Variation Generalized Feature Learning via Intra-view Variation Adaptation

Jiawei Li[1] , **Mang Ye**[1] , **Andy J Ma**[2] and **Pong C Yuen**[1]

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2] School of Data and Computer Science, Sun Yat-sen University, China
{jwli, mangye, pcyuen}@comp.hkbu.edu.hk, majh8@mail.sysu.edu.cn

Abstract

This paper addresses the variation generalized feature learning problem in unsupervised video-based person re-identification (re-ID). With advanced tracking and detection algorithms, large-scale intra-view positive samples can be easily collected by assuming that the image frames within the tracking sequence belong to the same person. Existing methods either directly use the intra-view positives to model cross-view variations or simply minimize the intra-view variations to capture the invariant component with some discriminative information loss. In this paper, we propose a Variation Generalized Feature Learning (VGFL) method to learn adaptable feature representation with intra-view positives. The proposed method can learn a discriminative re-ID model without any manually annotated cross-view positive sample pairs. It could address the unseen testing variations with a novel variation generalized feature learning algorithm. In addition, an Adaptability-Discriminability (AD) fusion method is introduced to learn adaptable video-level features. Extensive experiments on different datasets demonstrate the effectiveness of the proposed method.

1 Introduction

Person re-identification (re-ID), matching persons across camera views, is playing an important role in large-scale surveillance camera networks. When the person is represented by a video sequence, it is termed as video-based re-ID. Within this field, supervised learning methods [Deng *et al.*, 2018] [Chen *et al.*, 2018] [Ye *et al.*, 2019a] have achieved superior performance with large amount of annotated cross-view video sequences on different benchmarks. With the growth of robust visual tracking and detection algorithms, unlabelled video sequences can be easily obtained. The video sequence provides abundant intra-view positives by assuming that the image frames within the tracking sequence belong to the same person. These characteristics motivate us to investigate an unsupervised solution for video re-ID task, which does not require any cross-view(camera) annotated labels.

Figure 1: Intra-view invariant features (highlighted in red) are less discriminative than the variation parts (highlighted in green) for cross-view re-identification.

A popular approach in unsupervised video re-ID is pseudo-label estimation [Ye *et al.*, 2017; Liu *et al.*, 2017], i.e., iteratively estimate pseudo-labels and refine the model learning process. However, the performance relies heavily on a reliable model for initialization. To achieve a reliable initialized video representation, existing unsupervised feature learning methods usually extract the invariant components to intra-view variations [Khan and Bremond, 2016; Wu *et al.*, 2018]. However, it may lose discriminative information for cross-view re-identification as demonstrated in Fig. 1. Some other methods [Khan and Bremond, 2016] also directly use the intra-view variant representation for cross-view re-identification with the assumption that the intra-view and the cross-view variations are close. However, such an assumption is invalid in practical uncontrolled environments. It raises a issue about how to utilize the intra-view positive samples to simultaneously *address various cross-view variations* and *handle the challenging unseen testing variations*.

Inspired by pose-aware person re-identification, a possible solution is to design a variation-aware framework, in which multiple variation-specific models (model pool) are learned to extract the frame-level features. For precise variation-specific models, a huge number of intra-view variations should be considered and modeled, which is a very time-consuming. In this paper, we propose to learn the adaptable feature representation as shown in Fig. 2, which avoids learning a large model pool. And it can be generalized from the intra-view variation to cross-view and unseen testing variation. Specifically, we propose a novel video variation dictionary learning algorithm for intra-view variation modeling without using any manual labels. The popular variation dictionary learning methods usually assume that the variation-free images are available for all the persons in training stage [Yang *et al.*, 2013]. However,

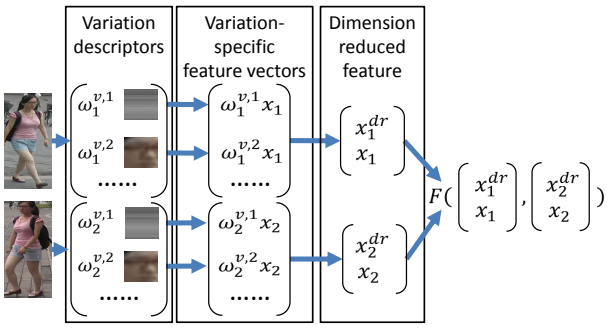

Figure 2: The variation generalized feature learning strategy, which only needs one matching model for all adaptable features.

this assumption is invalid in re-ID task due to unseen testing person identities with various unseen variations. To address this issue, a video variation dictionary is proposed to model the intra-view variation without any variation-free images for training. Secondly, a generalized variation modeling algorithm is proposed to learn frame level adaptable features, which guarantees that the learned model can be generalized to unseen variations. And the video level adaptable features are obtained by an Adaptability-Discriminability (AD) fusion of frame level adaptable features.

In summary, the contributions of this work are as follows.

• A variation generalized feature learning method is proposed by modeling intra-view variations. The learned representation could simultaneously handle the cross-view variations and generalize well on unseen testing variations.

• An Adaptability-Discriminability (AD) fusion method is proposed to generate video level adaptable feature representations using the learned frame-level adaptable features.

2 Related Works

2.1 Unsupervised Person Re-identification

Early unsupervised image based re-ID methods mainly focus on designing hand-crafted descriptors. In these works, the reliable features across frames are extracted to represent a video, i.e. the stable color region [Farenzena *et al.*, 2010], the recurrent structured region [Farenzena *et al.*, 2010], the motion-invariant local body-action features [Liu *et al.*, 2015], and the invariant subspace [Zhang *et al.*, 2016]. Above methods extracting the components invariant to intra-view variations will lose discriminative information as demonstrated in Fig 1.

To extract discriminative information in an unsupervised manner, a dictionary learning algorithm with implicit label mining is proposed [Kodirov *et al.*, 2016]. Along this direction, a label estimation algorithm using reciprocal nearest neighbor search and negative mining is presented to update the model iteratively [Liu *et al.*, 2017]. Incorporate with global matching, Ye *et al.*proposed a dynamic graph matching method to estimate the labels and learn the classification model iteratively [Ye *et al.*, 2017]. Since the label-estimation-based learning methods can extract the discriminative information only when such information is preserved in video level features, the adaptable feature can further improve the performance of label-estimation-based methods.

Some other methods adopt auxiliary labeled data for unsupervised domain adaptation and transfer learning [Lv *et al.*, 2018; Wang *et al.*, 2018b]. Besides, some methods adopt Generative Adversarial Network (GAN) to bridge the gap between source and target domains [Wei *et al.*, 2018] [Deng *et al.*, 2018]. In comparison, we do not need additional labeled data for unsupervised re-ID model learning.

2.2 Variation Dictionary Learning

Variation dictionary learning is a popular intra-class variation modeling approach for face recognition [Deng *et al.*, 2012] [Yang *et al.*, 2013] [Ding *et al.*, 2015]. By introducing an auxiliary intra-class variation dictionary, the samples are represented by the sum of a target-appearance component and an intra-class variation component [Deng *et al.*, 2012]. Since the intra-class variations of training data may not be the same to those of gallery data, the sparse variation dictionary is adaptive to the gallery set by learning a variation-model projection [Yang *et al.*, 2013]. To deal with the pose variation problem, a patch-based transformation dictionary is learned to connect corresponding patches across poses under the multitask learning scheme [Ding *et al.*, 2015]. The existing variation dictionary based methods are mainly designed for image based recognition problem and require a generic set for training, which make they suboptimal or even inapplicable in our video-based re-ID problem.

3 Variation Generalized Feature Learning

This section gives a detailed description of the proposed method. Suppose we have a collection of person image sequences $\{I^a_{i,j}\}$ and $\{I^b_{i,j}\}$, where $I^a_{i,j}$ ($I^b_{i,j}$) refers to jth frame of person i captured by camera a (b). $x^{c_{ID}}_{i,j}$ is a N_d-dimension feature vector extracted from image $I^{c_{ID}}_{i,j}$, $i = 1, 2, ..., N_{c_{ID}}$, $c_{ID} = \{a, b\}$. For simplification, camera ID a and b are skipped when all data are captured by the same camera. Then the feature of image $\{I_{i,j}\}$ is denoted as $\{x_{i,j}\}$, where $i = 1, 2, ..., N$, $j = 1, 2, ..., N_i$.

The proposed framework is shown in Fig. 3. At the training stage, three main blocks, i.e. variation model, structural bottleneck and AD estimator, are learned by minimizing the corresponding three losses. The adaptable component flow (red part) extracts a frame level adaptable feature, while the confidence of adaptable flow mines reliability of the frame level adaptable feature. The outputs of the two flows are then combined to generate video-level adaptable feature. Note that the frame level feature extractor is fixed for simplicity. At the testing stage, the three main blocks are fixed to extract the feature representations of the input image sequence.

The organization of this section is as follows. Section 3.1 introduces the variation modeling method via video variation dictionary learning. Section 3.2 introduces the generalized variation modeling. Section 3.3 introduces the frame level adaptable feature learning method. And the AD fusion for video level feature representation is shown in Section 3.4.

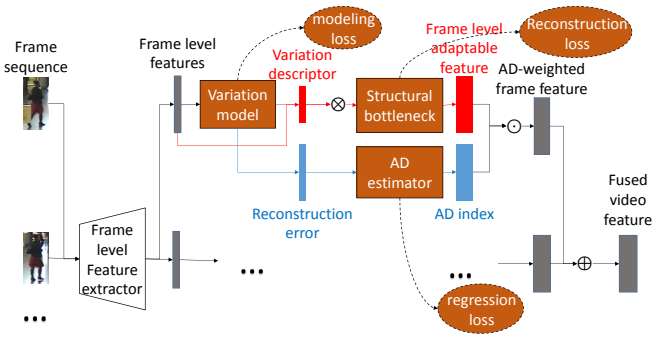

Figure 3: Framework of proposed variation generalized feature learning, where $\otimes$ denotes kronecker product, $\odot$ denotes element-wise Hadamard product, $\oplus$ denotes weighted mean, $\rightarrow$ denotes the (frame level) adaptable component flow and $\rightarrow$ denotes the confidence of adaptable flow. Three training losses in dash line are minimized at training stage.

3.1 Video Variation Dictionary Learning

For the variation dictionary learning, we first introduce the variation component decomposition of frame level feature vectors. Following [Deng *et al.*, 2012], the feature vector of a frame $x_{i,j}$ can be represented by the sum of the person-specific component x_i^p, representation of Variation Dictionary D^v and small dense noise z, i.e.

$$x_{i,j} = x_i^p + D^v \omega_{i,j}^v + z \tag{1}$$

where $\hat{D}^v = \left[\hat{D}_1^v, \hat{D}_2^v, ..., \hat{D}_{N_D}^v\right]$, each column of D^v models one particular variation and $\hat{\omega}_{i,j}^v$ denotes the responses of $x_{i,j}$ to the variations w.r.t the dictionary.

To learn a sparse variation dictionary $\hat{D}^{v*}$, the noise term z is minimized over all frames, i.e.

$$\begin{aligned} &\arg \min_{\hat{\omega}_{i,j}^v, \hat{D}^v} \sum_{i,j} \frac{1}{2} ||x_{i,j} - x_i^p - \hat{D}^v \hat{\omega}_{i,j}^v||_2^2 + \lambda ||\hat{\omega}_{i,j}^v||_1 \\ &s.t.\ ||D_k^v||_2 = 1, \forall k \end{aligned} \tag{2}$$

where $||\cdot||_1$ denotes ℓ_1 norm and λ is the weight of sparsity regularization term.

Since the person-specific component is usually unknown, we also estimate it in (2), i.e.

$$\begin{aligned} &\arg \min_{\hat{\omega}_{i,j}^v, \hat{D}^v, \hat{x}_i^p} \sum_{i,j} \frac{1}{2} ||x_{i,j} - \hat{x}_i^p - \hat{D}^v \hat{\omega}_{i,j}^v||_2^2 + \lambda ||\hat{\omega}_{i,j}^v||_1 \\ &s.t.\ ||D_k^v||_2 = 1, \forall k \end{aligned} \tag{3}$$

The inconsistent components in neighboring frames are considered as unadaptable features, which should be removed to handle the random noise. Following [Wang *et al.*, 2014], we introduce the temporal consistency of variation modelings across frames to remove the inconsistent components, i.e.

$$\begin{aligned} &||\hat{\omega}_{i,j_1}^{v*} - \hat{\omega}_{i,j_2}^{v*}||_2 < \mu_2 = (||\Omega_{i,j_1}^v||_2 + ||\Omega_{i,j_2}^v||_2)/2 \\ &\forall j_1 \in S_{i,m}, j_2 \in S_{i,m}, m = 1, 2, ..., M_i, \forall i \end{aligned} \tag{4}$$

where $S_{i,m}$ denotes the set of frames in the mth view-consistent subsequence. To obtain such subsequences $S_{i,m}$, we divide each sequence by employing an off-line change point detection algorithm [Basseville and Nikiforov, 1993], which can be solved via expectation-maximization (EM).

Combining constraint (4) with (3), we can estimate the person-specific component simultaneously, i.e.

$$\begin{aligned} &\min_{\omega_{i,j}^v, D^v, x_i^p} \lambda \sum_{i,j} ||\omega_{i,j}^v||_1 \\ &\quad + \sum_{i,j} \frac{1}{2} ||x_{i,j} - \left(x_i^p + D^v \omega_{i,j}^v\right)||_2^2 \\ &s.t.\ ||D_k^v||_2 = 1, \forall k \\ &\quad ||\omega_{i,j_1}^v - \omega_{i,j_2}^v||_2 < \mu_2, \forall j_1, j_2 \in S_{i,m}, \forall i \end{aligned} \tag{5}$$

Optimization method to solve (5) can be solved as searching for minimal reconstruction error with fixed sparsity following [Aharon *et al.*, 2006].

3.2 Generalized Variation Modeling

Given the variation modeling in (5), a straightforward method is to represent the video by an "invariant" component $\hat{x}_i^p$, which is also widely used technique in existing methods [Wang *et al.*, 2014] [McLaughlin *et al.*, 2016]. However, the person-specific component $\hat{x}_i^p$ is not "variation-free" feature for the cross-view variations and hence may not be discriminative for recognition across different camera views.

Alternatively, we use a set of aligned features $\left\{x_{i,j}^l | \forall j\right\}$ to represent a video, which is obtained by sum of the person-specific component x_i^p and the variation component represented by the dictionary D^{v*} learned in (5), i.e.

$$\begin{aligned} &x_{i,j}^l = D^{v*} \omega_{i,j}^{v*} + x_i^{p*} \\ &\left\{\omega_{i,j}^{v*}, x_i^{p*}\right\} = \arg \min_{\omega_{i,j}^v, D^v, x_i^p} \lambda \sum_{i,j} ||\omega_{i,j}^v||_1 \\ &\quad + \sum_{i,j} \frac{1}{2} ||x_{i,j} - \left(x_i^p + D^v \omega_{i,j}^v\right)||_2^2 \\ &s.t.\ ||\omega_{i,j_1}^v - \omega_{i,j_2}^v||_2 < \mu_2, \forall j_1, j_2 \in S_{i,m}, \forall i \\ &\quad \left(\omega_{i,j}^v\right)_k \geq 0, \forall k \end{aligned} \tag{6}$$

In (6), the components w.r.t. unseen variations are removed from the aligned feature as the reconstruction error.

To generalize the feature representation to unseen variations, we formulate it as a variation-specific domain generalization problem, where a variation-specific domain means a particular variation. In this manner, the learned model in training domains can be generalized to testing domains with unseen variations. In the following part, we will propose an adaptable matching method, in which the variation modeling is employed to construct the domains.

To deal with the large intra-view variations, we define variation feature sets using aligned feature vectors under the same variation condition, i.e. feature vectors x_{i_1,j_1}^l and x_{i_2,j_2}^l belong to the same variation feature set when $\omega_{i_1,j_1}^{v*} = \omega_{i_2,j_2}^{v*}$. Let $\Omega = \left(\omega_{i,j}^{a,v*}, \omega_{i',j'}^{b,v*}\right)$ denote the variation condition of a image pair $\left(x_{i,j}^{a,l}, x_{i',j'}^{b,l}\right)$, and

$\Delta_\Omega = \left\{ \left(x_{i,j}^{a,l}, x_{i',j'}^{b,l} \right) | \left(\omega_{i,j}^{a,v*}, \omega_{i',j'}^{b,v*} \right) = \Omega \right\}$ denotes a domain w.r.t. Ω. When the domain adaptable model f_Ω is learned, matching score of the pair $\left(x_{i,j}^{a,l}, x_{i',j'}^{b,l} \right)$ is given by $f_\Omega \left(x_{i,j}^{a,l}, x_{i',j'}^{b,l} \right)$. Our goal is to learn domain adaptable models $\{f_\Omega\}$ for different variation domains.

Due to the huge number of potential adaptable matching models, it is impossible to learn all the matching models one by one. To estimate such a large number of matching models efficiently, we assume that similar domains share similar matching models i.e.,

$$\lambda_{\Omega_1,\Omega_2} ||f_{\Omega_1} - f_{\Omega_2}|| < C_f \tag{7}$$

where $f_{\Omega_i}, i = 1, 2$ denotes the matching models on variation domains Δ_{Ω_1} and Δ_{Ω_2}. $\lambda_{\Omega_1,\Omega_2}$ denotes the similarity between domains Δ_{Ω_1} and Δ_{Ω_2}, and C_f is a positive number measuring the cross-domain recognition model dependency. For simplicity, the average correlation between two pairs of variation factors is employed to measure the domain similarity, i.e.

$$\begin{aligned} &\lambda_{\Omega_1,\Omega_2} = \left(\omega_{i_1,j_1}^{a,v*} \right)^T \omega_{i_2,j_2}^{a,v*} + \left(\omega_{i'_1,j'_1}^{b,v*} \right)^T \omega_{i'_2,j'_2}^{b,v*} \\ &s.t.\ \Omega_1 = \left(\omega_{i_1,j_1}^{a,v*}, \omega_{i'_1,j'_1}^{b,v*} \right), \Omega_2 = \left(\omega_{i_2,j_2}^{a,v*}, \omega_{i'_2,j'_2}^{b,v*} \right), \end{aligned} \tag{8}$$

Let $\{e_k\}$ denotes the N_D-dimensional standard basis vector set. According to (7), $f_{e_{n_1}}$ and $f_{e_{n_2}}$ are uncorrelated while all the matching models are correlated to the basis models $\{f_{e_n}\}$. So we use the basis domains $\{\Delta_{e_k}\}$ as source domains and represent matching models for all the other domains using the basis models. A matching model f_Ω^* for domain Δ_Ω can be estimated by minimizing the average matching model differences in (7) to the source domains, i.e.

$$f_\Omega^* = \arg\min_{f_\Omega} \sum \lambda_{\Omega,e_n} ||f_\Omega - f_{e_n}|| \tag{9}$$

According to (9), the adaptable model set $\{f_\Omega\}$ can be estimated when the basis models $\{f_{e_n}\}$ are obtained. Note that the number of basis models depend on the type of variations. When multiple types of variations occur simultaneously, the variation dictionary can contain over 100 atoms [Yang *et al.*, 2013]. So it is still difficult to learn all the basis models especially when multiple types of variations occur. Therefore, optimizing (9) is impractical for complicated environment with various variations.

3.3 Frame Level Adaptable Feature Learning

To address above issue, a frame level adaptable feature learning method will be proposed in the section as an efficient solution for generalized variation modeling rather than multiple matching models under a mild assumption. We transfer the complicated adaptable matching model learning problem to be an adaptable feature learning problem. Let $\hat{s}_{i,j,i',j'}^l$ denote the matching score of a feature vector pair $\left(x_{i,j}^{a,l}, x_{i',j'}^{b,l} \right)$ given by domain adaptable model f_Ω. When we estimate the adaptable feature representation of $x_{i,j}^{a,l}$, $x_{i',j'}^{b,l}$ is skipped for simple denotation of the matching score, i.e. the matching score $\hat{s}_{i,j,i',j'}^l = f_\Omega \left(x_{i,j}^{a,l}, x_{i',j'}^{b,l} \right)$ is simplified as $\hat{s}_{i,j}^l = f_\Omega \left(x_{i,j}^{a,l} \right)$.

When $\{f_\Omega\}$ are linear model and $||f_\Omega - f_{e_n}|| = ||f_\Omega - f_{e_n}||_2^2$, we obtain $f_\Omega = \sum_k \lambda_{\Omega,e_n} f_{e_n} / ||\Omega||_1$ by solving (9). The domain specific matching score $\hat{s}_{i,j,i',j'}^l$ for an aligned feature pair $\left(x_{i,j}^l, x_{i',j'}^l \right)$ is given by the weighted linear combination of the scores from basis models, i.e.

$$\begin{aligned} \hat{s}_{i,j}^l &= \sum_n \left(\omega_{i,j}^v \right)_n f_{e_n}^T x_{i,j}^l / ||\omega||_1 \\ &= \text{Tr} \left(F \cdot \omega_{i,j}^v \left(x_{i,j}^l \right)^T \right) / ||\omega_{i,j}^v||_1 \end{aligned} \tag{10}$$

where $F = \left[f_{e_1}, f_{e_2}, ..., f_{e_{D_n}} \right]$ is the summarized recognition model matrix and Tr $(\cdot)$ denotes the trace of input matrix.

The multiple adaptable classification problem is transferred to be a single domain classification problem on high dimensional feature space $x_{i,j}^l \left(\omega_{i,j}^v \right)^T$ in (10). However, such a high dimensional classifier F is usually noisy and unreliable when the labeled training data are limited. Therefore, we reduce the dimension of feature $\omega_{i,j}^v \left(x_{i,j}^l \right)^T$ to reduce computational cost of recognition model F.

It is reasonable to assume that there exists invariant structure in person re-ID. Therefore, we propose to utilize the similarity between models $\left\{ f_{e_n}^T \right\}$ to improve model learning. So feature $\omega_{i,j}^v \left(x_{i,j}^l \right)^T$ cannot be directly vectorized for dimension reduction until common part between $\left\{ f_{e_n}^T \right\}$ is removed. To maintain the common part after dimension reduction, we decompose the basis model f_{e_n} into a universal component f_0 and a domain specific component $f_{e_n}^\delta$, i.e. $f_n = f_0 + f_n^\delta$. Dimension reduction will be conducted only on the domain specific component. The recognition score $\hat{s}_{i,j}^l$ in (10) can be represented by the sum of score from the universal component and the domain specific components, i.e.

$$\hat{s}_{i,j}^l = f_0^T x_{i,j}^l + \text{Tr} \left(F^\delta \cdot \omega_{i,j}^v \left(x_{i,j}^l \right)^T \right) / ||\omega_{i,j}^v||_1 \tag{11}$$

where $F^\delta = \left[f_{e_1}^\delta, f_{e_2}^\delta, ..., f_{e_{D_n}}^\delta \right]$ is the summarized domain specific model matrix.

According to (11), F is decomposed into a low-dimensional common part f_0 and a domain specific part F^δ. So the dimension of feature $\omega_{i,j}^v \left(x_{i,j}^l \right)^T$ can be reduced after vectorization. Here we employ 1D Principal Component Analysis (PCA) on the vectorization of $x_{i,j}^l \left(\omega_{i,j}^v \right)^T$ for its simplicity.

Let $x_{i,j}^{dr}$ denote the reduced feature of $\omega_{i,j}^v \left(x_{i,j}^l \right)^T$ and F_{dr} denote the counterpart of F_δ, the summarized score $s_{i,j}^l$ can be estimated as follows.

$$s_{i,j}^l = \begin{pmatrix} F_{dr}^T & f_0^T \end{pmatrix} \begin{pmatrix} x_{i,j}^{dr} / ||\omega_{i,j}^v||_1 \\ x_{i,j}^l \end{pmatrix} + \delta_{dr} \tag{12}$$

where δ_{dr} is a small number, denoting the error derived from dimension reduction. From (12), the adaptable feature $x_{i,j}^a$ is given by the concatenation of the weighted reduced feature $x_{i,j}^{dr} / ||\omega_{i,j}^v||_1$ and the aligned feature $x_{i,j}^l$, i.e.

$$\begin{aligned} x_{i,j}^a &= \begin{pmatrix} x_{i,j}^{dr} / ||\omega_{i,j}^v||_1 \\ x_{i,j}^l \end{pmatrix} \\ s.t. \quad x_{i,j}^{dr} &= P_{PCA} \left(\omega_{i,j}^v \otimes x_{i,j}^l \right)^T \end{aligned} \tag{13}$$

where $\otimes$ denotes the Kronecker product and P_{PCA} denotes the projection matrix of PCA.

Note that above adaptable feature estimation method is given under the linear classifier assumption. It can be easily further extended to the non-linear cases by utilizing Taylor series of classification model and introducing higher order statistics in (10).

3.4 Adapatability-Discriminative Fusion

Following weighted sum-rule fusion scheme, we estimate the Adapatability-Discriminative (AD) index and propose a feature level fusion method to combine adaptable frame-level features for video representation. The video based adaptable feature vector s_i^{av} is given by the weighted sum of the frame level adaptable feature vectors $\{x_{i,j}^a\}$, i.e.

$$x_i^{av} = \sum_j \rho_{i,j} x_{i,j}^a \tag{14}$$

where $\rho_{i,j}$ is the AD index to be optimized in the following.

Since the reconstruction error of the dictionary measures the magnitude of unadaptable features and the variation component measures the magnitude of adaptable features, they both indicate the reliability of the adaptable feature. Thus, we introduce the adaptability feature vector $\epsilon_{i,j}$ as the concatenation of the reconstruction error and the variation component, i.e.

$$\epsilon_{i,j} = \left((x_{i,j} - x_i^{p*} - D^{v*}\omega_{i,j}^{v*})^T (D^{v*}\omega_{i,j}^{v*})^T\right) \tag{15}$$

It is reasonable to assume that the elements of the adaptability component may not be equally discriminative. Therefore, we learn discriminative weights to approximate a discriminability measure for the AD index. Inspired by Fisher's linear discriminant, we define the discriminability measure by the ratio $\rho_{i,j}^{d,c_{ID}}$ of standard deviation between false matching images (inter-class) and correct matching images (intra-class) to represent the discriminability, i.e.

$$\rho_{i,j}^{d,c_{ID}} = \frac{N_i^{c'_{ID}} \sum_{i' \neq i, j'} d\left(x_{i,j}^{c_{ID}}, x_{i',j'}^{c'_{ID}}\right)}{\left\{\sum_{i' \neq i} N_{i'}^{c'_{ID}}\right\}\left\{\sum_{j'} d\left(x_{i,j}^{c_{ID}}, x_{i,j'}^{c'_{ID}}\right)\right\}} \tag{16}$$

where $c'_{ID} \neq c_{ID}$ and $d(\cdot,\cdot)$ denotes a distance function between the two input vectors. The false matching and the correct matching image pairs are estimated by a pre-learned matching model [Ye *et al.*, 2017]. To approximate the discriminability measure $\rho_{i,j}^d$, we formulate the learning problem by linear regression on the adaptability component, i.e.

$$\alpha^* = \arg\min_\alpha ||\alpha^T \epsilon_{i,j} - \rho_{i,j}^d||_2^2 + \lambda_\alpha ||\alpha||_1 \tag{17}$$

where λ_α is the weight of regularization term and it is set to be 1 in our experiments.

With the learned weight vector α^*, the AD index is determined as the linear combination of elements in $x_{i,j}^p$ and $\epsilon_{i,j}$ using weight α^* learned in (17), i.e.

$$\rho_{i,j} = \epsilon_{i,j}^T \alpha^* \tag{18}$$

Dataset	iLIDS-VID			PRID-2011		
Rank	R=1	R=5	R=10	R=1	R=5	R=10
FAF+AD	50.33	78.50	87.83	82.02	96.07	98.88
FAF+Min	48.17	76.00	85.17	79.78	97.19	98.88
FAF+Avg	47.50	75.67	85.17	79.21	97.19	98.88
B + Min	47.17	73.83	84.00	76.40	93.82	97.19
B + Avg	44.83	73.00	82.67	79.21	94.94	98.88

Table 1: Top r ranked matching rate (%) with/without Frame level Adaptable Feature learning (FAF) and AD fusion (AD). B: Baseline. Red indicates the best performance while blue for second best.

Dataset	iLIDS-VID			PRID-2011		
Rank	R=1	R=5	R=10	R=1	R=5	R=10
Iter 0	44.83	73.00	82.67	79.21	94.94	98.88
Iter 1	50.67	78.33	87.33	81.46	96.07	98.88
Iter 10	50.33	78.83	87.67	81.46	96.07	98.88
Mean	50.62	78.67	87.75	81.52	96.07	98.88
Min	50.33	78.50	87.83	82.02	96.07	98.88

Table 2: Top r ranked accuracy (%) at different iterations. "Mean" represents the mean accuracy over 10 iterations. "Min" represents the accuracy with the minimum training loss over 10 iterations.

4 Experiment

4.1 Datasets and Settings

Datasets. We evaluate our method on three datasets, i.e., the iLIDS-VID dataset [Wang *et al.*, 2014], the PRID 2011 dataset [Hirzer *et al.*, 2011], the MARS dataset [Zheng *et al.*, 2016]. iLIDS-VID was captured by a multi-camera surveillance camera network at an airport arrival hall. It contains 300 person image sequences under each of the two camera views. The PRID-2011 dataset consists of person image sequences recorded from two static surveillance cameras outdoor. Following [Wang *et al.*, 2014], person video pairs with more than 27 frames are employed in the experiment. MARS dataset is a large-scale dataset captured by six cameras. It contains 20,715 different image sequences of 1261 persons.

Feature extraction. We employ a hand-crafted Local Maximal Occurrence (LOMO) feature [Liao *et al.*, 2015] fo two small datasets (PRID-2011 and iLIDS-VID) and a deep Multiple Granularities Network (MGN) feature [Wang *et al.*, 2018a] for the large-scale MARS dataset for evaluation. The LOMO feature analyzes the horizontal occurrence of local feature, while the MGN feature is extracted from a multi-branch deep network architecture with both global and local feature branches. In our proposed method, pseudo-labels are estimated following [Ye *et al.*, 2017] for training MGN.

Classifier. For the hand-crafted feature, a state-of-the-art unsupervised re-ID learning method, i.e. DGM [Ye *et al.*, 2017], is employed to learn the classification model in the scheme of unsupervised learning. For deep feature, we employ Euclidean distance following [Wang *et al.*, 2018a].

Evaluation protocol. For iLIDS-VID and PRID-2011 datasets, the sequence pairs are randomly separated into half for training and the other half for testing. For MARS dataset,

Dataset	iLIDS-VID			PRID2011			MARS			
Rank	R=1	R=5	R=10	R=1	R=5	R=10	R=1	R=5	R=10	mAP
GRDL [Kodirov *et al.*, 2016]	25.7	49.9	63.2	41.6	76.4	84.6	19.3	33.2	41.6	9.6
UnKISS [Khan and Bremond, 2016]	35.9	63.3	74.9	58.1	81.9	89.6	22.3	37.4	47.2	10.6
SMP [Liu *et al.*, 2017]	41.7	66.3	74.1	80.9	93.3	97.8	23.6	35.8	44.9	10.5
DGM [Ye *et al.*, 2017]	44.8	73.0	82.7	79.2	94.9	98.9	24.6	42.6	50.4	11.8
TAUDL [Li *et al.*, 2018]	26.7	51.3	82.0	49.4	78.7	98.9	43.8	59.9	-	29.1
DGM_+ + IDE [Ye *et al.*, 2019b]	38.6	64.2	74.6	62.7	90.8	96.0	48.1	64.7	71.1	29.2
VGFL	50.3	78.5	87.8	82.0	96.1	98.9	51.7	68.2	75.3	32.6

Table 3: Top r rank matching rate (%) and mAP (%)comparing with state-of-the-art **unsupervised** video-based person re-ID methods on three datasets.

the training and the testing separation follows the protocol suggested in [Zheng *et al.*, 2016]. The results are shown in Cumulated Matching Characteristics (CMC) curves. For evaluation in MARS dataset, mAP (mean average precision) value are also reported following [Zheng *et al.*, 2016]. For stable statistical results, the experiments was repeated 10 times and the mean accuracy is reported.

Implementation. 10 iterations are conducted with $\lambda = 0.5$ in DGM for all three datasets following [Ye *et al.*, 2017]. We set $t_{max,1} = 1$ for efficiency. $\lambda = 1$, $\lambda_\alpha = 1$. The dimension of $x_{i,j}^{dr}$ is set to be the same as $x_{i,j}$. For iLIDS-VID and PRID-2011 datasets, the subsequences $\{S_{i,m}\}$ are estimated follow [Basseville and Nikiforov, 1993]. For MARS, each tracking sequence is considered as a subsequence.

4.2 Self Evaluation

Effectiveness of each component. This part evaluates the effectiveness of frame level adaptable feature representation and the AD fusion, as shown Table 1. To evaluate the frame level adaptable feature (FAF, $x_{i,j}^{dr}$), we compare it with the original feature vectors (Baseline, $x_{i,j}$) with min/mean pooling. We observe that it achieves much better and stable performance than baseline feature representation. In addition, we also compare with the set-based distance by calculating the minimum (Min) and mean (Avg) of the frame-to-frame distances. The consistent improvements on two datasets demonstrate the effectiveness of the proposed AD fusion.

Iterative updating. We also report the performance at different iterations on PRID-2011 and iLIDS-VID datasets in Table 2. We observe that the iterative updating process improves the performance significantly on both datasets. And the performance is converged with less than 10 iterations.

Dictionary size. We also plot the rank-1 accuracy with different dictionary sizes on the PRID-2011 and iLIDS-VID datasets as shown in Fig. 4. According to the experimental results on two datasets, we find that the larger dictionary usually produces higher performance, especially on challenging iLIDS-VID dataset. However, larger dictionary size means higher computational cost in both training and test stage. For the efficiency considerations and avoid over-fitting, the dictionary size is set to 100 in all the experiments.

4.3 Comparison with the State-of-the-arts

Six state-of-the-art unsupervised multi-shot/video based person re-identification methods namely GRDL [Kodirov *et al.*, 2016], UnKISS [Khan and Bremond, 2016], SMP [Liu *et al.*, 2017], DGM [Ye *et al.*, 2017], DGM_+ [Ye *et al.*, 2019b] and TAUDL [Li *et al.*, 2018], are used for comparison. LOMO feature and MGN feature are employed on small-scale and large-scale datasets, respectively. The top-r matching accuracies and mAP on three datasets are shown in Table 3.

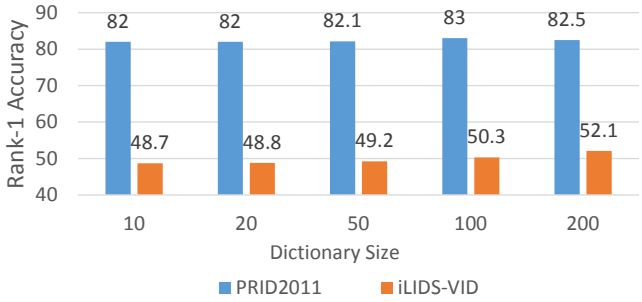

Figure 4: Rank-1 accuracy (%) of the proposed method with different dictionary sizes on PRID-2011 and iLIDS-VID datasets.

For the results on iLIDS-VID dataset, we can see that the proposed method achieves the best rank-1 accuracy and 5.5% improvement is achieved compare to the state-of-the-art methods. Similar trends can be observed in experimental results in PRID-2011 and MARS datasets. Note that the adaptable feature learning and AD fusion can be conducted on more advanced features to achieve better performance.

5 Conclusion

In this paper, we propose a variation generalized feature learning method for unsupervised video based person re-identification. Rather than minimizing the intra-view variations in existing methods, we propose to adapt the intra-view variations to unseen testing variations. Specifically, frame level adaptable feature is learned via domain adaptation and video variation dictionary learning. The video level adaptable feature is then estimated using AD fusion, which achieves much better performance than the widely used mean/min pooling strategy. Experimental results on three public person re-identification datasets show that the proposed method achieve better performance than existing methods.

Acknowledgments

This work is supported by Hong Kong RGC General Research Fund HKBU (12200518).

References

[Aharon *et al.*, 2006] Michal Aharon, Michael Elad, , and Alfred Bruckstein. k -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

[Basseville and Nikiforov, 1993] Michele Basseville and Igor V. Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. 1993.

[Chen *et al.*, 2018] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018.

[Deng *et al.*, 2012] Weihong Deng, Jiani Hu, and Jun Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Pattern Anal. Machine Intell.*, 34(9):1864–1870, 2012.

[Deng *et al.*, 2018] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, pages 994–1003, 2018.

[Ding *et al.*, 2015] Changxing Ding, Xu Chang, and Dacheng Tao. Multi-task pose-invariant face recognition. *IEEE Trans. Image Processing*, 24(3):980–993, 2015.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[Hirzer *et al.*, 2011] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. 2011.

[Khan and Bremond, 2016] Furqan M Khan and Francois Bremond. Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *AVSS*, pages 256–262, 2016.

[Kodirov *et al.*, 2016] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised l1 graph learning. In *ECCV*, pages 178–195, 2016.

[Li *et al.*, 2018] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, pages 737–753, 2018.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Liu *et al.*, 2015] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*, pages 3810–3818, 2015.

[Liu *et al.*, 2017] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, pages 2448–2457, 2017.

[Lv *et al.*, 2018] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *CVPR*, pages 7948–7956, 2018.

[McLaughlin *et al.*, 2016] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016.

[Wang *et al.*, 2014] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014.

[Wang *et al.*, 2018a] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018.

[Wang *et al.*, 2018b] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, pages 2275–2284, 2018.

[Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[Wu *et al.*, 2018] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

[Yang *et al.*, 2013] Meng Yang, Luc Van, and Lei Zhang. Sparse variation dictionary learning for face recognition with a single training sample per person. In *ICCV*, pages 689–696, 2013.

[Ye *et al.*, 2017] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, pages 5142–5150, 2017.

[Ye *et al.*, 2019a] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Trans. Inf. Forensics Secur*, 2019.

[Ye *et al.*, 2019b] Mang Ye, Jiawei Li, Andy J Ma, Liang Zheng, and Pong C. Yuen. Dynamic graph co-matching for unsupervised video-based person re-identification. In *IEEE Trans. Image Processing*, 2019.

[Zhang *et al.*, 2016] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016.

[Zheng *et al.*, 2016] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.