Multi-Task Learning with Low Rank Attribute Embedding for Multi-Camera Person Re-Identification

Chi Su, *Student Member, IEEE*, Fan Yang, *Member, IEEE*, Shiliang Zhang, *Member, IEEE*, Qi Tian, *Fellow, IEEE*, Larry Steven Davis, *Fellow, IEEE*, and Wen Gao, *Fellow, IEEE*

Abstract—We propose Multi-Task Learning with Low Rank Attribute Embedding (MTL-LORAE) to address the problem of person re-identification on multi-cameras. Re-identifications on different cameras are considered as related tasks, which allows the shared information among different tasks to be explored to improve the re-identification accuracy. The MTL-LORAE framework integrates low-level features with mid-level attributes as the descriptions for persons. To improve the accuracy of such description, we introduce the low-rank attribute embedding, which maps original binary attributes into a continuous space utilizing the correlative relationship between each pair of attributes. In this way, inaccurate attributes are rectified and missing attributes are recovered. The resulting objective function is constructed with an attribute embedding error and a quadratic loss concerning class labels. It is solved by an alternating optimization strategy. The proposed MTL-LORAE is tested on four datasets and is validated to outperform the existing methods with significant margins.

Index Terms—Multi-task learning, attribute, low rank, person re-identification

1 Introduction

Person re-identification aims to identify a query person by searching for the most similar instances in a gallery image or video set. Generally, the re-identification precision rate can be improved by acquiring more information from a larger amount of surveillance data. To ensure the recall rate, it is highly necessary to devise effective algorithms to cope with viewpoint variations, illumination conditions, and camera parameter differences across images. This is because that even for the same person appearing in various images, the low-level visual features could be inconsistent and unreliable. Furthermore, in real-world re-identification, images are often collected by a number of non-overlapping cameras with different settings and viewpoints, making person re-identification on multi-cameras a more challenging task.

Nonetheless, even though a person's appearance can be easily affected by many factors, his/her high-level semantic concepts could remain comparatively consistent and stable under different cameras. These semantic concepts, also known as attributes, have been used in many vision tasks like image classification and object detection, and have demonstrated promising robustness. For a person appearing in different cameras, his/her attributes are more stable and consistent than low-level features. For instance, if a person walking towards the camera has an attribute *short hair*, there is a high probability that this *short hair* attribute still could be detected even through this person turns his/her back to the camera. In addition, attributes exhibit substantial correlative relationships, i.e., some attributes tend to co-appear while some never show up at the same time. For example, *female* is more likely to be related with *long hair* than with *short hair*. Also, *long pants* and *short pants* are not likely to co-exist in one person.

By using attributes to describe an image, we can obtain a vector, where each dimension indicates the existence or absence (or the likelihood of existence) of the corresponding attribute. We also find that the above mentioned inter-attribute correlations could be utilized to map a person's attributes under different cameras into a low rank space. In this space, the original binary attributes can be represented by more accurate and informative continuous values. Additionally, this mapping enables us to eliminate noisy attributes and recover missing attributes, thus resulting in more accurate attributes. In order to take advantages of the inter-attribute correlations, the commonly used strategies model the relationships between camera pairs. However, it is unrealistic to do such modeling for large-scale data because of the quadratic complexity with respect to the number cameras. Therefore, most of conventional methods ignore the relationships in the scenarios containing more than two cameras, and thus show limited flexibility.

In Multi-Task Learning (MTL), multiple related tasks benefit each other and are jointly optimized. Because of its

- *C. Su, S. Zhang, and W. Gao are with Peking University, Beijing 100871, China. E-mail: {chisu, slzhang.jdl, wgao}@pku.edu.cn.*
- *F. Yang and L.S. Davis are with the Department of Computer Science, University of Maryland, College Park, MD 20740. E-mail: {fyang, lsd}@umiacs.umd.edu.*
- *Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249. E-mail: qitian@cs.utsa.edu.*

Manuscript received 17 Mar. 2016; revised 9 Feb. 2017; accepted 22 Feb. 2017. Date of publication 6 Mar. 2017; date of current version 10 Apr. 2018. Recommended for acceptance by T. Darell, C. Lampert, N. Sebe, Y. Wu, and Y. Yan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2679002

promising performance in uncovering latent relationships among tasks, MTL has been widely used in machine learning [1], [2] and computer vision [3], [4] tasks. Furthermore, MTL is also suitable for handling scenarios where only a small amount of training data is available for each task. In multi-camera person re-identification, persons appear in different cameras. In another word, different cameras share the same set of persons. Person re-identification task also easily suffers from the limited training data under each individual camera. Inspired by this, we leverage the MTL [5] to explore relationships between features and attributes in cross-camera person re-identification. By considering re-identifications from multiple cameras as related tasks, the MTL framework is well adapted to exploit features and attributes shared across cameras.

Based on the above considerations, we propose the **M**ulti-**T**ask **L**earning algorithm with **LO**w **R**ank **A**ttribute **E**mbedding (MTL-LORAE) algorithm for person re-identification. In our algorithm, we convert the person re-identification problem into a classification problem. Specifically, we use images from multiple cameras to learn a group of person-specific classifiers. A vector made up by outputs of these classifiers is created to represent each probe and gallery image. For training on each specific person, given his/her images from multiple cameras, we use MTL to learn a discriminative model so that the inter-camera relationships can serve to improve the learned model's quality. Our MTL objective function uses both attributes and low-level features. The low rank attribute embedding is also included in the objective function to discover relationships between attribute pairs. In the embedded space, a person's attributes under different cameras become similar while attributes of different people become more distinct from each other. The embedded space also helps to rectify inaccurate attributes and recover missing attributes. Its low rank structure allows only a small amount of latent attributes to contribute to the classification. An efficient alternating optimization method is proposed to solve the MTL-LORAE objective function. In this sense, our work is different from those algorithms performing distance metric learning [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Similar with representation learning algorithms [17], [18], [19], our goal is to acquire a more robust and informative descriptor, by which we use simple distance matching to do person re-identification.

We evaluate MTL-LORAE on four person re-identification datasets and demonstrate that MTL-LORAE has achieved satisfactory results. In [20], Su et al. has provided a preliminary version of this work. In this paper, we add in-depth theoretical discussions and more extensive experiments and detailed discussions. Besides that, motivated by the promising performance of deep learning in various visual tasks, we test deep learning features with our framework. New comparisons on four public datasets show deep learning features further boost the performance of our approach.

Our contributions can be summarized in the following three aspects:

- By regarding re-identification under multiple cameras as related tasks, we successfully exploit their inter-relationships to learn more discriminative classifiers for accurate person re-identification. To the best of our knowledge, multi-task learning based person re-identification is a rarely studied topic.
- We introduce low rank embedding into the MTL framework. This integrates complementary features, i.e., mid-level attributes and low-level visual features into the re-identification framework. Moreover, binary attributes are mapped into a continuous space based on the inter-attribute correlations inferred from the training data. This embedding process also rectifies inaccurate attributes and recovers missing attributes, resulting in more accurate attributes and more discriminative classifiers.
- We present a novel objective function that jointly learns task-specific classifiers and low rank attribute embedding. Although the objective function is difficult to solve, we successfully propose an efficient alternating optimization strategy with convergence guarantee.

2 Related Work

2.1 Person Re-Identification

Person re-identification is attracting more and more attentions nowadays. There are several surveys [21], [22], [23] on person re-identification. Traditional person re-identification works can be classified into three categories: (a) retrieving and encoding robust local features that represent a person's visual appearance; (b) learning a discriminative distance metric to narrow down the distance between features of the same person; (c) learning a new person representation, which should be more robust and informative than the low-level descriptor.

As for feature design, previous works design and use a variety of customized features, including histogram features from various color and texture channels [24], [25], symmetry-driven accumulation of local features [26], features from body parts with pictorial structures [27] to estimate human body configuration, covariance descriptor based on bio-inspired features [28] and space-time features from person tracklets [16], etc. In order to integrate multiple features, Gray et al. [24] select a subset of features by boosting for matching pedestrian images. To enhance the descriptive capability of multiple features, Liu et al. [29] fuse them by learning person-specific weights.

With respect to distance measurement, some works measure the similarity between images from two cameras by learning an optimized distance metric. Pairwise Constrained Component Analysis [7] and Relaxed Pairwise Metric Learning [8] learn a projection from high-dimensional input space to a low-dimensional space, where the distance between pairs of data points meets the pre-defined constraints. The Locally-Adaptive Decision Function in [14] learns a locally adaptive thresholding rule and a distance metric. The Probabilistic Relative Distance Comparison model [15] seeks to increase the possibility of finding a true match whose distance is smaller than a false match. In [11], Köstinger et al. propose a statistical inference perspective to address the problem of metric learning. Kernel-based distance learning [12] is used to handle linearly non-separable data. In [30], Li et al. present a deep learning framework to

learn filter pairs that are responsible for encoding photometric transforms. Bai et al. [31] investigate person re-identification task with manifold-based affinity learning and use an unconventional manifold-preserving algorithm to improve boost identification accuracy.

For representation learning, there are some works [17], [18], [19], [32], [33], [34] which learn new robust descriptors by model training. Matching images of faces from different imaging modalities is an essential step for Heterogeneous Face Recognition (HFR) [17]. For example, HFR matches a sketch or an infrared image with a photo. In HFR framework, probe and gallery images are both represented with respect to their nonlinear similarities to a group of prototype face images. AN et al. [19] propose a reference-based cross-camera person re-identification approach. During the training process, a subspace is learned, where Regularized Canonical Correlation Analysis (RCCA) is used to maximize the relationships between reference images captured by multiple cameras. Recently, Zhao et al. [32] propose to learn mid-level filters, which are designed to address cross-view invariance and use patch matching to infer the geometric configurations of body parts. Xiao et al. [34] propose a Domain Guided Dropout algorithm training on multiple domains with Convolutional Neural Networks (CNNs) to improve the deep feature learning procedure.

Similar to those algorithms, we train multiple person classifiers integrating both low-level features and attribute features to perform representation learning.

There are also approaches dedicated to person re-identification in large camera networks involving more than two cameras [35], [36], [37], [38], [39], [40], [41].

2.2 Attributes

Attributes are semantic concepts of objects and can be either learned from low-level features or manually defined. Previous works have studied the inter-attribute correlations in order to improve the performance of zero/one-shot learning for attribute-based classification [42], [43], [44], [45], [46], [47]. In person re-identification, attributes show promising performance in preserving consistent representations of the same person and identifying differences among different persons [33], [48], [49], [50]. However, attributes are often used as supplementary features for low-level features in previous person re-identification works, which also do not consider the correlations between attributes. Although several methods of object classification have managed to model correlations between attributes [51], [52], [53], as far as we know, no work has utilized both low-level features and attribute correlations across cameras for re-identification in a systematic manner. Our algorithm integrates both attributes and low-level features for training and acquire better attribute features through low rank attribute embedding.

2.3 Multi-Task Learning

There are some representative works concerning Multi-Task Learning, including clustered MTL [54], Robust MTL [55], trace norm regularization [56], and [57]. The modeling of information shared across tasks is often based on the assumption of a shared low rank structure [58], [59]. Kernel method has also been utilized to handle linearly non-separable features [60], [61]. Dictionary learning [62] and tree sparsity constraint [63] are also integrated with the standard MTL framework. Chen et al. [64] apply MTL to concurrently learn inter-attribute correlations and ranking functions for image ranking. By regarding attribute classifiers as auxiliary tasks for object classifiers, Hwang et al. [65] use MTL to learn a shared structure for improved classification and attribute prediction. Yang and Hospedales [57] provides a two sided neural network framework, one for original feature and one for associated semantic descriptor that addresses both multi-domain and multitask learning.

Both [64] and [65] assume attributes to be related tasks. In [66], the multi-task support vector ranks individuals by transferring information of matched or unmatched image pairs from the source domain to the target domain. Ma et al. [67] use multi-task learning to substitute multiple *Mahalanobis* distance metrics for the universal distance metric for all cameras. It should be noted that our approach is different from [66] in that, we directly model low-level features and inter-attribute correlations shared across cameras without using image pairs. Moreover, with respect to both attributes and low-level features, we seek for a shared structure across cameras, rather than learning a metric for each camera pair, which can be computationally expensive for real applications. Although the framework of [68] has a similar low-rank constraint with our work, it is not MTL based and adopts a different optimization method due to the additional l_1 and l_2 constraints. Robust MTL [55] can only be used to optimize W for MTL. Compared with the ones in [68] and [55], our formulation is more challenging by involving the optimization of both W for MTL and low rank matrix for attributes correlation. To address this formulation, we have proposed an efficient alternating optimization strategy with convergence guarantee.

3 Methodology

3.1 Problem Formulation

In this paper, learning a good representation for person re-identification is formulated as a problem of classification by learning a set of classifiers using images from multiple cameras, with each classifier corresponds to a specific person. Like [17], [18], [19], if the training set contains C persons, we use the MTL-LORAE to train C classifiers and then formulate each classifier learning as a regression problem. Each probe and gallery image is represented by a vector composed of outputs of these classifiers. Through the computation of the distance between vectors of probe and gallery images, we find and rank gallery images to perform person re-identification. Details of this procedure will be given in Section 3.5. For simplicity, no distinction is drawn between cameras and tasks, and we will use two terms interchangeably.

Given L learning tasks $\{\mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^L\}$ sharing the same feature space, we want to use information of all tasks to learn multi-class classifiers on a specific task. Typically, all tasks in a multi-class setting share the same set of C classes (persons). In a supervised one-versus-all manner, for the lth task $\mathcal{T}^l$, we begin with binary classification by considering images belonging to the cth class as positive samples and regarding those belonging to the rest of the classes as negative samples, where there are totally n_l labeled training

samples. We follow the standard supervised learning protocol, where labels of all training images are available.

By learning multiple tasks simultaneously, our method can perform effective task-to-task information transmission, which is a useful function when only a limited amount of training data from a task is available. For better clarity, we omit the class index c from all notations in the following text. For each training sample from the lth task $\mathcal{T}^l$, we have a low level feature vector $\mathbf{x}_i^l \in \mathbb{R}^d$ and a label $y_i^l \in \{-1, 1\}$, where 1 indicates this sample is from the cth class and -1 otherwise. Additionally, there is a binary attribute vector $\mathbf{a}_i^l \in \{0, 1\}^k$ for each sample, which may be semantic and manually labeled or correspond to learned binary codes as described in [69]. For each dimension of $\mathbf{a}_i^l$, 1 means that the corresponding attribute is present and 0 otherwise. Then, a predictor f_l with respect to the task $\mathcal{T}^l$ will be learned.

The discriminative and generalization ability of predictors can be enhanced by exploiting the relationship amongst tasks. Hence, information from task $\mathcal{T}^i$ can be transmitted to another task $\mathcal{T}^j$, where there may be only a limited number of training samples available. In this manner, the learning of the predictor f_j will benefit from the learning on both $\mathcal{T}^i$ and $\mathcal{T}^j$ simultaneously. This motivates the usage of MTL to match images taken by different cameras. Furthermore, the learned predictors can be improved if we integrate attributes and find the correlations among them. The following sections introduce the low rank attribute embedding (LORAE), complete MTL formulation, optimization algorithm, and re-identification process.

3.2 Low Rank Attribute Embedding

A simple approach of combining low-level features and attributes is to concatenate feature vectors and original attribute vectors. However, considering the possible inconsistency between human annotators and that it is difficult to obtain exhaustive semantic concepts, attributes tend to be inaccurate or incomplete in most cases. Actually, the absence of an attribute for an instance does not necessarily means that the instance does not have that attribute, which is a fact that could be misinterpreted by the learning algorithm. Similarly, the presence of a wrongly annotated attribute may constitute noise. All this make it difficult for the learned model based on the original attributes to accurately describe the instance. As there are many attributes, they are normally related to each other, meaning that some of them often co-occur across different tasks. Consequently, from the presence of one attribute, we can infer presence of its closely related attributes, which is helpful in recovering missing attributes. Likewise, some attributes are highly mutually exclusive, so that they never occur simultaneously, which serves as a clue to remove noisy attributes.

Following [68], we learn a low rank attribute space to embed the original binary attributes into continuous attributes using attribute dependencies. Specifically, in the low rank space, there exists a transformation matrix $\mathbf{Z}$ of each specific person, which is responsible for converting each of the original attribute vector of one person (class) into a new vector with continuous values. The transformation matrix should capture correlations between attributes pairs since an attribute can be affected by other attributes. The refined attributes of one person are able to discover the correlations of related

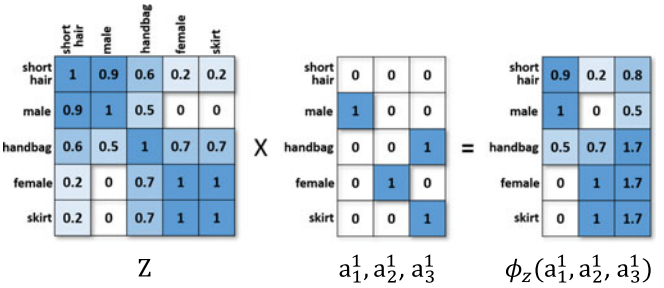


Fig. 1. Illustration of low rank attribute embedding with three attribute vectors from task $\mathcal{T}_1$ as examples. With the learned transformation matrix, the original binary attributes are converted to continuous attributes. Semantically related attributes are recovered even though they are absent in the original attribute vectors, i.e., the attribute *female* is nonzero in the embedded attribute vector due to the presence of both *skirt* and *handbag*, even though its value is 0 in the original attribute vector $\mathbf{a}_3^1$.

attributes and preserve more accurate information to recognize this person. Moreover, some attributes may show certain local patterns. For example, there usually exists groups of attributes like *shorts* and *barelegs*, which are strongly correlated with each other, while being independent with the rest. These local groups essentially imply a low-rank structure in matrix $\mathbf{Z}$. Therefore, $\mathbf{Z}$ should be a low-rank matrix to learn the potential correlations among attributes.

Formally, given an attribute vector $\mathbf{a}_i^l$ from task $\mathcal{T}^l$, the linear embedding is parameterized as

$$\phi_{\mathbf{Z}}(\mathbf{a}_i^l) = \mathbf{Z}^\top \mathbf{a}_i^l \quad \text{s.t.} \quad \text{rank}(\mathbf{Z}) \leq r, \tag{1}$$

where $\mathbf{a}_i^l$ and $\phi_{\mathbf{Z}}(\mathbf{a}_i^l) \in \mathbb{R}^k$ and $\mathbf{Z} \in \mathbb{R}^{k \times k}$. Although kernel methods are applicable here, we choose to focus on linear embeddings for easier learning. The rank constraint imposed on $\mathbf{Z}$ guarantees that $\mathbf{Z}$ is low rank. It means there exists a row $\mathbf{Z}_{i,:}$ (or a column $\mathbf{Z}_{:,i}$) that is a linear combination of other rows (or columns). Therefore, the number of parameters needed for a good embedding is smaller than $k \times k$. Hence, the computational complexity is decreased. In this way, we obtain a refined attribute vector with continuous values. It can precisely describes correlations between attributes while recovering missing values and reducing noises. Fig. 1 shows an intuitive illustration of the low rank embedding, where missing values are successfully recovered in the embedded continuous attributes.

3.3 Multi-Task Learning with Low Rank Attribute Embedding

MTL is designed to learn multiple task-specific predictors simultaneously by making use of the correlations among tasks, so that the shared information can be transmitted from one task to another. To obtain an accurate transformation matrix $\mathbf{Z}$ for the purpose of attribute embedding, we propose a unified MTL framework that can utilize inter-attribute correlations across multiple tasks and train task-specific predictors simultaneously. For the sake of simplicity, we assume a linear classifier for each learning task $\mathcal{T}^l$ to be represented by a weight vector $\mathbf{w}^l$. For notational convenience, we concatenate the embedded attribute vector $\phi_{\mathbf{Z}}(\mathbf{a}_i^l)$ with $\mathbf{x}_i^l$ to construct a new vector $\widetilde{\mathbf{x}}_i^l = [\mathbf{x}_i^l; \phi_{\mathbf{Z}}(\mathbf{a}_i^l)] \in \mathbb{R}^{d+k}$. Therefore, we have $\mathbf{w}^l \in \mathbb{R}^{d+k}$. In another word, while learning each person-specific classifier, we also learn a person-specific linear

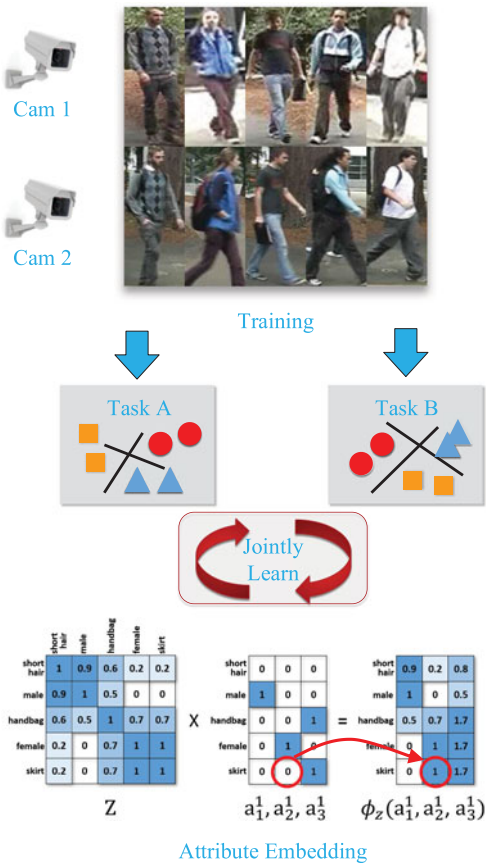


Fig. 2. Illustration of our MTL-LORAE framework.

projection of attributes as part of that classifiers feature space. We define the loss function as $\ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z})$ which can represent any smooth and convex function measuring the discrepancy between groundtruth and predictions from learning. The MTL-LORAE framework is shown in Fig. 2. We use a label $y_i^l \in \{-1, 1\}$ for classification with the goal of finding better feature descriptors instead of just solving the classification problem. In another word, we want to use the outputs of trained classifiers as a feature vector. Consequently, the task is formulated as a regression problem to learn feature vectors conveying the classification confidence scores. Therefore, we define the loss function as a regression problem, i.e.,

$$\ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) = \frac{1}{2}(||y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l||^2 + \gamma||\mathbf{a}_i^l - \mathbf{Z}^\top \mathbf{a}_i^l||^2). \quad (2)$$

The first term $||y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l||^2$ is the quadratic loss caused by applying the learned weight vector $\mathbf{w}^l$ to the newly constructed sample $\widetilde{\mathbf{x}}_i^l$. The second term $||\mathbf{a}_i^l - \mathbf{Z}^\top \mathbf{a}_i^l||^2$ is the attribute embedding error, which regularizes the difference between original attributes and refined attributes obtained from the linear embedding through $\mathbf{Z}$. For the results produced by the embedding, their deviation from the original attributes should be small. γ controls the contributions of the two terms.

We denote all the task-specific $\mathbf{w}^l$ as a single weight matrix $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^L] \in \mathbb{R}^{(d+k)\times L}$. Since information is shared among tasks and each task has a specific structure, similar to [59], we assume that $\mathbf{W}$ is composed of a low rank matrix shared by all tasks and a task-specific sparse component representing the incoherence caused by individual tasks. Formally, $\mathbf{W}$ can be decomposed into a low rank matrix $\mathbf{R} \in \mathbb{R}^{(d+k)\times L}$ and a sparse component $\mathbf{S} \in \mathbb{R}^{(d+k)\times L}$. Therefore, we have $\mathbf{W} = \mathbf{R} + \mathbf{S}$. Intuitively, non-zeros entries in $\mathbf{S}$ indicate the task-specific incoherence between the task and the shared low rank structure. The formulation of MTL-LORAE is then given by

$$\begin{aligned} \min_{\mathbf{R},\mathbf{S},\mathbf{Z}} \quad & \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) + \lambda||\mathbf{S}||_0 \\ \text{s.t.} \quad & \mathbf{W} = \mathbf{R} + \mathbf{S}, \text{rank}(\mathbf{R}) \le r_1, \text{rank}(\mathbf{Z}) \le r_2, \end{aligned} \quad (3)$$

where λ is a trade-off parameter controlling the importance of the regularization. r_1 and r_2 constrain the matrices $\mathbf{R}$ and $\mathbf{Z}$ to be low rank. $||\mathbf{S}||_0$ is the ℓ_0-norm of $\mathbf{S}$, which counts the number of non-zero entries of $\mathbf{S}$.

Solving Eq. (3) is NP-hard since it is non-convex and non-smooth towing to the sparse regularization and low rank constraints. It can be converted into a computationally solvable one through convex relaxation. First, since the ℓ_1-norm is a convex envelop of ℓ_0-norm, $||\mathbf{S}||_0$ is replaced by $||\mathbf{S}||_1$, which is the sum of all non-zero values. Second, the standard convex relaxation for the matrix rank is to use the nuclear norm (trace norm) $||\cdot||_* = \sum_i \sigma_i$, which is the sum of the singular values of a matrix. We then obtain

$$\begin{aligned} \min_{\mathbf{R},\mathbf{S},\mathbf{Z}} \quad & \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) + \lambda||\mathbf{S}||_1 \\ \text{s.t.} \quad & \mathbf{W} = \mathbf{R} + \mathbf{S}, ||\mathbf{R}||_* \le r_1, ||\mathbf{Z}||_* \le r_2, \end{aligned} \quad (4)$$

which is our complete MTL-LOREA formulation. For the convenience of notation, the value of the objective function is denoted as F. By minimizing Eq. (4), the desired weight matrix $\mathbf{W}$ and transformation matrix $\mathbf{Z}$ can be obtained.

3.4 Optimization

The optimization of Eq. (4) is difficult because $\mathbf{W}$ (i.e., $\mathbf{R}$ and $\mathbf{S}$) and $\mathbf{Z}$ are coupled together by $\widetilde{\mathbf{x}}_i^l$. However, the problem becomes solvable when we alternate between the tasks of optimizing the objective function with respect to one variable and fixing the other one. During the process of fixing $\mathbf{Z}$, $||\mathbf{a}_i^l - \mathbf{Z}^\top \mathbf{a}_i^l||^2$ becomes a constant so it can be omitted. $\widetilde{\mathbf{x}}_i^l$ is also constant with respect to $\mathbf{w}^l$, so that it can be regarded as an ordinary training sample. By removing the nuclear norm constraint on $\mathbf{Z}$, Eq. (4) reduces to the standard MTL formulation under the assumption of shared low rank structure plus incoherent sparse values

$$\begin{aligned} \min_{\mathbf{W}} \quad & \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell'(y_i^l, \widetilde{\mathbf{x}}_i^l) + \lambda||\mathbf{S}||_1 \\ \text{s.t.} \quad & \mathbf{W} = \mathbf{R} + \mathbf{S}, ||\mathbf{R}||_* \le r_1, \end{aligned} \quad (5)$$

where $\ell'(y_i^l, \widetilde{\mathbf{x}}_i^l) = \frac{1}{2}||y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l||^2$. Eq. (5) can be solved by using the *MixedNorm* approach as described in detail in [59]. Details can be found in [59].

In the process of fixing $\mathbf{W}$, both $\mathbf{R}$ and $\mathbf{S}$ become constant, so we can remove the constraints related to them. Therefore, we obtain the objective function

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) \\ \text{s.t.} \quad & ||\mathbf{Z}||_* \le r_2. \end{aligned} \quad (6)$$

After relaxing the constraint as a regularization term, we obtain

$$\min_{\mathbf{Z}} \quad \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) + \beta||\mathbf{Z}||_*. \tag{7}$$

With the nuclear norm regularization, the optimal transformation matrix $\mathbf{Z}$ will not degenerate to a trivial solution, i.e., an identity matrix $\mathbf{I}$. However, in the presence of the non-smooth nuclear constraint on $\mathbf{Z}$, it is difficult to optimize Eq. (7). For notational clarity, the loss function with respect to $\mathbf{Z}$ is denoted as $\ell_{\mathbf{Z}}$, and the regularization term as $h_{\mathbf{Z}} = ||\mathbf{Z}||_*$. Eq. (7) is then rewritten as

$$\min_{\mathbf{Z}} \quad \ell_{\mathbf{Z}} + \beta h_{\mathbf{Z}}. \tag{8}$$

$\ell_{\mathbf{Z}}$ is convex, differentiable and Lipschitz continuous. $h_{\mathbf{Z}}$ is convex but non-differentiable. Thus, Eq. (8) can be solved by the proximal gradient method iteratively.

First, we represent the gradient of $\ell_{\mathbf{Z}}$ with respect to $\mathbf{Z}$ as $\partial_{\mathbf{Z}}\ell$. According to the proximal gradient algorithm, at each iteration step j, we then have $\mathbf{Z}_j = \mathbf{prox}_{t_j}(\mathbf{Z}_{j-1} - t_j\partial_{\mathbf{Z}_{j-1}}\ell)$, where $t_j > 0$ is the step size and j is the iteration index. $\mathbf{prox}_{t_j}$ is a proximal operator, defined as

$$\begin{aligned}\arg\min_{\mathbf{Z}} \quad & \ell_{\mathbf{Z}_{j-1}} + \langle \partial_{\mathbf{Z}_{j-1}}\ell, \mathbf{Z} - \mathbf{Z}_{j-1}\rangle \\ & + \tfrac{1}{2t_j}||\mathbf{Z} - \mathbf{Z}_{j-1}||_F^2 + \beta h_{\mathbf{Z}},\end{aligned} \tag{9}$$

where $\langle\cdot,\cdot\rangle$ is the inner product. Eq. (9) finds the $\mathbf{Z}$ that minimizes the surrogate of the loss function ℓ at point $\mathbf{Z}_{j-1}$ plus a quadratic proximal regularization term and the non-smooth regularization term. Eq. (9) can be further simplified to

$$\arg\min_{\mathbf{Z}} \quad \tfrac{1}{2t_j}||\mathbf{Z} - (\mathbf{Z}_{j-1} - t_j\ell_{\mathbf{Z}_{j-1}})||_F^2 + \beta h_{\mathbf{Z}}. \tag{10}$$

It is clear that Eq. (10) can be effectively solved by performing SVD on $\mathbf{Z}_{j-1} - t_j\ell_{\mathbf{Z}_{j-1}}$ and then soft-thresholding the singular values.

In practice, we use the Accelerated Gradient Method (AGM) [56] to achieve faster optimization. AGM adaptively estimates the step size and introduces the search point $\widetilde{\mathbf{Z}}_j$ that is a linear combination of the latest two approximations $\mathbf{Z}_{j-1}$ and $\mathbf{Z}_{j-2}$, $\widetilde{\mathbf{Z}}_j = \mathbf{Z}_{j-1} + (\frac{\alpha_{j-1}-1}{\alpha_j})(\mathbf{Z}_{j-1} - \mathbf{Z}_{j-2})$. Here, α_{j-1} and α_j control the combination weights of the previous two approximations, which are also updated iteratively by $\alpha_j = \frac{1+\sqrt{1+4\alpha_{j-1}^2}}{2}$ with $\alpha_0 = 1$. The gradient in the jth iteration is then performed on $\widetilde{\mathbf{Z}}_j$ instead of $\mathbf{Z}_j$, where $\widetilde{\mathbf{Z}}_1 = \mathbf{Z}_0$.

The gradient $\partial_{\mathbf{Z}}\ell$ is explicitly computed as

$$\begin{aligned}\partial_{\mathbf{Z}}\ell = \quad & (y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l)\frac{\partial \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l}{\partial \mathbf{Z}} + \gamma\frac{\partial \mathbf{Z}^\top\mathbf{a}_i^l}{\partial \mathbf{Z}}(\mathbf{a}_i^l - \mathbf{Z}^\top\mathbf{a}_i^l)^\top \\ = \quad & (y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l)\frac{\partial \mathbf{w}_\phi^{l\top}\mathbf{Z}^\top\mathbf{a}_i^l}{\partial \mathbf{Z}} + \gamma\frac{\partial \mathbf{Z}^\top\mathbf{a}_i^l}{\partial \mathbf{Z}}(\mathbf{a}_i^l - \mathbf{Z}^\top\mathbf{a}_i^l)^\top \\ = \quad & (y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l)\mathbf{a}_i^l\mathbf{w}_\phi^{l\top} + \gamma\mathbf{a}_i^l(\mathbf{a}_i^l - \mathbf{Z}^\top\mathbf{a}_i^l)^\top \\ = \quad & \mathbf{a}_i^l[\mathbf{w}_\phi^{l\top}(y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l) + \gamma(\mathbf{a}_i^l - \mathbf{Z}^\top\mathbf{a}_i^l)^\top],\end{aligned} \tag{11}$$

where $\mathbf{w}_\phi^l \in \mathbb{R}^k$ is part of the weight vector $\mathbf{w}^l$ corresponding to the embedded attribute $\phi_{\mathbf{Z}}(\mathbf{a}_i^l)$. When the optimization for $\mathbf{Z}$ converges, we update $\mathbf{Z}$, fix it and minimize the objective function for $\mathbf{W}$. The optimization will stop once a pre-defined iteration number P or once the difference $\Delta F = F_{j-1} - F_j > 0$ between consecutive values of the objective function falls below a threshold. The detailed steps of the optimization are shown in Algorithm 1.

Algorithm 1. Multi-Task Learning with Low Rank Attribute Embedding (MTL-LORAE)

Input: training data samples $\{\mathbf{x}_i^l, \mathbf{a}_i^l, y_i^l\}$ for all L tasks, initial $\mathbf{Z}_0$ and $\mathbf{W}_0$, iteration number P and threshold $th > 0$ to control iteration step.
Output: Learned $\mathbf{Z}$ and $\mathbf{W}$.
$\mathbf{Z} \leftarrow \mathbf{Z}_0$, $\mathbf{W} \leftarrow \mathbf{W}_0$;
Evaluate objective function F_0 using $\mathbf{Z}$ and $\mathbf{W}$;
for $j = 1$ **to** P **do**
  Optimize Eq. (5) when fixing $\mathbf{Z}$ by *MixedNorm*;
  Update $\mathbf{W} \leftarrow \mathbf{W}_j$;
  Optimize Eq. (6) when fixing $\mathbf{W}$ by AGM algorithm;
  Update $\mathbf{Z} \leftarrow \mathbf{Z}_j$;
  Evaluate objective function F_j;
  Calculate $\Delta F = F_{j-1} - F_j$;
  if $\Delta F < th$
    break;
  end if
end for

3.5 Re-Identification Process

With C training classes (persons), we obtain C class-specific weight matrices and transformation matrices, each of which is denoted as $\mathbf{W}_{(c)} = [\mathbf{w}_{(c)}^1, \mathbf{w}_{(c)}^2, \ldots, \mathbf{w}_{(c)}^L]$ and $\mathbf{Z}_{(c)}$, respectively, by performing the optimization with respect to each class. Note that, since different persons may have different sensitivities to attribute correlations, we trained a transformation matrix $\mathbf{Z}_{(c)}$ for the cth specific person to enhance the recognition of this specific person. Therefore, there are C different transformation matrixes $\mathbf{Z}$ for re-identification instead of one global transformation matrix. Given an image taken by the l'th camera, $l' = 1, 2, \ldots, L$, which either comes from the gallery or the probe set, we first extract low level feature $\mathbf{x}^{l'}$ and attribute vector $\mathbf{a}^{l'}$. By utilizing the transformation matrices, we convert our feature and attribute vectors to a new set of vectors, denoted as $\widetilde{\mathbf{X}}^{l'} = [\widetilde{\mathbf{x}}_{(1)}^{l'}, \widetilde{\mathbf{x}}_{(2)}^{l'}, \ldots, \widetilde{\mathbf{x}}_{(C)}^{l'}] \in \mathbb{R}^{(d+k)\times C}$, where the cth column $\widetilde{\mathbf{x}}_{(c)}^{l'} = [\mathbf{x}^{l'}; \mathbf{Z}_{(c)}^\top\mathbf{a}^{l'}]$ is the concatenation of the feature vector and the embedded attribute vector using the cth transformation matrix $\mathbf{Z}_{(c)}$. Furthermore, we select weight vectors with respect to l'th task from C weight matrices, and multiply them with the new vectors to obtain a score vector $\mathbf{s}$ as

$$\mathbf{s} = \left[\mathbf{w}_{(1)}^{l'\top}\widetilde{\mathbf{x}}_{(1)}^{l'}, \mathbf{w}_{(2)}^{l'\top}\widetilde{\mathbf{x}}_{(2)}^{l'}, \ldots, \mathbf{w}_{(C)}^{l'\top}\widetilde{\mathbf{x}}_{(C)}^{l'}\right], \tag{12}$$

where $\mathbf{w}_{(c)}^{l'}$ is the column weight vector extracted from $\mathbf{W}_{(c)}$ corresponding to the l'th task $\mathcal{T}^{l'}$ trained for the cth class. Therefore, each image is finally represented by a C-dimensional score vector $\mathbf{s}$. Then the Euclidean distance between two score vectors is used to measure the similarity between a gallery image and a probe image. It should be noted that the classes in the training set can be either the same as or different from those in the gallery and probe sets.

In multi-shot cases, a number of images are presented for each probe/gallery identity. Given a probe image set

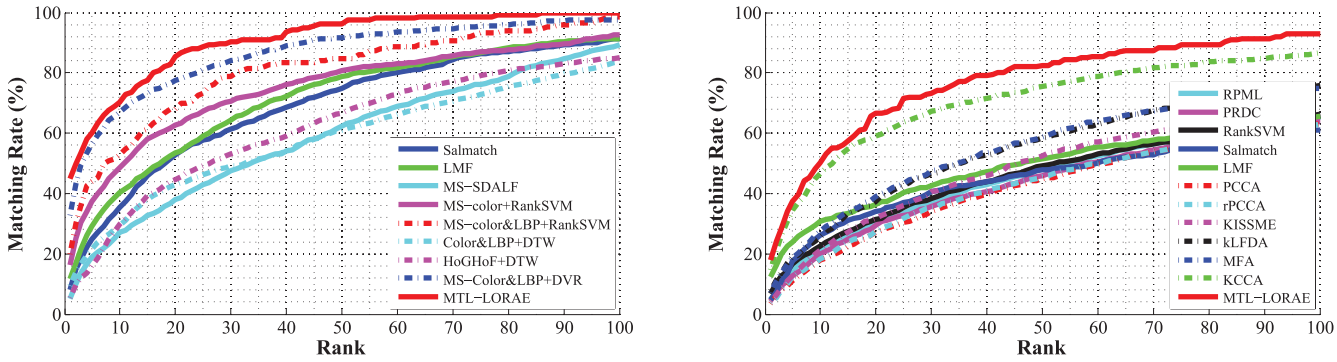


Fig. 3. CMC curves of our approach and state-of-the-art approaches on the *iLIDS-VID* dataset (left) and *PRID* dataset (right).

containing m_p images, the re-identification process ranks the gallery image sets by aggregating image-level similarities Therefore, the voting scheme below is used. First, we calculate the distances between m_p probe images and all gallery images. Then, we use a Gaussian kernel to convert the distances into similarities. In order to obtain a single similarity between the probe and a gallery image set of m_g images, we sum up all $m_p \times m_g$ similarities and divide the sum by the number of gallery images, m_g, to discount the effect of a gallery set including a large number of images.

4 Experiments

4.1 Datasets

We evaluate our approach on four public datasets, *iLIDS-VID* [16], *PRID* [70], *VIPeR* [71] and *SAIVT-SoftBio* [35]. The *iLIDS-VID* dataset consists of 600 image sets of 300 persons from two cameras at an airport. This dataset is designed for multi-shot re-identification. Each person has two image sets from the two cameras respectively, where each image set contains 23 to 192 images, sampled from a short video taken within a few seconds. The *PRID* dataset is used for single-shot scenario. It contains images of different people from two cameras A and B, under different illumination and background conditions. There are 385 and 749 persons appearing in cameras A and B, respectively, of which 200 appear in both cameras. The *VIPeR* dataset contains 632 persons from two cameras, with only one image per person in each camera. The *SAIVT-SoftBio* dataset is also designed for multi-shot re-identification, where images are also extracted from a short video containing a person. There are 152 persons from eight different cameras. Since not every person appears in all cameras, following the evaluation setting in [38], we select those appearing in three cameras (i.e., cameras #3, #5 and #8) as our evaluation set.

4.2 Implementation Details

We use a 2,784-dimensional color and texture descriptor [24] as our low level feature, which is composed of 8 color channels (RGB, HSV and YCbCr[1]) and 19 texture channels (Gabor and Schmid). As for attributes, using this descriptor as x, we learn binary SVMs referring to [49] to predict the same 20-bit attributes in [49] for *PRID* and 90-bit attributes in [72] for *VIPeR*. For other datasets, we learn attribute functions by [73] in an unsupervised manner on the training set and generate 32-bit attributes with the descriptor as x. This overall representation is generated by concatenating the feature x and the output of attribute classifiers.

Following the standard evaluation protocols, we randomly select 150, 100 and 316 persons appearing in all cameras as our training set for *iLIDS-VID*, *PRID* and *VIPeR*, respectively. The remaining 150, 649 and 316 persons serve as the test set (galleries and probes). All the results are averaged over 10 random training/test splits. Parameters for learning are empirically set via cross-validation and fixed for all experiments. $r_1 = 2$, $r_2 = 5$ and $\lambda = 0.3$ in Eq. (3). $\gamma = 0.5$ in Eq. (2). Iteration number $P = 500$ and threshold $th = 10^{-5}$ in Algorithm 1. If a classifier for a specific person is to be trained in the experiment, all images of this person are used as positive samples while images of other people are used as negative samples. Consequently, the positive/negative data ratio is highly imbalanced. We thus randomly select negative samples for training according to a 1:4 positive/negative ratio. Note that, this learning procedure is independent for each person. Therefore, all the classes (persons) can be trained in parallel.

4.3 Experimental Results

4.3.1 *iLIDS-VID*

Among 150 persons in the test set, images from one camera are used as the probe set, while those from another camera serve as the gallery set. We first compare our approach with eight competing learning-based methods for multi-shot re-identification: Salience Matching (Salmatch) [74], Learning Mid-level Filters (LMF) [32], Multi-shot Symmetry-driven Accumulation of Local Features (MS-SDALF) [26], Multi-shot color with RankSVM (MS-color+RSVM) [16], Multi-shot color&LBP with RankSVM (MS-color&LBP+RSVM) [16], color&LBP with Dynamic Time Warping (Color&LBP +DTW) [8], HoGHoF with DTW (HOGHOF+DTW) [75], color&LBP with Discriminative Video fragments selection and Ranking (MS-color&LBP+DVR) [16]. Note that, all of the above methods are trained by person IDs with supervised learning strategy. We use Cumulative Match Characteristic (CMC) curves to evaluate performance, and show experimental results in Fig. 3 and Table 1.

Table 1 clearly shows that our MTL-LOREA approach produces the best results consistently at different ranks. When inspecting the matching rate at ranks 1 and 5, we find a relatively large improvement compared to the MS-color&LBP+DVR approach that achieves the best performance among all the compared algorithms. Specifically, our

1. Only one of the luminance channels (V and Y) is used.

TABLE 1
CMC Scores of Ranks from 1 to 50 on the *iLIDS-VID* Dataset

Rank	1	5	10	20	30	50
Salmatch [74]	8.0	24.8	35.4	52.9	61.3	74.8
LMF [32]	11.7	29.0	40.3	53.4	64.3	78.8
MS-SDALF [26]	5.1	19.0	27.1	37.9	47.5	62.4
MS-color+RSVM [16]	16.4	37.3	48.5	62.6	70.7	80.6
MS-color&LBP+RSVM [16]	20.0	44.0	52.7	68.0	78.7	84.7
Color&LBP+DTW [16]	9.3	21.6	29.5	43.0	49.1	61.0
HoGHoF+DTW [16]	5.3	16.0	29.7	44.7	53.1	66.7
MS-color&LBP+DVR [16]	34.5	56.4	67.0	77.4	84.0	91.7
MTL-LOREA	**43.0**	**60.0**	**70.2**	**85.3**	**90.2**	**96.3**

Numbers indicate the percentage (%) of correct matches within a specific rank.

TABLE 2
CMC Scores of Ranks from 1 to 50 on the *PRID* Dataset

Rank	1	5	10	20	30	50
RPML [8]	4.8	14.3	21.6	30.2	37.2	48.1
PRDC [15]	4.5	12.6	19.7	29.5	35.8	46.0
RSVM [76]	6.8	16.5	22.7	31.5	38.4	49.3
Salmatch [74]	4.9	17.5	26.1	33.9	40.5	47.8
LMF [32]	12.5	23.9	30.7	36.5	42.6	51.6
PCCA [7]	3.5	10.9	17.9	27.1	34.2	45.0
rPCCA [12]	3.8	12.3	18.3	27.5	35.2	45.4
KISSME [11]	4.1	12.8	21.1	31.8	40.7	52.5
kLFDA [12]	7.6	18.9	25.6	37.4	46.7	58.5
MFA [12]	7.2	18.7	27.6	39.1	47.4	58.7
KCCA [77]	14.5	34.3	46.7	59.1	67.2	75.4
MTL-LOREA	**18.0**	**37.4**	**50.1**	**66.6**	**73.1**	**82.3**

Numbers indicate the percentage (%) of correct matches within a specific rank.

method successfully improves the rank 1 accuracy from 34.5 to 43.0 percent, resulting in an 8.5 percent increase. In addition, we obtain nearly 100 percent matching rate at rank 50, while most of the compared methods only achieve 80 percent matching rate or even less at the same rank.

4.3.2 PRID

Following the protocol in [70], we use images of 100 persons from camera A as the probe set, and 649 persons in camera B as the gallery set, excluding all training samples. We compare our algorithm with 11 supervised learning-based methods: Relaxed Pairwise Metric Learning (RPML) [8], Probabilistic Relative Distance Comparison (PRDC) [15], RankSVM (RSVM) [76], Salmatch [74], LMF [32], Pairwise Constrained Component Analysis (PCCA) [7], regularized PCCA (rPCCA) [12], Keep It Simple and Straightforward MEtric (KISSME) [11], kernel Local Fisher Discriminant Classifier (kLFDA) [12], Marginal Fisher Analysis (MFA) [12] and Kernel Canonical Correlation Analysis (KCCA) [77]. Among these compared methods, PRDC, PCCA, rPCCA, kLFDA and MFA use the same 2784-D low-level feature as our work. Note that, we do not compare with DVR [16] because DVR only uses 89 persons for testing, which does not follow the same protocol used by the aforementioned methods. We also use CMC curves to evaluate performance, as shown in Fig. 3 and Table 2.

The experimental results show that our MTL-LOREA approach outperforms all existing methods by a large margin. In particular, our approach achieves 50 percent matching rate at rank 10, while the matching rates of compared approaches are mostly less than 30 percent. Except for our approach and KCCA, all other methods are only able to obtain a 50 percent matching rate at rank 55. Our approach also consistently outperforms KCCA, which shows the best performance at various ranks among the compared algorithms. Specifically, our absolute improvement of matching rate over KCCA is about 6 percent on average. The margin grows larger as we move from lower ranks to higher ranks. In terms of the accuracy at rank 1 and rank 5, our approach achieves a matching rate 18 percent at rank 1 and 37.4 percent at rank 5, respectively, leading to a 3.5 and 3.1 percent performance gain at ranks 1 and 5 over KCCA. When evaluated with more retrieved samples, our approach still secures the best performance. It thus can be seen that pairwise distance metric learning based on camera pairs is clearly not as powerful as our approach. Although using kernel tricks, without fully investigating the relationships of features and attributes from multiple cameras, KCCA cannot improve the performance substantially. The experiments further verify that MTL-LOREA, which learns attribute correlations in an MTL setting with low rank embedding, successfully exploits relationships among attributes, thus produces a more discriminative model.

Since all the competing methods only use low level features while MTL-LOREA uses both low level features and attributes, we conduct additional experiments on the *PRID* dataset, where semantic attributes are provided, to verify that the performance boost of MTL-LOREA results from our learning framework rather than attributes only. We collect publicly available implementations of five existing approaches, which are Salmatch [74], LMF [32], rPCCA [12], kLFDA [12] and MFA [12]. We concatenate the original binary attribute vectors and low level features used by each approach to form a set of new feature vectors, while keeping other parts of each implementation unchanged. For fair comparison, we use the default parameter settings provided by original authors for each implementation. The comparisons are shown in Fig. 4 and Table 3.

With attribute added, all the five compared methods produce better results, justifying the use of attributes. Nevertheless, the performance of the five compared methods is still worse than that of our MTL-LOREA approach. This again verifies that our learning framework with MTL and low rank attribute embedding is effective in utilizing shared information amongst tasks, as well as exploiting attribute correlations, to improve the re-identification accuracy.

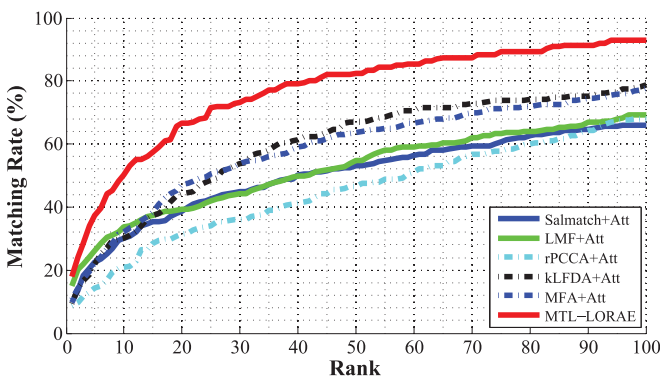


Fig. 4. CMC curves of our approach and 5 state-of-the-art approaches with attributes added on the *PRID* dataset.

TABLE 3
CMC Scores of Our Approach and Five State-of-the-Art Approaches with Attributes Added at Ranks from 1 to 50 on the *PRID* Dataset

Rank	1	5	10	20	30	50
Salmatch [74]	4.9	17.5	26.1	33.9	40.5	47.8
Salmatch+Att	9.6	22.6	30.2	38.8	44.8	53.1
LMF [32]	12.5	23.9	30.7	36.5	42.6	51.6
LMF+Att	15.0	26.2	33.6	39.3	44.1	54.7
rPCCA [12]	3.8	12.3	18.3	27.5	35.2	45.4
rPCCA+Att	8.7	14.4	20.8	31.5	36.0	46.7
kLFDA [12]	7.6	18.9	25.6	37.4	46.7	58.5
kLFDA+Att	9.4	22.0	30.2	44.1	53.9	66.8
MFA [12]	7.2	18.7	27.6	39.1	47.4	58.7
MFA+Att	10.7	22.1	32.0	47.3	53.8	63.7
MTL-LOREA	**18.0**	**37.4**	**50.1**	**66.6**	**73.1**	**82.3**

Numbers indicate the percentage (%) of correct matches within a specific rank. "Att" indicates attributes are added to the original features.

4.3.3 VIPeR

We apply data augmentation to generate more training samples for MTL-LORAE. Specifically, for each training image, we apply horizontal and vertical translation $t \in \{-6, -3, 0, 3, 6\}$ pixels and clockwise rotation $r \in \{-5, 0, 5\}$ degrees, resulting in totally 75 images per original training image.

We compare MTL-LORAE with some recent supervised learning-based methods, including KISSME [11], kLFDA [12], KCCA [77], LOMO+XQDA [78], TSR [79], EPKFM [80]and MLAPG [81], as shown in Table 4. Our MTL-LORAE achieves the best accuracy at rank 1 and rank 5, outperforming existing methods by a large margin, and comparable results at rank 10 and rank 20.

4.3.4 SAIVT-SoftBio

We use half of the persons as the training set and the remaining as the test set. In the test set, each image set serves as the probe while all the remaining image sets are regarded as the gallery. For fair comparison, we evaluate the performance using precision, recall, and F_1-score by regarding the identification problem as a classification problem as [38] does. We do not use the CMC score because it is not applicable to the scenario with more than two cameras. We compare our algorithm to RSVM [76], KISSME [11], RSVM with Conditional Random Field (R-CRF) [38], and KISSME with Conditional Random Field (K-CRF) [38]. All of these compared methods use the same 2,784-D low-level feature as our work. Results

TABLE 4
CMC Scores of Ranks from 1 to 20 on the VIPeR Dataset

Rank	1	5	10	20
KISSME [11]	19.6	47.5	62.2	77.0
kLFDA [12]	32.2	65.8	79.7	90.9
KCCA [77]	37.3	71.4	**84.6**	92.3
LOMO + XQDA [78]	40.0	68.9	81.5	91.1
TSR [79]	31.6	68.6	82.8	**94.6**
EPKFM [80]	36.8	70.4	83.7	91.7
MLAPG [81]	40.7	69.9	82.3	92.4
MTL-LORAE	**42.3**	**72.2**	81.6	89.6

Numbers indicate the percentage (%) of correct matches within a specific rank.

TABLE 5
Comparison of Precision, Recall and F_1-Score (in %) by Existing Methods and Our Approach on *SAIVT-SoftBio* Dataset

	RSVM [76]	KISSME [11]	R-CRF [38]	K-CRF [38]	MTL-LOREA
Precision	22.0	19.7	**53.7**	50.3	45.2
Recall	42.1	**66.1**	39.4	49.8	63.7
F_1-score	26.2	29.5	42.0	48.3	**52.9**

are averaged over all possible camera pairs of the three cameras, and are presented in Table 5.

The table shows that our MTL-LOREA is able to achieve the best F_1-score, outperforming the best of existing method, K-CRF, by 4.6 percent. In addition, MTL-LOREA achieves the second best recall rate and comparable precision rate. We also note that our learning framework can learn the models for all cameras simultaneously regardless of the number of cameras, which is more computationally efficient than existing methods that explicitly deal with all pairs of cameras.

In addition to the above comparisons, we further show comparisons of our approach and other competing methods with respect to each pair of cameras separately in Table 6. Compared with four competing methods, our MTL-LOREA approach achieves better or comparable precision and recall, and the best F_1-score on all the camera pairs, showing its outstanding capability of discovering and identifying a person accurately.

4.4 Performance Using Deep Features

As deep learning has shown promising performance and generalization ability in vision tasks, we also consider to incorporate deep features into our MTL-LOREA algorithm as another type of feature representation. In this experiment, we use the output of VGG-16 network [82] pretrained on the ImageNet image classification task [83] as deep features. The VGG-16 network includes 13 convolutional layers, followed by three fully-connected layers.

To improve the performance of deep features, we finetune the VGG-16 network using more than 40,000 samples

TABLE 6
Comparison of Precision, Recall and F_1-Score (in %) Regarding All Camera Pairs by Existing Methods and Our Approach on *SAIVT-SoftBio* Dataset

	RSVM [76]	KISSME [11]	R-CRF [38]	K-CRF [38]	MTL-LOREA
C3-C5					
Precision	14.9	15.9	37.2	38.0	**38.1**
Recall	24.7	50.3	15.5	28.5	**75.1**
F_1-score	15.9	23.4	18.2	30.3	**50.5**
C3-C8					
Precision	27.7	20.7	**55.4**	48.4	41.0
Recall	29.4	70.1	43.1	51.1	**65.6**
F_1-score	20.1	31.0	43.4	47.6	**50.4**
C5-C8					
Precision	25.7	19.9	45.2	**47.1**	36.8
Recall	43.4	**65.4**	30.8	44.7	53.8
F_1-score	24.6	29.6	32.4	**43.7**	**43.7**

C3, C5 *and* C8 *represent cameras #3, #5 and #8.*

TABLE 7
CMC Scores of MTL-LOREA with Deep Features, i.e., Percentage (%) of Correct Matches, of Ranks 1, Rank 5, Rank 10 and Rank 20 on the *iLIDS-VID* Dataset, *PRID* Dataset and *VIPeR* Dataset

Methods		Rank 1	Rank 5	Rank 10	Rank 20
iLIDS-VID	VGG-FC7	24.1	43.6	52.8	65.6
	MTL-LOREA	43.0	60.0	70.2	85.3
	MTL-LOREA-FC7	56.4	69.0	78.4	87.4
PRID	VGG-FC7	19.8	28.5	42.4	53.9
	MTL-LOREA	18.0	37.4	50.1	66.6
	MTL-LOREA-FC7	21.0	44.0	55.9	68.7
VIPeR	VGG-FC7	25.1	39.8	48.5	60.6
	MTL-LOREA	42.3	72.2	81.6	89.6
	MTL-LOREA-FC7	45.4	76.6	85.3	91.7

of 152 persons, which are taken by cameras #1, #2, #4, #6 and #7 on *SAIVT-SoftBio* dataset. The fine-tuning procedure is conducted in a classification task, where the person IDs are used as class labels. All parameters are the same as those in [82]. A total of 40,000 iterations are performed. Then, we use the fine-tuned VGG-16 model to extract features of images from cameras #3, #5 and #8 of *SAIVT-SoftBio* dataset and other datasets as our low-level features x. We select the 4,096-dim output of FC7 layer (the second fully-connected layer) as the deep feature, and denote it as VGG-FC7.

We run our MTL-LOREA with deep features on the four datasets. The results are summarized in Tables 7 and 8, respectively. In the tables, VGG-FC7 means that we directly use the output of FC7 layer in the VGG-16 network as feature representation and match them using *Euclidean* distance.

We do not use the feature of the output layer, i.e., score vector, because the score vector is more related with the training data. As the training data and test data do not share any overlap, the FC7 feature outperforms the score vector. MTL-LOREA-VGG means that the features of the FC7 layer are used as a substitution for low-level features in multi-task learning under MTL-LOREA.

Table 7 shows the results on the *iLIDS-VID* dataset, *PRID* dataset and *VIPeR* dataset, where the CMC scores of VGG-FC7 at rank 1 are 24.1, 19.8 and 25.1 percent, respectively. This means that deep features are not discriminative enough to distinguish different persons without being fine-tuned on the target datasets. On the other hand, the CMC scores of MTL-LOREA-VGG at rank 1 are 56.4, 21.0 and 45.4 percent on the three datasets, which are 13.4, 3.0 and 3.1 percent higher than those of MTL-LOREA, respectively. It demonstrates that our framework further boosts the person re-identification performance by integrating with deep features. The above experiments clearly verify that our MTL-LOREA framework is effective in correctly matching images from the same person, and is not dependent upon specific features.

Since VGG-16 is fine-tuned on *SAIVT-SoftBio*, it can be seen from the Table 8 that VGG-FC7 has higher Precision, Recall and F_1-score than MTL-LOREA. This is reasonable because deep features fine-tuned on the same dataset commonly perform better. However, the F_1-score of MTL-LOREA-VGG shows improvements of 5.8 percent on *C3-C5-C8*, 7.9 percent on *C3-C5*, 1.8 percent on *C3-C8*, and 12.7 percent *C5-C8*, respectively, compared to the results of VGG-FC7. Moreover, MTL-LOREA-VGG also produces significantly higher F_1-score than MTL-LOREA, i.e., about 15 percent improvement.

TABLE 8
Comparison of Precision, Recall and F_1-Score (in %) Regarding All Camera Pairs by Our Approach with Deep Features on *SAIVT-SoftBio* Dataset

	VGG-FC7	MTL-LOREA	MTL-LOREA-FC7
C3-C5-C8			
Precision	54.4	45.2	57.5
Recall	69.0	63.7	79.5
F_1-score	60.9	52.9	66.7
C3-C5			
Precision	45.9	38.1	50.8
Recall	69.1	75.1	79.5
F_1-score	54.8	50.5	62.7
C3-C8			
Precision	57.9	41.0	55.9
Recall	73.0	65.6	81.9
F_1-score	64.5	50.4	66.3
C5-C8			
Precision	32.2	36.8	44.7
Recall	66.0	53.8	74.6
F_1-score	43.2	43.7	55.9

C3, C5 *and* C8 *represent cameras #3, #5 and #8.*

4.5 Discussion

In this section, we conduct more experiments to show the characteristics and some interesting aspects of the proposed methods.

4.5.1 Convergence Analysis

Our original formulation in Eq. (4) is difficult to optimize. We solve this problem by alternatively optimizing the objective function with respect to one variable and fixing the other one. When fixing $\mathbf{Z}$, we obtain the Eq. (5), which can be solved by *MixedNorm* approach in [59]. The optimization algorithm of *MixedNorm* approach [59] guarantees the global convergence with a convergence rate $\mathcal{O}(1/k^2)$, where k is the iteration number. On the other hand, when fixing $\mathbf{W}$, both the loss function $\ell_{\mathbf{Z}}$ and regularization term $h_{\mathbf{Z}}$ in Eq. (8) are convex, so that a global optimal solution can be acquired. By adopting the Accelerated Gradient Method (AGM) in [56], we can achieve a convergence rate as $\mathcal{O}(1/k^2)$. Proofs of the convergence rate can be found in [56], [59] and [84]. Therefore, our approach will find the global optimal via alternating optimization.

To investigate the convergence rate of MTL-LOREA, we show the values of objective function during the optimization in Fig. 5. The optimization is conducted on the training samples of a person randomly selected from *iLIDS-VID* and *PRID*. The figure shows that the objective function value quickly decreases and reaches its minimal after a few iterations, verifying the effectiveness of our optimization strategy.

4.5.2 Analysis on Transformation Matrix

Based on the assumption that attributes are correlated, we propose to learn the low rank matrix $\mathbf{Z}$ to preserve

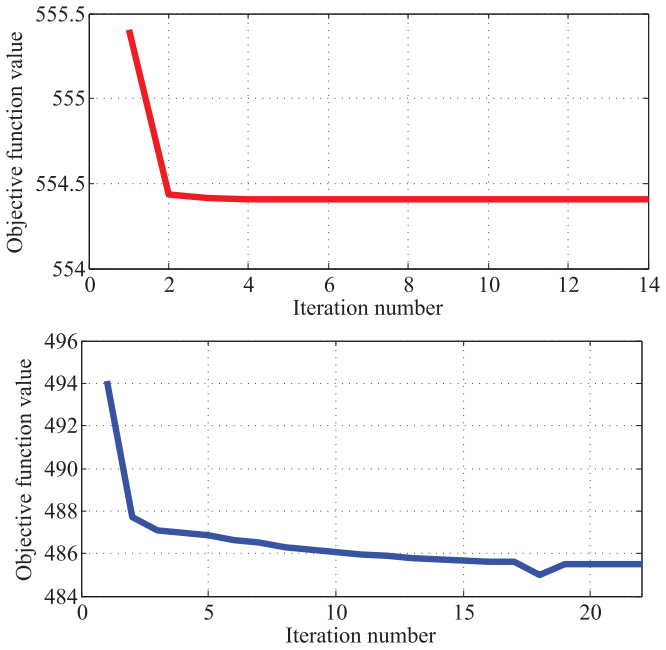


Fig. 5. The values of objective function during optimization on the *iLIDS-VID* dataset (top) and *PRID* dataset (bottom).

attribution correlations. In Figs. 6 and 7, we show some representative examples of learned attribute relations by **Z**. In these figures, the averaged correlation and mean absolute error over all persons are shown on the *PRID* and *VIPeR* datasets.

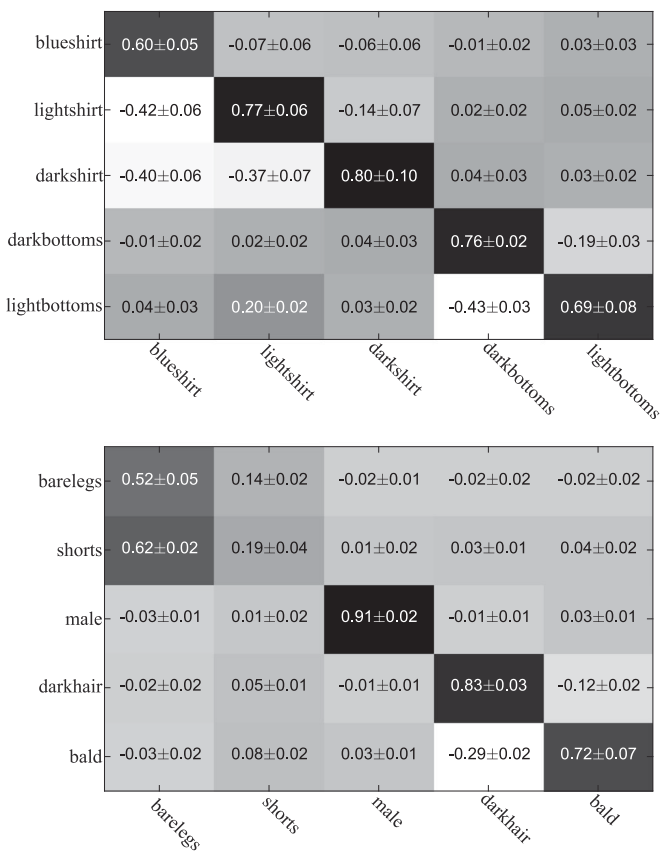


Fig. 6. Examples of attribute correlations learned on the *PRID* dataset. The averaged correlation and mean absolute error across different persons are shown in each example.

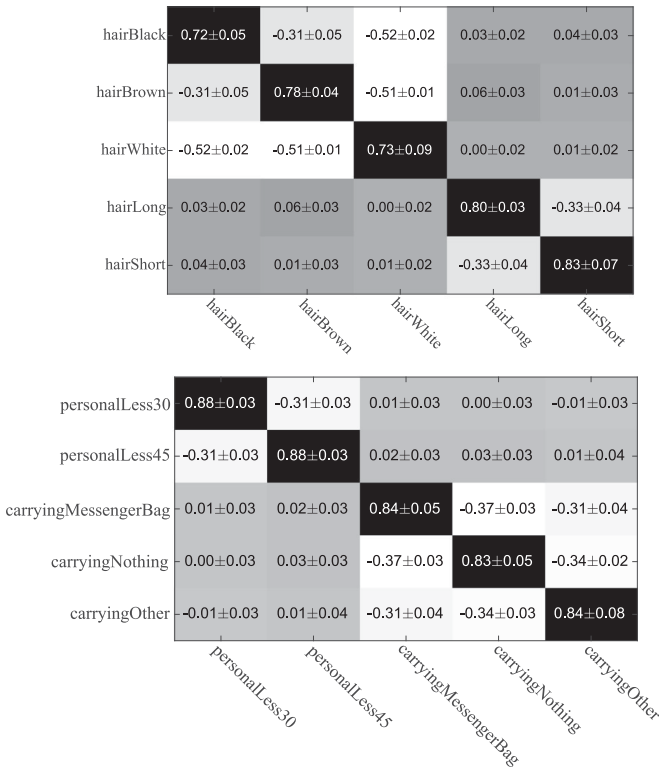


Fig. 7. Examples of attribute correlations learned on the *VIPeR* dataset. The averaged correlation and mean absolute error across different persons are shown in each example.

Because **Z** is trained for each specific person, the mean absolute error in Figs. 6 and 7 essentially represents the variability in attribute-projection that is required to support a good target person-classifier.

It can be seen from Figs. 6 and 7 that, the values of averaged correlation are generally larger than the ones of mean absolute error. It means that the learned correlations between attributes stay relatively stable across different persons. Meanwhile, the figures also demonstrate that different persons do have different sensitivities to attribute correlations. For example, the correlations between attributes like *darkhair* and *barelegs* show large mean absolute error across different persons. This means the correlation between *darkhair* and *barelegs* has diverse impact in identifying different persons. Therefore, it is necessary to train the low rank matrix **Z** for each person to encode the character of each person. We use the trained matrix **Z** on each person rather than a global **Z** also because such person-specific **Z** is easier to optimize and could better avoid under-fitting.

It can be observed from Figs. 6 and 7 that, some attributes are closely related and frequently co-occur in the same image. They reasonably have higher averaged correlation scores, e.g., the attributes *shorts* and *barelegs*. In contrast, a person cannot wear *light bottoms* (or *light shirt*) and *dark bottoms* (or *dark shirt*) at the same time. It is reasonable to observe that such attributes have negative correlations. Attributes *hairlong* and *hairshort* are also negatively correlated.

Similarly, the attribute *carryingNothing* has negative correlations with both the attributes *carryingMessengerBag* and *carryingOther* because a person is unlikely to carry different bags simultaneously. Therefore, the learned attribute correlations are reasonable. We thus use the learned correlations to update the initial attributes to improve their accuracy.

TABLE 9
CMC Scores of Ranks from 1 to 50 on the *iLIDS-VID*, *PRID* and *VIPeR* Datasets by STL, MTL-Att, MTL-FR and the Complete MTL-LOREA

	iLIDS-VID					
Rank	1	5	10	20	30	50
STL	14.7	42.7	41.8	58.5	83.5	91.7
MTL-FR	37.7	54.0	47.4	64.9	85.3	92.5
MTL-Att	40.5	54.9	47.5	64.2	84.2	91.2
MTL-LOREA	**43.0**	**60.0**	**70.2**	**85.3**	**90.2**	**96.3**
	PRID					
Rank	1	5	10	20	30	50
STL	11.3	27.9	41.8	53.0	68.5	74.6
MTL-FR	11.3	34.1	47.4	61.1	69.8	79.0
MTL-Att	12.2	34.7	47.5	61.7	70.9	79.8
MTL-LOREA	**18.0**	**37.4**	**50.1**	**66.6**	**73.1**	**82.3**
	VIPeR					
Rank	1	5	10	20	30	50
STL	13.3	27.4	32.8	42.7	56.2	68.3
MTL-FR	35.3	63.3	75.6	83.8	89.9	94.4
MTL-Att	37.2	64.2	76.3	84.9	91.4	95.3
MTL-LOREA	**42.3**	**72.2**	**81.6**	**89.6**	**93.1**	**97.4**

Numbers indicate the percentage (%) of correct matches within a specific rank.

4.5.3 Evaluation of Individual Components

To verify the effects of individual components in our framework and show that each of them contributes to the performance boost, we evaluate three variants of our approach. Instead of using multi-task learning, we assume tasks are independent and learn classifiers for each task separately. The resulting classifiers are acquired based on Single Tasks Learning (STL). In this way, STL trains its classifiers under different cameras separately, without sharing any data. We also use the original attributes without embedding. We thus have another variant denoted as MTL-Att by discarding the embedding error term in the objective function in Eq. (2). It means during training and testing of MTL-Att, $\widetilde{\mathbf{x}}$ is set as $[\mathbf{x}; \mathbf{a}]$. In addition, we remove the low rank constraint on $\mathbf{Z}$ in Eq. (4), which embeds original attributes to a possible full rank space by making attributes highly uncorrelated. We denote this variant as MTL-FR. The three variants are respectively evaluated on *iLIDS-VID*, *PRID* and *VIPeR* datasets to see how each component affects the performance.

We show CMC scores in Table 9. The results by STL are always worse than those by MTL-LOREA and the other two MTL-based variants. This indicates that learning related tasks simultaneously successfully exploits shared information amongst tasks and increases the discriminative ability of the learned model.

We also find that MTL-FR is inferior to MTL-Att. This suggests that assuming attributes are uncorrelated is unreasonable and degrades the performance of original attributes. However, only using the original attributes without investigating their correlations, MTL-Att cannot produce the best results. The experiments reveal that each of the above mentioned components is important in improving the performance. By integrating all of them, our approach exhibits the best performance.

TABLE 10
CMC Scores at Ranks 1, 5 and 10 by MTL-LOREA with Varying γ (Importance of Attribute Embedding Error Term) and λ (Sparse Regularization) on *iLIDS-VID* Dataset

γ	1	5	10	λ	1	5	10
0.1	42.8	56.9	69.1	0.05	43.2	59.6	69.3
0.3	42.5	56.1	69.9	0.1	42.4	57.5	67.3
0.5	43.0	60.0	70.2	0.3	43.0	60.0	70.2
1	42.9	60.1	68.3	0.5	42.7	59.7	69.9
2	42.4	57.3	68.0	1	42.5	59.1	68.5

Numbers indicate the percentage (%) of correct matches within a specific rank.

TABLE 11
CMC Scores at Ranks 1, 5 and 10 by MTL-LOREA with Varying γ (Importance of Attribute Embedding Error Term) and λ (Sparse Regularization) on *PRID* Dataset

γ	1	5	10	λ	1	5	10
0.1	12.0	38.2	49.3	0.05	16.6	36.9	49.1
0.3	17.5	36.1	50.6	0.1	15.4	37.0	48.6
0.5	18.0	37.4	50.1	0.3	18.0	37.4	50.1
1	16.3	36.7	47.2	0.5	17.2	37.4	50.0
2	18.2	40.1	53.7	1	19.1	37.3	51.2

Numbers indicate the percentage (%) of correct matches within a specific rank.

4.5.4 Evaluation of Parameter Sensitivity

There are two important parameters in our formulation, γ in Eq. (2) which controls the contribution of attribute embedding error term and λ in Eq. (3) which controls the importance of sparse regularization. To demonstrate that our MTL-LOREA approach is not sensitive to these parameters, we conduct experiments by changing the parameters and evaluate the performance of MTL-LOREA on three datasets. During our experiments, we vary one parameter while keeping another one and all the other parts fixed. Results are shown in Tables 10, 11, and 12.

Even though the parameters change within a relatively large range, i.e., the maximal value is 20 times of the minimal value, the performance by MTL-LOREA only slightly changes, i.e., the largest absolute change is no more than 7 percent. Actually, the absolute change is less than 3 percent for most cases, which is negligible given the significant improvement over existing approaches. The experimental results clearly demonstrate that the proposed MTL-LOREA is robust enough and not sensitive to the above mentioned parameters. Therefore, our MTL-LOREA approach does not rely on parameter tuning to obtain outstanding performance, making it suitable for practical applications.

TABLE 12
Precision (P), Recall (R) and F_1-Score (in %) by MTL-LOREA with Varying γ (Importance of Attribute Embedding Error Term) and λ (Sparse Regularization) on *SAIVT-SoftBio* Dataset

γ	P	R	F_1	λ	P	R	F_1
0.1	43.2	64.1	51.7	0.05	43.9	64.4	52.2
0.3	46.0	63.7	53.4	0.1	44.0	63.4	51.9
0.5	45.2	63.7	52.9	0.3	45.2	63.7	52.9
1	44.3	63.2	52.1	0.5	44.1	63.8	52.4
2	43.6	64.0	51.9	1	43.4	64.3	51.8

4.5.5 Weakness

Our approach performs fairly well in dealing with multi-shot datasets, as shown from the experimental results, but there still are several issues to address.

1) When the model is trained under single-shot scenario, it must use data augmentation to generate more training samples. For example, we need to expand one image into 75 images through rotation and translation on the *VIPeR* dataset.
2) When performing person re-identification, our method uses person-specific classifiers. Therefore, we need to train again for each additional person added to the dataset, which takes more time.
3) It is difficult to encode shared information across cameras for every person because there are few cameras but many images of persons. In the future work, we will consider two alternatives: (a) cameras are treated independently, while all persons from a camera share common characteristics; (b) information is shared across both persons and cameras, which could be a more effective solution.
4) Attributes provide auxiliary information apart from low-level features. However, it is usually expensive and impractical to obtain attributes from manual annotations. Even though data driven attribute can be learned, it still requires additional training and annotation efforts. Therefore, it is hard to perform person re-identification tasks with limited training data. Our future work will also work on one-shot attribute learning algorithm to address this problem. We will also combine better features to further improve the performance of our method.
5) Due to its powerful feature learning ability, deep model has shown promising performance on person Re-ID. Previous work [57] has implement a two-side neural network for multi-task and multi-domain learning. It is possible to propose a deep neural network structure that implements multi-task learning and attribute embedding. This will be explored in our future work.

5 Conclusion

We have proposed a multi-task learning formulation with low rank attribute embedding for person re-identification. Multiple cameras are treated as related tasks, whose relationships are decomposed as a low rank structure shared by all tasks and task-specific sparse components for individual tasks by MTL. Both low level features and semantic/data-driven attributes are used. We have further proposed a low rank attribute embedding that learns attributes correlations to convert original binary attributes to continuous attributes, where incorrect and incomplete attributes are rectified and recovered. Our objective function can be effectively solved by an alternating optimization under proper relaxation. Experiments on four datasets have demonstrated the outstanding performance and robustness of the proposed approach.

Acknowledgments

This work was supported in part by National Science Foundation of China 61572050, 91538111, 61620106009, 61429201, and the National 1000 Youth Talents Plan, in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. Chi Su and Fan Yang contributed equally to this work.

References

[1] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.
[2] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, "Multi-task learning for classification with Dirichlet process priors," *J. Mach. Learn. Res.*, vol. 8, pp. 35–63, 2007.
[3] X. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.
[4] M. Lapin, B. Schiele, and M. Hein, "Scalable multitask representation learning for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1434–1441.
[5] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 41–48.
[6] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3567–3574.
[7] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. 10th Asian Conf. Comput. Vis.*, 2011, pp. 501–512.
[8] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 780–793.
[9] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3318–3325.
[10] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[11] 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.*, 2012, pp. 2288–2295.
[12] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.
[13] C. Liu, C. C. Loy, S. Gong, and G. Wang, "POP: Person re-identification post-rank optimisation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 441–448.
[14] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3610–3617.
[15] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.
[16] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 688–703.
[17] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.
[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, 2009, pp. 365–372.
[19] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveillance*, 2013, pp. 244–249.
[20] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3739–3747.
[21] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, 2014.
[22] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surveys*, vol. 46, no. 2, 2013, Art. no. 29.

[23] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: Problem overview and current approaches," *J. Ambient Intell. Humanized Comput.*, vol. 2, no. 2, pp. 127–151, 2011.

[24] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

[25] L. Zheng, L. Sheng, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[26] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2360–2367.

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

[28] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image Vis. Comput.*, vol. 32, no. 6, pp. 379–390, 2014.

[29] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 391–401.

[30] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159.

[31] B. Song, B. Xiang, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[32] R. Zhao, W. Ouyang, and X. Wang, "Learning midlevel filters for person reidentification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 144–151.

[33] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. British Mach. Vis. Conf.*, 2012, pp. 24.1–24.11.

[34] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1249–1258.

[35] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes, "A database for person re-identification in multi-camera surveillance networks," in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl.*, 2012, pp. 1–8.

[36] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPeS: 3D people dataset for surveillance and forensics," in *Proc. Joint ACM Workshop Human Gesture Behavior Understanding*, 2011, pp. 59–64.

[37] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 330–345.

[38] B. Cancela, T. M. Hospedales, and S. Gong, "Open-world person re-identification by multi-label assignment inference," in *Proc. British Mach. Vis. Conf.*, 2014.

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[40] L. Zheng, et al., "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[41] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[42] G. Wang and D. A. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 537–544.

[43] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 155–168.

[44] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 127–140.

[45] D. K. Mahajan, S. Sellamanickam, and V. Nair, "A joint learning framework for attribute models and object descriptions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1227–1234.

[46] T. Mensink, J. J. Verbeek, and G. Csurka, "Tree-structured CRF models for interactive image labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 476–489, Feb. 2013.

[47] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 819–826.

[48] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 402–412.

[49] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. Berlin, Germany: Springer, 2014, pp. 93–117.

[50] R. Layne, T. M. Hospedales, and S. Gong, "Re-id: Hunting attributes in the wild," in *Proc. British Mach. Vis. Conf.*, 2014.

[51] S.-J. Huang, Z.-H. Zhou, and Z. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 949–955.

[52] J. Petterson and T. S. Caetano, "Submodular multi-label learning," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1512–1520.

[53] M.-L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 999–1008.

[54] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 702–710.

[55] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 895–903.

[56] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[57] Y. Yang and T. M. Hospedales, "A unified perspective on multi-domain and multi-task learning," in *Proc. Int. Conf. Learn. Representations*, 2015.

[58] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning shared structures from multiple tasks," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 137–144.

[59] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1179–1187.

[60] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, pp. 615–637, 2005.

[61] Q. Gu, Z. Li, and J. Han, "Learning a kernel for multi-task clustering," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 368–373.

[62] P. Ruvolo and E. Eaton, "Online multi-task learning via sparse dictionary optimization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2062–2068.

[63] L. Han, Y. Zhang, G. Song, and K. Xie, "Encoding tree sparsity in multi-task learning: A probabilistic framework," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1854–1860.

[64] L. Chen, Q. Zhang, and B. Li, "Predicting multiple attributes via relative multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1027–1034.

[65] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1761–1768.

[66] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3567–3574.

[67] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.

[68] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen, "Learning low-rank label correlations for multi-label classification with missing labels," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 1067–1072.

[69] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1042–1050.

[70] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Berlin, Germany: Springer, 2011, pp. 91–102.

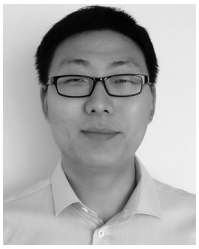
[71] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Performance Eval. Tracking Surveillance*, vol. 3, no. 5, 2007.

[72] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 789–792.

[73] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discriminative binary codes," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 876–889.

[74] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2528–2535.
[75] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[76] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. British Mach. Vis. Conf.*, 2010.
[77] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proc. Int. Conf. Distrib. Smart Cameras*, 2014, Art. no. 10.
[78] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2197–2206.
[79] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4184–4193.
[80] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1565–1573.
[81] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3685–3693.
[82] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[83] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[84] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

Chi Su is working toward the PhD degree in the Institute of Digital Media, EECS, Peking University. His research include computer vision and machine learning, with focus on object detection, object tracking, and human identification and recognition. He is a student member of the IEEE.



Fan Yang (M'10) received the PhD degree in computer science from the University of Maryland, College Park, Maryland, in 2016, where he was also affiliated with the University of Maryland Institute for Advanced Computer Studies (UMIACS). His broad interests include computer vision and machine learning, such as visual search, object tracking and detection, face recognition, person re-identification and deep learning. He has more than 20 publications in premier conferences and journals. He is a member of the IEEE.

Shiliang Zhang received the PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2012. He is currently an assistant professor in the School of Electronic Engineering and Computer Science, Peking University. He was a postdoctoral scientist in NEC Labs America and a postdoctoral research fellow with the University of Texas at San Antonio. His research interests include large-scale image retrieval, person re-identification, and computer vision for autonomous driving. He was awarded the National 1000 Youth Talents Plan of China, Outstanding Doctoral Dissertation Awards from both Chinese Academy of Sciences and Chinese Computer Federation (CCF), President Scholarship by Chinese Academy of Sciences, NEC Laboratories America Spot Recognition Award, and the Microsoft Research Fellowship. He is the recipient of Top 10 percent Paper Award in IEEE MMSP 2011. He is a member of the IEEE.

Qi Tian received the PhD degree in ECE from the University of Illinois at Urbana-Champaign (UIUC), in 2002. He is currently a professor in the Department of Computer Science, University of Texas at San Antonio (UTSA). He was a tenured associate professor from 2008-2012 and a tenure-track assistant professor from 2002-2008. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics and published more than 350 refereed journal and conference papers. He received the Best Paper Awards in ACM ICMR 2015, PCM 2013, MMM 2013, and ACM ICIMCS 2012, a Top 10 percent Paper Award in MMSP 2011, a Student Contest Paper Award in ICASSP 2006. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, Blippar, SALSI, Akiira Media Systems, and UTSA, etc. He received 2010 ACM Service Award and 2016 UTSA Innovation Award in the first category and 2014 Research Achievement Award from College of Science, UTSA. He is an associate editor of the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *ACM Transactions on Multimedia Computing, Communications and Application*, the *Multimedia System Journal*, and in the editorial board of the *Journal of Multimedia* and the *Journal of Machine Vision and Applications*. He is a fellow of the IEEE.

Larry Steven Davis received the BA degree from Colgate University, in 1970 and the MS and PhD degrees in computer science from the University of Maryland, in 1974 and 1976, respectively. From 1977 to 1981, he was an assistant professor in the Department of Computer Science, University of Texas, Austin. He returned to the University of Maryland as an associate professor, in 1981. From 1985 to 1994, he was the director of the University of Maryland Institute for Advanced Computer Studies. From 1999 to 2012, he was the chair of the Computer Science Department in the institute. He is currently a professor in the institute and in the Computer Science Department. He is a fellow of the ACM, the IEEE, and the IAPR.

Wen Gao received the PhD degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He was a professor of computer science in the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a professor in the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is currently a professor of computer science with Peking University, Beijing. He has authored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. He served or serves on the editorial board for several journals, such as the *IEEE Transactions on Circuits and Systems for Video Technology*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Autonomous Mental Development*, the *EURASIP Journal of Image Communications*, and the *Journal of Visual Communication and Image Representation*. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE International Conference on Multimedia & Expo and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. He is a fellow of the IEEE.