Semi-Supervised Ranking for Re-Identification with Few Labeled Image Pairs

Andy J Ma and Ping Li

Department of Statistics, Department of Computer Science,
Rutgers University, Piscataway, NJ 08854, USA

Abstract. In many person re-identification applications, typically only a small number of labeled image pairs are available for training. To address this serious practical issue, we propose a novel semi-supervised ranking method which makes use of unlabeled data to improve the re-identification performance. It is shown that low density separation or graph propagation assumption is not valid under some conditions in person re-identification. Thus, we propose to iteratively select the most confident matched (positive) image pairs from the unlabeled data. Since the number of positive matches is greatly smaller than that of negative ones, we increase the positive prior by selecting positive data from the top-ranked matching subset among all unlabeled data. The optimal model is learnt by solving a regression based ranking problem. Experimental results show that our method significantly outperforms state-of-the-art distance learning algorithms on three publicly available datasets using only few labeled matched image pairs for training.

1 Introduction

Person re-identification under non-overlapping camera views has become an active research topic due to its important applications in video surveillance systems, such as criminal detection, human tracking and behavior understanding across camera views. This problem can be extremely challenging because variations of illumination condition, background, human pose, scale, etc., are usually significant among disjoint camera views. Many research works [1]-[12] have been developed to extract robust features invariant to deal with these variations. To take advantage of label information of persons, discriminative learning methods were employed in [13]-[20]. With person labels for training, matched (positive) and unmatched (negative) image pairs are generated to learn the discriminative models for the query image. Although the re-identification performance is improved by supervised learning, these methods require a large number of positive image pairs for training.

In large-scale camera networks containing (e.g.) hundreds of thousands of cameras, it is extremely time-consuming and expensive to collect the label information of numerous training subjects from every camera. In this context, a domain transfer support vector ranking method was proposed in [21] by adapting the classifier learnt from the source domain with plenty of label information to the target domain without any labels. To align the distribution mismatch between the source and target domains, this domain transfer learning method assumes that the target positive (matched image pairs) distribution can be represented by the target positive mean. While this assumption can simplify the problem, it may degrade the performance when the assumption is not valid.

In this paper, we address the problem that only a small number of persons are labeled to generate few positive image pairs for training. Under this scenario, we develop a novel semi-supervised ranking algorithm which make use of the unlabeled data to boost the re-identification performance. By analyzing the data distribution of absolute difference vectors, we show that the widely used low density separation and graph propagation assumptions in many semi-supervised algorithms [22] [23] are not valid under some conditions in person re-identification. Therefore, we follow the self training direction to iteratively label the most confident positive image pairs from the unlabeled data. Since the number of positive matches is much smaller than that of negative ones, it is difficult to correctly select the true positive image pairs with a small amount of positive data. Therefore, we take advantages of properties in person re-identification and increase the positive prior by selecting potential positive data from the rank-one matching subset in all the unlabeled data. The optimal classification model is learnt by solving a regression based ranking problem with the selected positive data. The contributions of this paper are two-fold.

- We propose a new method to select positive image pairs for semi-supervised learning in person re-identification under data imbalance problem. It is shown that the positive prior in the rank-one matching subset is much larger than that in all the unlabeled data due to properties in re-identification. Thus, we propose to select the most confident positive matches from the rank-one matching subset for higher positive prior, which gives higher precision in selecting positive image pairs. On the other hand, we define a more robust confidence measure using negative data generated under non-overlapping cameras to select the potential positive data more accurately.

- We develop a novel semi-supervised ranking algorithm for person re-identification using only a small number of positive image pairs for training. Based on the potential positive image pairs selected from the unlabeled data, we formulate the ranking problem by least-square regression and propose an efficient updating method to determine the optimal solution. Since the proposed method updates the classification model iteratively, the classification model becomes more discriminative with iteration to better select the potential positive data.

The rest of the paper is organized as follows. Section 2 provides a brief review on person re-identification and semi-supervised learning. Section 3 reports the proposed Semi-Supervised Ranking method with Increased Positive Prior (SSR-IPP). Experimental results are given in Section 4. Finally, Section 5 concludes the paper.

2 Related Works

2.1 Supervised and Semi-Supervised Person Re-Identification

To take advantages of person labels, many existing supervised re-identification algorithms [14, 16–18, 20] convert the multi-class person identification problem into a two-class matching problem by training a unified classification model for different individuals. In [14], the Ranked Support Vector Machines (RSVM) model was employed to assign higher confidence to the positive image pairs and vice versa. To exploit higher-order correlations among features, Zheng *et al.* [17] proposed a Relative Distance Comparison (RDC) method using second-order distance learning. For solving the computational complexity issue, a Relaxed Pairwise Metric Learning (RPML) method was proposed in [16] by relaxing the original hard constraints, which leads to a simpler problem that can be solved more efficiently. On the other hand, more recently, there have been some research works on semi-supervised learning for person identification or re-identification [24–26]. While these methods employed the concept of semi-supervised learning, they did not address the problem that only a small number of matched image pairs are available to train a discriminative re-identification model.

2.2 Semi-Supervised Learning

Semi-supervised learning attempts to train a better classification model by incorporating a small amount of labeled data with a large amount of unlabeled data. Many semi-supervised learning algorithms were developed based on low density separation or graph propagation assumption [22, 23]. Under low density separation assumption, it is believed that the classification boundary lies in the low density region within which there are few data points. For the graph propagation approach, a regularization term is added to the objective function for the smoothness of the classification model. Besides classification, semi-supervised learning has been employed for ranking in information retrieval. Semi-supervised ranking methods were proposed in [27] based on low density separation assumption and [28] based on graph propagation approach. However, they do not take full advantages of the available information and it is shown by our analysis that these assumptions are not valid under some conditions in person re-identification.

3 Proposed Method

For clear presentation, we consider the re-identification task for images from a pair of cameras a and b. For multiple cameras, multiple classification models can be trained for each camera pair. As indicated in [17], the absolute difference space shows some advantages over the common difference space, so we use the Absolute Difference Vector (ADV) as the feature representation method for both positive and negative image pairs. Given two feature vectors $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_j^b$ representing two images under cameras a and b, respectively, the ADV $\boldsymbol{z}_{ij}$ is defined by

$$\boldsymbol{z}_{ij} = \boldsymbol{d}(\boldsymbol{x}_i^a - \boldsymbol{x}_j^b) = (|\boldsymbol{x}_i^a(1) - \boldsymbol{x}_j^b(1)|, \cdots, |\boldsymbol{x}_i^a(R) - \boldsymbol{x}_j^b(R)|)^T \tag{1}$$

where $\boldsymbol{x}(r)$ is the r-th element of feature vector $\boldsymbol{x}$ and R is the dimension of $\boldsymbol{x}$.

Given a small number of labeled person images under both cameras a and b for training, positive image pairs can be constructed for $y_i^a = y_j^b$, where y_i^a and y_j^b are person labels of feature vectors $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_j^b$, respectively. Denote positive ADVs as $\boldsymbol{z}_{ij}^+$. Similarly, negative ADVs $\boldsymbol{z}_{ik}^-$ can be obtained for $y_i^a \neq y_k^b$. On the other hand, we are given a large number of unlabeled person images under both cameras and their ADVs are denoted by $\boldsymbol{z}_{mn}^u$. Since the same person cannot be presented at the same instant under different non-overlapping cameras a and b, negative image pairs can be obtained for each unlabeled feature vector $\boldsymbol{x}_m^a$ or $\boldsymbol{x}_n^b$. This means we can easily get some negative ADVs from the unlabeled images and denote them as $\boldsymbol{z}_{mk}^-$ and $\boldsymbol{z}_{ln}^-$. Therefore, the key problem is to determine the potential positive image pairs from the unlabeled ones.

3.1 Data Distribution Analysis in Person Re-Identification

Let us consider the distance between a positive ADV $\boldsymbol{z}_{ij}^+$ and an unlabeled one $\boldsymbol{z}_{mn}^u$. According to the definition given by (1), we have

$$\|\boldsymbol{z}_{ij}^+ - \boldsymbol{z}_{mn}^u\|_p = \left(\sum_{r=1}^{R} \left| |\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r)| - |\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)| \right|^p \right)^{\frac{1}{p}} \tag{2}$$

where $\|\cdot\|_p$ denotes l_p norm. To show that the low density assumption may not be valid, we consider the unlabeled ADV $\boldsymbol{z}_{mn}^u$ for $y_m^a \neq y_n^b$. In this case, $\boldsymbol{z}_{mn}^u$ is negative. And, the difference between the r-th elements of feature vectors $\boldsymbol{x}_m^a$ and $\boldsymbol{x}_n^b$ could be large, i.e., $|\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)|$ is a large number. If the difference between $\boldsymbol{x}_i^a(r)$ and $\boldsymbol{x}_j^b(r)$ is small for positive image pair, we have the conclusion that the distance between $\boldsymbol{z}_{ij}^+$ and $\boldsymbol{z}_{mn}^u$ is large by (2) for $y_m^a \neq y_n^b$. However, it cannot be guaranteed that $|\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r)|$ for $y_i^a = y_j^b$ is significantly smaller than $|\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)|$ for $y_m^a \neq y_n^b$, since feature vectors $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_j^b$ are extracted from images under non-overlapping camera views. Thus, $|\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r)|$ could be large. Due to the large amount of negative image pairs, it is likely that there exists $\boldsymbol{z}_{mn}^u$ for $y_m^a \neq y_n^b$ such that the distance between $\boldsymbol{z}_{ij}^+$ and $\boldsymbol{z}_{mn}^u$ is small, i.e.,

$$\exists \boldsymbol{z}_{mn}^u, \text{s.t. } \|\boldsymbol{z}_{ij}^+ - \boldsymbol{z}_{mn}^u\|_p \leq \varepsilon, y_m^a \neq y_n^b \tag{3}$$

where ε is a small positive number. This equation means that for each positive ADV $\boldsymbol{z}_{ij}^+$ there are probably some negative ones around them. Therefore, the low density region separating the positive and negative data does not exist. This means the low density separation assumption in many semi-supervised learning methods is not valid under this condition in person re-identification.

On the other hand, for the positive ADVs from the unlabeled data, i.e., $y_m^a = y_n^b$, we expand the element-wise difference in (2) as follows,

$$\begin{aligned} &\big||\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r)| - |\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)|\big| \\ =&\begin{cases} |(\boldsymbol{x}_i^a(r) - \boldsymbol{x}_m^a(r)) + (\boldsymbol{x}_n^b(r) - \boldsymbol{x}_j^b(r))|, (\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r))(\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)) \geq 0 \\ |(\boldsymbol{x}_i^a(r) - \boldsymbol{x}_n^b(r)) + (\boldsymbol{x}_m^a(r) - \boldsymbol{x}_j^b(r))|, (\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r))(\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)) < 0 \end{cases} \end{aligned} \tag{4}$$

Let us consider the first case in (4), i.e., $(\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r))(\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)) \geq 0$. If there exists r_0 such that the signs of $\boldsymbol{x}_i^a(r_0) - \boldsymbol{x}_m^a(r_0)$ and $\boldsymbol{x}_n^b(r_0) - \boldsymbol{x}_j^b(r_0)$ are the same, i.e., $(\boldsymbol{x}_i^a(r_0) - \boldsymbol{x}_m^a(r_0))(\boldsymbol{x}_n^b(r_0)) - \boldsymbol{x}_j^b(r_0)) \geq 0$, we have

$$\begin{aligned} &|(\boldsymbol{x}_i^a(r_0) - \boldsymbol{x}_m^a(r_0)) + (\boldsymbol{x}_n^b(r_0) - \boldsymbol{x}_j^b(r_0))| \\ =&|\boldsymbol{x}_i^a(r_0) - \boldsymbol{x}_m^a(r_0)| + |\boldsymbol{x}_n^b(r_0) - \boldsymbol{x}_j^b(r_0))| \end{aligned} \tag{5}$$

Denote the value of (5) as λ. Since persons in the unlabeled set are likely to be different from the ones in the labeled set using few labeled image pairs, the absolute differences $|\boldsymbol{x}_i^a(r_0) - \boldsymbol{x}_m^a(r_0)|$ and $|\boldsymbol{x}_n^b(r_0) - \boldsymbol{x}_j^b(r_0)|$ could be large due to different identities (though the differences are calculated for feature vectors from the same camera). Therefore, the element-wise difference λ of (5) is a large number, which implies the distance between $\boldsymbol{z}_{ij}^+$ and $\boldsymbol{z}_{mn}^u$ for $y_m^a = y_n^b$ is large by (2). Similarly, for the second case that $(\boldsymbol{x}_i^a(r) - \boldsymbol{x}_j^b(r))(\boldsymbol{x}_m^a(r) - \boldsymbol{x}_n^b(r)) < 0$, the norm $\|\boldsymbol{z}_{ij}^+ - \boldsymbol{z}_{mn}^u\|_p$ is large, if there exists r_0 such that $(\boldsymbol{x}_i^a(r_0) - \boldsymbol{x}_n^b(r_0))(\boldsymbol{x}_m^a(r_0) - \boldsymbol{x}_j^b(r_0)) \geq 0$. Under this condition, the distances between $\boldsymbol{z}_{ij}^+$ and $\boldsymbol{z}_{mn}^u$ are large for any $y_i^a = y_j^b$ in the labeled set and $y_m^a = y_n^b$ in the unlabeled set, i.e.,

$$\|\boldsymbol{z}_{ij}^+ - \boldsymbol{z}_{mn}^u\|_p \geq \lambda, \forall y_i^a = y_j^b, y_m^a = y_n^b, y_i^a \neq y_m^a \tag{6}$$

In this case, the positive information from the labeled data cannot be propagated to the unlabeled data. As a results, the graph propagation assumption cannot be employed under this condition in person re-identification.

Based on the above analysis, we follow the self training approach to iteratively label the most confident positive image pairs from the unlabeled data which will be discussed in the following sections.

3.2 Selecting Potential Positive Data by Increasing Positive Prior

Given a classification model f on the ADVs, one way to determine the positive data is to label the potential positive image pairs with very high scores, i.e.,

$$\hat{E}^+ = \{\boldsymbol{z}_{mn}^u | f(\boldsymbol{z}_{mn}^u) \geq \theta\} \tag{7}$$

where $\hat{E}^+$ denotes the set of potential positive ADVs selected from the unlabeled data and θ is a threshold for the selection. However, according to (3), the region with high confidence may contain both positive and negative image pairs. On the other hand, according to (6), the scores of positive image pairs do not change continuously. This means not all the positive ADVs give very high confidence scores. Consequently, it may not be a good strategy to label positive image pairs from the unlabeled data using (7).

To deal with this problem, we propose to take advantages of properties in person re-identification and define a better confidence measure ρ by both the classification function f and negative data $\boldsymbol{z}_{mk}^-$ and $\boldsymbol{z}_{ln}^-$ generated under non-overlapping camera views. If the score difference between $\boldsymbol{z}_{mn}^u$ and the negative data is larger, $\boldsymbol{z}_{mn}^u$ is more likely to be a positive ADV. Thus, we normalize the scores and define a new confidence measure ρ for the unlabeled ADVs as,

$$\rho(\boldsymbol{z}_{mn}^u) = \frac{f(\boldsymbol{z}_{mn}^u)}{\max\left(\max_k f(\boldsymbol{z}_{mk}^-), \max_l f(\boldsymbol{z}_{ln}^-)\right)}, \tag{8}$$

On the other hand, with information about the cameras, we can group the unlabeled data according to the camera indexes, i.e.

$$G_{m\cdot} = \{\boldsymbol{z}_{mn}^u = \boldsymbol{d}(\boldsymbol{x}_m^a - \boldsymbol{x}_n^b)|\forall \boldsymbol{x}_n^b\}, G_{\cdot n} = \{\boldsymbol{z}_{mn}^u = \boldsymbol{d}(\boldsymbol{x}_m^a - \boldsymbol{x}_n^b)|\forall \boldsymbol{x}_m^a\} \tag{9}$$

To reduce the proportion of negative matches, we select only one ADV from each $G_{m\cdot}$ or $G_{\cdot n}$ to obtain a set E_1, i.e.

$$E_1 = \{\boldsymbol{z}_{mn'}^u = \arg \max_{\boldsymbol{z}_{mn}^u \in G_{m\cdot}} \rho(\boldsymbol{z}_{mn}^u)\} \cup \{\boldsymbol{z}_{m'n}^u = \arg \max_{\boldsymbol{z}_{mn}^u \in G_{\cdot n}} \rho(\boldsymbol{z}_{mn}^u)\} \tag{10}$$

According to the definition in (10), E_1 contains the best match for each $\boldsymbol{x}_m^a$ under camera a or $\boldsymbol{x}_n^b$ under camera b by the classification function f. Although there may be more than one positive ADVs in each group $G_{m\cdot}$ or $G_{\cdot n}$, the selected one can be representative for others due to the following reasons. Denote two positive ADVs in $G_{m\cdot}$ as $\boldsymbol{z}_{mn_1}^+$ and $\boldsymbol{z}_{mn_2}^+$. According to the definition of difference vector given by (1) and the expanded difference in (4), it has

$$\|\boldsymbol{z}_{mn_1}^+ - \boldsymbol{z}_{mn_2}^+\| \leq \|\boldsymbol{x}_{n_1}^b - \boldsymbol{x}_{n_2}^b\| \tag{11}$$

Since both $\boldsymbol{z}_{mn_1}^+$ and $\boldsymbol{z}_{mn_2}^+$ are positive, the person labels $y_{n_1}^b$ and $y_{n_2}^b$ are equal to y_m^a. This means feature vectors $\boldsymbol{x}_{n_1}^b$ and $\boldsymbol{x}_{n_2}^b$ are extracted from the same person under the same camera view b. Since the variation under the same camera view must be small, the difference between $\boldsymbol{z}_{mn_1}^+$ and $\boldsymbol{z}_{mn_2}^+$ is small according to (11). This implies any positive ADV in a group $G_{m\cdot}$ is representative for others in this group. Similarly, this conclusion is also true for group $G_{\cdot n}$. Thus, it is good enough to select only one positive ADV from each group.

More importantly, we further show that the positive prior in E_1 is much larger than that in all the ADVs. Let c_1 be the rank one accuracy obtained by f, J be the number of persons under both camera views, $J^a(\geq J)$ and $J^b(\geq J)$ be the numbers of persons under cameras a and b, respectively. It has (See Appendix)

$$\tau \approx \frac{J}{J^a J^b}, \tau_1 \geq \frac{J c_1}{\max(J^a, J^b)}. \textit{ If } \max(\frac{1}{J^a}, \frac{1}{J^b}) \ll c_1, \textit{ then } \tau \ll \tau_1 \tag{12}$$

where τ is the percentages of the positive data in all the image pairs and τ_1 is the positive prior in E_1. Since the number of persons is usually very large in person re-identification, both $1/J^a$ and $1/J^b$ are very small numbers. On the other hand, using a classification function to obtain the rank one accuracy c_1 should be much better than a random guess with rank one accuracy $1/J^a$ or $1/J^b$. Therefore, the condition in (12) can be satisfied easily. This means the positive prior τ can be increased to τ_1 by only considering rank one matches in E_1. And, it is easier to correctly label a positive image pair from the unlabeled data with higher positive prior.

Since the rank one accuracy c_1 is not very large in person re-identification, there are still many negative ADVs in E_1. Consequently, we select only one potential positive ADV $\hat{\boldsymbol{z}}_{mn}^{+}$ in E_1 with the highest score in each iteration, i.e.

$$\hat{\boldsymbol{z}}_{mn}^{+} = \arg \max_{\boldsymbol{z}_{mn}^{u} \in E_1} \rho(\boldsymbol{z}_{mn}^{u}) \tag{13}$$

3.3 Ranking by Regression

Since each positive ADV $\boldsymbol{z}_{ij}^{+}$ should be ranked before its corresponding negative ones $\boldsymbol{z}_{ik}^{-}$ and $\boldsymbol{z}_{lj}^{+}$, we learn a weight vector $\boldsymbol{w}$ such that $\boldsymbol{w}^T \boldsymbol{z}_{ij}^{+} > \boldsymbol{w}^T \boldsymbol{z}_{ik}^{-}$ and $\boldsymbol{w}^T \boldsymbol{z}_{ij}^{+} > \boldsymbol{w}^T \boldsymbol{z}_{lj}^{-}$. To preserve the ranking relationship, we set $\boldsymbol{w}^T(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{ik}^{-}) \approx 1$ and $\boldsymbol{w}^T(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{lj}^{-}) \approx 1$ for regression. Then, the optimal weight vector $\boldsymbol{w}$ can be learnt by solving the following least square regression problem,

$$\min_{\boldsymbol{w}} \sum_{i,j,k} \left(\boldsymbol{w}^T(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{ik}^{-}) - 1\right)^2 + \sum_{j,i,l} \left(\boldsymbol{w}^T(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{lj}^{-}) - 1\right)^2 + \mu \boldsymbol{w}^T \boldsymbol{w} \tag{14}$$

where μ is a positive parameter for the regularization term to prevent from overfitting. This optimization problem can be solved by taking the first derivative to zero, and hence the optimal solution $\boldsymbol{w}^*$ is given by

$$\begin{aligned} \boldsymbol{w}^* &= (H + \mu I)^{-1}\boldsymbol{h}, \boldsymbol{h} = \sum_{i,j,k}(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{ik}^{-}) + \sum_{j,i,l}(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{lj}^{-}) \\ H &= \sum_{i,j,k}(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{ik}^{-})(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{ik}^{-})^T + \sum_{j,i,l}(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{lj}^{-})(\boldsymbol{z}_{ij}^{+} - \boldsymbol{z}_{lj}^{-})^T \end{aligned} \tag{15}$$

where I denotes the unit matrix. According to the solution given by (15), we do not need to save all the positive and negative ADVs. Once a potential positive ADV is selected from the unlabeled data, we can simply update H and $\boldsymbol{h}$, which will be described in the following section. This ensures that the proposed regression based ranking method is computationally efficient.

Algorithm 1 Training procedure of SSR-IPP

```
Input: Positive ADVs $\boldsymbol{z}_{ij}^{+}$, negative ADVs $\boldsymbol{z}_{ik}^{-}, \boldsymbol{z}_{mk}^{-}, \boldsymbol{z}_{ln}^{-}$, unlabeled ADVs $\boldsymbol{z}_{mn}^{u}$,
    parameter $\mu$, number of selected positive ADVs $Q$, unsupervised classifier $g$;
Compute $H$, $\boldsymbol{h}$ and $\boldsymbol{w}$ in (15) by $\boldsymbol{z}_{ij}^{+}$, $\boldsymbol{z}_{ik}^{-}$ and $\boldsymbol{z}_{lj}^{-}$;
Calculate confidence scores for each unlabeled ADV $\boldsymbol{z}_{mn}^{u}$ by $\boldsymbol{w}$ and $g$;
Construct the rank one matching set $E_1$;
for $t = 1, \cdots, Q$ do
    Calculate confidence scores for each $\boldsymbol{z}_{mn}^{u}$ in $E_1$ by $\boldsymbol{w}$ and $g$;
    Select one potential positive ADV $\hat{\boldsymbol{z}}_{mn}^{+}$ by (13);
    Update $H$, $\boldsymbol{h}$ by (16) and $\boldsymbol{w}$ by (15);
    Delete $\hat{\boldsymbol{z}}_{mn}^{+}$ from $E_1$;
end for
Output: Optimal weight vector $\boldsymbol{w}^{*}$.
```

3.4 Iterative Semi-Supervised Ranking

According to the analysis in Section 3.1, we follow the self training approach to iteratively label potential positive ADVs and re-train the weight vector $\boldsymbol{w}$. At iteration t, we have calculated H_t, $\boldsymbol{h}_t$ and $\boldsymbol{w}_t$. With $\boldsymbol{w}_t$, we can determine the classification function f_t by $f_t(\boldsymbol{z}_{mn}^{u}) = \boldsymbol{w}_t^T \boldsymbol{z}_{mn}^{u}$. Since $\boldsymbol{w}_t$ may over-fit the training data when the number of positive image pairs is very small, we propose to define f_t by adding an unsupervised classification model g, i.e., $f_t(\boldsymbol{z}_{mn}^{u}) = \boldsymbol{w}_t^T \boldsymbol{z}_{mn}^{u} + g(\boldsymbol{z}_{mn}^{u})$. Then, we select one potential positive ADV $\hat{\boldsymbol{z}}_{m_t n_t}^{+}$ in E_1 by (13) and H_t, $\boldsymbol{h}_t$ can be updated by the following equations,

$$\begin{aligned} H_{t+1} &= H_t + \sum_k (\hat{\boldsymbol{z}}_{m_t n_t}^{+} - \boldsymbol{z}_{m_t k}^{-})(\hat{\boldsymbol{z}}_{m_t n_t}^{+} - \boldsymbol{z}_{m_t k}^{-})^T \\ &\quad + \sum_l (\hat{\boldsymbol{z}}_{m_t n_t}^{+} - \boldsymbol{z}_{l n_t}^{-})(\hat{\boldsymbol{z}}_{m_t n_t}^{+} - \boldsymbol{z}_{l n_t}^{-})^T \\ \boldsymbol{h}_{t+1} &= \boldsymbol{h}_t + \sum_k (\hat{\boldsymbol{z}}_{m_t n_t}^{+} - \boldsymbol{z}_{m_t k}^{-}) + \sum_l (\hat{\boldsymbol{z}}_{m_t n_t}^{+} - \boldsymbol{z}_{l n_t}^{-}) \end{aligned} \tag{16}$$

With H_{t+1} and $\boldsymbol{h}_{t+1}$, we can compute $\boldsymbol{w}_{t+1}$ by (15). After that, $\hat{\boldsymbol{z}}_{m_t n_t}^{+}$ is deleted from E_1 for the next iteration. Algorithm 1 summarizes the proposed Semi-Supervised Ranking method with Increased Positive Prior (SSR-IPP).

4 Experiments

4.1 Datasets

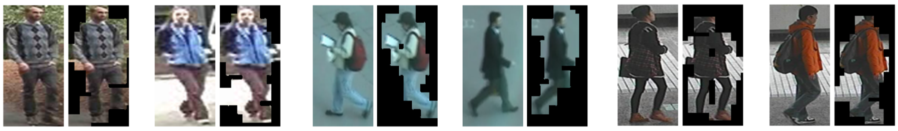
Three publicly available datasets, namely VIPeR[1] [29], PRID[2] [30] and CUHK[3] [18], are used for evaluation of the proposed method. Example images in these three

[1] http://soe.ucsc.edu/~dgray/VIPeR.v1.0.zip

[2] https://lrs.icg.tugraz.at/datasets/prid/

[3] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

datasets are shown in Figure 1(a), Figure 1(b) and Figure 1(c), respectively. VIPeR is a re-identification dataset containing 632 person image pairs captured by two cameras outdoor. In this dataset, 632 image pairs are randomly separated into half for training and the other half for testing. 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. 100 out of the 200 image pairs are randomly selected as the training set, and the others for testing. CUHK dataset contains five camera pairs. Under each camera view, there are two images for each person. Following the single shot setting in [18], images from camera pair one with 971 persons are used for experiments. On this dataset, 971 persons are randomly split as 485 for training and 486 for testing. For the testing data in VIPeR, PRID or CUHK, the evaluation is performed by searching the 316, 100 or 486 persons in one camera view from another view. These experiments were performed ten times and the average results are reported.



(a) VIPeR (b) PRID (c) CUHK

Fig. 1. Sample images and masked results on three datasets: (a) VIPeR [29], (b) PRID [30] and (c) CUHK [18].

For each image in these datasets, we concatenate two types of features as the input feature vector. The first type of feature is constructed by dividing a person image into six horizontal stripes and compute the RGB, YCbCr, HSV color features and two types of texture features extracted by Schmid and Gabor filters on each stripe as reported in [13, 14, 17]. For the second type of feature, we perform foreground detection to detect the human pixels by the spatial hierarchy pose estimation method [31] with source code online[4]. Example masked results are shown in Figure 1(a), Figure 1(b) and Figure 1(c) for VIPeR, PRID and CUHK datasets, respectively. Then, the masked person image is divided into 3×1 vertically overlapped boxes and the code[5] in [11] are used to extract color histogram and SIFT features on each box.

[4] `http://www.cs.cmu.edu/~ILIM/projects/IM/humanpose/humanpose.html`

[5] `http://mmlab.ie.cuhk.edu.hk/projects/project_salience_reid/index.html`

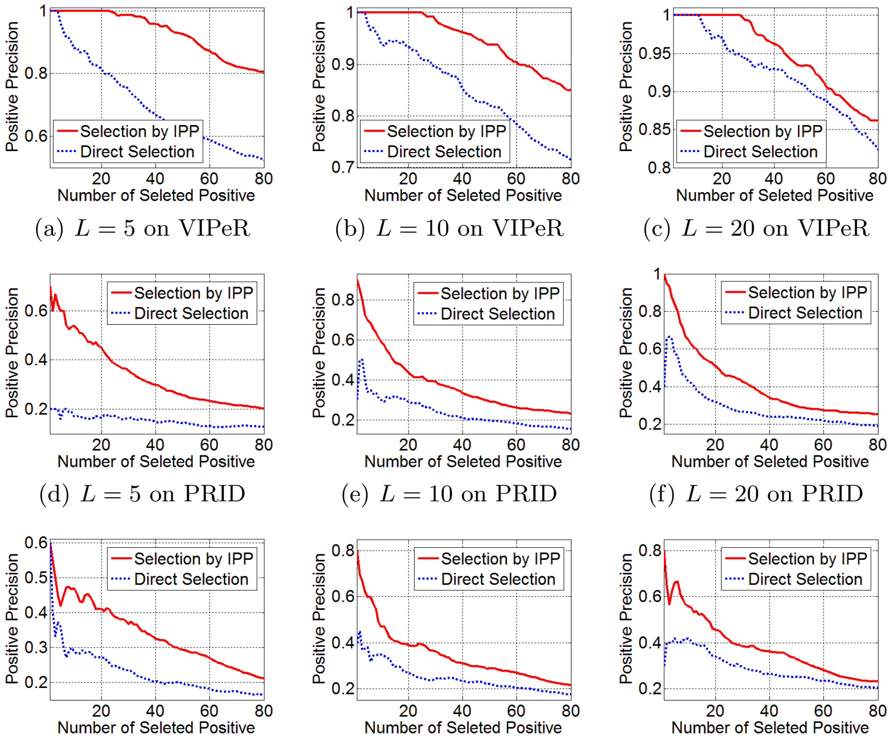
4.2 Evaluation of SSR-IPP

In our experiments, we use l_1 distance in the unsupervised classification model g and empirically set $\mu = 1$. Without the time acquisition information in the PRID, VIPeR and CUHK datasets, negative image pairs from non-overlapping cameras are generated by simulating the synchronization using label information. Ten negative image pairs are randomly generated for each unlabeled person image. We first show the precisions for labeling the positive data, i.e., the number of true positive ADVs divided by the number (Q) of selected potential positive ADVs. The results are shown in Figures 2(a)-2(c) for VIPeR, Figures 2(d)-2(f) for PRID and Figures 2(g)-2(i) for CUHK dataset. For each dataset, we use different numbers ($L = 5, 10, 20$) of labeled positive image pairs to evaluate the performance. Our method by Increasing Positive Prior (IPP) is compared with the direct selection approach given by (7) which selects the ADVs with top classification scores as positive. From Figures 2(a)-2(i), we can see that our method remarkably outperforms the direct selection approach with different numbers (L) of labeled positive image pairs on the three datasets. Thus, our method can achieve better re-identification performance by correctly selecting more (true) positive ADVs for training compared with the direct selection approach.

On the other hand, from Figures 2(a)-2(i), we can see that the positive labeling precision drops when the number (Q) of selected potential positive ADVs increases. This means more ADVs are wrongly labeled when Q is large. However, if Q is too small, we may not have enough labeled data to train a robust model for re-identification. To evaluate the relationship between Q and the re-identification performance, we plot the rank one accuracy for varying Q on the three datasets in Figures 2(a)-2(i), respectively. From these figures, we can see that when Q is large, the rank one accuracy does not drop very much, though the precision for selecting potential positive ADVs drops significantly as shown in Figures 2(a)-2(i). This may be due to that the corresponding negative ADVs are correctly labeled under non-overlapping cameras. Moreover, Figures 3(a)-3(c) show that the rank one accuracy can be increased by training with the potential positive ADVs selected from the unlabeled data. For example, when the number (L) of labeled image pairs is equal to five, the improvement by selecting potential positive ADVs is extremely significant. The rank one accuracies on these three datasets for $L = 5$ and $Q = 20$ are over two times higher than those using only few labeled image pairs for training. These results indicate that it becomes more important to learn from unlabeled data for person re-identification when only a small number of persons are labeled for training.

4.3 Comparison with Existing Methods

In this section, we compare the proposed method with two state-of-the-art distance learning methods for person re-identification, namely Ranked Support Vector Machines (RSVM) [14] and Relative Distance Comparison (RDC) [17]. We have re-implemented these two methods. In our implementation, the parameter

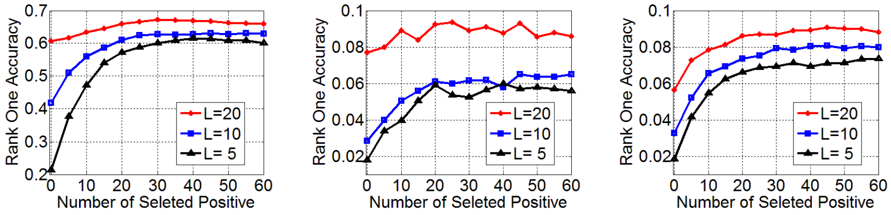


(a) $L = 5$ on VIPeR (b) $L = 10$ on VIPeR (c) $L = 20$ on VIPeR

(d) $L = 5$ on PRID (e) $L = 10$ on PRID (f) $L = 20$ on PRID

(g) $L = 5$ on CUHK (h) $L = 10$ on CUHK (i) $L = 20$ on CUHK

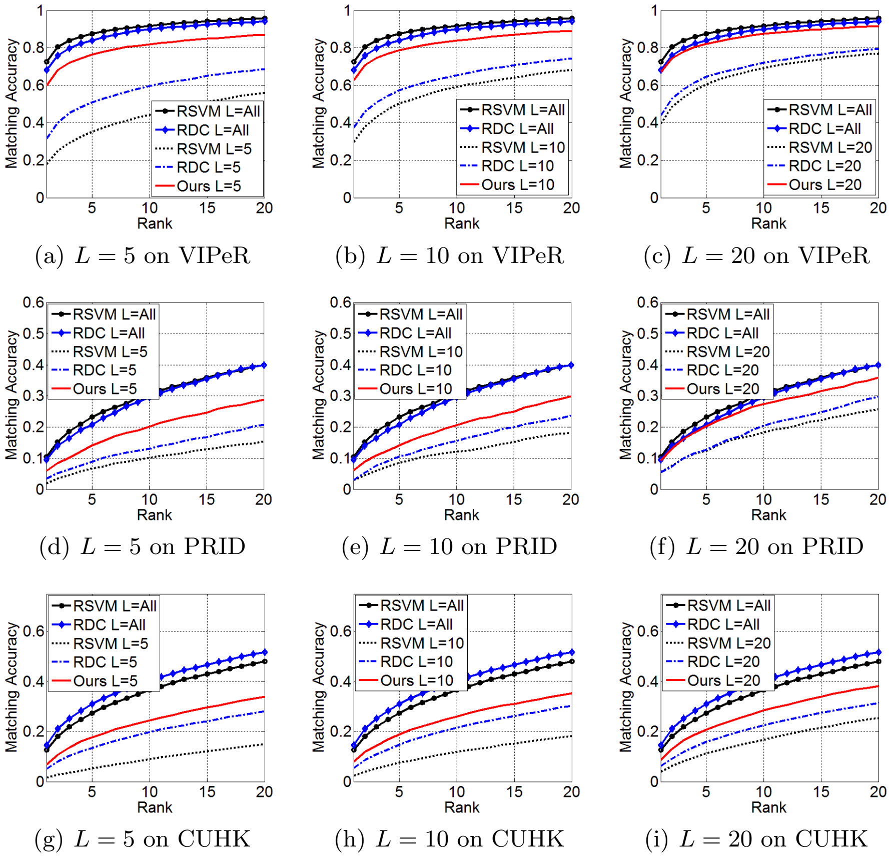
Fig. 2. Precisions for labeling positive ADVs by Increasing Positive Prior (IPP) and Direct Selection with varying numbers (Q) of selected potential positive ADVs on (a)-(c) VIPeR [29], (d)-(f) PRID [30] and (g)-(i) CUHK [18]



(a) VIPeR (b) PRID (c) CUHK

Fig. 3. Rank one accuracy for varying numbers (Q) of selected potential positive ADVs on 3 data sets: (a) VIPeR [29], (b) PRID [30] and (c) CUHK [18].

C in RSVM is empirically set as 1 for robust performance. According to the results shown in Figures 3(a)-3(c), we set the number of potential positive ADVs as $Q = 20$ on PRID, $Q = 30$ on VIPeR and CUHK datasets.



(a) $L = 5$ on VIPeR (b) $L = 10$ on VIPeR (c) $L = 20$ on VIPeR

(d) $L = 5$ on PRID (e) $L = 10$ on PRID (f) $L = 20$ on PRID

(g) $L = 5$ on CUHK (h) $L = 10$ on CUHK (i) $L = 20$ on CUHK

Fig. 4. CMC curves with different numbers ($L = 5, 10, 20$) of labeled image pairs on 3 datasets: (a)-(c) VIPeR [29], (d)-(f) PRID [30] and (g)-(i) CUHK [18].

The CMC curves on VIPeR, PRID and CUHK datasets are shown in Figures 4(a)-4(c), Figures 4(d)-4(f) and Figures 4(g)-4(i), respectively. From these figures, we can see that the re-identification performance degrades significantly when only few labeled positive image pairs, e.g., $L = 5, 10, 20$, are used for training. When all the training data are labeled, i.e., $L = All$, RSVM and RD-C achieve around 70% rank one accuracy on VIPeR dataset as shown in Figure 4(a)[6]. However, when using only five labeled positive image pairs for training,

[6] Note that the feature used in our experiments is different from those in existing methods. It is very discriminative for VIPeR dataset, so it can achieve 70% rank one accuracy using 316 matched image pairs for training. Such good performance may be due to the combination of foreground detection and global?feature extraction (on a large region of an image) which is very effective for VIPeR dataset. It is interesting to conduct further investigation on this issue, but it is not the focus of this paper.

the rank one accuracy degrades to about 20% by RSVM and 30% by RDC. This means the degradation of rank one accuracy can be up to about 50% by RSVM and 40% by RDC. The reason for these results is, the distance learning methods are over-fitted for the small amount of labeled data. Although both RSVM and RDC have a significant performance degradation, it is interesting to see in Figures 4(a)-4(i) that RDC outperforms RSVM on the three datasets when the number of available labels is small. This indicates that RDC has better generalization ability by utilizing the advantages of second-order distance when few labeled data are available for training.

Comparing the proposed method with RSVM and RDC for $L = 5, 10, 20$, our method achieves significantly better ranking performance as shown in Figures 4(a)-4(i). On VIPeR dataset, the rank one accuracy of our method is above 20% higher than that of RSVM or RDC when the number of labeled positive image pairs are less than or equal to ten, i.e., $L = 5, 10$. This indicates that our method can significantly improve the performance for re-identification by using unlabeled data. On the other hand, Figure 4(c) on VIPeR and Figure 4(f) on PRID show that the CMC curves of our method using only 20 labeled positive image pairs are close to those of RSVM and RDC using all the labeled training data. These results indicate that our method can achieve convincing performance for person re-identification only with few labeled positive image pairs, which helps reduce the expensive cost needed for manual labeling.

5 Conclusion

In this paper, we have developed a novel Semi-Supervised Ranking method with Increased Positive Prior (SSR-IPP) for person re-identification using only few labeled positive image pairs. By analyzing the data distribution properties in person re-identification, we show that the widely used low density separation and graph propagation assumptions are not valid under certain conditions. In this context, we propose to iteratively add the most confident potential positive Absolute Difference Vector (ADV) from the unlabeled data for training. Since it suffers from a severe data imbalance problem in person re-identification, i.e., the number of positive image pairs is much smaller than that of negative ones, it is more likely to select a negative ADV from the unlabeled data. To increase the positive prior, we select the potential positive ADVs from the rank one matching subset in all the unlabeled data. Adding the selected potential positive ADVs to the regression based ranking problem, the confidence measure and weight vector are updated iteratively for the optimal solution.

Experimental results demonstrate that our method significantly outperforms state-of-the-art distance learning methods using only a small number of labeled positive image pairs for training. For example, the rank one accuracy of SSR-IPP is above 20% higher than that of RSVM or RDC when the number of labeled positive image pairs are less or equal to ten. On the other hand, it is shown that the re-identification performance deteriorates dramatically when the number of labels is very small for training by existing methods. Since our method achieves convincing performance for re-identification with few matched image pairs, it can help reduce the expensive efforts needed for manual labeling. Moreover, our experiments also show that the second-order distance based Relative Distance Comparison (RDC) [17] method has better generalization ability than the first-order distance based Ranking Support Vector Machines (RSVM) [14] when the number of labeled positive image pairs is small. Since the proposed method is based on first-order distance, it is promising to study the development of second-order distance based semi-supervised ranking method for person re-identification.

Acknowledgement

The work is supported in part by ONR-N00014-13-1-0764, NSF-III-1360971, AFOSR-FA9550-13-1-0137, and NSF-Bigdata-1419210.

Appendix: Proof of Equation (12)

Suppose there are N_i^a images for person i under camera a and N_j^b images for person j under b. The number of positive matches for person i in both camera views is $N_i^a N_i^b$. Since the total numbers of images are $\sum_{i=1}^{J^a} N_i^a$ under camera view a and $\sum_{j=1}^{J^b} N_j^b$ under camera view b, the positive prior τ is calculated by

$$\tau = \frac{\sum_{i=1}^{J} N_i^a N_i^b}{\sum_{i=1}^{J^a} N_i^a \sum_{j=1}^{J^b} N_j^b} \tag{17}$$

The total number of image pairs in E_1 is equal to the number of groups $G_{m\cdot}$ and $G_{\cdot n}$, i.e., $\sum_{i=1}^{J^a} N_i^a + \sum_{j=1}^{J^b} N_j^b$. There are $\sum_{i=1}^{J} N_i^a$ groups $G_{m\cdot}$ and $\sum_{j=1}^{J} N_j^b$ groups $G_{\cdot n}$ containing at least one positive ADV. However, the classification function f may wrongly select a negative ADV from $G_{m\cdot}$ or $G_{\cdot n}$ that contains positive ADV(s). Thus, the number of ADVs in E_1 is $(\sum_{i=1}^{J} N_i^a + \sum_{j=1}^{J} N_j^b)c_1$, where c_1 is the rank one accuracy measuring the performance of f. Then, the positive prior τ_1 in E_1 is given by the following equation,

$$\tau_1 = \frac{(\sum_{i=1}^{J} N_i^a + \sum_{j=1}^{J} N_j^b)c_1}{\sum_{i=1}^{J^a} N_i^a + \sum_{j=1}^{J^b} N_j^b} \tag{18}$$

Since it is difficult to compare τ and τ_1 by (17) and (18) directly, we approximate them by assuming $N_i^a \approx \sum_{i'=1}^{J^a} N_{i'}^a / J^a$ and $N_j^b \approx \sum_{j'=1}^{J^b} N_{j'}^b / J^b$. Substituting the approximations of N_i^a and N_j^b into (17) and (18), respectively, τ and τ_1 become

$$\tau = \frac{J}{J^a J^b}, \qquad \tau_1 = \frac{(\frac{J}{J^a}\sum_{i=1}^{J^a} N_i^a + \frac{J}{J^b}\sum_{j=1}^{J^b} N_j^b)c_1}{\sum_{i=1}^{J^a} N_i^a + \sum_{j=1}^{J^b} N_j^b} \geq \frac{Jc_1}{\max(J^a, J^b)} \tag{19}$$

If $\max(1/J^a, 1/J^b) \ll c_1$, multiplying $J^a J^b$ on both sides, we obtain $\max(J^a, J^b) \ll J^a J^b c_1$. Thus, $\tau \ll \tau_1$, which leads to (12).

References

1. Bąk, S., Corvée, E., Brémond, F., Thonnat, M.: Boosted human re-identification using riemannian manifolds. Image and Vision Computing **30** (2010) 443–452
2. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. (2010)
3. Bauml, M., Stiefelhagen, R.: Evaluation of local features for person re-identification in image sequences. In: AVSS. (2011)
4. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC. (2011)
5. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. JAIHC **2** (2011) 127–151
6. Jungling, K., Arens, M.: View-invariant person re-identification with an implicit shape model. In: AVSS. (2011)
7. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters **33** (2012) 898–903
8. Bąk, S., Charpiat, G., Corvée, E., Brémond, F., Thonnat, M.: Learning to match appearances by correlations in a covariance metric space. In: ECCV. (2012)
9. Ma, B., Su, Y., Jurie, F.: BiCov: a novel image representation for person re-identification and face verification. In: BMVC. (2012)
10. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. TPAMI **35** (2013) 1622–1634
11. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR. (2013)
12. Xu, Y., Lin, L., Zheng, W.S., Liu, X.: Human re-identification by matching compositional template with cluster sampling. In: ICCV. (2013)
13. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV. (2008)
14. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC. (2010)
15. Avraham, T., Gurvich, I., Lindenbaum, M., Markovitch, S.: Learning implicit transfer for person re-identification. In: ECCV Workshop. (2012)
16. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: ECCV. (2012)
17. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. TPAMI **35** (2013) 653–668
18. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR. (2013)
19. Liu, C., Loy, C.C., Gong, S., Wang, G.: POP: Person re-identification post-rank optimisation. In: ICCV. (2013)
20. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: ICCV. (2013)
21. Ma, A.J., Yuen, P.C., Li, J.: Domain transfer support vector ranking for person re-identification without target camera label information. In: ICCV. (2013)
22. Chapelle, O., Schölkopf, B., Zien, A., et al.: Semi-Supervised Learning. Volume 2. MIT Press Cambridge (2006)
23. Zhu, X.: Semi-supervised learning literature survey. Computer Science, University of Wisconsin - Madison (2008)

24. Figueira, D., Bazzani, L., Minh, H.Q., Cristani, M., Bernardino, A., Murino, V.: Semi-supervised multi-feature learning for person re-identification. In: AVSS. (2013)
25. Bäuml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: CVPR. (2013)
26. Iqbal, U., Curcio, I.D.D., Gabbouj, M.: Who is the hero? - semi-supervised person re-identification in videos. In: VISAPP. (2014)
27. Amini, M.R., Truong, T.V., Goutte, C.: A boosting algorithm for learning bipartite ranking functions with partially labeled data. In: SIGIR. (2008)
28. Hoi, S.C., Jin, R.: Semi-supervised ensemble ranking. In: AAAI. (2008)
29. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance. (2007)
30. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: SCIA. (2011)
31. Tian, Y., Zitnick, C., Narasimhan, S.: Exploring the spatial hierarchy of mixture models for human pose estimation. In: ECCV. (2012)