larger than 300×300 pixels while keeping their Aspect ratios. Then, we extract the gradient KDES features for each image using the code[4] from [47]. We use the default setting in their code, in which we extract the low-level KDESs on the 16×16 image patches using a step of eight pixels. Then, the extracted KDESs are quantized into a feature vector using a codebook with 1000 codewords. We also employ three levels of pyramids (i.e., 1×1, 2×2, and 4×4 for the RGB images and 1×1, 2×2, and 3×3 for the depth images) for spatial pooling. Finally, the feature vectors from each region of the pyramids are concatenated into a single feature vector (21 000-dim for the RGB images and 14 000-dim for the depth images). We extract the KDES features from both RGB images and depth images in the training set, while we only extract the KDES features from the RGB images in the validation set and two test sets. Similarly as in the face verification task, we perform PCA on both RGB features and depth features to reduce the feature dimensions as 150, respectively.

3) Experimental Results: From the results in Table III, we observe that the distance-metric learning algorithms are generally better than the baseline method (i.e., $L2$ distance) in terms of the mean Rank-1 recognition rate. The classification methods SVM and SVM+ are worse than the $L2$ distance-based method, which indicates that the classification methods are not effective for person re-identification. Our proposed ITML+ method is better than ITML as well as other baseline methods, which again show the effectiveness of our proposed ITML+ method to utilize additional depth information in the training set. We observe that the recognition rates for the Still case are much better than those for the Walking case, because there are more variations in the test set walking.

D. Experimental Analysis

In this section, we conduct the experiments to analyze our proposed ITML+ algorithm. We first investigate partial ITML+ using different percentages of training pairs with privileged information, and study the performance change of our ITML+ method using different numbers of training pairs. We also analyze the learned distance metrics, and compare the running time of our method with other baseline methods.

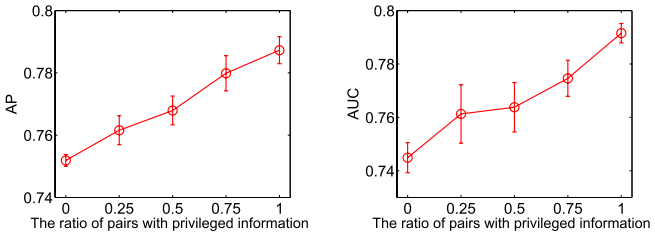
1) Evaluating Partial ITML+ Using Different Percentages of Training Pairs With Privileged Information: In real-world applications, privileged information may be hard to be obtained. Therefore, it is also possible that some training samples are not associated with privileged information. We evaluate our partial ITML+ method discussed in Section III-D using different percentages of training pairs with privileged information.

We take the CurtinFaces data set as an examples and use the partial ITML+ formulation to learn the distance metric by varying the percentage of the training pairs with privileged information. We use the first 0%, 25%, 50%, 75%, and 100% of positive training pairs and negative training pairs with privileged information and the remaining 100%, 75%, 50%, 25%, and 0% training samples are not associated with privileged information. Then, we train our partial ITML+ model to learn a distance metric on the main features, which is used on the testing set for performance evaluation.

We report AP and AUC on the CurtinFaces data set in Fig. 2(a) and (b), respectively. We can observe that the results are the same with those of ITML (resp., ITML+) when the ratio is set to 0% (resp., 100%). Note our partial ITML+ incorporates ITML and ITML+ as two special cases according to the formulation in (5). By varying the ratio in the range of {0%, 25%, 50%, 75%, and 100%}, we observe that the performances are improved when more training pairs are with privileged information.

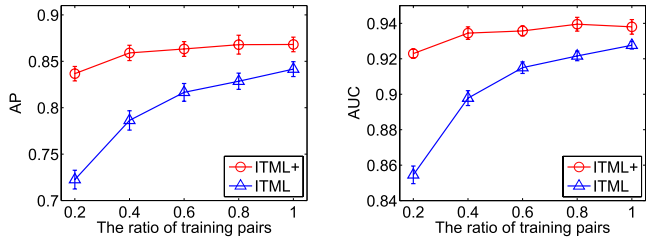
2) Evaluating ITML+ Using Different Percentages of Training Pairs: We take the EUROCOM Face data set as an example to study the performance changes of our proposed

[4]http://mobilerobotics.cs.washington.edu/projects/kdes/.



(a) (b)

Fig. 2. Performances on the CurtinFaces data set using different percentages of training pairs with privileged information. (a) AP. (b) AUC.



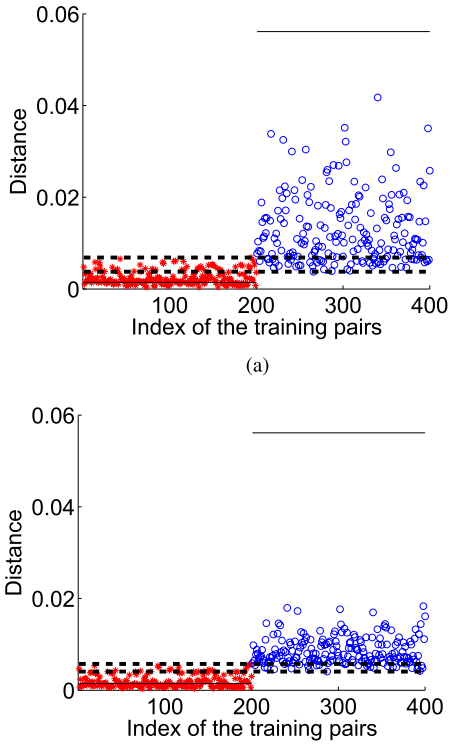
(a) (b)

Fig. 3. Performance comparison between ITML+ and ITML on the EUROCOM data set using different percentages of training pairs. (a) AP. (b) AUC.

ITML+ algorithm with respect to the number of training pairs. We compare ITML+ with the baseline method ITML using 20%, 40%, 60%, 80%, and 100% of the 10 000 training pairs used in Section V-B. The APs and AUCs of ITML+ and ITML when using different numbers of training pairs are reported in Fig. 3(a) and (b), respectively. We observe the AP and AUC of each method generally become higher when the number of training pairs increases, which shows that both methods can be benefited by using more training pairs. Moreover, we also observe that the performance improvement of our ITML+ method over the baseline ITML method is larger when using less training pairs.

3) Analyzing the Learned Distance Metric: We take the BIWI RGBD-ID data set as an example to analyze the learned distance metric. In particular, we analyze the distance metrics learned using ITML and ITML+ for classifying the first 200 positive training pairs as well as the first 200 negative training pairs.

Note the KEDS-RGB features are used as the main features in the testing processes. We show the distances of these 400 pairs of RGB images based on the learned distance metrics from ITML and ITML+ in Fig. 4(a) and (b), respectively. In the two figures, each red star indicates one positive pair, while each blue circle indicates one negative pair. The two horizontal lines are the predefined parameters l (i.e., $l = 1.5 \times 10^{-3}$) and u (i.e., $u = 5.6 \times 10^{-2}$). As shown in Fig. 4(b), we observe that there are less points in the area between the two dashed lines when compared with the results in Fig. 4(a). Note in Fig. 4(a) and (b), the top dash line denotes the maximum distance from the positive pairs, while the bottom dashed line denotes the minimum distance from the negative pairs. The results show the positive and negative pairs are better separated if the distances are calculated based on



(a)

(b)

Fig. 4. Distances between 200 positive pairs of images and 200 negative pairs of images based on the distance metrics learned using ITML and our ITML+. Red star: one positive pair. Blue circle: one negative pair. (a) ITML. (b) ITML+.

TABLE IV
TRAINING TIME (SECONDS) OF DIFFERENT DISTANCE-METRIC LEARNING ALGORITHMS ON THE EUROCOM DATA SET

	LMNN	ITML	ITML-S	ITML+
Time	37.36±2.27	58.40±0.93	108.68±14.55	109.95±3.48

the metric from ITML+. Thus, we conclude that the distance metric learned using ITML+ is better than ITML by exploiting the additional depth features in the training stage. In our new constraints [see (4) and (6)], the slack variables in ITML+ are defined based on the distances using privileged information. In contrast, there are no such constraints for the slack variables in ITML. Therefore, ITML+ could reduce the overfitting problem by imposing new constraints based on the distances using privileged information.

4) Comparison of Training Time Between ITML+ and Other Baselines: We use the EUROCOM data set as an example to report the training time of our proposed ITML+ algorithm as well as the related distance-metric learning methods LMNN, ITML, and ITML-S. All the experiments are conducted on an IBM workstation (2.79-GHz CPU with 32-GB RAM). We report the average training times and standard deviations from five rounds of experiments in Table IV. It can be observed that the LMNN method is the most efficient one among the four methods. Our proposed ITML+ method takes about two times the training time when compared with ITML, because we need to learn an additional metric $\mathbf{P}$ for

privileged information. This is also consistent with our analysis on computational complexity (Section IV-F). Moreover, the computational time of our ITML+ method is comparable with that of ITML-S, which uses ITML twice.

VI. Conclusion

In this paper, we have studied the face verification and person re-identification tasks in the RGB images using the RGB-D data with side information. We formulate a new problem called distance metric learning with privileged information, where the distance metric is learned with extra information that is available only in the training data but unavailable in the test data. We take the ITML method as an example, and propose a new method called ITML+ for distance metric learning by additionally using privileged information. An efficient cyclic projection method based on the analytical solutions for updating all the variables is also developed to solve the new objective function in our proposed ITML+. Extensive experiments are conducted on the real-world EUROCOM, CurtinFaces, and BIWI RGBD-ID data sets. The results demonstrate the effectiveness of our newly proposed ITML+ algorithm for learning the distance metric from RGB-D data for the face verification and person re-identification tasks in the RGB images. It is worth mentioning that our proposed (partial) ITML+ is a general distance-metric learning method using privileged information. It can be used for more real-world applications, which will be studied in the future. Moreover, it is also interesting to consider the kernelization [48] of the proposed ITML+ algorithm.

References

[1] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 539–546.

[2] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[3] M. Kan, D. Xu, S. Shan, W. Li, and X. Chen, "Learning prototype hyperplanes for face verification in the wild," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3310–3316, Aug. 2013.

[4] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 498–505.

[5] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *Proc. 9th Asian Conf. Comput. Vis.*, Xi'an, China, Sep. 2009, pp. 88–97.

[6] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[7] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 1817–1824.

[8] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Clearwater, FL, USA, Jan. 2013, pp. 186–192.

[9] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *Proc. Workshops 11th Asian Conf. Comput. Vis.*, Daejeon, Korea, Nov. 2012, pp. 133–145.

[10] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Annu. Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 209–216.

[11] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, nos. 5–6, pp. 544–557, 2009.

[12] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2002, pp. 505–512.

[13] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[14] J. Wang, A. Kalousis, and A. Woznica, "Parametric local metric learning for nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1610–1618.

[15] Y.-K. Noh, B.-T. Zhang, and D. D. Lee, "Generative local metric learning for nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 1822–1830.

[16] L. Yang, "Distance metric learning: A comprehensive survey," Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, Tech. Rep., May 2006.

[17] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2013.

[18] G. Lebanon, "Metric learning for text documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 497–508, Apr. 2006.

[19] P. Xie and E. P. Xing, "Multi-modal distance metric learning," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Beijing, China, Aug. 2013, pp. 1806–1812.

[20] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3554–3561.

[21] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *J. Mach. Learn. Res.*, vol. 12, pp. 491–523, Feb. 2011.

[22] D. Pechyony and V. Vapnik, "On the theory of learnining with privileged information," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 1894–1902.

[23] W. Li, L. Duan, I. W.-H. Tsang, and D. Xu, "Co-labeling: A new multi-view learning approach for ambiguous problems," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 419–428.

[24] W. Li, L. Niu, and D. Xu, "Exploiting privileged information from web data for image categorization," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 437–452.

[25] L. Chen, W. Li, and D. Xu, "Recognizing RGB images by learning from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1418–1425.

[26] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider, "Incorporating privileged information through metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1086–1098, Jul. 2013.

[27] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *Proc. Faces Real-Life Images Workshop Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 2008, pp. 1–14.

[28] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 365–372.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[31] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

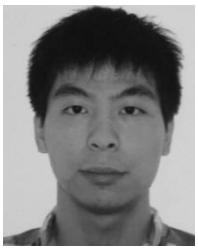
[32] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, Dundee, U.K., Sep. 2011, pp. 68.1–68.11.

[33] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf.*, Surrey, U.K., Sep. 2012, pp. 1–11.

[34] S. Bak, E. Corvée, F. Brémond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal-Based Surveill.*, Boston, MA, USA, Aug./Sep. 2010, pp. 435–440.

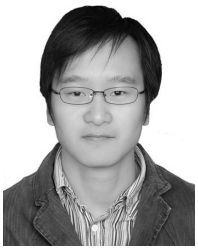
[35] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3586–3593.

[36] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 1528–1535.
[37] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.
[38] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vis. Conf.*, Aberystwyth, U.K., Sep. 2010, pp. 21.1–21.11.
[39] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2288–2295.
[40] B. Kulis, M. Sustik, and I. Dhillon, "Learning low-rank kernel matrices," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 505–512.
[41] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications*. Oxford, U.K.: Oxford Univ. Press, 1997.
[42] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *Ann. Math. Statist.*, vol. 21, no. 1, pp. 124–127, 1950.
[43] F. Deutsch and H. Hundal, "The rate of convergence for the cyclic projections algorithm I: Angles between convex sets," *J. Approx. Theory*, vol. 142, no. 1, pp. 36–55, 2006.
[44] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[45] P. Dago-Casas, D. Gonzalez-Jimenez, L. L. Yu, and J. Alba-Castro, "Single- and cross- database benchmarks for gender classification under unconstrained settings," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Barcelona, Spain, Nov. 2011, pp. 2152–2159.
[46] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, "3D reconstruction of freely moving persons for re-identification with a depth sensor," in *Proc. IEEE Int. Conf. Robot. Autom.*, Hong Kong, May/Jun. 2014, pp. 4512–4519.
[47] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 244–252.
[48] X. Xu, I. W. Tsang, and D. Xu, "Soft margin multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 749–761, May 2013.



Xinxing Xu received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Computer Engineering, Nanyang Technological University, Singapore.

His current research interests include kernel learning and its applications to computer vision.



Wen Li (M'12) received the B.S. and M.Eng. degrees from Beijing Normal University, Beijing, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Engineering, Nanyang Technological University, Singapore.

His current research interests include weakly supervised learning, domain adaptation, and multiple kernel learning.

Dong Xu (M'07–SM'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively.

He was with Microsoft Research Asia, Beijing, China, and the Chinese University of Hong Kong, Hong Kong, for over two years, while pursuing the Ph.D. degree. He was a Post-Doctoral Research Scientist with Columbia University, New York, NY, USA, for one year. He also worked as a faculty member in the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently a faculty member in the School of Electrical and Information Engineering, The University of Sydney, Australia.. His current research interests include computer vision, statistical learning, and multimedia content analysis.

Dr. Xu co-authored a paper that received the Best Student Paper Award in the IEEE International Conference on Computer Vision and Pattern Recognition in 2010. His another coauthored paper also won the IEEE Transactions on Multimedia (T-MM) prize paper award in 2014.