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

Chi Su[1*] Fan Yang[2*] Shiliang Zhang[1] Qi Tian[3] Larry S. Davis[2] Wen Gao[1]
[1]Peking University [2]University of Maryland College Park [3]University of Texas at San Antonio

Abstract

We propose a novel Multi-Task Learning with Low Rank Attribute Embedding (MTL-LORAE) framework for person re-identification. Re-identifications from multiple cameras are regarded as related tasks to exploit shared information to improve re-identification accuracy. Both low level features and semantic/data-driven attributes are utilized. Since attributes are generally correlated, we introduce a low rank attribute embedding into the MTL formulation to embed original binary attributes to a continuous attribute space, where incorrect and incomplete attributes are rectified and recovered to better describe people. The learning objective function consists of a quadratic loss regarding class labels and an attribute embedding error, which is solved by an alternating optimization procedure. Experiments on four person re-identification datasets have demonstrated that MTL-LORAE outperforms existing approaches by a large margin and produces promising results.

1. Introduction

The aim of person re-identification is to identify a person in a probe image/video by searching for the most similar instances from a gallery set. Although one can take advantage of increasing amounts of surveillance data to obtain more information to improve re-identification accuracy, it is non-trivial to design an effective re-identification algorithm due to large appearance, pose and illumination changes across images. Additionally, images are usually from multiple cameras with different parameters and viewpoints, making accurate and efficient person re-identification even harder.

Nevertheless, even though the appearance of a person greatly changes, high-level semantic concepts with respect to the person are relatively stable and consistent across different cameras. Such semantic concepts, referred to as attributes, have been widely applied to various vision applications, such as image classification and object detection, and shown promising results. When we describe an image or object by attributes, we obtain a vector in which each dimension indicates whether the corresponding attribute is present or not (or, more generally, its likelihood). In addition, it is intuitive that some attributes frequently co-occur, leading to a few subsets which contain related attributes while are mutually independent. For example, the attribute *female* is likely to be highly related to the attribute *long hair* rather than *short hair*. We show that by utilizing correlations of attributes, attributes of the same person from different cameras can be embedded into a low rank space, where embedded attributes are more accurate and informative for matching. Through the low rank attribute space, we can better match samples of the same person from one camera to another. Additionally, using this low rank embedding, we can prune noisy attributes and recover missing attributes that are introduced by inaccurate human annotation.

Nevertheless, it is computationally expensive to infer attribute correlations using pairs of cameras, which also ignores the relationship of more than two cameras. To utilize relationships of features and attributes more efficiently for matching instances across cameras, we employ the Multi-Task Learning (MTL) [5] algorithm, where one jointly learns solutions to multiple related tasks which benefit each other. MTL has been shown successful in discovering latent relationships among tasks, which cannot be found by learning each task independently. It has been widely applied to machine learning [2, 44] and computer vision [45, 24]. In addition, MTL is particularly suitable for the situation in which only a limited amount of training data is available for each task. By considering re-identifications from multiple cameras as tasks, the MTL framework can be naturally adapted to exploit features and attributes shared across cameras by learning from multiple cameras simultaneously.

In the remainder of the paper we will present a **M**ulti-**T**ask **L**earning algorithm with **LO**w **R**ank **A**ttribute **E**mbedding (MTL-LORAE) for person re-identification. We aim to discover shared information amongst cameras that are treated as related tasks. Given images of people from multiple cameras, we learn a discriminative model using MTL, so that the relationships among images from these cameras can be utilized to improve the quality of the learned model. Both low level features and attributes are used in our

* indicates equal contribution.

MTL objective function. Our low rank attribute embedding is included into the objective function as well to discover relationships of attributes from multiple cameras jointly. In the embedded space, attributes of the same person from different cameras become closer, while attributes of different people become more distinct. Inaccurate and incomplete attributes can be rectified and recovered as well. The low rank structure of the embedding ensures that only a small number of "latent" attributes contribute to the classification. We present an efficient alternating optimization method to solve the MTL-LORAE objective function. We evaluate MTL-LORAE on four person re-identification datasets and demonstrate that MTL-LORAE produces promising results.

Our contributions are four-fold. First, we propose a multi-task learning framework which utilizes standard MTL for person re-identification. Second, we incorporate attributes, which are complementary to low level features, into the re-identification framework by introducing a low rank embedding into the MTL framework to increase the discriminative ability of the learned classifiers. Third, we present a novel objective function including both low level features and attributes, where the task-specific classifiers and low rank attribute embedding are jointly learned by an alternating optimization. Finally, our MTL-LORAE approach outperforms existing approaches by a large margin.

2. Background and Related Work

Person re-identification is an important research topic for video surveillance. Feature design and distance measure are two key components in solving this problem. As for feature design, different kinds of features have been tailored and employed in previous work, including histogram features from various color and textture channels [15, 49], symmetry-driven accumulation of local features [12], features from body parts with pictorial structures [9] to estimate human body configuration, and space-time features from person tracklets [41], *etc*. To use multiple features, Gray *et al*. [15] select a subset of features by boosting for matching pedestrian images, while Liu *et al*. [34] learn person-specific weights to fuse multiple features to improve the description power of multiple features.

Considering distance measures, some works focus on learning an optimal distance metric to measure the similarity between images from two cameras. Pairwise Constrained Component Analysis [11] and Relaxed Pairwise Metric Learning [17] learn a projection from high-dimensional input space to a low-dimensional space, where the distance between pairs of data points satisfies predefined constraints. The Locally-Adaptive Decision Function in [31] jointly learns a distance metric and a locally adaptive thresholding rule. A Probabilistic Relative Distance Comparison model [50] attempts to maximize the likelihood of a true match which has a relatively smaller distance than a false match. A statistical inference perspective is applied in [21] to address the metric learning problem. Kernel-based distance learning has also been used [42] to handle linearly non-separable data. More recently, Zhao *et al*. [48] propose learning mid-level filters, which mainly focuses on cross-view invariance and considers geometric configurations of body parts through patch matching. A deep learning framework to learn filter pairs that encode photometric transforms is presented in [30].

Attributes are semantic concepts of objects, which are manually defined or directly learned from low level features. For person re-identification, attributes are powerful in preserving consistent representations of the same person and capturing differences among different people [29, 26, 27, 28]. However, attributes are mostly used as additional information in conjunction with low level features without considering their correlations. Although a few approaches to object classification have modeled attribute correlations [18, 37, 46], to the best of our knowledge, no work has utilized both low level features and attribute correlations across cameras for re-identification in a principled way.

Multi-Task Learning has been extensively studied. Representative work includes clustered MTL [51], Robust MTL [13] and trace norm regularization [20]. To model the shared information across tasks, a shared low rank structure is widely assumed [7, 6]. Chen *et al*. [8] apply MTL to jointly learn attribute correlations and ranking functions for image ranking. Hwang *et al*. [19] consider attribute classifiers as auxiliary tasks to object classifiers and adopt MTL to learn a shared structure for better classification and attribute prediction. Both [8] and [19] assume attributes are related tasks while we regard cameras as tasks and infer attribute correlations by low rank embedding. For person re-identification, the multi-task support vector ranking adopted in [35] ranks individuals by transferring information of matched/unmatched image pairs from source domain to target domain. Ma *et al*. [36] also apply multi-task learning to replace the universal distance metric for all cameras by multiple Mahalanobis distance metrics, which are different, but related, for camera pairs. We note that our approach is fundamentally different from [35] in that we explicitly model attribute correlations shared by multiple cameras, as well as low level features, without using image pairs. In addition, we seek a shared structure in terms of both low level features and attributes across multiple cameras rather than learning a metric for each pair of cameras, which can be computationally expensive.

3. Methodology

3.1. Problem Formulation

We formulate re-identification as a classification problem by learning a set of classifiers using images from multiple cameras, where a classifier corresponds to a specific

person. Each gallery and probe image is then represented by a vector composed of outputs of these classifiers. By computing distance between vectors of probe and gallery images, we find and rank gallery images to complete re-identification. For simplicity, we do not distinguish between cameras and tasks, and use them interchangeably.

We are given L learning tasks $\{\mathcal{T}^1, \mathcal{T}^2, ..., \mathcal{T}^L\}$ sharing the same feature space. Our goal is to learn multi-class classifiers on a specific task using information from all tasks. In a typical multi-class setting, all tasks have the same set of C classes (persons). In a supervised one-vs-all manner, for the l-th task $\mathcal{T}^l$, we start from binary classification by considering images belonging to the c-th class as positive samples and images from all the other classes in this task as negative samples, where there are totally n_l labeled training samples. By simultaneously learning multiple tasks, our method is able to effectively transfer information from one task to another task, which is particularly desirable when training data from a task is limited. In the following, we omit the class index c from all notation for clarity. For each training sample from the l-th 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 c-th class and -1 otherwise. In addition, each sample has a binary attribute vector $\mathbf{a}_i^l \in \{0, 1\}^k$, which may be semantic and labeled by humans or correspond to learned binary codes such as [22]. For each dimension of $\mathbf{a}_i^l$, 1 denotes that the corresponding attribute is present and 0 otherwise. A predictor f_l with respect to the task $\mathcal{T}^l$ will then be learned.

We can improve the discriminative and generalization ability of predictors by exploiting the relationship amongst tasks. In this way, information from task $\mathcal{T}^i$ is transferred to some other task $\mathcal{T}^j$, where training samples may be limited, so that learning the predictor f_j will benefit from learning on both $\mathcal{T}^i$ and $\mathcal{T}^j$ simultaneously. This motivates us to adopt MTL to address the problem of matching images from different cameras. In the subsequent sections, we will first introduce the low rank attribute embedding (LORAE), followed by the complete MTL formulation, the optimization algorithm and re-identification process.

3.2. Low Rank Attribute Embedding

A simple approach to combine low level features and attributes is to concatenate the feature vectors and original attribute vectors. However, attributes are usually inaccurate or incomplete due to the difficulty of obtaining exhaustive semantic concepts and possible inconsistency between human annotators. The absence of an attribute for an instance does not necessarily indicate that the instance does not have that attribute, which could be incorrectly interpreted by the learning algorithm. Similarly, the presence of an attribute may be noise due to incorrect annotation. Therefore, the learned model based on the original attributes may not describe the instance accurately. Since there are a large number of attributes, they are typically related, which means some attributes often co-occur across different tasks. In this way, the presence of an attribute implies the presence of other attributes that are closely related, which helps to recover missing attributes. On the other hand, some attributes are highly independent, so that they do not occur simultaneously, which helps to remove noisy attributes.

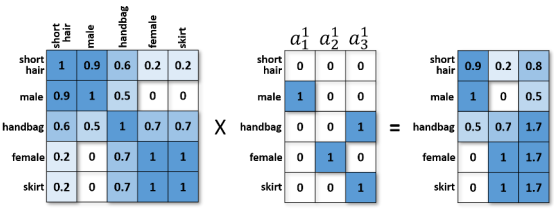

Figure 1. Illustration of low rank attribute embedding with three attribute vectors from task $\mathcal{T}_1$ as examples. With the learned transformation matrix, the original binary attributes are converted to continuous attributes. Semantically related attributes are recovered even though they are absent in the original attribute vectors, *i.e.*, the attribute *female* is non-zero 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 a_3^1.

Following [43], we learn a low rank attribute space to embed the original binary attributes into continuous attributes using attribute dependencies. In particular, there exists a transformation matrix $\mathbf{Z}$ in the low rank space converting an original attribute vector into a new vector with continuous values. The transformation matrix should capture correlations between all attributes pairs since an attribute can be affected by multiple pairs of other attributes globally. Moreover, groups of attributes can be independent from each other, suggesting the low rank property of the transformation matrix.

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 $\mathbf{Z} \in \mathbb{R}^{k \times k}$ is the transformation matrix, and $\text{rank}(\mathbf{Z})$ is the rank of $\mathbf{Z}$. We use linear embeddings although kernel methods can also be applied. The rank constraint imposed on $\mathbf{Z}$ ensures that $\mathbf{Z}$ is low rank, which 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 parameters required for a good embedding are fewer than $k \times k$, which reduces the computational complexity. In this way, we obtain a refined attribute vector with continuous values, which better describes attribute correlations with missing values recovered and noise reduced. An intuitive illustration of the low rank embedding is presented in Figure 1, where missing values are successfully recovered in the embedded continuous attributes.

3.3. Multi-Task Learning with Low Rank Attribute Embedding

The goal of MTL is to learn task-specific predictors simultaneously using the correlations among tasks, so that the shared information can be transferred among tasks. To obtain an accurate transformation matrix $\mathbf{Z}$ for attribute embedding, we propose a unified MTL framework that utilizes attribute correlations across multiple tasks, as well as training task-specific predictors at the same time. For simplicity, we assume a linear classifier for each learning task $\mathcal{T}^l$ 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 form 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 $\mathbf{w}^l \in \mathbb{R}^{d+k}$. We define the loss function as $\ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z})$ which can be any smooth and convex function measuring the discrepancy between groundtruth and predictions from learning. Specifically, we define the loss function as

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

The first term $||y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l||^2$ is the quadratic loss from 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 $\mathbf{Z}$. The results from the embedding should not deviate from the original attributes too much. γ controls the contributions of the two terms.

We denote all the task-specific $\mathbf{w}^l$ as a single weight matrix $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^L] \in \mathbb{R}^{(d+k)\times L}$. Since tasks have shared information and each task also has specific structure, similar to [6], we assume $\mathbf{W}$ is composed of a low rank matrix shared by all tasks and a task-specific sparse component representing the incoherence introduced 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

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

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 Problem (3) is NP-hard since it is non-convex and non-smooth due to the sparse regularization and low rank constraints. It can be converted into a computationally tractable one by 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

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

which is our complete MTL-LORAE formulation. For notational convenience, we denote the value of the objective function as F. By minimizing (4), we obtain the desired weight matrix $\mathbf{W}$ and transformation matrix $\mathbf{Z}$.

3.4. Optimization

The optimization of Problem (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, by alternating between optimizing the objective function with respect to one variable and fixing the other one, the problem is solvable. When 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 $\mathbf{Z}$, Problem (4) reduces to the standard MTL formulation under the assumption of shared low rank structure plus incoherent sparse values

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

where $\ell'(y_i^l, \widetilde{\mathbf{x}}_i^l) = \frac{1}{2}||y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l||^2$. Problem (5) can be solved by the *MixedNorm* approach from [6]. Details can be found in [6].

When 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

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

Relaxing the constraint as a regularization term, we obtain

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

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, due to the non-smooth nuclear constraint on $\mathbf{Z}$, it is not easy to optimize (7). For clarity of notation, we denote the loss function with respect to $\mathbf{Z}$ as $\ell_{\mathbf{Z}}$, and the regularization term as $h_{\mathbf{Z}} = ||\mathbf{Z}||_*$. Problem (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, (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. (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. (9) can be 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 (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 adopt the Accelerated Gradient Method (AGM) [20] to accelerate the 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. The gradient in the j-th 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

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

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 after a pre-defined iteration number P or when the difference $\Delta F = F_{j-1} - F_j > 0$ between consecutive values of the objective function is below a threshold. The entire optimization process is summarized in Algorithm 1.

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 $\mathbf{W}_{(c)} = [\mathbf{w}_{(c)}^1, \mathbf{w}_{(c)}^2, \cdots, \mathbf{w}_{(c)}^L]$ and $\mathbf{Z}_{(c)}$, respectively, by performing the optimization with respect to each class. Given an image taken by the l'-th camera, $l' = 1, 2, \cdots, L$, which is either from the gallery or the probe set, we first extract low level feature $\mathbf{x}^{l'}$ and attribute vector $\mathbf{a}^{l'}$. By applying the transformation matrices, we convert our feature and attribute vectors to a new set of vectors, denoted as $\widetilde{\mathbf{X}}^{l'} = [\widetilde{\mathbf{x}}_{(1)}^{l'}, \widetilde{\mathbf{x}}_{(2)}^{l'}, \cdots, \widetilde{\mathbf{x}}_{(C)}^{l'}] \in \mathbb{R}^{(d+k)\times C}$, where the c-th 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 c-th transformation matrix $\mathbf{Z}_{(c)}$. We further 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

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

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 c-th class. Therefore, each image is finally represented by a C-dimensional score vector $\mathbf{s}$, similar to the reference coding method in [23] and [1]. The similarity between a gallery image and a probe image is then measured by the Euclidean distance between two score vectors. Note that the classes in the training set can be the same as or disjoint from those in the gallery and probe sets.

For multi-shot scenarios, multiple images are presented for each probe/gallery. Given a probe image set containing m_p images, the re-identification process needs to aggregate image-level similarities to rank the gallery image sets. To this end, we adopt the following voting scheme. We first compute the distances between m_p probe images and all gallery images, and then apply a Gaussian kernel to convert the distances to similarities. 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 affect of a gallery set that contains many images.

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 (5) when fixing $\mathbf{Z}$ by *MixedNorm*;
  Update $\mathbf{W} \leftarrow \mathbf{W}_j$;
  Optimize (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

4. Experiments

4.1. Datasets

We evaluate our approach on 4 public datasets, *iLIDS-VID* [41], *PRID* [16] and *VIPeR* [14] and *SAIVT-SoftBio* [3]. The *iLIDS-VID* dataset consists of 600 image sets for 300 people from two cameras at an airport, which 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 people 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 people from 8 different cameras. Since not every person appears in all cameras, following the evaluation setting in [4], we select those appearing in three cameras (#3, #5 and #8) as our evaluation set.

4.2. Implementation Details

We use a 2784-dimensional color and texture descriptor [15] as our low level feature representation, which is composed of 8 color channels (RGB, HSV and YCbCr [1]) and 19 texture channels (Gabor and Schmid). As for attributes, we learn binary SVMs as in [27] to predict the same 20-bit attributes in [27] for *PRID* and 90-bit attributes in [10] for *VIPeR*. For other datasets, we learn attribute functions by [39] in an unsupervised manner on the training set and generate 32-bit attributes. 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, while 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 (3). $\gamma = 0.5$ in (2). Iteration number $P = 500$ and threshold $th = 10^{-5}$ in Algorithm 1.

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 8 completing methods for multi-short re-identification: Salience Matching (Salmatch) [47], Learning Mid-level Filters (LMF) [48], Multi-short Symmetry-driven Accumulation of Local Features (MS-SDALF) [12], Multi-short color with RankSVM (MS-color+RSVM) [41], Multi-short color&LBP with RankSVM (MS-color&LBP+RSVM) [41], color&LBP with Dynamic Time Warping (Color&LBP+DTW) [17], HoGHoF with DTW (HOGHOF+DTW) [25], color&LBP with Discriminative Video fragments selection and Ranking (MS-color&LBP+DVR) [41]. We use cumulative match characteristic (CMC) curves to evaluate performance, and show experimental results in Figure 2 and Table 1.

Our MTL-LORAE approach produces the best results consistently in terms of matching rate with respect to varying ranks. Specifically, when inspecting the matching rate at rank 1 and rank 5, we find a relatively large improvement compared to the best existing method, MS-color&LBP+DVR. Specifically, our method successfully increases the rank 1 accuracy from 34.5% to 43.0%, resulting in an 8.5% improvement. In addition, we obtain nearly 100% matching rate at rank 50, while most compared methods can only achieve 80% matching rate or even less.

4.3.2 *PRID*

Following the protocol in [16], 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 learning-based methods [2]: Relaxed Pairwise Metric Learning (RPML) [17], Probabilistic Relative Distance Comparison (PRDC) [50], RankSVM (RSVM) [38], Salmatch [47], LMF [48], Pairwise Constrained Component Analysis (PCCA) [11], regularized PCCA (rPCCA) [42], Keep It Simple and Straightforward MEtric (KISSME) [21], kernel Local Fisher Discriminant Classifier (kLFDA) [42], Marginal Fisher Analysis (MFA) [42] and Kernel Canonical Correlation Analysis (KCCA) [33]. We again use CMC curves to evaluate performance, as shown in Figure 2 and Table 2.

Our MTL-LORAE approach outperforms all existing methods by a large margin. In particular, our approach achieves 50% matching rate at rank 10, while the matching rate of most other approaches is less than 30%. Except for our approach and KCCA, all other methods are only able to obtain a 50% matching rate as far as rank 55. Our approach also consistently outperforms KCCA, which currently holds state-of-the-art performance, from the beginning. Specifically, on average the absolute improvement in terms of matching rate by our approach over KCCA is 6%, where the margin gradually increases as we move from lower ranks to higher ranks. Notably, the relative improvement by our approach over KCCA is nearly 10%. In terms

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

[2] We do not compare with DVR [41] that uses 89 persons for testing.

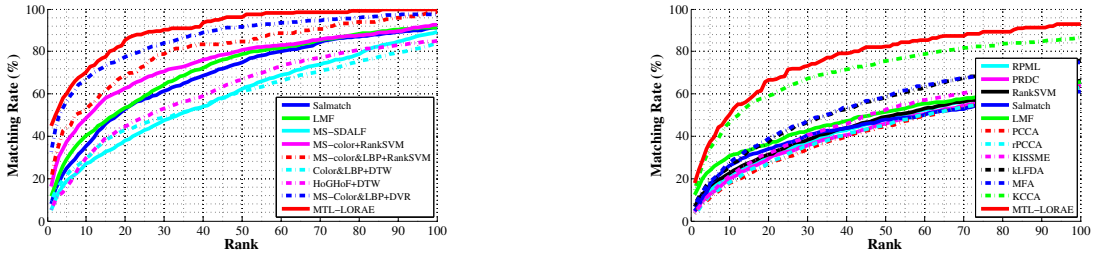

Figure 2. CMC curves of our approach and state-of-the-art approaches on the *iLIDS-VID* dataset (left) and *PRID* dataset (right).

Table 1. CMC scores of ranks from 1 to 50 on the *iLIDS-VID* dataset. Numbers indicate the percentage (%) of correct matches within a specific rank.

Rank	1	5	10	20	30	50
Salmatch [47]	8.0	24.8	35.4	52.9	61.3	74.8
LMF [48]	11.7	29.0	40.3	53.4	64.3	78.8
MS-SDALF [12]	5.1	19.0	27.1	37.9	47.5	62.4
MS-color+RSVM [41]	16.4	37.3	48.5	62.6	70.7	80.6
MS-color&LBP+RSVM [41]	20.0	44.0	52.7	68.0	78.7	84.7
Color&LBP+DTW [41]	9.3	21.6	29.5	43.0	49.1	61.0
HoGHoF+DTW [41]	5.3	16.0	29.7	44.7	53.1	66.7
MS-color&LBP+DVR [41]	34.5	56.4	67.0	77.4	84.0	91.7
MTL-LORAE	**43.0**	**60.0**	**70.2**	**85.3**	**90.2**	**96.3**

Table 2. CMC scores of ranks from 1 to 50 on the *PRID* dataset. Numbers indicate the percentage (%) of correct matches within a specific rank.

Rank	1	5	10	20	30	50
RPML [17]	4.8	14.3	21.6	30.2	37.2	48.1
PRDC [50]	4.5	12.6	19.7	29.5	35.8	46.0
RSVM [38]	6.8	16.5	22.7	31.5	38.4	49.3
Salmatch [47]	4.9	17.5	26.1	33.9	40.5	47.8
LMF [48]	12.5	23.9	30.7	36.5	42.6	51.6
PCCA [11]	3.5	10.9	17.9	27.1	34.2	45.0
rPCCA [42]	3.8	12.3	18.3	27.5	35.2	45.4
KISSME [21]	4.1	12.8	21.1	31.8	40.7	52.5
kLFDA [42]	7.6	18.9	25.6	37.4	46.7	58.5
MFA [42]	7.2	18.7	27.6	39.1	47.4	58.7
KCCA [33]	14.5	34.3	46.7	59.1	67.2	75.4
MTL-LORAE	**18.0**	**37.4**	**50.1**	**66.6**	**73.1**	**82.3**

of the accuracy at rank 1 and rank 5, our approach achieves a matching rate 18% at rank 1 and 37.4% at rank 5, respectively, leading to a 3.5% and 3.1% performance gain over KCCA. When evaluated with more retrieved samples, our approach still secures the best performance. Pairwise distance metric learning based on camera pairs is clearly not powerful enough to obtain good results. Although using kernel tricks, without fully investigating the relationships of features and attributes, KCCA cannot improve the performance much. The experiments further verify that MTL-LORAE, which learns low rank attribute embedding in an MTL setting, successfully exploits relationships among attributes and produces a more discriminative model.

4.3.3 *VIPeR*

We apply data augmentation [3] to generate more training samples for MTL-LORAE. We compare MTL-LORAE with 4 best-performing methods, including 2 recent ones: LOMO+XQDA (LX) [32] and TSR [40], as shown in Table 3. 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 people as the training set and the remaining half 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 [4] does, instead of CMC score that is not applicable to the scenario with more than two cameras. We compare our algorithm to RSVM [38], KISSME [21], RSVM with Conditional Random Field (R-CRF) [4], and KISSME with Conditional Random Field (K-CRF) [4]. Results are averaged over all possible camera pairs of the three cameras, and presented in Table 4. Our MTL-LORAE is able to achieve the best F_1-score, outperforming the best existing method, K-CRF, by 4.6%. In addition, MTL-LORAE 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 ex-

Table 3. CMC scores of ranks from 1 to 20 on the *VIPeR* dataset. Numbers indicate the percentage (%) of correct matches within a specific rank.

Rank	kLFDA [42]	KCCA [33]	LX [32]	TSR [40]	MTL-LORAE
1	32.2	37.3	40.0	31.6	**42.3**
5	65.8	71.4	68.9	68.6	**72.2**
10	79.7	**84.6**	80.5	82.8	81.6
20	90.9	92.3	91.1	**94.6**	89.6

[3]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.

Table 4. Comparison of precision, recall and F_1-score (in %) by existing methods and our approach on *SAIVT-SoftBio* dataset.

	RSVM [38]	KISSME [21]	R-CRF [4]	K-CRF [4]	MTL-LORAE
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**

Table 5. CMC scores of ranks from 1 to 50 on the *iLIDS-VID* and *PRID* datasets by STL, MTL-Att, MTL-FR and the complete MTL-LORAE.

	iLIDS-VID					*PRID*				
Rank	1	5	10	20	50	1	5	10	20	50
STL	14.7	42.7	41.8	58.5	91.7	11.3	27.9	41.8	53.0	74.6
MTL-FR	37.7	54.0	47.4	64.9	92.5	11.3	34.1	47.4	61.1	79.0
MTL-Att	40.5	54.9	47.5	64.2	91.2	12.2	34.7	47.5	61.7	79.8
MTL-LORAE	**43.0**	**60.0**	**70.2**	**85.3**	**96.3**	**18.0**	**37.4**	**50.1**	**66.6**	**82.3**

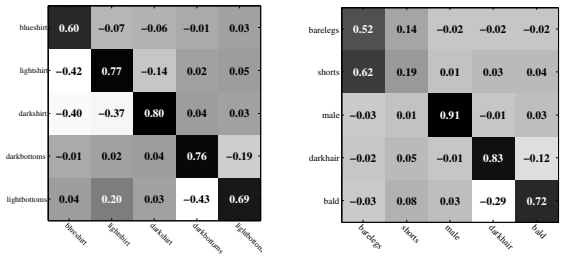

Figure 3. Attribute correlations learned on the *PRID* dataset. Larger values indicate two attribute are more positively correlated. We only show representative examples rather than the whole matrix $\mathbf{Z}$.

isting methods that explicitly deal with all pairs of cameras.

4.4. Discussion

We conduct further experiments to better understand the characteristics of our MTL-LORAE formulation and analyze the contribution of individual components.

Analysis on transformation matrix Z. Based on the assumption that attributes are usually correlated, the learned low rank matrix $\mathbf{Z}$ should preserve attribution correlations well. We show in Figure 3 some representative examples of attribute relations from the learned $\mathbf{Z}$ (averaged over all persons) on the *PRID* dataset since the attributes are manually defined and have semantic meaning. Clearly, some attributes are closely related so that they have higher correlation score, *i.e.*, the attributes *shorts* and *barelegs*, since they should frequently co-occur. In contrast, a person cannot wear *light bottoms* (or *light shirt*) and *dark bottoms* (or *dark shirt*) at the same time so that these two attributes have negative correlation. Similar relationships of other attributes can also be seen. The learned transformation matrix captures the correlations amongst attributes well and thus improves the quality of the original attributes, which justifies the effectiveness of the low rank structure of the embedding space and our learning framework.

Evaluation of individual components. To verify the effect 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 MTL, we assume tasks are independent and learn classifiers for each task separately while keeping other components unchanged, so that the learning is based on single tasks (STL). We also use the original attributes without embedding, and discard the embedding error term in the objective function in (2) to have another variant, MTL-Att. In addition, we remove the low rank constraint on $\mathbf{Z}$ in (4), which embeds original attributes to a possible full rank space by making attributes highly uncorrelated. We denote this variant as MTL-FR. We then evaluate the three variants on *iLIDS-VID* and *PRID* to see how each component affects the performance.

We show CMC scores at some ranks in Table 5. The results by STL are always worse than those by MTL-LORAE and other two MTL-based variants, which indicates that learning related tasks simultaneously successfully exploits shared information amongst tasks and thus increases the discriminative ability of the learned model. We also find that MTL-FR is inferior to MTL-Att, suggesting that assuming attributes are uncorrelated is unreasonable and even hurts performance. However, only using the original attributes without investigating their correlations, MTL-Att cannot produce the best results, although it already outperforms most existing approaches. The experiments reveal that individual components, *i.e.*, MTL and low rank embedding, are integrated into our formulation in a principled way and together improve the performance.

5. Conclusion

We have proposed a multi-task learning (MTL) 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.

Acknowledgement. This research was partially supported by ONR MURI grant N000141010934 and National Science Foundation of China (NSFC) 61429201. Q. Tian was supported in part by ARO grants W911NF-15-1-0290 and W911NF-12-1-0057 and Faculty Research Awards by NEC Laboratories of America. S. Zhang was supported in part by National Science Foundation of China (NSFC) 61572050, the National 1000 Youth Talents Plan of China.

References

[1] L. An, M. Kafai, S. Yang, and B. Bhanu. Reference-based person re-identification. In *AVSS*, pages 244–249, 2013. 5

[2] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005. 1

[3] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes. A database for person re-identification in multi-camera surveillance networks. *DICTA*, 2012. 6

[4] B. Cancela, T. M. Hospedales, and S. Gong. Open-world person re-identification by multi-label assignment inference. 2014. 6, 7, 8

[5] R. Caruana. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, 1993. 1

[6] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. In *KDD*, 2010. 2, 4

[7] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *ICML*, 2009. 2

[8] L. Chen, Q. Zhang, and B. Li. Predicting multiple attributes via relative multi-task learning. In *CVPR*, pages 1027–1034, 2014. 2

[9] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 2

[10] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, pages 789–792, 2014. 6

[11] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2011. 2, 6, 7

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2, 6, 7

[13] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *KDD*, 2012. 2

[14] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007. 6

[15] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 2, 6

[16] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011. 6

[17] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*. 2012. 2, 6, 7

[18] S.-J. Huang, Z.-H. Zhou, and Z. Zhou. Multi-label learning by exploiting label correlations locally. In *AAAI*, 2012. 2

[19] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *CVPR*, pages 1761–1768, 2011. 2

[20] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML*, 2009. 2, 5

[21] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *CVPR*, 2012. 2, 6, 7, 8

[22] B. Kulis and T. Darrell. Learning to hash with binary reconstructive embeddings. In *NIPS*, 2009. 3

[23] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372, 2009. 5

[24] M. Lapin, B. Schiele, and M. Hein. Scalable multitask representation learning for scene classification. In *CVPR*, 2014. 1

[25] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 6

[26] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *ECCV Workshops*, 2012. 2

[27] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014. 2, 6

[28] R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *BMVC*. 2014. 2

[29] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, 2012. 2

[30] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[31] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 2

[32] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 7

[33] G. Lisanti, I. Masi, and A. Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *ICDSC*, 2014. 6, 7

[34] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012. 2

[35] A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013. 2

[36] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014. 2

[37] J. Petterson and T. S. Caetano. Submodular multi-label learning. In *NIPS*, 2011. 2

[38] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 6, 7, 8

[39] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. 2012. 6

[40] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, pages 4184–4193, 2015. 7

[41] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*. 2014. 2, 6, 7

[42] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014. 2, 6, 7

[43] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *ICDM*, 2014. 3

[44] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007. 1

[45] X. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012. 1

[46] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *KDD*, 2010. 2

[47] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 6, 7

[48] R. Zhao, W. Ouyang, and X. Wang. Learning midlevel filters for person reidentification. In *CVPR*, 2014. 2, 6, 7

[49] L. Zheng, L. Sheng, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2

[50] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. In *CVPR*, 2013. 2, 6, 7

[51] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, 2011. 2