Domain Transfer Support Vector Ranking for Person Re-Identification without Target Camera Label Information

Andy J Ma[1] Pong C Yuen[1,2] Jiawei Li[1]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2]BNU-HKBU United International College, Zhuhai, China
{jhma, pcyuen, jwli}@comp.hkbu.edu.hk

Abstract

This paper addresses a new person re-identification problem without the label information of persons under non-overlapping target cameras. Given the matched (positive) and unmatched (negative) image pairs from source domain cameras, as well as unmatched (negative) image pairs which can be easily generated from target domain cameras, we propose a Domain Transfer Ranked Support Vector Machines (DTRSVM) method for re-identification under target domain cameras. To overcome the problems introduced due to the absence of matched (positive) image pairs in target domain, we relax the discriminative constraint to a necessary condition only relying on the positive mean in target domain. By estimating the target positive mean using source and target domain data, a new discriminative model with high confidence in target positive mean and low confidence in target negative image pairs is developed. Since the necessary condition may not truly preserve the discriminability, multi-task support vector ranking is proposed to incorporate the training data from source domain with label information. Experimental results show that the proposed DTRSVM outperforms existing methods without using label information in target cameras. And the top 30 rank accuracy can be improved by the proposed method upto 9.40% on publicly available person re-identification datasets.

1. Introduction

1.1. Background and Motivation

In recent years, person re-identification across a camera network comprising multiple cameras with non-overlapping views has become an active research topic due to its importance in many camera-network-based computer vision applications. The goal of person re-identification is to re-identify a person when he/she disappears from the field-of-view of a camera and appears in another. Matching individuals over disjoint cameras views can be substantially challenging when variations in illumination condition, background, human pose and scale are significant among those views. Moreover, the temporal transition time between cameras varies greatly for each individual and such an uncertainty makes the person re-identification task even harder.

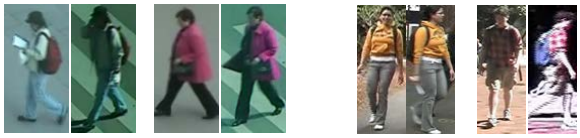
To address this problem, existing schemes focus on developing either robust feature representations [2] [3] [4] [5] [6] [8] [9] [11] [12] [18] [19] [21] [23] or discriminative learning models [1] [15] [17] [26] [30]. For the discriminative learning methods, it is generally assumed that the label information of persons is available for training. With the person labels, matched (positive) and unmatched (negative) image pairs are generated to train the discriminative distance model. While these methods could achieve encouraging re-identification performance, the assumption that label information is available for all the cameras, could only be practically feasible in a small-scale camera network.

Contrarily, in the case of large-scale camera network, collecting the label information of every training subject from every camera in the network can be extremely time-consuming and expensive. Therefore, labels of the training subjects may not be able to be collected from certain cameras. This renders existing approaches inapplicable, since the person labels are not available. Apart from this reason, significant inter-camera variations as exemplified in Fig. 1 would also lead to dramatic performance deterioration, when the distance model learnt from other camera set with label information is directly applied to the cameras missing person labels.

These setbacks pose a need for a new method to handle the afore-described person re-identification issue in the large-scale camera network setting.

1.2. Problem Definition

Motivated by domain transfer learning approach (see [25] for a review), we consider data from the camera set with label information as the source domain; while data from camera set missing label information as the tar-



(a) PRID (b) VIPeR

Figure 1. Comparison between person images from different datasets (better viewed in color): matched image pairs in (a) PRID [16] dataset and (b) VIPeR [14] dataset. There are three major differences between these two image sets: 1) different backgrounds; 2) smaller viewpoint changes in PRID than that in VIPeR; 3) different illumination conditions.

get domain. Here, we denote the source and target domains as s and t, respectively. Due to non-trivial inter-camera variations as indicated in Fig. 1, the source and target joint distributions of the positive or negative tag y and feature vector $\boldsymbol{z}$ for an image pair are supposed to be different, i.e. $\Pr_s(y,\boldsymbol{z}) \neq \Pr_t(y,\boldsymbol{z})$. In order to overcome the mismatch problem of the marginal distributions, a mapping Φ, s.t. $\Pr_s(\Phi(\boldsymbol{z})) \approx \Pr_t(\Phi(\boldsymbol{z}))$, can be learnt via domain adaptation techniques [13] [24]. In general, existing domain adaptation schemes assume that, after projection, such Φ also satisfies the equal conditional probability condition, i.e. $\Pr_s(y|\Phi(\boldsymbol{z})) \approx \Pr_t(y|\Phi(\boldsymbol{z}))$. Since the conditional probability $\Pr_s(y|\Phi(\boldsymbol{z}))$ or $\Pr_t(y|\Phi(\boldsymbol{z}))$ can be interpreted as classification score, the condition that $\Pr_s(y|\Phi(\boldsymbol{z})) \approx \Pr_t(y|\Phi(\boldsymbol{z}))$ implies an equivalence of the distance models in source and target domains. In this case, existing person re-identification algorithms, e.g. [1] [15] [17] [26] [30], can be employed to learn the distance model in the source domain (consisting of projected data with positive and negative image pairs generated by the label information), which can be applied to the target domain without significant performance degradation.

However, it is almost impossible to verify the validity of the assumption that $\Pr_s(y|\Phi(\boldsymbol{z})) \approx \Pr_t(y|\Phi(\boldsymbol{z}))$ in practice. As a result, there is no way to guarantee that the distance model learnt from the projected data in source domain is equivalent to the target one. Thus, we propose to learn the target distance model using data from both source and target domains. In this context, multi-task learning (see [25] for more details) can be employed to learn the distance model for the target cameras by discovering the relationship between the tasks in source and target domains, but the task in the target domain cannot be defined using existing person re-identification methods directly, due to the absence of label information under target cameras. Therefore, the key problem is how to define the learning task in target domain and incorporate the data from source domain for training.

1.3. Contributions

The contributions of this paper are two-folds.

• We develop a new method to train the target discriminative model based on the negative image pairs generated from non-overlapping target cameras. Without positive image pairs generated by the label information of persons, we propose to relax the discriminative constraint into a necessary condition to it, which only relies on the mean of positive pairs. Since source and target domains must be related, we estimate the positive mean in the target domain by assuming that the difference between the positive and negative means in the source domain is close to that in the target domain. With the estimated mean of positive pairs in the target domain, the target learning problem is defined by maximizing the difference between the estimated mean and the negative image pairs, similar to RankSVM [26].

• We propose a novel multi-task support vector ranking method to rank the individuals for person re-identification. Since the learning task in the target domain depends on a necessary condition to the discriminative constraint, the learnt classification model may not be discriminative enough for classifying the detected persons in target cameras. On the other hand, the performance may deteriorate due to the estimation error of the positive mean. Therefore, we propose to incorporate the data in the source domain to learn the classification model for the target domain. Inspired by multi-task SVM [10], we propose to learn the optimal models for the source and target tasks, simultaneously.

1.4. Organization of This Paper

The rest of this paper is organized as follows. We will first give a brief review on existing person re-identification methods. Section 3 will report the proposed Domain Transfer Ranked Support Vector Machines (DTRSVM) method. Experimental results and conclusion are given in Section 4 and Section 5, respectively.

2. Related Work on Person re-identification

In order to ensure that feature representation of the person image is less sensitive to large inter-camera variations, many existing methods focus on extracting robust features. Popular ones include SIFT [2] [19], texture [3] [4] [5] [6] [11], color distribution [21] [23], space-time methods [12] [18] and pictorial structures [8].

Besides feature extraction, discriminative distance learning methods are proposed to further improve the re-identificaiton performance. In [26], person re-identification was formulated as a ranking problem and the RankSVM model is learnt by assigning higher confidence to the positive image pairs and vice versa. Denote $\boldsymbol{x}_j$ as the feature vector for person j, $\boldsymbol{x}_{ji}^+$ for $i = 1, \cdots, n_j^+$ as feature vectors of its matched observations, and $\boldsymbol{x}_{ji}^-$ for $i = 1, \cdots, n_j^-$ as feature vectors of its unmatched observations, where n_j^+ (n_j^-) is the number of the matched (unmatched) observations. And, the absolute difference vector for the positive

(resp. negative) image pair of $\boldsymbol{x}_j$ and $\boldsymbol{x}_{ji}^+$ (resp. $\boldsymbol{x}_{ji}^-$) is calculated by $\boldsymbol{z}_{ji}^+ = \boldsymbol{d}(\boldsymbol{x}_j - \boldsymbol{x}_{ji}^+)$ (resp. $\boldsymbol{z}_{ji}^- = \boldsymbol{d}(\boldsymbol{x}_j - \boldsymbol{x}_{ji}^-)$), where $\boldsymbol{d}$ is an entry-wise function of absolute values, i.e. $\boldsymbol{d}(\boldsymbol{x}) = (|\boldsymbol{x}(1)|, \cdots, |\boldsymbol{x}(R)|)^T$, $\boldsymbol{x}(r)$ is the r-th element of the input vector $\boldsymbol{x}$ and R is the dimension of $\boldsymbol{x}$. The weight vector $\boldsymbol{w}$ in RankSVM is obtained by solving the following optimization problem,

$$\begin{aligned} &\min_{\boldsymbol{w},\xi} \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{j,i_+,i_-} \xi_{ji_+i_-} \\ &\text{s.t. } \boldsymbol{w}^T(\boldsymbol{z}_{ji_+}^+ - \boldsymbol{z}_{ji_-}^-) \geq 1 - \xi_{ji_+i_-}, \\ &\quad \xi_{ji_+i_-} \geq 0, j = 1, \cdots, J, \\ &\quad i_+ = 1, \cdots, n_j^+, i_- = 1, \cdots, n_j^- \end{aligned} \tag{1}$$

where $\xi_{ji_+i_-}$ is the slack variable, C is a positive parameter, and J is the number of detected persons. Similar to RankSVM, Zheng *et al*. [30] proposed a Relative Distance Comparison (RDC) method using a second-order distance learning model. This method is able to exploit higher-order correlations among different features, compared with RankSVM. In order to solve the computational complexity issues in RankSVM and RDC, a Relaxed Pairwise Metric Learning (RPML) method [17] was proposed by relaxing the original hard constraints, which leads to a simpler problem that can be solved much more efficiently.

Besides the supervised distance learning methods [17] [26] [30], Kuo *et al*. [20] proposed an on-line learnt appearance affinity model to decrease the required number of labeled samples under some specific assumptions. On the other hand, an adaptive feature weighting method was proposed in [22] under the observation that the universal model may not be good for all individuals. Different from traditional per-individual identification scheme, Zheng *et al*. [29] addressed a watch list (set) based verification problem and proposed to transfer the information from non-target person data to mine the discriminative information for the target people in the watch list.

3. Domain Transfer Support Vector Ranking

As indicated in [30], the absolute difference space shows some advantages over the common difference space, so we follow [30] to use the absolute difference vector as the feature representation method for both positive and negative image pairs. Similar to the symbol definition as presented in Section 2, denote $\boldsymbol{z}_{j_si}^{s+}$ and $\boldsymbol{z}_{j_si}^{s-}$ as the difference vectors of positive and negative image pairs in the source domain, respectively. For the target domain, the label information of persons is not available, so positive image pairs cannot be generated. *However, negative image pairs can be easily generated, because same person cannot be presented at the same instant under different non-overlapping cameras.* Denote the difference vectors for the negative image pairs as $\boldsymbol{z}_{j_ti}^{t-}$. With $\boldsymbol{z}_{j_si}^{s+}$ and $\boldsymbol{z}_{j_si}^{s-}$ in the source domain and $\boldsymbol{z}_{j_ti}^{t-}$ in the target domain, we first propose a new method to learn the target distance model in Section 3.1. Then, a multi-task support vector ranking method is presented in Section 3.2.

3.1. Leaning without Positive Image Pairs in Target Cameras

Since feature entries give different importance in identifying a person, we follow [26] to use the weighted summation of the absolute difference vector to calculate the confident score for the image pairs in the target domain, i.e. $\boldsymbol{w}_t^T \boldsymbol{z}_{j_ti}^{t-}$, where $\boldsymbol{w}_t^T$ denotes the transposition of the weight vector. If positive image pairs are available, the scores of positive image pairs must be larger than those of the negative ones for a discriminative weight vector $\boldsymbol{w}_t$, i.e.

$$\boldsymbol{w}_t^T \boldsymbol{z}_{j_ti}^{t+} > \boldsymbol{w}_t^T \boldsymbol{z}_{j_t'i'}^{t-}, \forall j_t, i, j_t', i' \tag{2}$$

However, positive image pairs are not available in the target domain, so we cannot obtain the absolute difference vectors $\boldsymbol{z}_{j_ti}^{t+}$ in practice. One way to solve this problem is to determine the weight vector $\boldsymbol{w}_t$ by assigning smaller values to the difference vectors $\boldsymbol{z}_{j_ti}^{t-}$ of negative image pairs using one-class SVM [27]. Nevertheless, it is possible that the scores of positive image pairs also decrease when minimizing those of the negative ones. Thus, it cannot be guaranteed that the learnt weight vector $\boldsymbol{w}_t$ satisfies the discriminative constraint (2). In order to deal with this problem, we propose to learn the weight vector $\boldsymbol{w}_t$ by a necessary condition to constraint (2).

Taking the summation of constraint (2) over the difference vectors $\boldsymbol{z}_{j_ti}^{t+}$ of positive image pairs for all j_t and i, it has

$$\boldsymbol{w}_t^T \boldsymbol{m}_{t+} > \boldsymbol{w}_t^T \boldsymbol{z}_{j_t'i'}^{t-}, \forall j_t', i' \tag{3}$$

where $\boldsymbol{m}_{t+}$ denote the mean of positive image pairs in the target domain. Therefore, a necessary condition to constraint (2) is given by equation (3) such that the score of the positive mean is larger than those of the negative image pairs.

Denote the mean calculated by the difference vectors of all the image pairs in the target domain as $\boldsymbol{m}_t$ and the mean of negative image pairs estimated by the available data $\boldsymbol{z}_{j_ti}^{t-}$ of unmatched pairs as $\widetilde{\boldsymbol{m}}_{t-}$. Then, the mean of positive image pairs can be estimated by the following equation,

$$\widetilde{\boldsymbol{m}}_{t+} = (N_t \boldsymbol{m}_t - N_{t-} \widetilde{\boldsymbol{m}}_{t-})/N_{t+} \tag{4}$$

where N_t, N_{t+} and N_{t-} denote the number of the overall, positive and negative image pairs, respectively. However, N_{t+} and N_{t-} are difficult to be computed, if target positive samples are not available. On the other hand, the estimation error for the positive mean with equation (4) can be

very large, since the number of negative pairs is much larger than that of positive pairs, i.e. $N_{t-} \gg N_{t+}$. Denote the genuine means of the positive and negative image pairs as $\boldsymbol{m}_{t+}$ and $\boldsymbol{m}_{t-}$, respectively. The estimation error for the positive mean with equation (4) is given by the following equation,

$$\|\boldsymbol{m}_{t+} - \widetilde{\boldsymbol{m}}_{t+}\| = \frac{N_{t-}}{N_{t+}}\|\boldsymbol{m}_{t-} - \widetilde{\boldsymbol{m}}_{t-}\| \tag{5}$$

Since $N_{t-} \gg N_{t+}$, the division value of N_{t-}/N_{t+} is very large. Therefore, according to equation (5), the estimation error for the positive mean is very large, even though the error for the negative mean is small.

In order to solve this problem, we propose to incorporate the data with label information of persons in the source domain. With the label information, the true means of positive and negative image pairs in the source domain can be calculated and denoted as $\boldsymbol{m}_{s+}$ and $\boldsymbol{m}_{s-}$, respectively. Since the source and target domains are related, the positive and negative distributions in the source domain must be related to those in the target domain. We suppose the relationship can be modeled in a way that the difference between the positive and negative means in the source domain is close to that in the target domain, i.e.

$$\boldsymbol{m}_{t+} - \boldsymbol{m}_{t-} \approx \boldsymbol{m}_{s+} - \boldsymbol{m}_{s-} \tag{6}$$

With equation (6), the positive mean in the target domain can be estimated by the following equation,

$$\widetilde{\boldsymbol{m}}_{t+} = \widetilde{\boldsymbol{m}}_{t-} + \boldsymbol{m}_{s+} - \boldsymbol{m}_{s-} \tag{7}$$

And the upper bound of the estimation error is given by

$$\begin{aligned}\|\boldsymbol{m}_{t+} - \widetilde{\boldsymbol{m}}_{t+}\| \leq \|\boldsymbol{m}_{t-} - \widetilde{\boldsymbol{m}}_{t-}\| \\ + \|(\boldsymbol{m}_{t+} - \boldsymbol{m}_{t-}) - (\boldsymbol{m}_{s+} - \boldsymbol{m}_{s-})\|\end{aligned} \tag{8}$$

Since lots of negative image pairs can be obtained from the non-overlapping target cameras, the estimated mean of negative pairs is close to the true one. Under the assumption given by equation (6), the upper bound of the estimation error for the positive mean in the target domain is small. From the results in the experiment section, we can see that the estimation error of the positive mean using equation (7) is much smaller than that using equation (4).

With the estimated positive mean $\widetilde{\boldsymbol{m}}_{t+}$ in the target domain and the necessary condition given by equation (3) to the discriminative constraint, we define the learning task in the target domain similar to the optimization problem (1) in RankSVM [26] as the following equation,

$$\begin{aligned}&\min_{\boldsymbol{w}_t} \frac{1}{2}\|\boldsymbol{w}_t\|^2 + C\sum_{j_t,i} \xi_{j_t i} \\ &\text{s.t. } \boldsymbol{w}_t^T(\widetilde{\boldsymbol{m}}_{t+} - \boldsymbol{z}_{j_t i}^{t-}) \geq 1 - \xi_{j_t i}, \\ &\quad \xi_{j_t i} \geq 0, j_t = 1, \cdots, J_t, i = 1, \cdots, n_{j_t}^-\end{aligned} \tag{9}$$

where $n_{j_t}^-$ denotes the number of negative image pairs for person j_t and J_t is the number of detected persons in the target domain.

3.2. Multi-Task Support Vector Ranking

Since equation (3) is not a sufficient condition to the discriminative constraint (2), the weight vector $\boldsymbol{w}_t$ learnt by the derived optimization problem (9) may not be discriminative enough for classifying the detected persons in target cameras. On the other hand, the performance may deteriorate by directly using the weight vector learnt from optimization problem (9), since the assumption on the relation between the source and target domains introduces the estimation error to some extent. Therefore, we employ the concept of multi-task learning to incorporate the source domain data with label information of persons for the determination of the weight vector $\boldsymbol{w}_t$. Following multi-task SVM [10], the weight vectors $\boldsymbol{w}_s$ for the source domain and $\boldsymbol{w}_t$ for the target domain are related in a way that they are closed to a common model $\boldsymbol{w}_0$ and can be written as

$$\boldsymbol{w}_s = \boldsymbol{w}_0 + \boldsymbol{v}_s, \boldsymbol{w}_t = \boldsymbol{w}_0 + \boldsymbol{v}_t \tag{10}$$

where $\boldsymbol{v}_s$ and $\boldsymbol{v}_t$ are vectors measuring the differences between the specific models and the common one. The learning task in the source domain is defined by employing the learning problem (1) in RankSVM [26]. Combining the source task similar to (1) and target task (9), all the vectors $\boldsymbol{w}_0$, $\boldsymbol{v}_s$ and $\boldsymbol{v}_t$ can be estimated simultaneously by solving the following optimization problem,

$$\begin{aligned}&\min_{\boldsymbol{w}_0,\boldsymbol{v}_s,\boldsymbol{v}_t} \frac{1}{2}\|\boldsymbol{w}_0\|^2 + \frac{\mu}{2}(\|\boldsymbol{v}_s\|^2 + \|\boldsymbol{v}_t\|^2) \\ &\qquad + C(\sum_{j_s,i_{s+},i_{s-}} \xi_{j_t i_{s+} i_{s-}} + \sum_{j_t,i_t} \xi_{j_t i_t}) \\ &\text{s.t. } (\boldsymbol{w}_0 + \boldsymbol{v}_s)^T(\boldsymbol{z}_{j_s i_{s+}}^{s+} - \boldsymbol{z}_{j_s i_{s-}}^{s-}) \geq 1 - \xi_{j_s i_{s+} i_{s-}}, \\ &\quad \xi_{j_s i_{s+} i_{s-}} \geq 0, j_s = 1, \cdots, J_s, \\ &\quad i_{s+} = 1, \cdots, n_{j_s}^+, i_{s-} = 1, \cdots, n_{j_s}^-, \\ &\quad (\boldsymbol{w}_0 + \boldsymbol{v}_t)^T(\widetilde{\boldsymbol{m}}_{t+} - \boldsymbol{z}_{j_t i_t}^{t-}) \geq 1 - \xi_{j_t i_t}, \\ &\quad \xi_{j_t i_t} \geq 0, j_t = 1, \cdots, J_t, i_t = 1, \cdots, n_{j_t}^-\end{aligned} \tag{11}$$

where $n_{j_s}^+$ ($n_{j_s}^-$) denotes the number of positive (negative) image pairs for person j_s and J_s is the number of detected persons in the source domain. In optimization problem (11), the positive regularization parameter μ controls the difference between the weight vector $\boldsymbol{w}_s$ ($\boldsymbol{w}_t$) in the source (target) domain and the common model $\boldsymbol{w}_0$. Intuitively, for a fixed value of C, if $\mu \to \infty$, $\boldsymbol{v}_s$ and $\boldsymbol{v}_t$ tend to zero, which means that the source and target models become the same. On the other hand, if $\mu \to 0$, $\boldsymbol{v}_s$ and $\boldsymbol{v}_t$ can be very large compared with $\boldsymbol{w}_0$. In this case, the common model $\boldsymbol{w}_0$

takes little effect on the source and target models, which implies that the source task and target task are unrelated.

Inspired by multi-task SVM [10], we solve the optimization problem (11) by converting it to the standard RankSVM formulation as in equation (1). Denote the column concatenation of $\boldsymbol{w}_0$, $\sqrt{\mu}\boldsymbol{v}_s$ and $\sqrt{\mu}\boldsymbol{v}_t$ as $\boldsymbol{w}$, i.e.

$$\boldsymbol{w} = (\boldsymbol{w}_0; \sqrt{\mu}\boldsymbol{v}_s; \sqrt{\mu}\boldsymbol{v}_t) \tag{12}$$

On the other hand, we construct the feature map as

$$\begin{aligned} \boldsymbol{a}_{l_s} &= (\boldsymbol{z}^{s+}_{j_s i_{s+}} - \boldsymbol{z}^{s-}_{j_s i_{s-}}; \frac{\boldsymbol{z}^{s+}_{j_s i_{s+}} - \boldsymbol{z}^{s-}_{j_s i_{s-}}}{\sqrt{\mu}}; \boldsymbol{0}) \\ \boldsymbol{b}_{l_t} &= (\widetilde{\boldsymbol{m}}_{t+} - \boldsymbol{z}^{t-}_{j_t i_t}; \boldsymbol{0}; \frac{\widetilde{\boldsymbol{m}}_{t+} - \boldsymbol{z}^{t-}_{j_t i_t}}{\sqrt{\mu}}) \end{aligned} \tag{13}$$

where l_s and l_t denote the indexes of the feature vectors. With equations (12) and (13), it has the following equations,

$$\begin{aligned} \|\boldsymbol{w}\|^2 &= \|\boldsymbol{w}_0\|^2 + \mu(\|\boldsymbol{v}_s\|^2 + \|\boldsymbol{v}_t\|^2) \\ \boldsymbol{w}^T\boldsymbol{a}_{l_s} &= (\boldsymbol{w}_0 + \boldsymbol{v}_s)^T(\boldsymbol{z}^{s+}_{j_s i_{s+}} - \boldsymbol{z}^{s-}_{j_s i_{s-}}) \\ \boldsymbol{w}^T\boldsymbol{b}_{l_t} &= (\boldsymbol{w}_0 + \boldsymbol{v}_t)^T(\widetilde{\boldsymbol{m}}_{t+} - \boldsymbol{z}^{t-}_{j_t i_t}) \end{aligned} \tag{14}$$

Therefore, the optimization problem (11) is rewritten as

$$\begin{aligned} &\min_{\boldsymbol{w}_0, \boldsymbol{v}_s, \boldsymbol{v}_t} \frac{1}{2}\|\boldsymbol{w}\|^2 + C(\sum_{l_s}\xi_{l_s} + \sum_{l_t}\xi_{l_t}) \\ &\text{s.t. } \boldsymbol{w}^T\boldsymbol{a}_{l_s} \geq 1 - \xi_{l_s}, \xi_{l_s} \geq 0, l_s = 1, \cdots, L_s, \\ &\quad \boldsymbol{w}^T\boldsymbol{b}_{l_t} \geq 1 - \xi_{l_t}, \xi_{l_t} \geq 0, l_t = 1, \cdots, L_t \end{aligned} \tag{15}$$

where L_s and L_t represent the number of inequality constraints in source and target domains, respectively.

In order to solve the optimization problem (15) more efficiently, we reformulate (15) by the square hinge loss as

$$\begin{aligned} \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C(&\sum_{l_s}[\max(0, 1 - \boldsymbol{w}^T\boldsymbol{a}_{l_s})]^2 \\ &+ \sum_{l_t}[\max(0, 1 - \boldsymbol{w}^T\boldsymbol{b}_{l_t})]^2) \end{aligned} \tag{16}$$

Employing the primal Newton method based efficient algorithm [7] to solve the optimization problem (16), the optimal combined weight vector $\boldsymbol{w}$ can be obtained. Then, the target weight vector $\boldsymbol{w}_t$ is calculated by equations (10) and (12). At last, the algorithmic procedure for training the proposed Domain Transfer Ranked Support Vector Machines (DTRSVM) model is presented in Algorithm 1.

Algorithm 1 Training DTRSVM

Input: Difference vectors $\boldsymbol{z}^{s+}_{j_s i}$ and $\boldsymbol{z}^{s-}_{j_s i}$ in source domain, $\boldsymbol{z}^{t-}_{j_t i}$ in target domain, parameters C and μ;

1: Compute the means $\boldsymbol{m}_{s+}$ by $\boldsymbol{z}^{s+}_{j_s i}$, $\boldsymbol{m}_{s-}$ by $\boldsymbol{z}^{s-}_{j_s i}$, and $\widetilde{\boldsymbol{m}}_{t-}$ by $\boldsymbol{z}^{t-}_{j_t i}$;

2: Estimate the relevant mean in the target domain $\widetilde{\boldsymbol{m}}_{t+}$ by equation (7);

3: Construct the feature map by equation (13);

4: Solve the optimization problem (16) by the efficient method [7] and obtain the weight vector $\boldsymbol{w}$;

5: Calculate the target domain weight vector $\boldsymbol{w}_t$ by equations (10) and (12);

Output: Weight vector $\boldsymbol{w}_t$.

4. Experiments

In this section, we first give an introduction to the datasets and settings used for evaluation of the proposed method. Then, the comparison results with state-of-the-art methods are reported in Section 4.2.

4.1. Datasets and Settings

Three datasets, namely VIPeR [14], PRID [16] and i-LIDS [28], are used for evaluation of the proposed method. VIPeR is a re-identification dataset containing 632 person image pairs captured by two cameras outdoor. PRID dataset consists of person images from two static surveillance cameras. Total 385 persons were captured by camera A, while 749 persons captured by camera B. The first 200 persons appeared in both cameras, and the remainders only appear in one camera. In our experiments, the single-shot version is used, in which at most one image of each person from each camera is available. The i-LIDS Multiple-Camera Tracking (MCT) dataset contains a number of video clips captured by five cameras indoor. In re-identification application, total 476 person images from 119 persons are used for experiments as in [30]. We follow the procedures as reported in [15] [26] [30] to extract feature vectors for the detected person images in these datasets.

In our experiments, we use VIPeR or PRID as the target domain. Without the time acquisition information in the PRID and VIPeR datasets, negative image pairs from non-overlapping cameras are generated by simulating the synchronization using label information. Since the i-LIDS dataset does not provide the camera information, negative image pairs from non-overlapping cameras cannot be generated to simulate the real situation. Thus, we do not use the i-LIDS dataset as the target domain. Fixing the target domain dataset as VIPeR or PRID, one of the other two datasets is used as the source domain to train the proposed DTRSVM. Therefore, experiments of four transfer scenarios with different source or target domain are performed. If VIPeR is used as the the target dataset, 632 image pairs are randomly separated into half for training and the other half for testing. When PRID is used as the the target dataset, 100 out of the 200 image pairs are randomly selected as the training set, and the others for testing set. For the testing data in VIPeR or PRID, the evaluation is performed by

	Estimated by (4)	Estimated by (7)
i-LIDS to PRID	71.11	2.95
VIPeR to PRID	71.11	2.22
i-LIDS to VIPeR	1120.24	2.46
PRID to VIPeR	1120.24	2.18

Table 1. Estimation errors of the positive mean introduced by equations (4) and (7)

searching the 316 or 100 persons of one camera view from another view. Each experiment was repeated ten times and the mean accuracy is reported.

Three state-of-the-art distance learning methods for person re-identification, namely Rank Support Vector Machines (RankSVM) [26], Relative Distance Comparison (RDC) [30], and Relaxed Pairwise Metric Learning (RPML) [17], as well as two commonly used non-learning based metrics namely L_1 and L_2 norms are employed for the comparison with the proposed method. Following [30], one positive and one negative image pair for each person in the source dataset are used for training, while the training data in the target domain contains only one negative image pair for each person. Since the label information of persons is supposed to be not available in the target dataset, cross-validation cannot be performed to select the best parameters. Thus, we empirically set the parameters in existing methods and the proposed DTRSVM. The PCA dimension in RPML is set as 80. The parameter C in RankSVM and the proposed method is set as 1, while μ in the proposed DTRSVM is set as 1 as well.

4.2. Results

Estimation errors of the positive mean: We use L_1 norm to calculate the difference between the estimated mean and the true one. The estimation errors using equations (4) and (7) are recorded in Table 1. From Table 1, we can see that the estimation error introduced by equation (4) does not change with different source domains, since the estimated mean using equation (4) is only based on the information in the target domain. On the other hand, equation (7) gives much smaller estimation errors. This convinces that the assumption given by equation (6) is reasonable for person re-identification.

Comparing with state-of-the-art re-identification methods: The CMC curves of the learning based and non-learning based methods are shown in Fig. 2. We also plot the CMC curves of the learning based methods training with the label information in the target domain as the baseline of the upper bound performance. From Fig. 2, we can see that all the learning based methods have a dramatic deterioration of performance, when the classification model is trained with the data in the source domain. On the other hand, as shown in Fig. 2(a), when the source domain is i-LIDS and the target domain is PRID, the proposed DTRSVM achieves convincing performance closed to that of the upper bound using the label information in the target domain for training. For the transfer scenario from VIPeR to PRID, DTRSVM also clearly outperforms the other methods without using the label information in the target domain as indicated in Fig. 2(b). Although the performance of DTRSVM for VIPeR as target domain is close to RankSVM, RDC, and the non-learning based methods when using PRID as the source domain, DTRSVM achieves obvious improvement when using i-LIDS as the source domain. This implies that the selection of source data is one of the factors with influence on the performance of the proposed DTRSVM.

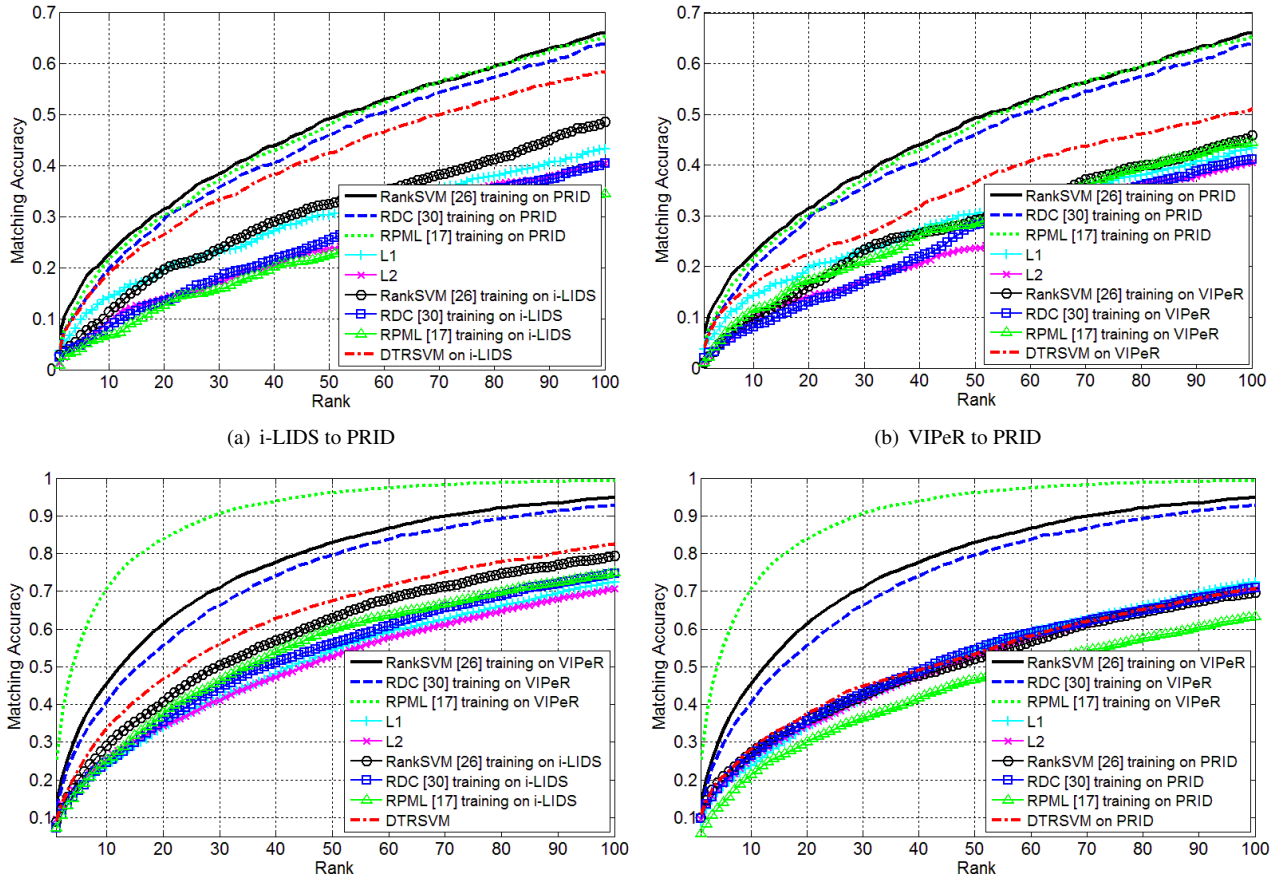
Method	Source	r=1	r=10	r=20	r=30
DTRSVM	i-LIDS	3.95	**18.85**	**26.60**	**33.20**
	VIPeR	**4.60**	17.25	22.90	28.10
RankSVM [26]	i-LIDS	2.95	11.40	19.65	23.80
	VIPeR	1.05	9.70	16.20	23.35
RDC [30]	i-LIDS	2.35	8.35	13.40	18.00
	VIPeR	1.95	8.05	12.90	17.05
RPML [17]	i-LIDS	0.90	6.80	12.65	15.85
	VIPeR	1.10	11.85	17.40	21.00
L1	—	3.65	14.25	17.90	23.15
L2	—	1.35	9.55	14.00	17.25

Table 2. Top r ranked matching accuracy (%) on PRID dataset

Method	Source	r=1	r=10	r=20	r=30
DTRSVM	i-LIDS	8.26	**31.39**	**44.83**	**53.88**
	PRID	**10.90**	28.20	37.69	44.87
RankSVM [26]	i-LIDS	8.94	29.11	40.60	50.44
	PRID	10.00	27.37	35.71	42.33
RDC [30]	i-LIDS	7.23	24.53	35.41	44.22
	PRID	9.70	26.46	35.63	42.85
RPML [17]	i-LIDS	7.14	25.17	37.71	46.46
	PRID	5.65	21.34	30.09	36.33
L1	—	8.86	23.80	33.84	41.55
L2	—	9.53	25.60	34.38	41.17

Table 3. Top r ranked matching accuracy (%) on VIPeR dataset

In order to further compare the performance without label information of persons in the target domain, we summarize the top r ranked matching accuracies (%) of different methods and different source domains for PRID in Table 2 and VIPeR in Table 3. From these two tables, we can see that the best rank one accuracies in these two datasets are obtained by the proposed method using VIPeR for PRID and PRID for VIPeR as source domain, while the DTRSVM from i-LIDS improves the top 30 ranked accuracy by 9.40% for PRID and 3.44% for VIPeR dataset. This convinces that the unmatched image pairs generated from non-overlapping cameras help to improve the re-identification performance, when the label information is not available.



(a) i-LIDS to PRID

(b) VIPeR to PRID

(c) i-LIDS to VIPeR

(d) PRID to VIPeR

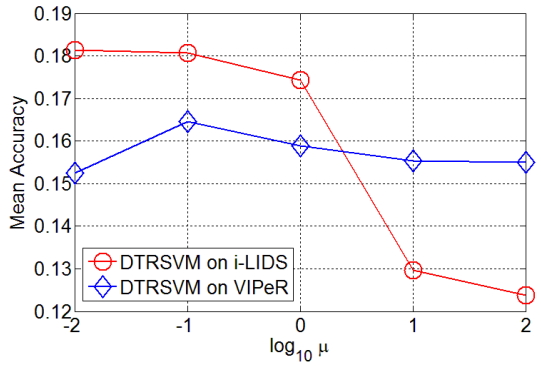
Figure 2. CMC curves of all the four source and target domain combinations (better viewed in color)

Performance influence of parameter μ: As mentioned in Section 3.2, the regularization parameter μ measures the degree of relevance of the source and target domains. The mean values from rank one to top 30 ranked accuracies with different μ of $10^{-2}, \cdots, 10^{2}$ are presented in Fig. 3(a) for PRID and Fig. 3(b) for VIPeR dataset. As shown in these two figures, the best performance is achieved by different values of μ under different transfer learning scenarios. This implies that the degree of relevance differs with different source or target domain. Therefore, if the degree of relevance between source and target domains can be discovered, the parameter μ in the proposed method can be determined more accurately to further improve the proposed re-identification performance.

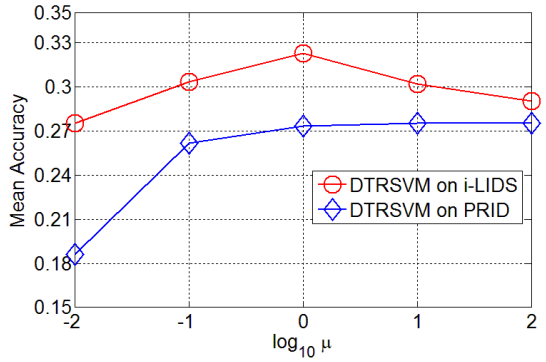
5. Conclusions

In this paper, we propose a novel Domain Transfer Ranked Support Vector Machines (DTRSVM) method to deal with the problem that label information of persons is not available under target cameras. Without positive image pairs generated by the label information, the learning task in the target domain is defined by a necessary condition to the discriminative constraint, which only relies on the mean of positive pairs. In order to estimate the positive mean in the target domain, we assume that the difference between the positive and negative means in source domain is close to the one in the target domain. After defining the learning problem in the target domain, a multi-task support vector ranking method is developed to incorporate the data in source domain with label information to train the classification model for the target domain.

Experimental results show that the estimation error of the positive mean is small, which indicates that the proposed assumption is suitable in person re-identification applications. Comparing state-of-the-art discriminative learning methods using the source and target domain data for training, respectively, it is shown that the performance deteriorates dramatically when using the learnt model trained on source domain to target domain. With the help of the negative image pairs generated from non-overlapping target cameras, the proposed DTRSVM outperforms existing methods without using the label information in the target domain for training. Since the experiments also indicate that different source do-



(a) Results on PRID



(b) Results on VIPeR

Figure 3. Mean accuracy with different values of parameter μ in the DTRSVM

mains have an effect on the performance of the proposed DTRSVM, we will further investigate how to select or combine different source domains to train the DTRSVM in the future.

Acknowledgments

This project was partially supported by the Science Faculty Research Grant of Hong Kong Baptist University, Hong Kong Research Grants Council General Research Fund 212313, National Science Foundation of China Research Grants 61128009 and 61172136. The authors would like to thank the reviewers for their helpful comments improving the quality of this paper and thank Dr. M.-H. Lim for his help in improving the writing of this paper.

References

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *ECCV*, 2012.

[2] M. Bauml and R. Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *AVSS*, 2011.

[3] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898 – 903, 2012.

[4] S. Bąk, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, 2012.

[5] S. Bąk, E. Corvée, F. Brémond, and M. Thonnat. Boosted human re-identification using riemannian manifolds. *Image Vision Comput.*, 30(6-7):443–452, 2010.

[6] S. Bąk, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using Haar-based and DCD-based signature. In *AVSS*, 2010.

[7] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010.

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

[9] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *J. Ambient Intell. Humanized Comput.*, 2(2):127–151, 2011.

[10] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *ACM SIGKDD*, 2004.

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

[12] N. Gheissari, T. B. Sebastian, P. H. Tu, and J. Rittscher. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[13] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[14] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007.

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

[16] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

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

[18] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *ICCV*, 2003.

[19] K. Jüngling and M. Arens. View-invariant person re-identification with an implicit shape model. In *AVSS*, 2011.

[20] C.-H. Kuo, C. Huang, and R. Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *ECCV*, 2010.

[21] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person re-identification. *TPAMI*, 35(7):1622–1634, 2013.

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

[23] C. Madden, E. D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Mach. Vision Appl.*, 18(3):233–247, 2007.

[24] S. J. Pan, J. T. K. Ivor W. Tsang, and Q. Yang. Domain adaptation via transfer component analysis. *TNN*, 22(2):199–210, 2011.

[25] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010.

[26] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[27] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

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

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

[30] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *TPAMI*, 35(3):653–668, 2013.