Unsupervised Cross-Dataset Transfer Learning for Person Re-identification

Peixi Peng[1,5], Tao Xiang[2], Yaowei Wang[3*], Massimiliano Pontil[4],
Shaogang Gong[2], Tiejun Huang[1],Yonghong Tian[1,5*]
[1]National Engineering Laboratory for Video Technology, Peking University, Beijing, China
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[3]Department of Electronic Engineering, Beijing Institute of Technology, China
[4]Italian Institute of Technology, Italy
[5]Cooperative Medianet Innovation Center, China
{pxpeng, yhtian, tjhuang}@pku.edu.cn, yaoweiwang@bit.edu.cn,
{t.xiang, s.gong}@qmul.ac.uk, massimiliano.pontil@iit.it

Abstract

Most existing person re-identification (Re-ID) approaches follow a supervised learning framework, in which a large number of labelled matching pairs are required for training. This severely limits their scalability in real-world applications. To overcome this limitation, we develop a novel cross-dataset transfer learning approach to learn a discriminative representation. It is unsupervised in the sense that the target dataset is completely unlabelled. Specifically, we present an multi-task dictionary learning method which is able to learn a dataset-shared but target-data-biased representation. Experimental results on five benchmark datasets demonstrate that the method significantly outperforms the state-of-the-art.

1. Introduction

Person re-identification (Re-ID) is the problem of matching people across non-overlapping camera views. It has become one of the most studied problems in video surveillance due to its great potentials for security and safety management applications. Despite the best efforts from the computer vision researchers [33, 10], it remains an unsolved problem. This is because a person's appearance often changes dramatically across camera views due to changes in body pose, view angle, occlusion and illumination conditions.

To address these challenges, most existing research efforts on Re-ID [25, 15, 16, 4, 20, 37, 39, 36, 23, 31, 2] are based on supervised learning. Specifically, they require a large number of labelled matching pairs across each two camera views to learn a representation or matching function that is invariant to the appearance changes. However, relying on manually labelled data for each camera-pair leads to poor *scalability*. This is due to two reasons: (1) For each pair of cameras, eye-balling the two views to annotate *correctly* matching pairs among hundreds of imposters is a tough job even for humans. (2) Given a camera network of even a moderate size, e.g. one installed in an underground station, there can be easily over one hundred cameras and thousands of camera pairs. Since hundreds of labelled image pairs are typically needed from each camera pair for supervised learning, the labelling cost would be prohibitively high. This scalability issue thus severely limits the applicability of the existing methods.

In order to make a person Re-ID model scalable, one solution is to utilise the unlabelled data, which are abundant in the context of Re-ID – in a busy public space, thousands of people pass by in each camera view everyday. There are a few existing efforts on exploiting unlabelled data for unsupervised Re-ID modelling [24, 34, 6]. However, compared to supervised learning approaches, the matching performance of unsupervised models are typically much weaker, rendering them less effective. The reason is that without labelled matching pairs across camera views, existing unsupervised models are unable to learn what makes a person recognisable under severe appearance changes.

Different from existing unsupervised Re-ID methods, we propose to solve the Re-ID problem without any labelled matching pairs of target data using cross-dataset transfer learning. The idea is that labelling matching pairs for a set of given target camera views is tedious for practical applications. However, there already exist labelled datasets collected elsewhere from other camera networks; it is therefore possible to learn a representation that captures the

*Corresponding author: Yaowei Wang and Yonghong Tian (email:yaoweiwang@bit.edu.cn and yhtian@pku.edu.cn).

view-invariant aspects of a person's appearance and transfer it to the target dataset for matching. Since the target views/dataset contains no label, this is an unsupervised learning problem [7, 9, 11]. It is thus an extremely challenging problem because not only the source and target domains are different (different camera views), more critically they also have different recognition tasks (different sets of person identities to be matched in each domain), in contrast to most existing transfer learning assumptions.

To solve the above unsupervised cross-dataset transfer learning problem, we propose a novel asymmetric multi-task learning approach which is able to transfer a view-invariant representation from a number of existing labelled source datasets, each consisting of camera pairs with different viewing conditions, to an unlabelled target dataset containing people who never appeared in the source datasets. Our method is based on dictionary learning, that is, we assume that a person's appearance can be represented as a linear combination of latent factors each corresponding to a dictionary atom. Furthermore, we assume that some of the atoms are view/dataset-independent, thus shared across different datasets/tasks, whilst others are unique to each dataset and may or may not be useful for Re-ID in a *new* unlabelled target dataset. This results in three types of dictionaries being jointly learned using all datasets.

The key strength of our method, which also distinguishes it from existing multi-task learning methods [30], is that it is able to learn from unlabelled target data. This is precisely why a dictionary learning model is adopted – it is originally designed for unsupervised learning and can thus be naturally reformulated for unsupervised transfer learning. To this end, graph Laplacian regularisation terms with iterative updating are introduced in our formulation in order to learn from both the labelled information from the source data and the unlabelled data from the target data. In addition, to make the learned dictionary biased towards the target dataset, different decompositions of dictionaries are introduced for the source and target datasets respectively to reflect the fact that our multi-task learning model is asymmetric, i.e. the multi-task joint learning only aims to benefit the target task.

2. Related Work

Most existing person Re-ID models are supervised, based on either distance metric learning [18, 25, 15, 16, 4, 37, 39], discriminative subspace learning [25, 23], learning to rank [31], or deep learning [2]. These models are thus unscalable as they need a large number of labelled data (cross-view matched image pairs) to train for each given camera pair. In particular, each learned model is camera-pair-specific thus cannot be directly applied to another new camera pair due to the view condition changes, as verified by our experiments (Sec. 4).

To address the scalability issue, there have been a number of unsupervised Re-ID methods proposed in the literature, including two types: those designing hand-crafted appearance features [24, 6, 28] and those modelling localised saliency statistics [38, 34]. However, compared to supervised learning approaches, both approaches yield much weaker performance, since without pairwise identity labels they cannot exploit cross-view identity-discriminative information that is critical for matching. To strike a balance between scalability and matching accuracy, a semi-supervised approach [26] is proposed. Nevertheless, it still requires a fairly substantial amount of pairwise labelling which is not possible for large scale deployment in real-world applications.

Recently, cross-dataset transfer learning has been adopted for Re-ID in the hope that labelled data from other camera views/datasets can provide transferable identity-discriminative information for a given target dataset. Note that this cross-dataset transfer learning problem is very different from the same-dataset cross-identity or same-dataset cross-view problems tackled in some early transfer Re-ID works [22, 44]. When both the dataset/domain and the identities are different, the transfer learning problem considered in this work is much harder. Among the existing cross-dataset transfer learning works, [19] adopted an SVM multi-kernel learning transfer strategy, and both [29] and [35] employed multi-task metric learning models. All of theses works are supervised and they need labelled data in the target dataset.

As far as we know, the only existing unsupervised cross-dataset transfer learning model for Re-ID is the work in [27]. The model proposed in [27] utilises cross-domain ranking SVMs. Unlike the dictionary learning model employed in this work, an SVM-based model does not naturally learns from completely unlabelled data. As a result, their target dataset is not exactly unlabelled: it is assumed that negative image pairs are given for the target dataset. Therefore, strictly speaking, the model in [27] is a weakly-supervised rather than unsupervised model. In contrast, our model is completely *unsupervised* without requiring any labelled data from the target dataset. Our experiments show that our method significantly outperforms that of [27], even with less supervision.

Beyond person Re-ID, dictionary learning for sparse coding has been extensively studied [17, 1]. Graph Laplacian regularisation has also been explored in a sparse coding formulation before, for problems such as unsupervised clustering [8, 42], or supervised face verification/recognition [13]. Our model differs in that (1) dictionary learning is performed under an asymmetric multi-task learning framework, hence the unique design of different decompositions of dictionaries for the source and target tasks; and (2) the Laplacian regularisation terms are updated iteratively to adapt transferable knowledge learned from the labelled

source data to the unlabelled target data.

Note that a number of works [41, 40, 32] have exploited domain adaptation for cross-view classification or verification of faces/actions, based on dictionary learning and/or sparse representation models. They are thus related to our work. But there are significant differences. In particular, some of them [41, 40] are supervised and require labelled training data from the target domains. The work in [32] is unsupervised and based on unsupervised domain adaptation [7, 9, 11]. Nevertheless they tackle a within-dataset cross-camera view domain adaptation problem. This is fundamentally different to our cross-dataset transfer learning problem: the domain change is much greater across datasets, and importantly the images from cross-domain/view but same dataset contain people of the same identities, whilst a completely different set of people are captured in different datasets. In our experiments, we demonstrate that these unsupervised domain adaptation methods do not work on our cross-dataset transfer learning task because the target dataset contains different classes.

Contributions The main contributions of this work are: (1) We formulate the Re-ID problem as an unsupervised cross-dataset transfer learning problem and do not require any labelled data from the target dataset, and (2) a novel asymmetric multi-task dictionary learning model is proposed to learn view-invariant and identity-discriminative information from unlabelled target data. Extensive experiments are carried out on five widely used Re-ID benchmark datasets. The results show that the proposed model significantly outperforms the state-of-the-art Re-ID approaches, as well as existing transfer learning models, under both unsupervised and semi-supervised settings.

3. Methodology

3.1. Problem Definition

Assume a number of source datasets are collected from different camera networks each consisting of two camera views[1]. The images in the source datasets (domains) are paired across camera views by the person's identity, i.e. they are pairwise labelled. An unlabelled target dataset is captured from an entirely different domain (camera view/location) and contains a completely different set of identities/classes. Therefore, the unsupervised transfer learning for Re-ID problem is defined as the problem of learning the optimal representation/matching function for the target dataset/domain using knowledge transferred from the labelled source datasets.

3.2. Formulation

Taking a multi-task learning approach, we consider learning a Re-ID model for each dataset as a task. We wish to learn all tasks jointly so that they can benefit each other. Importantly, since we are only concerned with the target task, the multi-task model is asymmetric and biased towards the target task. Formally, assume $X_t \in \mathbb{R}^{M \times N_t}$ is a feature matrix with each column $x_{t,i}$ corresponding to an M-dimensional feature vector representing the i-th person's appearance in the dataset t ($t = 1, ..., T$) consisting of N_t samples.

Assume task T is the target task and the others are source tasks. Adopting a dictionary learning model, for each task/dataset, the goal is to learn a shared dictionary $D \in \mathbb{R}^{M \times k}$ using all datasets $\{X_1, ..., X_T\}$. With this dictionary, each M-dimensional feature vector, regardless which view it comes from, is projected into a lower k-dimensional subspace spanned by the k dictionary atoms (columns of D) so that the corresponding coefficients (code vectors) can be matched by the cosine distance in this subspace. The idea is that each atom or the dimension of the subspace corresponds to a latent appearance attribute which is invariant to the camera view changes, thus useful for cross-view matching. Figure 1 shows some examples of latent attributes learned by the proposed model.

In a multi-task dictionary learning model, it is necessary to decompose the dictionary into two parts: the one shared between the tasks, which captures latent attributes that are invariant against any view changes, and a task-specific one that captures dataset-unique aspects of human appearance [5]. In addition, it is important to note that apart from the latent attributes that can contribute to Re-ID, there are also other aspects of appearance that are variant to view changes. These appearance aspects must be modelled as part of the dictionary as well. Furthermore, the decomposition should be different for the source and target datasets as we only care about the target one. Based on these consider-

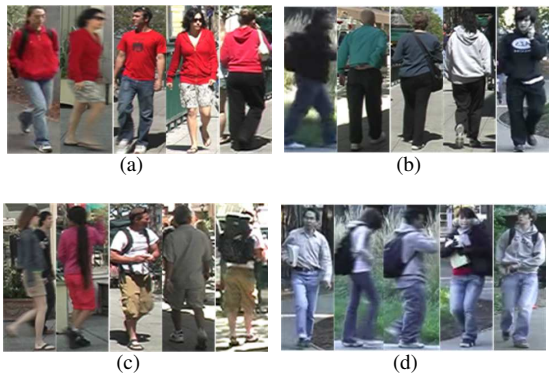


Figure 1: Some samples of latent attributes discovered by the proposed Unsupervised Multi-task Dictionary Learning (UMDL) model. It is clear that the latent attributes are visually and semantically meaningful. (a) Upper body red. (b) Lower body black. (c) Short trousers. (d) Jeans.

[1] This is for simplification of notations; datasets of more than two views can be easily modelled by our model.

ations, three types of dictionaries are introduced in our Unsupervised Multi-task Dictionary Learning (UMDL) model: (1) Task-shared dictionary D^s which is used to encode the dataset/view invariant latent attributes, and is shared by all tasks, (2) dictionary unique to the target task D_T^u that is view-invariant , and (3) task-specific residual dictionary D_t^r $(t = 1, ..., T)$ which is task-specific and used to encode the residual parts of features which cannot be modelled by D^s (source tasks) or D^s and D_T^u (target task). It is clear that the source and target tasks are treated differently: for the target task, an additional third dictionary D_T^u is needed to account for view-invariant but dataset-variant latent attributes unique to the target views.

Now we can formulate our UMDL method as:

$$
\begin{aligned}
&[D^s, D_T^u, D_1^r, \cdots, D_T^r] = \arg\min \\
&\sum_{t=1}^{T-1} \eta_t^2 \{\|X_t - D^s A_t^s\|_F^2 + \|X_t - D^s A_t^s - D_t^r A_t^r\|_F^2\} + \\
&\|X_T - D^s A_T^s\|_F^2 + \|X_T - D^s A_T^s - D_T^u A_T^u\|_F^2 + \\
&\|X_T - D^s A_T^s - D_T^u A_T^u - D_T^r A_T^r\|_F^2 + \\
&\lambda \sum_{t=1}^{T} \sum_{i,j=1}^{N_t} w_{t,i,j} \left\|a_{t,i}^s - a_{t,j}^s\right\|^2 + \lambda \sum_{i,j=1}^{N_T} w_{T,i,j} \left\|a_{T,i}^u - a_{T,j}^u\right\|^2, \\
&s.t.\ \|d_i^s\|_2^2 \le 1,\ \left\|d_{T,i}^u\right\|_2^2 \le 1,\ \left\|d_{t,i}^r\right\|_2^2 \le 1,\ \forall i, t
\end{aligned}
\tag{1}
$$

where matrices A_t^s, A_t^r and A_T^u are codes corresponding to dictionaries D^s, D_t^r and D_T^u respectively; d_i^s, $d_{t,i}^r$ and $d_{T,i}^u$ are the i^{th} column of D^s, D_t^r and D_T^u respectively; $a_{t,i}^s$ is the i^{th} column of A_t^s and $a_{T,i}^u$ is the i^{th} column of A_T^u; η_t and λ are weights of various cost function terms; and W_t is the affinity matrix for the task t indicating the relationships among different training samples. Specifically, for the labelled source datasets, $w_{t,i,j} = 1$ if $x_{t,i}$ and $x_{t,j}$ are of the same person across views and $w_{t,i,j} = 0$ otherwise. For the target task, W_T is initialised as a zero matrix because the target task are unlabelled.

There are seven terms in this cost function and they fall into two categories: the first five are reconstruction error terms that make sure that the learned dictionaries can encode the feature matrices well, and the last two are graph Laplacian regularisation terms that enforce that similar codes are obtained for instances of the same person across camera views. Note that these two regularisation terms are put on the codes obtained using D^s and D_T^u only. As for those obtained using D_t^r, they are not subject to the graph Laplacian regularisation because they are either untransferrable to the target task or are view-variant thus useless for Re-ID. Note that since W_T is a zero matrix, it seems to make no sense to have the seventh term for the target task T. However, we shall see later that W_T will be updated iteratively once a better representation for Re-ID is learned. The last two terms can be rewritten using the Laplacian matrix as

$$
\sum_{i,j=1}^{N_t} w_{t,i,j} \left\|a_{t,i}^s - a_{t,j}^s\right\|^2 = \mathrm{Tr}(A_t^s L_t A_t^{s\prime}),
$$

where $L_t = D_t - W_t$ and D_t is a diagonal matrix whose diagonal elements are the sums of the row elements of W_t.

Now we explain how the first five reconstruction error terms are designed. First, we note that the reconstruction error terms are formulated stepwise by the priority of different dictionaries. Let us consider the first two terms $\|X_t - D^s A_t^s\|_F^2 + \|X_t - D^s A_t^s - D_t^r A_t^r\|_F^2$ for the source task t. The minimisation of the first reconstruction error term enables D^s to encode X_t as much as possible while the minimisation of the second reconstruction error term enables D_t^r to encode and align the residual part of X_t that cannot be coded using D^s. This stepwise reconstruction formulation is also applied to the target task T resulting in terms 3–5. However, as an asymmetric multi-task learning model, the target task is biased with three dictionaries rather than two, hence the three terms. We shall see in our experiments that both the stepwise reconstruction terms and asymmetric design contribute positively to the final performance of the model.

Note that unlike conventional dictionary learning for sparse coding models, in our model, there is no L_1 sparsity penalty term. This is because (1) empirically, we find that less-sparse codes contain richer information for Re-ID, and (2) removing these L_1 terms leads to a simpler optimisation problem.

3.3. Optimisation

Next we describe how the optimisation problem in (1) can be solved. This optimisation problem is divided into the following subproblems:

(1) Computing A_t^s and A_T^u Given fixed D^s, D_t^r, A_t^r and D_T^u, the coding problem of the task t $(t = 1, ..., T)$ becomes:

$$
\min \left\|\tilde{X}_t - \tilde{D}\tilde{A}_t\right\|_F^2 + \lambda \mathrm{Tr}(\tilde{A}_t L_t \tilde{A}_t'), \tag{2}
$$

where, for the target task:

$$
\tilde{X}_T = \begin{bmatrix} X_T \\ X_T \\ X_T - D_T^r A_T^r \end{bmatrix}, \tilde{D} = \begin{bmatrix} D^s, \mathbf{0} \\ D^s, D_T^u \\ D^s, D_T^u \end{bmatrix}, \tilde{A}_T = \begin{bmatrix} A_T^s \\ A_T^u \end{bmatrix},
$$

and for the source tasks:

$$
\tilde{X}_t = \begin{bmatrix} \eta_t X_t \\ \eta_t \left(X_t - D_t^r A_t^r\right) \end{bmatrix}, \tilde{D} = \begin{bmatrix} \eta_t D^s \\ \eta_t D^s \end{bmatrix}, \tilde{A}_t = \begin{bmatrix} A_t^s \end{bmatrix}.
$$

Let the derivative of (2) equals to 0 and the analytical solution of $\tilde{a}_{t,i}$ (the i^{th} column of $\tilde{A}_t$) can be obtained as:

$$
\tilde{a}_{t,i} = \left(\tilde{D}'\tilde{D} + 2\lambda l_{t,i,i} I\right)^{-1} \left(\tilde{D}'\tilde{x}_{t,i} - 2\lambda \sum_{k \neq i} \tilde{x}_{t,k} l_{t,k,i}\right),
$$

where $l_{t,k,i}$ is the (k,i) element of L_t. I is the identity matrix and $\tilde{x}_{t,i}$ is the i^{th} column of $\tilde{X}_t$.

(2) Computing A_t^r Fix other terms and A_t^r is solved as:

$$\begin{aligned} &\text{For the target task:} \\ &\min \|X_T - D^s A_T^s - D_T^u A_T^u - D_T^r A_T^r\|_F^2, \\ &\text{and for the source tasks:} \\ &\min \|X_t - D^s A_t^s - D_t^r A_t^r\|_F^2. \end{aligned} \tag{3}$$

Let the derivative of (3) equals to 0 and the analytical solution of A_t^r can be obtained as:

$$\begin{aligned} &\text{For the target task:} \\ &A_T^r = \left(D_T^{r\prime} D_T^r\right)^{-1} D_T^{r\prime} \left(X_T - D^s A_T^s - D_T^u A_T^u\right), \\ &\text{and for the source tasks:} \\ &A_t^r = \left(D_t^{r\prime} D_t^r\right)^{-1} D_t^{r\prime} \left(X_t - D^s A_t^s\right). \end{aligned}$$

(3) Updating dictionaries When D_t^r, A_t^s, A_t^r $(t = 1, ..., T)$, D_T^u and A_T^u are given, D^s is optimised as:

$$\min \|\mathcal{X} - D^s \mathcal{A}\|_F^2, \ \ s.t. \ \|d_i^s\|_2^2 \leq 1, (\forall i), \tag{4}$$

where

$$\begin{aligned} &\mathcal{X} = [\eta_1 X_1, ..., \eta_{T-1} X_{T-1}, \eta_1 (X_1 - D_1^r A_1^r), ..., \eta_{T-1}(X_{T-1} - \\ &D_{T-1}^r A_{T-1}^r), X_T, X_T - D_T^u A_T^u, X_T - D_T^u A_T^u - D_T^r A_T^r], \\ &\mathcal{A} = [\eta_1 A_1^s, ..., \eta_{T-1} A_{T-1}^s, \eta_1 A_1^s, ..., \eta_{T-1} A_{T-1}^s, A_T^s, A_T^s, A_T^s]. \end{aligned} \tag{5}$$

(4) can be optimised by the Lagrange dual and the analytical solution of D^s can be computed as: $D^s = (\mathcal{X}\mathcal{A}')\left(\mathcal{A}\mathcal{A}' + \Lambda\right)^{-1}$, where Λ is a diagonal matrix constructed from all the dual variables.

Then, for the target task, fix the dictionaries D^s, D_T^r and codes A_T^s, A_T^u, A_T^r, then D_T^u can be updated by:

$$\min \|\mathcal{X}_T^u - D_T^u \mathcal{A}_T^u\|_F^2, \ \ s.t. \ \left\|d_{T,i}^u\right\|_2^2 \leq 1, (\forall i), \tag{6}$$

where

$$\begin{aligned} &\mathcal{X}_T^u = [X_T, X_T - D^s A_T^s, X_T - D^s A_T^s - D_T^r A_T^r], \\ &\mathcal{A}_T^u = [A_T^u, A_T^u, A_T^u]. \end{aligned} \tag{7}$$

At last, fix D^s, D_T^u , A_t^s, A_T^u and A_t^r, the objective function to solve D_t^r is $(t = 1, ..., T)$:

$$\min \|\mathcal{X}_t^r - D_t^r A_t^r\|_F^2, \ \ s.t. \ \left\|d_{t,i}^r\right\|_2^2 \leq 1 (\forall i), \tag{8}$$

where

$$\begin{aligned} &\text{For the target task:} \\ &\mathcal{X}_T^r = X_T - D^s A_T^s - D_T^u A_T^u, \\ &\text{and for the source tasks:} \\ &\mathcal{X}_t^r = X_t - D^s A_t^s. \end{aligned} \tag{9}$$

Algorithm 1: Unsupervised Multi-task learning

Input: X_t; initialise D^s, D_t^r and D_T^u randomly; $A_t^r \rightarrow \mathbf{0}$;
Output: D^s, D_T^u, D_t^r, A_t^s, A_T^u and A_t^r; $(t = 1, ..., T)$.
while *Non-convergence* **do**
  for $t = 1 \rightarrow T$ **do**
    if *Source tasks* **then**
      Fix D^s,D_t^r, and A_t^r, then calculate A_t^s by (2).
      Fix D^s,D_t^r, and A_t^s, then calculate A_t^r by (3).
    if *Target task* **then**
      Fix D^s,D_T^r, $D_T^u$and A_T^r, then calculate A_T^s and A_T^u by (2).
      Fix D^s,D_T^r, D_T^u, A_T^s and A_T^u, then calculate A_T^r by (3).
  Fix other terms, update D^s by (4).
  for $t = 1 \rightarrow T$ **do**
    if *Source tasks* **then**
      Update D_t^r with fixed D^s, A_t^s and A_t^r by (8).
    if *Target task* **then**
      Update D_T^u with fixed D^s,D_T^r, A_T^s, A_T^r and A_T^u by (6).
      Update D_T^r with fixed D^s,D_T^u, A_T^s, A_T^r and A_T^u by (8).

(6) and (8) can be solved similarly as (4):

$$\begin{aligned} D_T^u &= \left(\mathcal{X}_T^u \mathcal{A}_T^{u\prime}\right)\left(\mathcal{A}_T^u \mathcal{A}_T^{u\prime} + \Lambda\right)^{-1}, \\ D_t^r &= \left(\mathcal{X}_t^r A_t^{r\prime}\right)\left(A_t^r A_t^{r\prime} + \Lambda\right)^{-1}. \end{aligned}$$

Alg. 1 summarises our optimisation algorithm. It converges after a few (< 30) iterations in our experiments.

Iterative Updating W_T After running Alg. 1, each training sample $x_{T,i}$ from the target task will be coded by (10) (detailed below) and the code is $a_{T,i} = \left[a_{T,i}^s, a_{T,i}^u, a_{T,i}^r\right]$. With this code, we can measure the similarity between each pair of target data samples across views and recompute W_T. This matrix now captures the soft relationships among the training samples from the target tasks which we aim to preserve in the lower dimensional space spanned by the dictionary columns. Specifically, if $a_{T,j}$ is among the k-nearest-neighbours of $a_{T,i}$ and $a_{T,i}$ is among the k-nearest-neighbours of $a_{T,j}$, $w_{T,i,j} = \frac{a_{T,i} \cdot a_{T,j}}{\|a_{T,i}\| \|a_{T,j}\|}$, otherwise, $w_{T,i,j} = 0$. With the updated W_T, we re-run Alg. 1 to enter the next iteration. The iterations terminate when a stopping criterion is met, and the number of iterations is typically < 5 in our experiments.

3.4. Application to Re-ID

Re-ID for the Target Task After training the model, each test sample $x_{T,i}$ from the target task T can be encoded via D^s, D_T^u and D_T^r as $\left[a_{T,i}^s, a_{T,i}^u, a_{T,i}^r\right]$ by solving the fol-

lowing problem:

$$\left[a^s_{T,i}, a^u_{T,i}, a^r_{T,i}\right] = \min \left\|x_{T,i} - D^s a^s_{T,i} - D^u_T a^u_{T,i} - D^r_T a^r_{T,i}\right\|_2^2 + \gamma \left\|a^s_{T,i}\right\|_2^2 + \gamma \left\|a^u_{T,i}\right\|_2^2 + \gamma \left\|a^r_{T,i}\right\|_2^2, \quad (10)$$

which can be solved easily by a linear system. With this new representation, Re-ID is done simply by computing the cosine distance between the code vectors of a probe and a gallery sample.

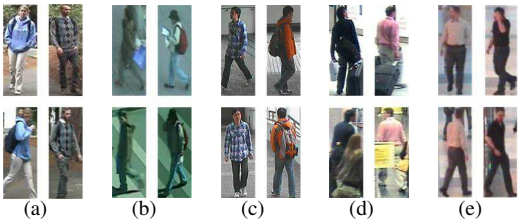
Extension to Semi-Supervised Re-ID If the target task are partially labelled, our model can be readily extended with minimal modification. Specifically, for the labelled data, $w_{T,i,j}$ will be set to 1 if $x_{T,i}$ and $x_{T,j}$ are from same individual, otherwise $w_{T,i,j} = 0$. For the unlabelled data, the corresponding part of W_T is initialised and iteratively updated as in the unsupervised setting.

4. Experiments

4.1. Datasets and Settings

Datasets Five widely used benchmark datasets are chosen in our experiments. **The VIPeR dataset** [12] contains 1,264 images of 632 individuals from two distinct camera views (two images per individual) featured with large viewpoint changes and varying illumination conditions (Fig. 2(a)). **The PRID dataset** [14] consists of images extracted from multiple person trajectories recorded from two surveillance static cameras (Fig. 2(b)). Camera view A contains 385 individuals, camera view B contains 749 individuals, with 200 of them appearing in both views. **The CUHK01 dataset** [21] contains 971 individuals captured from two camera views in a campus environment (Fig. 2(c)). Each person has two images in each camera view. We follow the single-shot setting, that is, we randomly select one image for each individual in each camera view for both training and testing in our experiments. **The iLIDS dataset** [43] has 476 images of 119 individuals captured in an airport terminal from three cameras with different viewpoints (Fig. 2(d)). It contains large occlusions caused by people and luggage. **The CAVIAR dataset** [3] includes 1,220 images of 72 individuals from two cameras in a shopping mall (Fig. 2(e)). Each person has 10 to 20 images. The image sizes of this dataset vary significantly (from 141 $\times$ 72 to 39 $\times$ 17). By examining Fig. 2, it is clear that the obvious variations of visual scenes and cross-view conditions between the five benchmark datasets make the transfer learning task extremely challenging.

Settings A single-shot experiment setting is adopted similar to [25, 36]. In each experiment, one dataset is chosen as the target dataset and the other four are used as the source datasets. All the individuals in the source data are labelled and used for model training, while the individuals in the target dataset are randomly divided into two equal-sized subsets as the training and test sets, with no overlapping on person identities. This process is repeated 10 times, and the averaged performance is reported as the final results. For datasets with only two camera views (VIPeR, PRID and CHUK01), we randomly select one view as probe and the other as gallery. While for the multi-view dataset (iLIDS), one image of each individual in the test set is randomly selected as the gallery image and the rest of the images are used as probe images. Results are obtained by averaging with 10 trials. For the CAVIAR dataset, the same setting as iLIDS is used in the unsupervised setting. However, for fair comparison with published results under the semi-supervised setting, we follow [26] and randomly choose 14 of the 50 individuals appearing in two cameras as the labelled training data, and the remaining 36 individuals as testing data. The 22 people appearing in one camera are used as the unlabelled training data. Also, final results are obtained by averaging with 10 trials. All images are normalized to 128 $\times$ 48 pixels and the colour+HOG+LBP histogram-based 5138-D feature representation in [25] is used. As for the number of dictionary atoms, the size of the task-shared dictionary D^s is the same as the residual dictionary D^r_t $(t = 1, 2..., T)$, which is half of the size of the unique dictionary D^u_T. The size of D^s is set to 150 for all experiments. We found that the model's performance is insensitive to the different dictionary sizes. Other parameters (η_t and λ in Eq. (1)) in our model are set automatically using four-fold cross-validation with one of the four source datasets as the validation set and the other three as training set[2].



(a) (b) (c) (d) (e)

Figure 2: Image samples of the five datasets. Images in the same column are from the same person across two views. Better viewed in colour.

4.2. Unsupervised Re-ID Results

Under this setting, the target dataset is unlabelled. The compared methods can be categorised into two groups:

(1) *Single-task methods*. Without transfer learning, the training data of these unsupervised methods are

[2]The code and features can be downloaded at http://pkuml.com/resources/code.html.

	VIPeR	PRID	CUHK01	CAVIAR	iLIDS
SDALF	19.9	16.3	-	-	29.0
eSDC	26.7	-	-	-	36.8
GTS	25.2	-	-	-	42.4
ISR	27.0	17.0	-	29.0	39.5
Ours_S	24.3	14.1	13.8	33.5	45.7
kLFDA_N	12.9	8.5	7.6	32.8	36.9
SA_DA+kLFDA	11.6	8.1	6.8	32.1	35.8
AdaRSVM	10.9	4.9	5.8	-	-
Ours	**31.5**	**24.2**	**27.1**	**41.6**	**49.3**

Table 1: Results on unsupervised Re-ID. '-' means no implementation code or reported result is available.

	VIPeR	PRID	CUHK01	CAVIAR	iLIDS
Fea_AdaRSVM	9.5	3.7	3.8	-	-
AdaRSVM	10.9	4.9	5.8	-	-
Fea_Ours	23.4	13.5	12.3	32.4	38.7
Ours	**31.5**	**24.2**	**27.1**	**41.6**	**49.3**

Table 2: More detailed comparisons with AdaRSVM.

only the unlabelled data from the target dataset. Some state-of-the-art unsupervised Re-ID methods are selected for comparison, including the hand-crafted-feature-based method SDALF [6], the saliency-learning-based eSDC [38], the graphical-model-based GTS [34] and the sparse-representation-classification-based ISR [24]. We also report results of the single-task version of proposed model by removing all source data related terms in Eq. (1), denoted as Ours_S.

(2) *Multi-task methods*. There are few multi-task learning methods, or unsupervised transfer learning methods in general, available for the unsupervised setting. AdaRSVM [27] is the only unsupervised cross-data transfer learning work that we are aware of, and it is also designed for person Re-ID. As discussed in Sec. 2, the main difference is that they assume the availability of negative pairs in the target dataset, thus using more supervision than our method. We also use the subspace alignment based unsupervised domain adaptation method SA_DA [7] to align the data distributions of the source and target datasets first. Then a supervised Re-ID model, kLFDA [36], is trained on the labelled source datasets and applied to the aligned target dataset. This method is denoted as SA_DA+kFLDA. Note that as an unsupervised domain adaptation method, SA_DA assumes that the source and target domains have the same classes, which is invalid for cross-dataset transfer learning. In addition, we compare with a naive transfer approach, that is, learning kFLDA on source datasets first, and applying it directly to the target dataset without any model adaptation. This is denoted as kLFDA_N.

Table 1 reports the results measured with the Rank 1 matching accuracy (%)[3]. From these results, it is evident that: (1) Compared with existing unsupervised methods including SDALF, eSDC, GTS and ISR, our model is significantly better. This shows that transfer learning indeed helps for unsupervised Re-ID. (2) The difference in performance between "Ours_S" and "Ours" models shows exactly how much the target dataset has benefited from the source datasets using our unsupervised asymmetric multi-task transfer learning. (3) The results of kLFDA_N is very poor, showing that the knowledge learned from the labelled source datasets cannot be directly used to help match target data. This is due to the drastically different viewing conditions and condition changes across views in the target dataset compared to those in the source (see Fig. 2). A naive transfer learning approach such as kLFDA_N would not be able to cope with the domain shift/difference of this magnitude. (4) Importantly it is noted that when an existing unsupervised domain-adaptation based transfer learning model is applied to alleviate the domain shift problem (SA_DA+kLFDA), the result is even worse. This is not surprising as existing unsupervised domain adaptation methods are designed under the assumption that the source and target domains have the same recognition tasks (i.e. having the same set of classes) – an invalid assumption for our unsupervised Re-ID problem as different datasets contain different person identities. (5) The results of the only existing cross-dataset unsupervised Re-ID method AdaRSVM is actually the worst. Note that since the code is not available, these are the reported results in [27]. Since different feature representation and two instead of four source datasets were used, this comparison is only indicative. However, by examining some additional results in Table 2, we can still conclude that AdaRSVM is able to transfer very little useful information from the source datasets even when they use more supervision on the target dataset than our model. More specifically, in Table 2, Fea_AdaRSVM (Fea_Ours) means the matching accuracy by L1-Norm distance of the features used in AdaRSVM (Ours). The results in Table 2 show that our transfer model can improve 8%-15% matching accuracy over non-learning based L1-Norm distance. In contrast, the increase for AdaRSVM is 1%-2%. (6) It is noted that on three of five datasets (PRID,CAVIAR and iLIDS), our unsupervised results is close or surpasses the best reported results using existing *supervised* methods[25, 36, 31]. This shows the clear advantage of our unsupervised transfer learning model over existing models (supervised and unsupervised) on both scalability and accuracy.

4.3. Semi-supervised Re-ID Results

In this experiment, one third of the training set from the target dataset is labelled as in [26]. Again, we compare

[3]The CMC curves of the proposed method can be found at http://pkuml.com/resources/code.html

with two groups of methods:

(1) *Single-task methods*. SSCDL [26] is the most relevant semi-supervised Re-ID method because it is also based on dictionary learning. In addition, with the target data partially labelled, we can now compare with the existing fully-supervised models by training them using the labelled target data only. These include kLFDA [36] and KCCA [25]. The same features are used for fair comparison.

(2) *Multi-task methods*. cAMT [35] is the latest multi-task learning method for person Re-ID to our knowledge. Based on a constrained asymmetric multi-task discriminant component analysis model, it also attempts to transfer knowledge from source tasks to target task. However the key difference is that it needs labelled data in both source datasets and the target dataset; it thus can only be compared under this semi-supervised setting. We also compare with a naive transfer learning method denoted as kLFDA_N, that is, we learn kFLDA using a mix of the labelled source data and the labelled target data. Again, the same features are used for fair comparison.

Dataset	VIPeR	PRID	CUHK01	CAVIAR	iLIDS
SSCDL	25.6	-	-	**49.1**	-
kLFDA	27.5	14.1	25.2	35.7	41.6
KCCA	24.6	5.3	24.8	-	-
kLFDA_N	18.4	12.4	20.6	34.8	38.4
cAMT	16.2	13.5	14.6	29.1	33.6
Ours	**34.1**	**25.3**	**32.1**	47.3	**50.3**

Table 3: Semi-supervised Re-ID results

The results are shown in Table 3, from which we note that: (1) Compared to our results in Table 1, all results improve, albeit by a moderate margin. This means on one hand, our model does benefit from additional labelled data from the target task; on the other hand, they are the ice on the cake as the transferred knowledge from the source task is already very discriminative for the target task. (2) Compared to SSCDL, our result is much better on VIPeR and slightly worse on CAVIAR. Overall, our model is better because as a transfer learning model it can take advantage more labelled data from the source datasets. (3) The results of supervised models (kLFDA and KCCA) are much weaker than ours indicating that they require much more labelled data than the one-third to function properly[4]. (4) The naive transfer model kLFDA_N failed miserably. Again this is due to the untreated domain shift problem. (5) The existing multi-task transfer Re-ID method cAMT fares even worse. This shows that a dictionary learning based multi-task model is more appropriate. This is because being originally designed for unsupervised learning, dictionary learning can exploit the unlabelled target data more naturally than the discriminant component analysis model in [35] which is originally designed for supervised learning.

[4]We note that when trained using fully labelled target set, their results are close to ours under the same setting showing the advantage of being a transfer model diminishes when labels are abundant.

4.4. Further Evaluations

Contributions of Model Components The two key model design components are evaluated: (1) the asymmetric treatment of the target task by including an additional dictionary D_T^u; and (2) the stepwise reconstruction error formulation. For the former, we remove D_T^u in Eq. (1), and for the latter, we remove Terms 1, 3 and 4 in Eq. (1). Table 4 shows clearly that both components contribute positively to the final performance of the model.

Dataset	VIPeR	PRID	CUHK01	CAVIAR	iLIDS
Without D_T^u	27.2	22.3	24.5	38.1	46.5
Without stepwise	23.8	18.9	21.8	35.7	44.2
Our Full Model	**31.5**	**24.2**	**27.1**	**41.6**	**49.3**

Table 4: Evaluation under unsupervised setting on the model components

Running Cost On a desktop PC with two 3.20 GHz CPUs and 4G RAM running in Matlab, our model takes 12 minutes to train and 0.78 seconds to match 312 images against 312 images when VIPeR is used as the target dataset. It is thus extremely efficient during testing as a linear model.

5. Conclusion

We have developed a novel unsupervised cross-dataset transfer learning approach based on *asymmetric* multi-task dictionary learning. It differs significantly from existing methods in that it can exploit labelled datasets collected elsewhere whilst requiring no labelling on a target dataset. Extensive experiments show that our model is superior to existing Re-ID methods with or without transfer learning and has great potentials for real-world applications due to its high scalability, low running cost, and high matching accuracy.

Acknowledgements

This work was partially supported by the National Basic Research Program of China under grant No. 2015CB351806, the National Natural Science Foundation of China under contract No. 61425025, No. 61390515, No. 61471042, and No. 61421062, the National Key Technology Research and Development Program under contract No. 2014BAK10B02, and the Shenzhen Peacock Plan.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 2006.

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

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

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

[5] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *SIGKDD*, 2004.

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

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[8] S. Gao, I. Tsang, L. Chia, and P. Zhao. Local features are not lonely laplacian sparse coding for image classification. In *Proc. CVPR*, 2010.

[9] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[10] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales. The re-identification challenge. *Person Re-Identification*, pages 1–20, 2014.

[11] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006, Nov 2011.

[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3. Citeseer, 2007.

[13] H. Guo, Z. Jiang, and L. S. Davis. Discriminative dictionary learning with pairwise constraints. In *Proc. ACCV*, 2014.

[14] 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.

[15] M. Hirzer, M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *Proc. AVSS*, 2012.

[16] M. Hirzer, M. Roth, M. Koestinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proc. ECCV*, 2012.

[17] K. Kenneth, M.Joseph, R. Bhaskar, E. Kjersti, L. Te-Won, and S. Terrence. Dictionary learning algorithms for sparse representation. *Neural Computing*, 15(2), Feb. 2003.

[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, 2012.

[19] R. Layne, T. Hospedales, and S. Gong. Domain transfer for person re-identification. In *ARTEMIS*, 2013.

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

[21] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[22] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[23] 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.

[24] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[25] G. Lisanti, I. Masi, and A. Del Bilmbo. Matching people across camera views using kernel canonical correlation analysis. In *Proc. ICDSC*, 2014.

[26] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proc. CVPR*, 2014.

[27] A. J. Ma, J. Li, P. C. Yuen, and P. Li. Cross-domain person reidentification using domain adaptation ranking svms. *Image Processing, IEEE Transactions on*, 24(5):1599–1613, 2015.

[28] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *Proc. BMVC*, 2012.

[29] 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.

[30] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013.

[31] S. Paisitkriangkrai, C. Shen, and A. ven den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

[32] Q. Qiu, J. Ni, and R. Chellappa. Dictionary-based domain adaptation methods for the re-identification of faces. In *Person Re-Identification*, pages 269–285. Springer, 2014.

[33] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.

[34] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *Proc. BMVC*, 2014.

[35] X. Wang, W.-S. Zheng, X. Li, and J. Zhang. Cross-scenario transfer person re-identification. 2015.

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

[37] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. L. S. Salient color names for person re-identification. In *Proc. ECCV*, 2014.

[38] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proc. CVPR*, 2013.

[39] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Proc. CVPR*, 2014.

[40] J. Zheng and Z. Jiang. Learning view-invariant sparse representations for cross-view action recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3176–3183. IEEE, 2013.

[41] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa. Cross-view action recognition via a transferable dictionary pair. In *BMVC*, volume 1, page 7, 2012.

[42] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[43] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.

[44] W. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012.