Research Article

Co-Metric Learning for Person Re-Identification

Qingming Leng

School of Information Science and Technology, Jiujiang University, Jiujiang 332000, China

Correspondence should be addressed to Qingming Leng; lengqingming@126.com

Received 17 April 2018; Revised 8 June 2018; Accepted 12 June 2018; Published 15 July 2018

Academic Editor: Chen Gong

Copyright © 2018 Qingming Leng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Person re-identification, aiming to identify the same pedestrian images across disjoint camera views, is a key technique of intelligent video surveillance. Although existing methods have developed both theories and experimental results, most of effective ones pertain to fully supervised training styles, which suffer the small sample size (SSS) problem a lot, especially in label-insufficient practical applications. To bridge SSS problem and learning model with small labels, a novel semisupervised co-metric learning framework is proposed to learn a discriminative Mahalanobis-like distance matrix for label-insufficient person re-identification. Different from typical co-training task that contains multiview data originally, single-view person images are firstly decomposed into pseudo two views, and then metric learning models are produced and jointly updated based on both pseudo-labels and references iteratively. Experiments carried out on three representative person re-identification datasets show that the proposed method performs better than state of the art and possesses low label sensitivity.

1. Introduction

Person re-identification (re-id), namely, seeking occurrences of a query person (probe) from person candidates (gallery), is a hot-spot and challenging topic of intelligent video surveillance [1, 2], which also underpins many crucial multimedia applications, such as person retrieval [3, 4], long-term pedestrian tracking [5, 6], and cross-view action analysis [7]. The main challenge of re-id can be concluded as intrapersonal visual variations across multicamera views even larger than interpersonal ones, due to the significant changes in view-points, illuminations, body poses, and background clutters (see Figure 1). Moreover, traditional biometrics, such as gait and face, are unreliable to be exploited especially in uncontrolled practical environment; thus researchers always carry out person re-identification task based on body appearance characteristics. Current person re-identification methods have been primarily introduced to two aspects: feature construction and learning, or subspace/metric learning. Due to more and more attention from computer vision and machine learning fields in recent years, researchers bring great improvements on both theories and experimental results of person re-identification study.

- (i) *Feature construction and learning* aim at designing or studying discriminative appearance descriptions

[8–20] that are robust for distinguishing different pedestrians across arbitrary cameras. However, hand-crafted feature construction is extremely challenging due to miscellaneous and complicated variations. Therefore, feature learning based on salience model, deep neural network, and so forth becomes popular approaches to practice better feature representation.

- (ii) *Subspace and metric learning* aims at seeking a proper subspace or distance measure by Mahalanobis-like metric learning [21–30]. Given a set of person image pairs, metric learning based methods are to learn an optimal positive semidefinite matrix for the validity of metric that maximizes the probability of true matches pair having smaller distance than wrong match pairs.

Whether feature learning or metric learning methods, state of the art usually exploits the characteristics of labelled training data as far as possible, which typically pertains to fully supervised method. However labels are always insufficient in practical applications, resulting in the number of labelled training samples even smaller than that of feature dimensions, namely, small sample size (SSS) problem [31] that is a core challenge of learning based person re-identification. To solve the SSS issue, there are many training styles designed

FIGURE 1: Illustration of the person re-identification task. (a) Person image samples of pedestrian derived from VIPeR dataset [35], in which each column represents the same person images and each row represents images observed from the same camera views, and appearance of the same person images changes severely in different camera views. (b) Example illustrating the characteristics of person re-identification in practical surveillance environment; as can be seen, gait and face are *infeasible* to be exploited by reason of low resolution and occlusion.

for noisy learning and inadequate supervision [32], and co-training is always one of the most important paradigms that is still vibrant for multiview learning [33]. Therefore, motivated by semisupervised co-training [34], we propose a novel co-metric learning framework for person re-identification to bridge the inadequate labelled data and metric learning model.

In a typical co-training work, training data is adopted to study classification models in two views separately, whereas the updates of models benefit from each other's views. However, different from applications where data is collected from multimodal sources, person re-identification datasets are commonly presented as single-view pedestrian images. In that case, the core difficulty of applying co-training paradigm in person re-identification community comes at learning and updating a model in single view. As we know, the features in higher dimension own more useful information but larger noise, such that dimension reduction is always necessary for feature extraction. If we decompose the high-dimension features into two views before dimension reduction, it is probably to produce different but effective descriptions in pseudo two views for our co-metric learning framework. Therefore, we firstly present a binary-weight learning method for splitting the single-view representation to pseudo two views automatically, and then two metric learning models are studied, respectively, in each view for matching the unlabelled training samples; finally metrics benefit each other and meanwhile are jointly updated based on the ranking list of unlabelled samples iteratively.

The main contributions of this paper can be summarized as follows: (1) An effective co-metric learning framework is presented for semisupervised person re-identification; it can learn a discriminative Mahalanobis-like distance matrix, even lacking adequate labelled data. (2) Pseudo two views of person data could be used for metrics generation based on self-adaptive feature decomposition. (3) Both

pseudo-labels and references on unlabelled dataset are adopted for acquiring discriminative metrics update. The rest of the paper is organized as follows. Section 2 introduces a brief review of related work for person re-identification. Section 3 explains our method in detail. Section 4 presents experimental results compared with state of the art on three datasets. Section 5 concludes this paper.

2. Related Work

In this section, we give a brief review of the studies most related to person re-identification task. Typically, current person re-identification research can be categorized into two classes: feature representation based methods and distance measure based methods.

Feature representation based methods pay attention to constructing discriminative visual descriptions by feature selection or learning. Gheissari et al. [8] generated salient edges based on a spatial-temporal segmentation algorithm and then obtained an invariant identity signature by combining normalized color and salient edge histograms. Wang et al. [9] designed a co-occurrence matrix based appearance model to capture the spatial distribution of the appearance relative to each of the object parts. Farenzena et al. [10] tried to combine multiple features from five body regions that are exploited by symmetry and asymmetry perceptual principles. Kviatkovsky et al. [11] found that color structure descriptors derived from different body parts turn out to be invariants under different lighting conditions. To improve the discriminative power of visual descriptions, feature selection technique is used to pick out more robust feature weightings, or dimensions, or patch salience. Gray et al. [12] transformed person re-identification into a classification problem and employed an ensemble of the localized features through AdaBoost algorithm. Zhao et al. [13] applied adjacency constrained patch matching to build dense correspondence between image pairs and assigned

FIGURE 2: Flowchart of the co-metric learning framework for person re-identification. Single-view features of training data are firstly decomposed into pseudo two views for learning corresponding metric models. And then ranking list of unlabelled dataset in each view could be generated via distance measurement. Finally, positive and negative pseudo labels that are the top- n and bottom- m samples of ranking list, respectively, and references that are the top- k neighbours of consensual pseudo-labels (marked red) are jointly utilized for metric update.

salience to each patch in an unsupervised manner. Some recent works introduce deep learning framework to acquire robust local feature representations and then encoding them. Li et al. [14] learned a unified deep filter by introducing a patch matching layer and a max-out grouping layer for person re-identification. Ahmed et al. [15] presented a deep convolutional architecture that captured local relationships between person images based on mid-level features. Generally, deep learning is usually utilized to learn feature representations by using deep convolutional features [14–17] or from the fully connected features [18–20] in person re-identification works.

Distance measure based methods aim at finding out a uniform distance measure by subspace learning or metric learning. Most successful metric learning algorithms demonstrate an obvious superiority based on supervised learning. Hizer et al. [21] and Dikmen et al. [22] utilized a classical metric learning method called LMNN to learn an optimal metric for person re-identification. Zheng et al. [23] learned a Mahalanobis distance metric with a probabilistic relative distance comparison method. Kostinger et al. [24] introduced a simpler metric function (KISSME) to fit pairwise samples based on Gaussian distribution hypothesis, and Tao et al. [25] got better estimation of the covariance matrices of KISS metric learning by seamlessly integrating smoothing and regularization. Mignon et al. [26] learn distance metric from sparse pairwise similarity/dissimilarity constraints in high dimensional space called pairwise constrained component analysis. Pedagadi et al. [27] conducted a metric-like work that combined unsupervised PCA dimensionality reduction and Local Fisher Discriminant Analysis. Li et al. [28] proposed to learn a decision function that joined distance metric and locally adaptive thresholding rule. Wang et al. [29] transformed the deep learning as the most popular machine learning paradigm is also adopted to learn the distance

metric. Wang et al. [30] put forward a data-driven distance metric method, re-exploiting the training data to adjust the metric for each query-gallery pair.

3. Methodology

This section presents the main procedures of our co-metric learning framework (see Figure 2), mainly including self-adaptive feature decomposition for pseudo two-view metric learning, semisupervised metric update based on pseudo-labels and references.

3.1. Problem Formulation. Under a semisupervised person re-identification setting, it considers a pair of cameras C_a and C_b with nonoverlapping field of views and training persons set $O = \{O_L, O_U\}$. Labelled training persons set $O_L = \{o_1, o_2, \dots, o_m\}$ is associated with the two cameras, where m is the number of persons. Images of persons captured from C_a and C_b are denoted as x_a^i and x_b^j , respectively, $x_a^i, x_b^j \in \mathbb{R}^d$. Two labelled training sets corresponding to C_a and C_b are represented by $X_{a,L} = \{x_a^1, \dots, x_a^i, \dots, x_a^m\}$, $1 \leq i \leq m$, and $X_{b,L} = \{x_b^1, \dots, x_b^j, \dots, x_b^m\}$, $1 \leq j \leq m$, where $i = j$ means the same person o_i . Then let unlabelled training persons set $O_U = \{o_{m+1}, o_{m+2}, \dots, o_{m+n}\}$, $X_{a,U} = \{x_a^{m+1}, \dots, x_a^{m+i}, \dots, x_a^{m+n}\}$, $1 \leq i \leq n$, and $X_{b,U} = \{x_b^{m+1}, \dots, x_b^{m+j}, \dots, x_b^{m+n}\}$, $1 \leq j \leq n$; however o_i and o_j may not be the same pedestrian here even if $i = j$.

A classical supervised metric learning algorithm [21] trains a Mahalanobis-like distance function based on $X_{a,L}$ and $X_{b,L}$. Given a pair of training samples x_a^i and x_b^j , their distance can be defined as

$$d_M(x_a^i, x_b^j) = (x_a^i, x_b^j)^T M (x_a^i, x_b^j). \quad (1)$$

$\mathbf{M}$ is a positive semidefinite matrix for the validity of metric. By performing matrix decomposition on $\mathbf{M}$ with $\mathbf{M} = \mathbf{L}^T \mathbf{L}$, (1) can be rewritten as

$$\begin{aligned} d_{\mathbf{M}}(x_a^i, x_b^j) &= (x_a^i, x_b^j)^T \mathbf{L}^T \mathbf{L} (x_a^i, x_b^j) \\ &= \|\mathbf{L} \cdot x_a^i - \mathbf{L} \cdot x_b^j\|^2. \end{aligned} \quad (2)$$

It is easy to see from the above derivation that the essence of the metric is to seek an optimal projection matrix $\mathbf{M}$ (or $\mathbf{L}$) under the supervised information generally containing two pairwise constraints, i.e., similar constraint and dissimilar constraint. However, access to labelled data is usually difficult or too expensive to obtain; comparatively unlabelled data is massive and easily acquired. Therefore learning based on both labelled and unlabelled samples is not only meaningful issue but also pressing for practical intelligent video surveillance.

3.2. Self-Adaptive Feature Decomposition. Given a set of single-view training samples, it aims at producing pseudo two-view representations that could be used for learning metric model in each view. In a typical co-training task, there is dataset O consisting of two feature views V^1 and V^2 , which satisfy two conditions [34]: (1) two hypotheses $h_1, h_2 \in H$ occur having low-error on V^1, V^2 ; (2) V^1 and V^2 need to be conditionally independent.

To achieve the above demands, a binary learning method based on binary-weight vectors ω_1, ω_2 is proposed to decompose single-view features $V \in \mathbb{R}^d$ with dimension d into two totally different but both effective views V^1, V^2 automatically, which could be treated as pseudo two-view features of training samples. $V_1^i = V^i * \omega_1^i, V_2^i = V^i * \omega_2^i$. V^i, V_1^i, V_2^i indicate the i th dimension of V, V_1, V_2 , respectively, $1 \leq i \leq d$. To make V^1, V^2 conditionally independent, a succinct way is that V^i can only be used by one of V_1^i, V_2^i . In other words, ω_1^i, ω_2^i can be both indicated as 0/1 weights as

$$\begin{aligned} \omega_1^i + \omega_2^i &= 1. \\ \omega_1^i * \omega_2^i &= 0. \end{aligned} \quad (3)$$

As can be seen, the values of ω_1^i, ω_2^i are still uncertain. Therefore, ω_1, ω_2 are trained together on the labelled samples set O_L , ensuring that feature representations generated from V^1, V^2 both perform well. x_a^i, x_b^j ($i \neq j$) are, respectively, positives and negatives of sample x_a^i on O_L , and ω_1 can be trained by the objective function as

$$\begin{aligned} \arg \min_{\omega_1} \ell(\omega_1) \\ = \sum \|1 + d(\omega_1^T x_a^i, \omega_1^T x_b^j) - d(\omega_1^T x_a^i, \omega_1^T x_b^j)\|. \end{aligned} \quad (4)$$

So x_a^i is similar to x_b^i and meanwhile dissimilar to x_b^j as much as possible by applying ω_1 . $d(\cdot)$ denotes the normalized distance of objects; here Euclidean distance is adopted. Similarly, $\ell(\omega_2)$ is constructed for ω_2 . And then, $\ell(\omega_1), \ell(\omega_2)$ are trained

with the constraints of (3) jointly through minimizing the maximum of the two as

$$f(\omega_1, \omega_2) = \arg \min_{\omega_1, \omega_2} \max(\ell(\omega_1), \ell(\omega_2)). \quad (5)$$

3.3. Semisupervised Metric Update. After acquiring pseudo two-view representations V^1, V^2 of person images, Mahalanobis-like metric model would be learned each from one view for matching the unlabelled training samples. Consider a pairwise difference $\Delta = x_i - x_j, x_i, x_j \in O$, where O is the person dataset and Δ is the intrapersonal difference if $y_i = y_j$, namely, $y_{ij} = 1$, while Δ is the interpersonal difference if $y_i \neq y_j$, namely, $y_{ij} = 0$. Mahalanobis-like metric can be learned via zero-mean Gaussian structure [24] of the difference space as

$$\delta(\Delta) = \log \left(\frac{(1/\sqrt{2\pi \sum_{y_{ij}=0}^{-1}}) e^{-(1/2)\Delta^T \sum_{y_{ij}=0}^{-1} \Delta}}{(1/\sqrt{2\pi \sum_{y_{ij}=1}^{-1}}) e^{-(1/2)\Delta^T \sum_{y_{ij}=1}^{-1} \Delta}} \right). \quad (6)$$

The above decision function can be simplified as (7) by the log-likelihood ratio test, and then distance between x_i and x_j can be written as (8):

$$\delta(\Delta) = \Delta^T \left(\sum_{y_{ij}=1}^{-1} - \sum_{y_{ij}=0}^{-1} \right) \Delta. \quad (7)$$

$$d(x_i, x_j) = (x_i - x_j)^T \left(\sum_{y_{ij}=1}^{-1} - \sum_{y_{ij}=0}^{-1} \right) (x_i - x_j). \quad (8)$$

The original semidefinite matrix $\mathbf{M}$ in Mahalanobis-like metric function is reflected by $\sum_{y_{ij}=1}^{-1} - \sum_{y_{ij}=0}^{-1}$. Since the ranking lists of unlabelled training samples are calculated based on (8), the core issue comes to how to use these ranking lists for metric update, and three observations could be helpful and important to answer the question. First, co-training style is promoting the models in two views teaching each other; thereby ranking list in one view should benefit model in another. Second, top- n samples in the ranking lists probably have more similar visual appearance as the probe, whereas the visual information of bottom- m samples is perhaps further dissimilar as that of the probe; thus the top- n and bottom- m samples could be treated as positive and negative pseudo-labels for iterative metric update of each other's view. Third, the aim of co-training is to reach an agreement between two views just as increasing consensual pseudo-labels from both views. In that case, top- k neighbours of consensual pseudo-labels on unlabelled samples set O_U may be also useful for metric update, and they could be regarded as special references. Therefore, we attempt to learn a generic model $f(\mathbf{M}_1)$ that updates metric learning model by discovering both pseudo-labels and references.

Assume that a metric model $\mathbf{M}_1$ is learned in view V^1 on labelled samples set O_L and unlabelled training samples $X = \{x_1, \dots, x_i, \dots, x_N\}, 1 \leq i \leq N$ on O_U . x_i^+, x_i^- are used to define the positive and negative pseudo-labels of x_i from metric model $\mathbf{M}_2$ in view V^2 , $x_i^+ = \{x_{i,1}^+, x_{i,2}^+, \dots, x_{i,n}^+\}$,

FIGURE 3: Some samples of two public datasets. Each column shows two images of the same person from two different cameras. (a) VIPeR dataset. (b) PRID2011 dataset.

$x_i^- = \{x_{i,1}^-, x_{i,2}^-, \dots, x_{i,m}^-\}$. r_i^+ , r_i^- , the positive and negative references of x_i , are indicated as $r_i^+ = \{r_{i,1}^+, r_{i,2}^+, \dots, r_{i,k}^+\}$, $r_i^- = \{r_{i,1}^-, r_{i,2}^-, \dots, r_{i,k}^-\}$. Firstly, $f_1(\mathbf{M}_1)$ is defined to pull the pseudo-positives x_i^+ to x_i as close as possible and meanwhile push the pseudo-negatives x_i^- away from x_i as far as possible, as

$$\arg \min_{\mathbf{M}_1} f_1(\mathbf{M}_1) = \|1 + d_{\mathbf{M}_1}(x_i, x_i^+) - d_{\mathbf{M}_1}(x_i, x_i^-)\|. \quad (9)$$

And then, $f_2(\mathbf{M}_1)$ is to both pull the pseudo-positives x_i^+ to referential-positives r_i^+ and pull the pseudo-negatives x_i^- to referential-negatives r_i^- close enough as

$$\arg \min_{\mathbf{M}_1} f_2(\mathbf{M}_1) = \|d_{\mathbf{M}_1}(x_i^+, r_i^+) + d_{\mathbf{M}_1}(x_i^-, r_i^-)\|. \quad (10)$$

Finally, metric update becomes an optimizing problem with the following objective function:

$$\arg \min_{\mathbf{M}_1} f(\mathbf{M}_1) = f_1(\mathbf{M}_1) + f_2(\mathbf{M}_1). \quad (11)$$

Gradient descent algorithm is adopted to optimize (11), and learning procedure of metric model $\mathbf{M}_2$ is similar to that of $\mathbf{M}_1$. The final $\mathbf{M}_1$, or $\mathbf{M}_2$, or combination after iterative update can be utilized for test dataset.

4. Experimental Results

In this section, the proposed method is validated by comparing with state-of-the-art person re-identification approaches

on three publicly available datasets: the VIPeR dataset [35], PRID2011 dataset [43], and PRID450s dataset [44]. The widely used VIPeR dataset contains 632 person image pairs obtained from two different cameras. Some example images are shown in Figure 3(a). All images of individuals are normalized to a size of 128×48 pixels. View changes are the most significant cause of appearance change with most of the matched image pairs containing a viewpoint change of 90 degrees. Other variations are also considered, such as illumination conditions and the image qualities. The PRID2011 is a challenge dataset from two surveillance cameras; particularly there is serious camera characteristics variation as shown in Figure 3(b). In particular, 385 persons' images are from one camera and 749 persons' images are from the other camera, with 200 common images in both views. All images are normalized to 128×48 pixels. The PRID450S is an extension of PRID2011; it has significant and consistent lighting changes and chromatic variation, and there are 450 single-shot image pairs captured over two spatially disjoint camera views. All images are normalized to 168 × 80 pixels.

4.1. Implementation Details. Both hand-crafted and deeply learned features are adopted as the original single-view representations in this paper. Hand-crafted feature employs salient color name [42], and deeply learned feature is produced by a typical Siamese convolutional neural network [45]. All the quantitative results are exhibited in standard Cumulated Matching Characteristics (CMC) curves [9], which are plots of the recognition performance versus the rank score and represent the expectation of finding the correct match inside

TABLE 1: CMC values (%) at top ranks on VIPeR dataset. Best results are shown in bold text.

	Method	Rank@1	5	10
Semi/unsupervised	Ours	32.9	60.3	73
	SDALF [10]	19.9	40	49.4
	eSDC [13]	26.7	50.7	62.4
	TSR [36]	27.7	55.3	68.3
	SSCDL [37]	25.6	53.7	68.2
	Null-semi [38]	31.7	59.4	72.8
Fully supervised	KISSME [24]	19.6	-	62.2
	kLDFA [39]	32.3	65.8	79.7
	DeepNN [15]	34.8	63.6	75.6
	XQDA [40]	40	68.1	80.5
	Null [38]	42.3	71.5	82.9

TABLE 2: CMC values (%) at top ranks on PRID2011. Best results are shown in bold text.

Method	Rank@1	5	10
Ours	25	47.5	58
kCCA [41]	5.8	16	24.7
kLFDA [39]	12	27.1	37.8
XQDA [40]	12.6	29.4	40.2
Null-semi [38]	24.7	46.8	58.2

top k matches. Following the evaluation protocol described by state of the art [23], dataset is randomly divided into two parts, a half for training and the other for testing. However, different from fully supervised methods that training data are all labelled, only one-third of labelled data are used in this semisupervised person re-identification evaluation while the remaining training data are unlabelled, similarly to [37]. All images from camera view A are treated as probes and those from camera view B as gallery set. For each probe image, there is one person image matched in the gallery set. With two different methods, we use the same configuration for experiments at each trial to get the ranking lists. To achieve stable statistics, we repeated the evaluation procedure for 10 times.

4.2. Experiments on VIPeR. We compare our co-metric learning (CML) based person re-identification method with ten most published unsupervised, semisupervised, and fully supervised results on the VIPeR dataset. Unsupervised/semisupervised approaches include SDALF [10], eSDC [13], TSR [36], SSCDL [37], Null-semi [38], and fully supervised baselines including KISSME [24], kLDFA [39], DeepNN [15], Null Space [38], and XQDA [40]. Semisupervised person re-identification usually assumes the availability of one-third of the training set, while the whole training set of fully supervised approaches is labelled and adopted in learning procedure. To show the quantized comparison results more clearly, we summarize the performance comparison (see Table 1). As can be seen, we make the following observations: (1) our method achieves 32.9% at rank@1 matching rate, which improves the previous best results over 1.1%, and matching rates at rank@5 and rank@10 also possess the highest performance compared with all unsupervised/semisupervised

results. (2) Compared with fully supervised baselines, our result is also competitive, especially at rank@1; e.g., the performances of KISSME and kLDFA are both lower than that of our CML. (3) Although there is still long way compared with best fully supervised result, our approaches only need one-third labelled training data, which is more suitable for label-insufficient practical environment.

4.3. Experiments on PRID2011. Compared with VIPeR dataset, the number of person images on PRID2011 is small, where training sample size may be much smaller than feature dimension; i.e., SSS problem can be worse. We compare the state-of-the-art semisupervised baselines kCCA [41], kLFDA [39], XQDA [40], and Null-semi [38] on PRID2011 with access to the implementation codes using the same LOMO features. It can be seen that (see Table 2) (1) except result at rank@10, rank@1, and rank@5, matching rate of our method is the best result compared with baselines, and there is only 0.2% margin below Null-semi that takes the best performance at rank@10. (2) Influenced by small sample size, our approach and baselines all yield much poorer results on PRID2011 dataset compared with results on VIPeR dataset.

4.4. Experiments on PRID450s. Many published unsupervised/semisupervised SDALF [10], eSDC [13], and TSR [36] and fully supervised KISSME [24] and SCNCD [42] are introduced as baselines on PRID450s. The performance of our method is much better than all unsupervised/semisupervised comparisons (see Table 3). It achieves 61.8% at rank@5 and 73.8% at rank@10, which improves the previous best results over 10%. Moreover, for verifying the label sensitivity of our CML method, we test SCNCD with metric learning and our method with 1/2, 1/3, 1/5 labelled training samples

TABLE 3: CMC values (%) at top ranks on PRID450s. Best results are shown in bold text.

	Method	Rank@1	5	10
Semi/unsupervised	Ours	32.9	61.8	73.8
	SDALF [10]	17.4	30.9	40.8
	eSDC [13]	25.5	40.6	48.4
	TSR [36]	29.0	49.4	58.4
Fully supervised	KISSME [24]	26.5	47.8	57.6
	SCNCD [42]	41.5	66.6	75.9

TABLE 4: CMC values (%) at different label-sizes on PRID450S.

Method	Rank@1	5	10
Ours_1/2	38.2	67.1	76
Ours_1/3	32.9	61.8	73.8
Ours_1/5	25.3	45.8	57.8
SCNCD_1/2	34.7	64.2	74.7
SCNCD_1/3	27.6	53.3	65.3
SCNCD_1/5	0	1	2

(see Table 4). And our results exceed significantly that of SCNCD at every labelled training size and decrease gently along with lower training samples; however SCNCD declines sharply especially with 1/5 training samples. That is because the within-class scatter matrix of traditional metric learning becomes singular, when the number of labels is smaller than the dimension of feature representation. Relatively speaking, our method combines labelled and unlabelled data for learning procedure, which is more robust and less sensitive about label size.

5. Conclusions

This paper proposes a novel semisupervised co-metric learning framework for label-insufficient person re-identification. To bridge the small sample size problem and learning model with small labels, motivated by co-training that is commonly used for insufficient/imperfect-label learning, we adopt binary-weight learning to decompose single-view person features into pseudo two views, which could be used to learn two metric models as a co-training style, and then metrics are jointly updated by discovering both pseudo-labels and references. Experiments on three representative person re-identification datasets show that proposed method performs better than state of the art with small labelled sample size and possesses low label sensitivity.

Data Availability

The three public datasets utilized in this work are freely acquired online, and readers can easily find the download links of datasets via searching references [35, 43, 44] on the Internet.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation Projects of China (no. 61562048 and no. 61562047).

References

- [1] S. Sunderrajan and B. S. Manjunath, "Context-aware hypergraph modeling for re-identification and summarization," *IEEE Transactions on Multimedia*, vol. 18, no. 1, pp. 51–63, 2016.
- [2] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 701–713, 2007.
- [3] L. Lo Presti, S. Sclaroff, and M. L. Cascia, "Path modeling and retrieval in distributed video surveillance databases," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 346–360, 2012.
- [4] M. Ye, "Specific person retrieval via incomplete text description," in *Proceedings of the 5th ACM International Conference on Multimedia Retrieval*, pp. 547–550, 2015.
- [5] K. Hariharakrishnan and D. Schonfeld, "Fast object tracking using adaptive block matching," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 853–859, 2005.
- [6] K.-W. Chen, C.-C. Lai, P.-J. Lee, C.-S. Chen, and Y.-P. Hung, "Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 625–638, 2011.
- [7] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 1, pp. 51–61, 2015.
- [8] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1528–1535, June 2006.
- [9] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proceedings of the IEEE*

- 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.
- [10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2360–2367, IEEE, San Francisco, Calif, USA, June 2010.
 - [11] I. Kviatkovsky, A. Adam, and E. Rivlin, “Color invariants for person reidentification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.
 - [12] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Proceedings of the European Conference on Computer Vision (ECCV '08)*, pp. 262–275, 2008.
 - [13] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3586–3593, IEEE, Portland, Ore, USA, June 2013.
 - [14] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: deep filter pairing neural network for person re-identification,” in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 152–159, June 2014.
 - [15] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 3908–3916, June 2015.
 - [16] M. Geng, Y. Wang, T. Xiang, and Y. Tian, “Deep transfer learning for person re-identification,” *Computer Science—Computer Vision and Pattern Recognition*, 2016.
 - [17] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*, pp. 1249–1258, Las Vegas, NV, USA, June 2016.
 - [18] L. Zheng, Z. Bie, Y. Sun et al., “MARS: a video benchmark for large-scale person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '16)*, 2016.
 - [19] D. Yi, Z. Lei, and S. Z. Li, “Deep metric learning for practical person re-identification,” in *IEEE International Conference on Pattern Recognition (ICPR '14)*, 2014.
 - [20] H. Shi, Y. Yang, X. Zhu et al., “Embedding deep metric for person re-identification: a study against large variations,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '16)*, vol. 9905, 2016.
 - [21] M. Hirzer, C. Beleznaï, M. Kostinger, P. M. Roth, and H. Bischof, “Dense appearance modeling and efficient learning of camera transitions for person re-identification,” in *Proceedings of the 19th IEEE International Conference on Image Processing (ICIP '12)*, pp. 1617–1620, Orlando, FL, USA, September 2012.
 - [22] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, “Pedestrian recognition with a learned metric,” in *Proceedings of the Asian Conference on Computer Vision*, vol. 6495, pp. 501–512, 2010.
 - [23] W. S. Zheng, S. Gong, and T. Xiang, “Person re-identification by probabilistic relative distance comparison,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 649–656, June 2011.
 - [24] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2288–2295, IEEE, Providence, RI, USA, June 2012.
 - [25] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, “Person re-identification by regularized smoothing kiss metric learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.
 - [26] A. Mignon and F. Jurie, “PCCA: a new approach for distance learning from sparse pairwise constraints,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2666–2672, June 2012.
 - [27] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, “Local fisher discriminant analysis for pedestrian re-identification,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3318–3325, June 2013.
 - [28] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, “Learning locally-adaptive decision functions for person verification,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3610–3617, June 2013.
 - [29] Y. Wang, R. HU, C. Liang, C. Zhang, and Q. Leng, “Camera compensation using feature projection matrix for person re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1350–1361, 2014.
 - [30] Z. Wang, R. Hu, C. Liang et al., “Zero-shot person re-identification via cross-view consistency,” *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.
 - [31] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, “A new LDA-based face recognition system which can solve the small sample size problem,” *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.
 - [32] B. Han, I. W. Tsang, L. Chen, C. P. Yu, and S. Fung, “Progressive stochastic learning for noisy labels,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2018.
 - [33] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, “Multi-modal curriculum learning for semi-supervised image classification,” *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3249–3260, 2016.
 - [34] M. F. Balcan, A. Blum, and K. Yang, “Co-training and expansion: towards bridging theory and practice,” *Advances in Neural Information Processing Systems*, vol. 17, pp. 89–96, 2004.
 - [35] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS '07)*, vol. 3, 2007.
 - [36] Z. Shi, T. M. Hospedales, and T. Xiang, “Transferring a semantic representation for person re-identification and search,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 4184–4193, June 2015.
 - [37] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, “Semi-supervised coupled dictionary learning for person re-identification,” in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 3550–3557, June 2014.
 - [38] L. Zhang, T. Xiang, and S. Gong, “Learning a discriminative null space for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*, pp. 1239–1248, Las Vegas, NV, USA, June 2016.
 - [39] F. Xiong, M. Gou, O. Camps, and M. Sznajder, “Person re-identification using kernel-based metric learning methods,” in *Proceedings of the European Conference on Computer Vision (ECCV '14)*, vol. 8695, pp. 1–16, 2014.
 - [40] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,”

- in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 2197–2206, June 2015.
- [41] G. Lisanti, I. Masi, and A. Del Bimbo, “Matching people across camera views using kernel canonical correlation analysis,” in *Proceedings of the 8th ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '14)*, November 2014.
 - [42] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, “Salient color names for person re-identification,” in *Proceedings of the European Conference on Computer Vision (ECCV '14)*, pp. 536–551, 2014.
 - [43] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Proceedings of the Seventh Scandinavian Conference of Image Analysis*, pp. 91–102, 2011.
 - [44] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, “Mahalanobis distance learning for person re-identification,” *Advances in Computer Vision and Pattern Recognition*, vol. 56, pp. 247–267, 2014.
 - [45] J. Bromley, J. W. Bentz, L. Bottou et al., “Signature verification using a siamese time delay neural network,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 669–688, 1993.