Consistent Iterative Multi-view Transfer Learning for Person Re-identification

Cairong Zhao[1*], Xuekuan Wang[1], Yipeng Chen[1], Can Gao[2], Wangmeng Zuo[3], Duoqian Miao[1]

[1]Department of Computer Science and Technology, Tongji University, Shanghai, China
[2]Institute of Textiles and Clothing, The Hong Kong Polytechnic University, China
[3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
zhaocairong@tongji.edu.cn

Abstract

*Inconsistent data distributions among multiple views is one of the most crucial aspects of person re-identification. To solve the problem, this paper presents a novel strategy called consistent iterative multi-view transfer learning model. The proposed model captures seven groups of multi-view visual words (**MvVW**) through an unsupervised cluster method (**K-means**) from human body. For each group of **MvVW**, a multi-view discriminative common subspace can be obtained by the fusion of transfer learning and discriminative analysis. In these common subspaces, the original samples can be reconstructed based on **MvVW** under the low-rank and sparse constraints. Then, we solve it via the inexact augmented Lagrange multiplier method. The proposed strategy is performed on three different challenging person re-identification databases (i.e., VIPeR, CUHK01 and PRID450S), which shows that our model outperforms several state-of-the-art models with improving of 6.36%, 7.7% and 4.0% respectively.*

1. Introduction

The central theme of person re-identification (Re-ID) is to match the same person undergoing a multiple non-overlapping system [14]. This is a challenging problem due to significant variations of human appearance with substantial changes in viewpoint, illumination and pose across camera views (e.g., Figure 1). One approach to address this challenge is to capture robust feature descriptor from human appearance [11, 7, 27]. However, this is not always possible when it comes to the complex and uncontrollable environment or small sample size (***SSS***). Other approaches pay more attention to the metric learning which tries to learn a similarity function or a robust distance to optimize the matching. According to the above two different ways of the treatment, many state-of-the-art methods have been proposed [28, 20, 8, 23, 24, 13, 30].

Figure 1. Examples of person re-identification databases.

Currently, most Re-ID methods based on appearance, focus mainly on the low-level visual features such as color [21] and texture [10]. To improve the performance of Re-ID, a wide variety of fusion methods have been designed, such as Hierarchical Gaussian Descriptor [11], local maximal occurrence representation [7], Structure Learning [17] and Salience Matching [29]. Apart from these methods, deep learning is also a noteworthy method which has exhibited an excellent performance in learning representation of data [22]. Unfortunately, it is extremely difficult to design a feature that is distinct, reliable and invariant to severe changes and misalignment across disjoint views.

Another interesting aspect of Re-ID is metric learning, and typical methods include Relative Distance Comparison (RDC) [30], Local Fisher Discriminant Analysis (LFDA) [13], Kernel-based Method [24], Cross-view Quadratic Discriminant Analysis (XQDA) [7], Dual-regularized KISS (DR-KISS) [20], Discriminative Null Space [28] and Deep Metric Learning [18]. There are still many other kinds of methods which try to slove the problem of Re-ID by ranking methods [12, 2]. Although these metric-based methods outperform the existing Re-ID benchmarks, they are nevertheless limited by some of classical problems, such as the inconsistent distributions for multiple views and small sample size (***SSS***) for model learning.

To address these problems, we propose a novel approach called Consistent Iterative Multi-view Transfer Learning (***CIMvTL***),by which seven groups of multi-view visual words (***MvVW***) can be captured including six groups of local features and one group of global features via an unsupervised cluster method (***K-means***) so that it is feasible to effectively describe the structure of human body. Also, the ***MvVW*** has the ability to integrate the multi-view information. Based on these ***MvVW*** groups, we can then reconstruct the original samples with the assistance of the transformation matrix, reconstruction coefficient matrix and noise matrix. Note that, for the sake of ensuring the consistent distributions of sample data, we utilize transfer learning [16] to obtain a common subspace, denoted as the transformation matrix. Meanwhile, we impose joint low-rank and sparse constraints on the reconstruction matrix and noise matrix in order that more relevant samples from different domains are interlaced, compared to irrelevant samples in these domains [9]. Furthermore, we apply discriminative analysis to restrict the reconstruction coefficient matrix that is defined as the mid-level features in our model. To get the consistent optimal solutions, we combine the discriminative analysis with the mid-level features and transfer learning, and then produce the solutions via the proposed method of consistent iterative multi-view transfer learning (***CIMvTL***) which can maintain the consistency of representation and metric learning [5]. In addition, by employing a simple weighted method, max operator and min operator, we can expand the samples to reduce the influence of the small sample size (***SSS***) problem for Re-ID.

2. Related Work

In the real-world, the data taken from different domains have different feature spaces and different data distribution characteristics [16]. To address the problem of inconsistent distributions, a lot of approaches based on transfer learning have been proposed and been applied for various visual tasks [25].

For person Re-ID systems, one of the essential requirements is to built a robust recognition model which can always work well from one type of scene to another under the challenges of camera viewing angles, posture variation, occlusion change, and so on [23]. Accordingly, the transfer learning methods have been exploited to address the challenges of cross-scenario transfer [1, 23, 19, 31]. In [1], Tamar *et al.* proposed the approach of Implicit Camera Transfer (ICT) to model the binary relation by training a (non-linear) binary classifier with concatenations of pairs of vectors captured from different camera views. Similarly, considering the consistency of cross-view, Wang *et al.* [23] combined the learning of the shared latent subspace and the learning of the corresponding task-specific subspace to get the similarity measurement for each task in cross-scenario transfer person Re-ID. Furthermore, Zheng *et al.* [31] formulated a novel transfer local relative distance comparison (t-LRDC) model to address the open-world person re-identification problem. In addition, Shi *et al.* [19] contributed a new framework to learn a semantic attribute model from the existing fashion databases, and adapted the resulting model to facilitate person Re-Id.

3. Method

3.1. Multi-view Visual Words by K-means

To capture structure information and multi-view information, we propose a novel descriptor called Multi-view Visual words (***MvVW***) using an unsupervised cluster method of ***K-means***. Firstly, we divide a person image (x_i) into six horizontal stripes, in view of the consistency of body-structures in vertical direction. Next, we define each low-level feature histogram as a visual word, and then capture six groups of local visual words from six horizontal stripes and one group of global visual words from the whole person images, as shown in Figure 2. Furthermore, we employ ***K-means*** to fuse the multi-view information and obtain seven-group multi-view visual words (***MvVW***). Note that, a simple weighted method, together with max operator and min operator, is employed to expand the sample data for reducing the influence of the small sample size (***SSS***) problem.

In what follows, define ***MvVW*** as $MvVW = \{D_i\}$, where D_i represents the i-th group of multi-view visual words, $\{D_1, D_2, ..., D_6\}$ are local multi-view visual words obtained from six horizontal stripes for person images and D_7 is global multi-view visual words obtained from the whole person images. Then, we use each group of ***MvVW*** to reconstruct the corresponding region of multi-view person sample data X. It is noticeable that the head of a human body is most probably represented by the other heads with similar structures. We can therefore formulate the reconstruction problem as:

$$X = DZ \tag{1}$$

where Z is the reconstruction coefficient matrix and can be captured from the low-level features, denoted as the mid-level features for person Re-ID.

3.2. Consistent Iterative Multi-view Transfer Learning

In the proposed method, we assume that the original samples can be linearly represented by ***MvVW*** in a common subspace. According to [16], we can reconstruct the original samples (X) using the coefficient matrix Z and transfer learning (ensuring the consistency of distributions), rewriting Eq.(1) as:

$$P^T X = P^T DZ \tag{2}$$

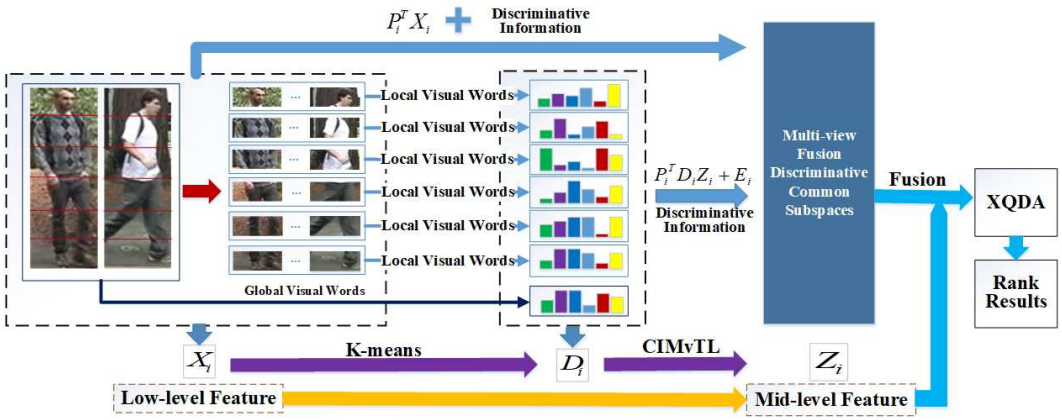


Figure 2. The framework of our proposed method (MvVW+CIMvTL)

where P denotes the transfer matrix, which can be used to obtain a common subspace and can minimize the divergence between the distributions of both domains. However, due to the fact that n samples belong to c different classes and $n \gg c$, these samples should be drawn from c different subspaces, and therefore, the coefficient matrix Z is expected to be low rank [25]. Plus the sparse constraint can be utilized to preserve the local structure of data such that each source sample can be well reconstructed by a few ***MvVW***. Therefore, Eq.(2) can be further written as

$$\min_{P,Z} rank(Z) + \alpha\|Z\|_F^2, s.t. P^T X = P^T DZ \quad (3)$$

where $\| \bullet \|_F$ is the Frobenius norm, $rank(\bullet)$ is a nonconvex function, and α is the penalty parameter. In order to alleviate the influence of noise, we use the matrix E within the sparse constraint to model the noise and replace Eq.(3) with the following

$$\begin{aligned} &\min_{P,Z,E} rank(Z) + \alpha\|Z\|_F^2 + \beta\|E\|_1, \\ &s.t. P^T X = P^T DZ + E \end{aligned} \quad (4)$$

We select nuclear norm to substitute the rank function [25], changing Eq.(4) into

$$\begin{aligned} &\min_{P,Z,E} \|Z\|_* + \alpha\|Z\|_F^2 + \beta\|E\|_1, \\ &s.t. P^T X = P^T DZ + E \end{aligned} \quad (5)$$

where $\|Z\|_*$ is the nuclear norm of matrix Z.

As for the label information, we design a discriminant metric learning function $\phi(Z_\hbar)$ based on the idea of Fisher criterion [4], that is $\phi(Z_\hbar) = Tr(S_B(Z_\hbar)) - Tr(S_w(Z_\hbar))$, where $Z_\hbar \in Z$ is the reconstruction coefficient matrix of label samples and $Tr(Z_\hbar)$ is the trace of matrix $Z_\hbar$. $S_B(Z_\hbar)$ and $S_W(Z_\hbar)$ are the between-class and within-class scatter matrices related to the label samples $X_\hbar$, defined respectively by $S_B(Z_\hbar) = \sum_{i=1}^{c} n_i(Z_{m_i} - Z_m)(Z_{m_i} - Z_m)^T$ and $S_W(Z_\hbar) = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(\widetilde{z}_{\hbar i,j} - Z_{m_i})(\widetilde{z}_{\hbar i,j} - Z_{m_i})^T$, where Z_{m_i} is the mean sample of the i-th person in $Z_\hbar$, Z_m is the overall mean sample of $Z_\hbar$, $\widetilde{z}_{\hbar i,j}$ is the j-th sample in the i-th person of $Z_\hbar$, and n_i is the number of samples in i-th person. To ensure consistency in model learning [5], the discriminant metric learning can be combined with the representation transfer learning, encoding Eq.(5) as

$$\begin{aligned} &\min_{P,Z,E} \lambda\phi(Z_\hbar) + \|Z\|_* + \alpha\|Z\|_F^2 + \beta\|E\|_1, \\ &s.t. P^T X = P^T DZ + E \end{aligned} \quad (6)$$

Notice that the first term $\phi(Z_\hbar)$ in Eq.(6) is not convex to $Z_\hbar$ [6], so we add an elastic term to ensure the convexity of $Z_\hbar$, which can be defined as

$$\begin{aligned} \phi(Z_\hbar) &= Tr(S_B(Z_\hbar)) - Tr(S_w(Z_\hbar)) + \eta\|Z_\hbar\|_F^2 \\ &= \|Z_\hbar(I - H_b)\|_F^2 - \|Z_\hbar(H_b - H_t)\|_F^2 \\ &+ \eta\|Z_\hbar\|_F^2 \end{aligned} \quad (7)$$

where η is a trade-off parameter, I is an identity matrix in $R^{p\times p}$, p is the number of different person, H_b and H_t are two constant coefficient matrices. In detail, $H_b(i,j) = 1/n_c$, where n_c is the number of samples in each class, and $H_t(i,j) = 0$, only if x_i and x_j belong to the same person, otherwise $H_b(i,j) = 0$ and $H_t(i,j) = 1/p$.

In addition, an orthogonal constraint $P^T P = I_p$ is incorporated into our framework, where $I_p \in R^{d\times d}$ is an identity matrix.

As the last point of this part, with the combination of Eq.(6) and Eq.(7), we can obtain the objective function:

$$\begin{aligned} &\min_{P,Z,E} \alpha\phi(Z_\hbar) + \|Z\|_* + \beta\|E\|_1, \\ &s.t. P^T X = P^T DZ + E, P^T P = I_p, Z_\hbar \in Z \end{aligned} \quad (8)$$

where the term $\alpha\phi(Z_\hbar)$ is formed by merging the optimization terms of $\lambda\phi(Z_\hbar)$ and $\alpha\|Z\|_F^2$.

3.3. Optimization

In the light of the non-convexity of Eq.(8), we adopt the inexact ALM (IALM) algorithm [25] to solve this optimization problem. IALM algorithm is an iterative method that can solve each variable in a coordinate descent manner. First, we introduce a variable Z_1 and impose a constraint on Z, i.e., $Z = Z_1$, to relax the original problem, yielding

$$\begin{aligned} &arg \min_{P,Z,Z_\hbar,Z_1,E} \alpha\phi(Z_\hbar) + \|Z_1\|_* + \beta\|E\|_1 \\ &s.t. P^T X = P^T DZ + E, \\ &P^T P = I_p, Z = Z_1, Z_\hbar = Z_* \end{aligned} \tag{9}$$

where Z_* is the reconstruction coefficient matrix of label sample in Z. More specifically, Eq.(9) can further be converted into the following problem:

$$\begin{aligned} &arg \min_{P,Z,Z_1,Z_\hbar,E,\mathcal{L}_1,\mathcal{L}_2,\mathcal{L}_3} \alpha\phi(Z_\hbar) + \|Z_1\|_* \\ &+ \beta\|E\|_1 + \langle \mathcal{L}_1, P^T X - P^T DZ - E\rangle \\ &+ \langle \mathcal{L}_2, Z_* - Z_\hbar\rangle + \langle \mathcal{L}_3, Z - Z_1\rangle \\ &+ \frac{\mu}{2}(\|P^T X - P^T DZ - E\|_F^2 \\ &+ \|Z - Z_\hbar\|_F^2 + \|Z - Z_1\|_F^2) + \gamma\|P^T P - I_p\|_1 \end{aligned} \tag{10}$$

where $\mu > 0$ and $\gamma > 0$ are penalty parameters. $\mathcal{L}_1 \in R^{m\times n}$, $\mathcal{L}_2 \in R^{m\times p}$ and $\mathcal{L}_3 \in R^{m\times n}$ are Lagrange multipliers. The main steps of solving Eq.(10) are given as follows and all steps have closed form solutions.

Step 1 (Update P): P can be updated by solving optimization problem of Eq.(11).

$$\begin{aligned} arg \min_P \frac{\mu}{2}\|P^T X - P^T DZ - E + \frac{\mathcal{L}_1}{\mu}\|_F^2 \\ + \gamma\|P^T P - I_p\|_1 \end{aligned} \tag{11}$$

Then, we can obtain the closed-form solution of Eq.(12)

$$P^* = (\mu G_1 G_1^T + 2\gamma I)^{-1}(\mu G_1 G_2^T) \tag{12}$$

where $G_1 = X - DZ$ and $G_2 = E - \frac{\mathcal{L}_1}{\mu}$.

Step 2 (Update Z): Z is updated by solving optimization problem of Eq.(13).

$$\begin{aligned} &arg \min_Z \|P^T X - P^T DZ - E + \frac{\mathcal{L}_1}{\mu}\|_F^2 \\ &+ \|Z_* - Z_\hbar + \frac{\mathcal{L}_2}{\mu}\|_F^2 + \|Z - Z_1 + \frac{\mathcal{L}_3}{\mu}\|_F^2 \end{aligned} \tag{13}$$

Then, we can obtain the closed-form solution of Eq.(13)

$$Z^* = (\mu D^T P P^T D + 2\mu I)^{-1}(G_4 + G_5 - D^T P G_3) \tag{14}$$

where $G_3 = P^T X - E + \frac{\mathcal{L}_1}{\mu}$, $G_4 = \psi(Z_\hbar - \frac{\mathcal{L}_2}{\mu})$ and $G_5 = Z_1 - \frac{\mathcal{L}_3}{\mu}$. ψ represents the transformation function and $Z_\hbar$ serves as the reconstruction coefficient matrix of labeled samples in Z to ensure the consistent dimensionality.

Step 3 (Update $Z_\hbar$): $Z_\hbar$ is updated by solving optimization problem of Eq.(15)

$$\begin{aligned} &arg \min_{Z_\hbar} \alpha(\|Z_\hbar(I - H_b)\|_F^2 - \|Z_\hbar(H_b - H_t)\|_F^2 \\ &+ \eta\|Z_\hbar\|_F^2) + \frac{\mu}{2}\|Z_\hbar - (Z_* + \frac{\mathcal{L}_2}{\mu})\|_F^2 \end{aligned} \tag{15}$$

Then, we can obtain the closed-form solution of Eq.(15)

$$Z_\hbar^* = (\mu G_8)(2\alpha G_6 G_6^T - 2\alpha G_7 G_7^T + (2\alpha\eta + \mu)I)^{-1} \tag{16}$$

where $G_6 = I - H_b$, $G_7 = H_b - H_t$ and $G_8 = Z_* + \frac{\mathcal{L}_2}{\mu}$

Step 4 (Update Z_1): Z_1 is updated by solving optimization problem of Eq.(17)

$$arg \min_{Z_1} \|Z_1\|_* + \frac{\mu}{2}\|Z - Z_1 + \frac{\mathcal{L}_3}{\mu}\|_F^2 \tag{17}$$

The closed-form solution of Eq.(17) is

$$Z_1^* = \theta_{\frac{1}{\mu}}(Z + \frac{\mathcal{L}_3}{\mu}) \tag{18}$$

where $\theta_\lambda(X) = US_\lambda(\Sigma)V^T$ is a thresholding operator with respect to a singular value λ; $S_\lambda(\Sigma_{i,j}) = sign(\Sigma_{i,j})max(0, |\Sigma_{i,j} - \lambda|)$ is the soft-thresholding operator. $X = U\Sigma V^T$ is the singular value decomposition of X.

Step 5 (Update E): E is updated by solving optimization problem of Eq.(19)

$$\begin{aligned} &arg \min_E \beta\|E\|_1 + \langle \mathcal{L}_1, P^T X - P^T DZ - E\rangle \\ &+ \frac{\mu}{2}\|P^T X - P^T DZ - E\|_F^2 \end{aligned} \tag{19}$$

According to the shrinkage operator [25], the above problem of Eq.(19) has the following closed form solution

$$E^* = shrink(P^T X - P^T DZ - E + \frac{\mathcal{L}_1}{\mu}, \frac{\beta}{\mu}) \tag{20}$$

where $shrink(x, a) = signmax(|x| - a, 0)$.

Step 6 Multipliers $\mathcal{L}_1, \mathcal{L}_2$ and $\mathcal{L}_3$ and iteration step-size $\rho(\rho > 1)$ are updated using Eq.(21),

$$\begin{cases} \mathcal{L}_1 = \mathcal{L}_1 + \mu(P^T X - P^T DZ - E) \\ \mathcal{L}_2 = \mathcal{L}_2 + \mu(Z_* - Z_\hbar) \\ \mathcal{L}_3 = \mathcal{L}_3 + \mu(Z - Z_1) \\ \mu = min(\rho\mu, \mu_{max}) \end{cases} \tag{21}$$

To close this section, the process of solving Eq.(8) is summarized in Algorithm 1.

Algorithm 1: Solving Problem of Eq.(8) by IALM

Input: $X, D, \lambda, \beta, \eta, \gamma, \rho, \mu, \mu_{max}$.
Initialization: $Z = Z_1 = 0, Z_\hbar = 0, E = 0$,
$\mathcal{L}_1 = 0, \mathcal{L}_2 = 0, \mathcal{L}_3 = 0, \alpha = 0.11, \beta = 1.0, \eta = 1.0$,
$\gamma = 0.05, \mu = 0.8, \mu_{max} = 10^7, \rho = 1.01, \epsilon = 10^{-5}$.
Begin:
While not converged end
Update P by sloving Eq.12, given others fixed.
Update Z by sloving Eq.14, given others fixed.
Update $Z_\hbar$ by sloving Eq.16, given others fixed.
Update Z_1 by sloving Eq.18, given others fixed.
Update E by sloving Eq.20, given others fixed.
Update the multipliers and parameters by sloving Eq.21, given others fixed.
Check the convergence condition:
$\|P^T X - P^T D Z - E\|_\propto < \epsilon, \|P^T P - I_p\|_\propto < \epsilon$
$\|Z - Z_1\|_\propto < \epsilon, \|Z_* - Z_\hbar\|_\propto < \epsilon$.
End while
Output: Z, P, E

3.4. Metric Learning

In our approach, we firstly get the low-level feature of Local Maximal Occurrence Feature($LOMO$) [7] and Hierarchical Gaussian Descriptor (GOG) [11]. Then, we obtain the mid-level feature via the aforementioned method (***MvVW+CIMvTL***), defined respectively as $\widehat{Z}_{LOMO}$ and $\widehat{Z}_{GOG}$, which all include seven reconstruction coefficient matrices. Furthermore, we combine the low-level features ($F_{LOMO} \in R^{d_{LOMO} \times n}, F_{GOG} \in R^{d_{GOG} \times n}$) and the mid-level feature ($\widehat{Z}_{LOMO} \in R^{m \times n}$, $\widehat{Z}_{GOG} \in R^{m \times n}$) to formulate our descriptor. Note that, in order to reduce the dimensionality of our descriptor, we define the new low-level features as $\widehat{F}_{LOMO} = F_{LOMO}^T F_{LOMO} \in R^{n \times n}$ and $\widehat{F}_{GOG} = F_{GOG}^T F_{GOG} \in R^{n \times n}$. Therefore, the ultimate dimensionality of our descriptor is $(2n + 2 \times 7m)$. Finally, we apply the metric learning method of XQDA [7] to the measure of the similarity for person Re-ID.

4. Experiments

We evaluate the proposed method on three benchmark databases: VIPeR [3], CUHK01 [29] and PRID450S [11]. All images are scaled to 128×48 pixels. For these databases, we divide all of the images randomly into half for training and the other half for testing, and repeat this procedure 10 times to get an average performance. Besides, we compare our proposed method (MvVW+CIMvTL) with the state-of-the-art methods, including GOG [11], LSSL [26], LOMO+XQDA [7], Semantic [19], SCNCD [27], kLFDA [24], KISSME [15] and SalMatch [29].

Table 1. The recognition results of our model and other the state-of-the-art methods on VIPeR database at Rank-1,10.

Method	Rank=1	Rank=10	Reference
CIMvTL(Fusion)	**56.04**	**91.01**	**Proposed**
CIMvTL(GOG)	**50**	**89.14**	**Proposed**
CIMvTL(LOMO)	**42.66**	**84.68**	**Proposed**
GOG+LOMO	**45.76**	**87.34**	**Fusion**
GOG+XQDA	49.68	88.67	CVPR2016 [11]
LSSL	47.86	87.63	AAAI2016 [26]
LOMO+XQDA	40	80.51	CVPR2015 [7]
kLFDA	22.17	47.23	ECCV2014 [24]
KISSME	22.53	49.57	Spring2014 [15]

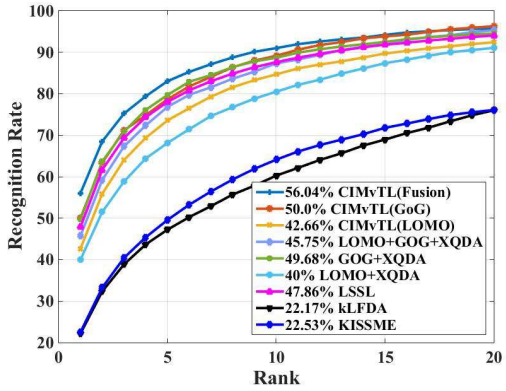


Figure 3. The CMC curves and rank-1 matching rates on the Viper database

4.1. Experiments on the VIPeR Dataset

VIPeR is a challenging person Re-ID database with great variations in background, illumination and viewpoint, containing 632 person image pairs from two cross-view. We choose randomly 316 pairs of images for training and the rest for testing.

4.1.1 Comparison to the State-of-the-Art Methods

We utilize the ***K-means*** method to obtain 7×100 multi-view visual words (***MvVW***) including 6-group local and 1-group global features. The results of Cumulative Matching Characteristic (***CMC***) curves are shown in Figure 3 and Table 1. It can been seen that our proposed method (***CIMvTL***) is obviously better than other state-of-the-art methods, achieving a rate of 56.04%, 91.01% with an improvement of 6.36% and 2.34%, compared with the method of ***GOG+XQDA***, at rank=1,10. From these results, we can see that the consideration of the multi-view information and applying the discriminative transfer learning to a common subspace under consistent contributions, is necessity for person Re-ID.

In addition, we compare the performances of our method with different low-level features (GOG, LOMO and Fusion

of them). As shown in Figure 3, our model achieves rank-1 match rates of 50%, 42.66%, with ***GOG*** and ***LOMO*** respectively, and all of them outperform the original models with improvements of (0.32% and 2.26%). Moreover, by the fusion of these two low-level features (***MvVW+CIMvTL***), we can obtain a superior results with an increase of 10.28%, compared with the method of ***GOG+LOMO+XQDA*** . It further proves our model, capturing the mid-level features, can improve effectively the performance of person Re-ID.

4.1.2 Comparison to the Metric Learning Methods

We evaluate the proposed method of ***MvVW+CIMvTL*** with different metric learning methods, including L_1-Norm distance, kLFDA and XQDA. The resulting Cumulative Matching Characteristic (***CMC***) curves are shown in Figure 4 and Table 2. It can be seen that the proposed method with ***XQDA*** is better than the other metric learning algorithms, with an improvement of 33.51%, compared with ***kLFDA***. This indicates that our model with ***XQDA*** can successfully learn a discriminant transfer subspace as well as an effective metric.

Table 2. The recognition results of our model with different metric methods on the VIPeR database at Rank-1,10,20.

Method	Rank=1	Rank=10	Rank=20
CIMvTL+XQDA	**56.04**	**91.01**	**95.75**
CIMvTL+kLFDA	22.53	49.57	34.49
CIMvTL+L1-Norm	9.18	24.68	60.75

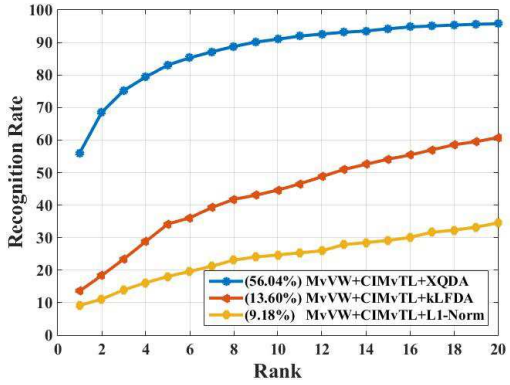


Figure 4. The CMC curves and rank-1 matching rates with different metric methods on the Viper database

4.1.3 Comparison with the Number of Multi-view Visual Words

In this part, we compare the performances with different numbers of multi-view visual words (***MvVW***) obtained by the cluster method of ***K-means***, and the results are shown in Figure 5 and Table 3. It is obvious that our method with the values of (100, 150 and 200) can do better than other models. It can also be observed that ***CIMvTL*** performs consistently best under all numbers of (***MvVW***). Especially, we can obtain the best result of 57.72% at rank-1 with $m = 150$, achieving an increase of recognition rate of 7.94%, compared with the visual words without ***K-means (All-MvVW)***. This indicates the original visual words have more redundant information and the ***MvVW***, fusing multi-view information with the ***K-means***, can achieve an excellent recognition rate. Of course, we should also ensure that the available information is sufficient, so we set $m = 150$ on VIPeR database.

Table 3. The results of comparison with different numbers of Multi-view visual words, ($m = 50, 100, 150, 200, All$).

Method	Rank = 1	Rank = 5	Rank = 10
CIMvTL (50-MvVW)	47.59	78.77	90.03
CIMvTL (100-MvVW)	56.04	**83.04**	**91.01**
CIMvTL (150-MvVW)	57.69	81.23	88.80
CIMvTL (200-MvVW)	**57.72**	80.57	87.85
CIMvTL (ALL-MvVW)	49.78	70.09	78.48

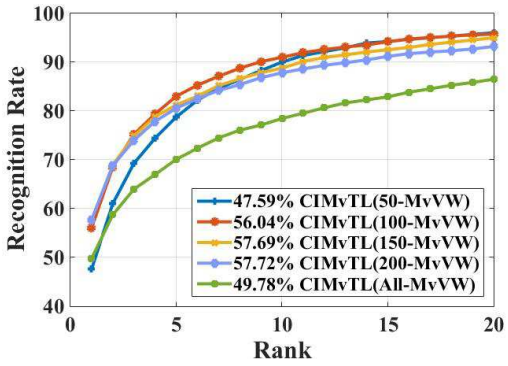
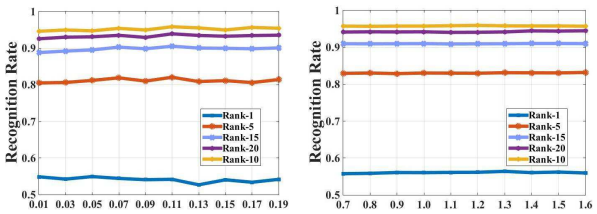


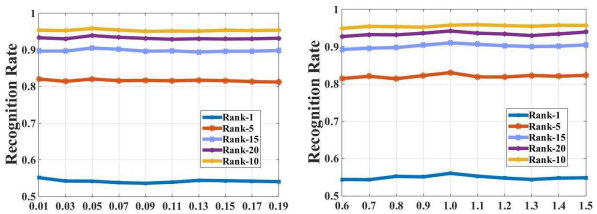
Figure 5. The CMC curves and rank-1 matching rates on the Viper database with $m = 50, 100, 150, 200$ and all

4.1.4 Comparison with Parameters Selection

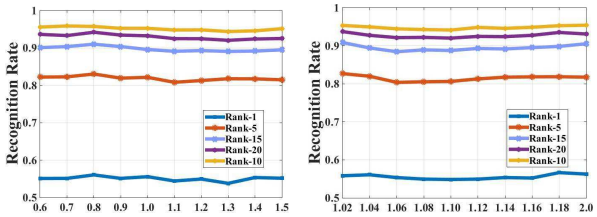
In this experiment, we compare the performances with different parameters and describe the method of parameters selection. In our model, the parameters include mainly $\alpha, \beta, \gamma, \eta$, μ and ρ. We provide the results of our model with different parameters at rank-1,5,10,20 in Figure 6. As we can see in this figure, these parameters are not sensitive, with the best performing on a small change for person Re-ID. In our model, the optimal parameters can be obtained through a method of adjusting one parameter while fixing other parameters, and by setting the values of $\alpha, \beta, \gamma, \eta$, μ and ρ as $0.11, 1.0, 0.05, 1.0, 0.8$ and 1.01 with $m = 100$ on . Note that, if a fast convergence speed is required, we can



(a) α=0.01, 0.02,..., 0.19. (b) β=0.7, 0.8,..., 1.6.



(c) γ=0.01, 0.03,..., 0.19. (d) η=0.6, 0.7,..., 1.5.



(e) μ=0.6, 0.7,..., 1.5. (f) ρ=1.02, 1.04,..., 2.0.

Figure 6. The matching rates of our model with different parameters at Rank-1,5,10,15,20.

Table 4. The recognition results of our model (***CIMvTL***) and other the state-of-the-art models with Rank-1,10 on the CUHK01 database.

Method	Rank = 1	Rank = 10	Reference
MvVW+CIMvTL	**65.5**	**91.6**	Proposed
GOG+XQDA	57.8	86.2	CVPR2016 [11]
LOMO+XQDA	49.2	84.2	CVPR2015 [7]
Semantic	32.7	64.4	ECCV2014 [19]
SalMatch	28.5	55	ICCV2013 [29]

set a large value for μ.

4.2. Experiments on the CUHK01 Database

The CUHK-01 database was captured from two camera views, with higher resolution, containing 971 persons, and each person has two images in each view. We choose randomly 486 pairs of images for training and the rest for testing. And we utilize the K-means method to obtain 7×200 ***MvVW***. The results are described in Table 4. Our method outperforms obviously other methods, achieving the best rank-1 matching rate of 65.5% and 91.6% with an improvement of 7.7% and 5.4%.

4.3. Experiments on the PRID450S Database

The PRID450S dataset contains 450 image pairs recorded from two different static surveillance cameras. In this experiment, we choose randomly 250 pairs of images for training and the rest for testing. And we utilize the K-means method to obtain 7×100 ***MvVW***. The results are reported in Table 5. As can be seen from this table, our proposed method improves the state-of-the-art rank-1,10 matching rates by 4.0% and 0.5%, respectively.

Table 5. The recognition results of our model (***CIMvTL***) and other the state-of-the-art models with Rank-1,10 on the PRID450S database.

Method	Rank=1	Rank=10	Reference
MvVW+CIMvTL	**71.6**	**94.9**	Proposed
GOG+XQDA	67.9	94.4	CVPR2016 [11]
LOMO+XQDA	52.3	84.6	CVPR2015 [7]
SCNCD	41.6	79.4	ECCV2014 [27]
Semantic	43.1	78.2	CVPR2015 [19]

5. Conclusion

In this paper, we have proposed a new model for person Re-ID, based on the integration of the mid-level feature representation and the metric learning. After formulating as a consistent iterative multi-view transfer learning optimal problem, we solved this model using ***IALM***. The obtained solution has been proven to be robust against inconsistent data distributions in terms of viewpoint changes and illumination variations. Meanwhile, we discussed the problem of parameters selection in our model, including $m, \alpha, \beta, \gamma, \eta, \mu$ and ρ, where m is the number of ***MvVW*** obtained by ***K-means***. Experiments on three challenging person Re-ID benchmark databases, VIPeR, CUHK01 and PRID450s, show that the proposed method improves the state-of-the-art rank-1 identification rates by 6.36%, 7.7% and 4.0%, respectively. The future work will try to address optimal choice of the number of ***MvVW*** instead of the use of ***K-means*** which has a high time complexity. In addition, the on-line learning with our model is also a valuable research issue.

Acknowledgement

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the China National Natural Science Foundation under Grant No. 61673299, 61203247, 61573259, 61573255.

References

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *European Conference on Computer Vision*, pages 381–390. Springer, 2012.

[2] S.-Z. Chen, C.-C. Guo, and J.-H. Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

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

[4] Y.-F. Guo, S.-J. Li, J.-Y. Yang, T.-T. Shu, and L.-D. Wu. A generalized foley–sammon transform based on generalized fisher discriminant criterion and its application to face recognition. *Pattern Recognition Letters*, 24(1):147–158, 2003.

[5] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013.

[6] S. Li and Y. Fu. Learning robust and discriminative subspace with low-rank constraints. 2015.

[7] 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*, pages 2197–2206, 2015.

[8] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1629–1642, 2015.

[9] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[10] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun. Orientation driven bag of appearances for person re-identification. *arXiv preprint arXiv:1605.02464*, 2016.

[11] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.

[12] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015.

[13] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[14] E. Poongothai and A. Suruliandi. Survey on colour, texture and shape features for person re-identification. *Indian Journal of Science and Technology*, 9(29), 2016.

[15] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014.

[16] L. Shao, F. Zhu, and X. Li. Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, 26(5):1019–1034, 2015.

[17] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3200–3208, 2015.

[18] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, pages 732–748. Springer, 2016.

[19] 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*, pages 4184–4193, 2015.

[20] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang. Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6):2726–2738, 2016.

[21] R. R. Varior, G. Wang, J. Lu, and T. Liu. Learning invariant color features for person re-identification. 2016.

[22] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang. Deeplist: Learning deep features with adaptive listwise constraint for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

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

[24] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[25] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang. Discriminative transfer subspace learning via low-rank and sparse representation. *IEEE Transactions on Image Processing*, 25(2):850–863, 2016.

[26] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[27] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *European Conference on Computer Vision*, pages 536–551. Springer, 2014.

[28] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. *arXiv preprint arXiv:1603.02139*, 2016.

[29] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535, 2013.

[30] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, 2013.

[31] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):591–606, 2016.