Efficient Video Summarization Using Principal Person Appearance for Video-Based Person Re-Identification

Seongro Yoon
yoonseongro@gmail.com
Furqan M. Khan
furqan.khan@inria.fr
Francois Bremond
francois.bremond@inria.fr

INRIA,
Sophia Antipolis, France

Abstract

In video-based person re-identification, while most work has focused on problems of person signature representation and matching between different cameras, intra-sample variance is also a critical issue to be addressed. There are various factors that cause the intra-sample variance such as detection/tracking inconsistency, motion change and background. However, finding individual solutions for each factor is difficult and complicated. To deal with the problem collectively, we assume that it is more effective to represent a video with signatures based on a few of the most stable and representative features rather than extract from all video frames. In this work, we propose an efficient approach to summarize a video into a few of discriminative features given those challenges. Primarily, our algorithm learns principal person appearance over an entire video sequence, based on low-rank matrix recovery method. We design the optimizer considering temporal continuity of the person appearance as a constraint on the low-rank based manner. In addition, we introduce a simple but efficient method to represent a video as groups of similar frames using recovered principal appearance. Experimental results show that our algorithm combined with conventional matching methods outperforms state-of-the-arts on publicly available datasets.

1 Introduction

Person re-identification (Re-ID) refers to resolving identity of a person in a camera using past observations of persons in other cameras with non-overlapping views. The task is often formulated as an information retrieval problem. Appearance of the concerned person, called *probe*, is matched against examples in a reference set, called *gallery*, to produce a ranked list of matching persons. Re-ID algorithms can be classified as either *single-shot*, where each example in both probe and gallery sets consists of only one image [7, 17, 21, 28, 42], or *multi-shot*, where each example has multiple images or is a video sequence of a person [8, 22, 24, 27, 30, 34, 36]. The objective of our work is to re-identify persons in video based multi-shot scenario, where examples are pedestrian tracks cropped from original videos.

High variance in appearance of a person makes Re-ID a challenging problem due to difficulty in learning discriminative appearance models. In video based scenario, a person's appearance does not only exhibit *inter-sample* variance between his different examples but also *intra-sample* variance between images in one example. If intra-sample variance is not adequately handled before the matching process, resulting errors accrue to further increase inter-sample variance. Consequently, even with a sophisticated matching method, it becomes difficult to handle overall ambiguity. Thus this paper focuses on learning robust appearance models in presence of high intra-sample variance to facilitate video based Re-ID process.

The main contributors to intra-sample variance are irrelevant pixels, detection/tracking errors and motion change. Probe and gallery examples for video based Re-ID are often obtained using an automated pedestrian detection/tracking system. Most conventional Re-ID approaches regard all pixels in each image example as relevant for person's appearance despite existence of a relatively high proportion of irrelevant pixels in each image. Complete removal or irrelevant pixels is non-trivial. One factor is that variance in spatiotemporal occurrence of irrelevant pixels is high and their visual pattern is inconsistent due to changes in background and its location, shading from others, and occlusion. Another factor is the errors induced by person detection/tracking system. Even for state-of-the-art detection/tracking systems, it cannot be guaranteed that the person is always in the same location, particularly in the center, within the target bounds. Articulated motion of pedestrians makes precise alignment difficult in comparison to other objects, such as face. This may lead to variance in some (spatially-aware) appearance models. Likewise, a person may show pose changes within a track or pass through spaces with distinctly different illuminations. Naturally, the factors that affect intra-sample variance also contribute to inter-sample variance. Inevitably, this makes it difficult to effectively learn a person's appearance.

In recent years, some algorithms have attempted to learn multi-modal appearance models to better deal with intra-sample variance [1, 8, 33]. However, their efforts generally setup a hypothesis using one of the causes mentioned above to learn multiple appearance models. Unfortunately in the real-world scenario, since the various factors that cause intra-sample variance appear irregularly and have complex interactions, it is difficult to pre-determine the cause of variance in each example. Additionally, conventional approaches that address intra-sample variance have endeavoured to solve the problem integrally in the representation or the matching processes, but have not shown a remarkable performance.

We believe that the intra-sample variance is still one of the major challenges affecting performance of video based Re-ID. The problem needs to be addressed intensively at early stage in a Re-ID framework. Thus in this paper, we introduce a novel approach to reduce intra-sample variance as a data analysis step before the conventional framework. Our ultimate goal is to summarize a video track to a few of the most representative appearance *signatures*, and use them as input for state-of-the-art matching algorithms.

Briefly, proposed algorithm first learns the most stable and common appearance model, called *Principal Person Appearance* (PPA), for the track of a person. Next, it represents each frame in the track using PPA, and then all frames are grouped into several sets based on their representation. Finally, it constructs an appearance *signature* for each image group. Since the algorithm discovers the principal appearance over all the frames and represents each frame using the model, it can effectively reduce the variance that occurs for certain periods regardless of the cause. The overall scheme for our approach is depicted in Figure 1.

To learn a person's PPA, we apply low-rank matrix recovery method. A number of solutions have been proposed to improve the performance of low-rank matrix recovery due to its use in a wide range of applications. However, existing low-rank optimizers are not

suitable for video, since they concentrate solely on reducing the total amount of sparse error in the data matrix without considering local (temporal) similarity. If it is used for video, it is prone to under-fitting and lose unique textures for the person. To avoid the problem, we design an algorithm that incrementally solves for principal appearance under the assumption that a frame in a track should be most similar to its temporally neighbouring frames.

To summarize main contributions of this paper, first, we suggest an independent process separate from the conventional Re-ID framework to effectively reduce the intra-sample variance for video based person Re-ID. Second, we introduce a novel low-rank optimization technique that can efficiently learn the principal appearance from the input track. Finally, our approach analyzes and classifies frames based on the Principal Person Appearance, so that it can efficiently summarize a video with a few of most representative person features without being affected by different causes of variance. Our algorithm shows remarkable performance on two state-of-the-art datasets for video based person Re-ID.

2 Related Work

To learn principal person appearance, we propose an optimization method based on low-rank matrix recovery. Due to its remarkable performance, low-rank recovery method has lately been applied to different computer vision topics. Jia *et al*. [14] use low-rank recovery method to find correspondence of common features for matching objects. It is further extended for establishing the correspondence of multiple images [37] and for associating images with proper captions [38]. Additionally, the low-rank is used for visual tracking. Zhang *et al*. [39] address a method to seek the common appearance of the target with an assumption that its appearance is almost same at a certain period in a video sequence. Sui *et al*. [31] learn the subspace of a target using tracklets in a certain period, and estimate the most proper region of the target in the next frame based on the subspace. However, the input of these methods is not a video sequence, or it is assumed that the targeted appearance does not change by limiting the application of the low-rank to a specific interval of the video. Low-rank theory is also utilized for person Re-ID [30, 32], but the purpose is not to learn the principal appearance in a video sequence instead to deal with the inter-sample variance problem for single-shot Re-ID. We design a method to use the low-rank property for video, and our optimizer focuses on the similarity of the frame appearance depending on the temporal continuity.

Moreover, solutions using Singular Value Decomposition (SVD) have been proposed to improve accuracy of low-rank recovery [20, 25]. However, reliance on SVD makes them computationally expensive for application such as Re-ID, because of large data matrix. ROSL [29] improves performance for the high-dimensional matrix, by avoiding SVD computation and by approximating the low-rank matrix with orthonormal subspace and its coefficients. We adopt ROSL to solve our Principal Person Appearance recovery problem.

3 Principal Appearance Learning

3.1 Problem Formulation

Given an input image sequence (track) of a person, cropped from video based on the target bound and scaled to a fixed size, we define a corresponding data matrix by concatenating features from images in their temporal order as columns of matrix $\mathbf{D} = [f_1, ..., f_{N_s}] \in \mathbb{R}^{d \times N_s}$,

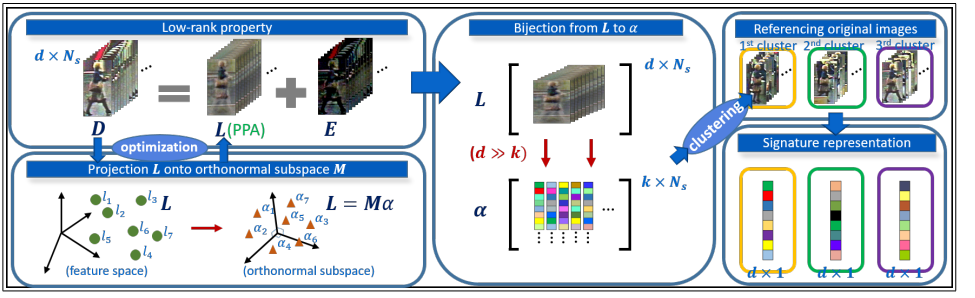

Figure 1: The overall scheme of our algorithm.

where f is a d-dimensional feature vector and N_s is the total number of frames for the s-th track. f can be any feature descriptors representing an image such as color histogram, SIFT [23] or HOG [4]. In our case, we utilize vectorized RGB color image as the feature.

Provided $\mathbf{D}$, our goal is to estimate the principal appearance of a person, which is distinguishable from the image variation generated by various factors. We exploit conventional low-rank recovery method [6] to solve the problem. We assume that the deviation among the vectors that make up the data matrix is caused by irrelevant pixels, detection/tracking errors and motion change, and it can be represented as sparse error. Thus, the data matrix should be rank-deficient. Specifically, $\mathbf{D}$ is equal to $\mathbf{L}+\mathbf{E}$, where $\mathbf{L}$ is a low-rank matrix and $\mathbf{E}$ is the sparse error term. We consider the columns of rank-deficient matrix $\mathbf{L}$ to be the principal appearance of corresponding images. $\mathbf{L}$ can be recovered by solving the following problem:

$$\min_{\mathbf{L},\mathbf{E}} \quad \|\mathbf{L}\|_* + \lambda_E \|\mathbf{E}\|_1, \quad s.t. \quad \mathbf{D} = \mathbf{L} + \mathbf{E}, \tag{1}$$

where $\|\mathbf{L}\|_*$ is the nuclear norm of the matrix $\mathbf{L}$.

Above problem can be solved reliably using [6], provided that $\mathbf{E}$ is *reasonably* sparse. However, in the real-world scenario, certain factors may not allow for a reasonably sparse error $\mathbf{E}$. Precisely, a track may include drastic changes of background, illumination and partial occlusions. In addition, The tracked person is often in motion and shows changes of pose and location. It makes alignment difficult and inaccurate. Thus the error is generally less sparse than in other applications of low-rank recovery methods, which are better aligned. This may lead to the divergence of the optimizer, and eventually reach to undesirable results.

As a solution, we consider temporal continuity of a person's appearance to improve robustness against a diverse range of variance inducing factors. Although different variations appear irregularly throughout a track, the appearance in each frame is similar to its adjacent frames. In particular, the principal appearance (columns of $\mathbf{L}$) of two adjacent frames should have the highest similarity. In order to enforce temporal consistency, we quantify the deviation in adjacent vectors of $\mathbf{L}$ as $\|\mathbf{LP}\|$ and add it to the minimization problem in (1), where the entries of $\mathbf{P} \in \mathbb{R}^{N_s \times (N_s-1)}$ are defined as:

$$p_{i,j} = \begin{cases} -1 & if\ i = j, \\ 1 & if\ i = j+1, \\ 0 & \text{otherwise}, \end{cases} \quad \forall i \in \{1,...,N_s\},\ \forall, j \in \{1,...,N_s-1\}. \tag{2}$$

Introduction of $\|\mathbf{LP}\|$ in (1) makes the problem intractable irrespective of the norm used. Thus, similar to [40], we introduce slack variables and equality constraints to decouple the problem into multiple independent sub-problems. Thus, our problem formulation becomes:

$$\min_{\mathbf{L}_1,\mathbf{L}_2,\mathbf{L}_3,\mathbf{E}} \|\mathbf{L}_1\|_* + \lambda_L \|\mathbf{L}_2\mathbf{P}\|_F^2 + \lambda_E \|\mathbf{E}\|_1, \quad s.t.\ \mathbf{D} = \mathbf{L}_3 + \mathbf{E},\ \mathbf{L}_3 = \mathbf{L}_1,\ \mathbf{L}_3 = \mathbf{L}_2, \tag{3}$$

where $\|.\|_F$ is the Frobenius norm of a matrix.

Low-rank optimization problems can be solved by Augmented Lagrange multipliers (ALM) based algorithms [3, 4, 10] and singular value thresholding operator [5]. However, considering the size of a normalized image (often 128×64 for Re-ID) and the number of frames in a video sequence, the computational cost of SVD required by [5] is quite high. Alternatively, [29] proposed ROSL to solve the problem efficiently by representing the matrix $\mathbf{L}$ through an ordinary orthonormal subspace $\mathbf{M}$ and its coefficients α and by using the approximation of SVD for the low-rank matrix. Thus we reformulate our problem as:

$$\min_{\mathbf{L}_1,\mathbf{L}_2,\mathbf{L}_3,\mathbf{E},\alpha} \|\alpha\|_{\text{row-1}} + \lambda_L \|\mathbf{L}_2\mathbf{P}\|_F^2 + \lambda_E \|\mathbf{E}\|_1, \quad s.t. \begin{cases} \mathbf{D} = \mathbf{L}_3 + \mathbf{E},\ \mathbf{L}_3 = \mathbf{L}_1,\ \mathbf{L}_3 = \mathbf{L}_2, \\ \mathbf{L}_1 = \mathbf{M}\alpha,\ \mathbf{M}^\top\mathbf{M} = \mathbf{I}_k, \end{cases} \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^{d\times k}$ is an orthogonal matrix composed of k orthonormal vectors of d-dimension, and $\alpha \in \mathbb{R}^{k\times N_s}$ is the coefficient matrix. The *row*-1 norm is defined as $\|\alpha\|_{\text{row-1}} = \sum_{i=1}^{k} \|\alpha_i\|_2$, where α_i is the i-th row. $\mathbf{I}_k$ is the k-dimensional identity matrix.

Finally, the principal appearance $\mathbf{L}$ of the tracked person in each frame is represented by the linear combination of basis in $\mathbf{M}$, referred as *Principal Person Appearance* (PPA) model. We solve equation (4) to find optimal decomposition of observations into principal appearance via discovery of its principal appearance model.

3.2 Optimization

To solve the problem using ALM method, the augmented Lagrangian function is given as:

$$\begin{aligned} &\mathcal{L}_{\mu_1,\mu_2,\mu_3}(\mathbf{L}_1,\mathbf{L}_2,\mathbf{L}_3,\mathbf{E},\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Y}_3) \\ &= \|\alpha\|_{\text{row-1}} + \lambda_L\|\mathbf{L}_2\mathbf{P}\|_F^2 + \lambda_E\|\mathbf{E}\|_1 + <\mathbf{Y}_3, \mathbf{D}-\mathbf{L}_3-\mathbf{E}> + \frac{\mu_3}{2}\|\mathbf{D}-\mathbf{L}_3-\mathbf{E}\|_F^2 \\ &\quad + <\mathbf{Y}_2, \mathbf{L}_3-\mathbf{L}_2> + \frac{\mu_2}{2}\|\mathbf{L}_3-\mathbf{L}_2\|_F^2 + <\mathbf{Y}_1, \mathbf{L}_3-\mathbf{L}_1> + \frac{\mu_1}{2}\|\mathbf{L}_3-\mathbf{L}_1\|_F^2, \\ &\quad s.t.\ \mathbf{L}_1 = \mathbf{M}\alpha,\ \mathbf{M}^\top\mathbf{M} = \mathbf{I}_k, \end{aligned} \tag{5}$$

where $\mathbf{Y}_{1:3} \in \mathbb{R}^{d\times N_s}$ are matrices of Lagrange multipliers and $\mu_{1:3}$ are positive scalars. We consider the inexact alternating direction method (ADM) which is widely used to find solution of diverse variables using coordinate descent [4, 29, 40]. It finds optimal solution by iteratively minimizing (5) with respect to one variable at a time while freezing others. The details of variable updates for the $(t+1)$-th iteration are discussed in the following.

Update $\mathbf{L}_1^{t+1}$: The minimization problem (5) with respect to $\mathbf{L}_1$ can be replaced by the one with respect to $\mathbf{M}$ and α as follows:

$$\begin{aligned} (\mathbf{M}^{t+1}, \alpha^{t+1}) &= \arg\min_{\mathbf{M},\alpha} \mathcal{L}_{\mu_1,\mu_2,\mu_3}(\mathbf{L}_1,\mathbf{L}_2^t,\mathbf{L}_3^t,\mathbf{E}^t,\mathbf{Y}_1^t,\mathbf{Y}_2^t,\mathbf{Y}_3^t) \\ &= \|\alpha\|_{\text{row-1}} + \frac{\mu_1}{2}\|\mathbf{L}_1 - \mathbf{L}_3^t - \frac{1}{\mu_1}\mathbf{Y}_1^t\|_F^2, \quad s.t.\ \mathbf{L}_1 = \mathbf{M}\alpha,\ \mathbf{M}^\top\mathbf{M} = \mathbf{I}_k. \end{aligned} \tag{6}$$

The problem can be solved by the ROSL method [29], which applies block coordinate descent (BCD) that allows shrinking $\|\alpha\|_{\text{row-1}}$ while sequentially updating $\mathbf{M}$ and α. For seeking orthonormal subspace $\mathbf{M}$, Gram-Schmidt orthonormalization is applied. In practice, the algorithm needs the initial value of k to define the dimension of the subspace. We use $k = 10$ in the implementation. Readers may refer to [2, 29] for more details.

Update $\mathbf{L}_2^{t+1}$: The minimization problem (5) with respect to $\mathbf{L}_2$ is given as follows:

$$\mathbf{L}_2^{t+1} = \arg\min_{\mathbf{L}_2} \lambda_L \|\mathbf{L}_2\mathbf{P}\|_F^2 + \frac{\mu_2}{2}\|\mathbf{L}_2 - \mathbf{L}_3^t - \frac{1}{\mu_2}\mathbf{Y}_2^t\|_F^2. \tag{7}$$

Since the first term ties different entries of $\mathbf{L}_2$ together by $\mathbf{P}$, the closed form solution for (7) is non-trivial in general. Thus, instead of updating $\mathbf{L}_2$ globally, we derive the element-wise equation. For the minimum solution, the necessary condition for element $l_{i,j}$ in the i-th row and the j-th column of $\mathbf{L}_2$ is:

$$0 = \frac{\lambda_L}{\mu_2} \cdot \frac{\partial}{\partial l_{i,j}} \left(\sum_{r=1}^{N_s-1} |\sum_{c=1}^{N_s} l_{i,c} \cdot p_{c,r}|^2 \right) + l_{i,j} - q_{i,j}, \tag{8}$$

where we use $\mathbf{L}_3^t + \frac{1}{\mu_2}\mathbf{Y}_2^t = \mathbf{Q}$ for the second term of (7), and $q_{i,j}$ means the (i,j) entry of $\mathbf{Q}$.

Now, given that all entries in $\mathbf{P}$ are zero except the two diagonals located near the center of the matrix, as described in (2), (8) can be solved to give the following update equation:

$$l_{i,j}^{(t+1)} = \begin{cases} \dfrac{2\tau \cdot l_{i,j+1}^{(t)} + q_{i,j}}{2\tau+1}, & if\ j = 1, \\ \dfrac{2\tau \cdot (l_{i,j-1}^{(t)} + l_{i,j+1}^{(t)}) + q_{i,j}}{4\tau+1}, & if\ 1 < j < N_s, \\ \dfrac{2\tau \cdot l_{i,j-1}^{(t)} + q_{i,j}}{2\tau+1}, & if\ j = N_s, \end{cases} \tag{9}$$

where we simplify $\frac{\lambda_L}{\mu_2} = \tau$. This is similar to general coordinate descent approach, the updates of $\mathbf{L}_2$ entries at the $(t+1)$-th iteration depend on values of $\mathbf{L}_2$ at the (t)-th iteration.

Update $\mathbf{L}_3^{t+1}$: The minimization problem (5) with respect to $\mathbf{L}_3$ is given as follows:

$$\mathbf{L}_3^{t+1} = \arg\min_{\mathbf{L}_3} \tfrac{\mu_3}{2}\|\mathbf{D} - \mathbf{L}_3 - \mathbf{E}^t + \tfrac{1}{\mu_3}\mathbf{Y}_3^t\|_F^2 + \tfrac{\mu_2}{2}\|\mathbf{L}_3 - \mathbf{L}_2^{t+1} + \tfrac{1}{\mu_2}\mathbf{Y}_2^t\|_F^2 + \tfrac{\mu_1}{2}\|\mathbf{L}_3 - \mathbf{L}_1^{t+1} + \tfrac{1}{\mu_1}\mathbf{Y}_1^t\|_F^2. \tag{10}$$

Since, each term in (10) is square of Frobenius norm, the update has a closed form solution:

$$\mathbf{L}_3^{t+1} = \frac{1}{\gamma} \times [\, \mu_3(\mathbf{Z}_3^t + \mathbf{D} - \mathbf{E}^t) + \mu_2(\mathbf{Z}_2^t - \mathbf{L}_2^{t+1}) + \mu_1(\mathbf{Z}_1^t - \mathbf{L}_1^{t+1}) \,], \tag{11}$$

where we substitute $\gamma = \mu_3 - \mu_2 - \mu_1$ and $\mathbf{Z}_i^t = \frac{1}{\mu_i}\mathbf{Y}_i^t$, $\forall i = 1,2,3$, for simplicity.

Update $\mathbf{E}^{t+1}$: The soft thresholding based closed form solution of the minimization problem in (5) with respect to $\mathbf{E}$ is as follows:

$$\mathbf{E}^{t+1} = \arg\min_{\mathbf{E}} \lambda_E\|\mathbf{E}\|_1 + \frac{\mu_3}{2}\|\mathbf{E} - \boldsymbol{\varepsilon}\|_F^2 = sign(\boldsymbol{\varepsilon}) \times \max(0, |\boldsymbol{\varepsilon}| - \frac{\lambda_E}{\mu_3}), \tag{12}$$

where $\boldsymbol{\varepsilon} = \mathbf{D} - \mathbf{L}_3^{t+1} + \frac{1}{\mu_3}\mathbf{Y}_3^t$ is to simplify the expression.

The entire algorithm including the update of Lagrange multipliers is summarized in Algorithm 1.

4 Track Representation via Principal Appearance Groups

The tracks need to be represented as signatures to perform Re-ID. A trivial solution is to mean or max pool principal appearance vectors $\mathbf{L}$ (or features computed from them) to construct corresponding signature. There are two issues with this approach. First, considerable

Algorithm 1 Principal Appearance Learning

Require: $\mathbf{D}$, λ_L, λ_E, $\rho_{1:3} > 1$
Ensure: $\mathbf{L}$, $\mathbf{M}$, α, $\mathbf{E}$
1: **Initialize:** $\mathbf{L}_{1:3} = \mathbf{D}$, $\mathbf{Y}_{1:3} = \mathbf{0}$, $\mathbf{M} = \mathbf{0}$, $\mathbf{E} = \mathbf{0}$
$\alpha = \text{rand}$, $k = 10$
2: **while** *not converged* **do**
3: Update $\mathbf{L}_1 = \mathbf{M}\alpha \leftarrow$ solving (6)
4: Update $\mathbf{L}_2 \leftarrow$ solving (9)
5: Update $\mathbf{L}_3 \leftarrow$ solving (11)
6: Update $\mathbf{E} \leftarrow$ solving (12)
7: Update Lagrange multipliers and parameters:
$\mathbf{Y}_1 = \mathbf{Y}_1 + \mu_1(\mathbf{L}_3 - \mathbf{L}_1)$
$\mathbf{Y}_2 = \mathbf{Y}_2 + \mu_2(\mathbf{L}_3 - \mathbf{L}_2)$
$\mathbf{Y}_3 = \mathbf{Y}_3 + \mu_3(\mathbf{D} - \mathbf{L}_3 - \mathbf{E})$
$\mu_1 = \mu_1\rho_1, \ \mu_2 = \mu_2\rho_2, \ \mu_3 = \mu_3\rho_3$
8: **end**

texture information is lost in recovery of matrix $\mathbf{L}$, which is useful for re-identification. Secondly, unconstrained motion of the person may result in significantly different observed and recovered images, thus a unimodal assumption about appearance (*e.g.* mean) may not be true representative. To cater multi-modality and bridge the information gap between principal and observed appearance, we group original images of the track into *Principal Appearance Groups* (PAGs) using corresponding principal appearances. Each group is then represented by its signature and the track is represented as a set of group signatures.

Recall that columns of principal appearance matrix $\mathbf{L}$ belong to the linear span of orthonormal subspace $\mathbf{M}$ and that columns of α are their corresponding coefficients. In other words, α represents projection of principal appearance $\mathbf{L}$ on orthonormal basis $\mathbf{M}$, thus the images which have similar principal appearance should have similar coefficients. Therefore, images can be grouped by applying K-means algorithm to the coefficient matrix α. More precisely, we consider i-th column of α as the feature vector of i-th image and use AIC to find optimal number (and memberships) of clusters - PAGs - in range $K \in (1, 10)$.

For each PAG, we use both vivid input and smooth principal appearance to compute its signature. Specifically, we compute feature descriptors for both principal appearance, $\{f_i^p | i = 1, ..., n_g\}$, and input images, $\{f_i^o | i = 1, ..., n_g\}$, belonging to the group, where n_g means the number of images in the g-th group. Next, we compute mean feature descriptor for the principal appearance, $\mu = \frac{1}{n_g}\sum_i f_i^p$ and weights $w_i = 1 - \frac{\|f_i^o - \mu\|}{\sum_i \|f_i^o - \mu\|}$. A group signature s is then computed as weighted sum of features: $s = \sum_i w_i f_i^o$.

5 Signature Matching

In order to find the most relevant gallery signature $g \in \mathcal{G}$ for a given query signature $q \in \mathcal{Q}$, we use RBF kernel based similarity measure:

$$sim(q, g) = \exp\left(-\frac{d(q,g) - \beta_g}{0.3 \times (\alpha_g - \beta_g)}\right), \tag{13}$$

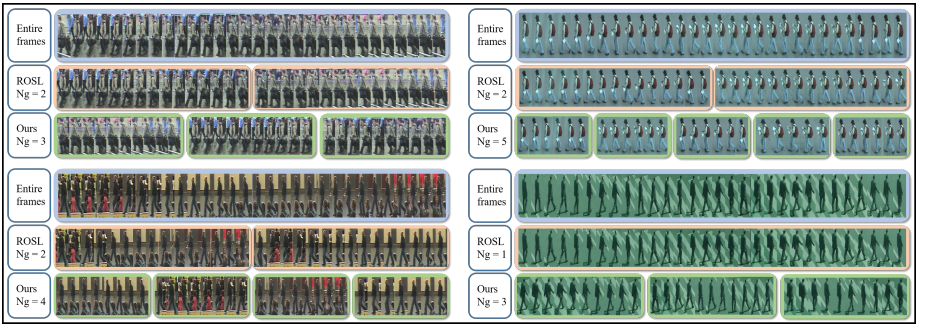

Figure 2: Visualization of principal appearance groups. Examples of our algorithm result are shown in comparison with ROSL [29] for the same process. The top and bottom on the left column are ID:218 and ID:016 of iLIDS-VID [33], and the top and bottom on the right column are ID:001 and ID:185 of PRID [12], respectively. N_g means the number of image groups (Section 4).

where, $d(q,g)$ is the distance between signatures q and g, $\alpha_g = \max_{q' \in \mathcal{Q}} d(q',g)$ and $\beta_g = \min_{q' \in \mathcal{Q}} d(q',g)$. As both q and g are sets of PAG signature vectors, $d(q,g)$ is defined as the average Mahalanobis distance between any pair $\{(u,v)|u \in q, v \in g\}$:

$$d(q,g) = \frac{1}{|q||g|} \sum_{u \in q, v \in g} (u-v)' \Psi (u-v). \tag{14}$$

The matrix Ψ can be learned using either KISSME[16] or XQDA[33] algorithm.

6 Experiments and Results

For evaluation of our proposed framework, we selected two popular publicly available datasets for multi-shot Re-ID task: iLIDS-VID [33] and PRID [12]. These datasets are chosen because they provide multiple images per individual collected in realistic visual surveillance settings using two cameras. During our experiments, we describe images in each PAG using LOMO [19] as it has shown to be effective for Re-ID task [11].

PRID dataset consists of 385 persons viewed from camera A and 750 persons viewed from camera B. However, only 200 persons appear in both cameras. Both cameras have very different viewpoint and color profile that make Re-ID challenging. For a fair comparison with other methods, we follow evaluation protocol of [33], *i.e.* 178 persons with at least 21 images available in each camera are selected and randomly divided into equal size train and test sets based on person identities with no overlap. Train set is used to learn adaptive metric and results are averaged for 10 trials.

6.1 Qualitative Analysis of Principal Appearance Groups

Table 1 presents statistics about the optimal number of PAGs found using algorithm presented in Section 4 for each input track on average. The mean number of groups per person is between 3 and 3.5, whereas the standard deviation varies between 0.6 and 1.0.

Due to absence of ground truth information about the number of groups and membership of images, we qualitatively analyzed a handful of samples. PAGs obtained using our algorithm by solving Eq. (1) and ROSL [29] (Eq. (3)) for some randomly selected tracks from

each dataset are shown in Figure 2. The optimal number of groups N_g varies for each input depending on the variance in appearance of the person. It can be observed that our algorithm is able to find distinct appearance groups for each track based on person's semantic attributes, such as pose and orientation. In addition, the segmentation appears to be robust to the variance of the background. Each group is also homogeneous in terms of human pose or orientation but may have varying illumination and occlusions. Our method often finds more groups with higher consistency than ROSL. These results provide significant confidence to assume that the our algorithm can successfully find a small set of homogeneous yet semantically distinct principal appearance groups for each person.

6.2 Comparison with State-of-the-Art

Table 2 reports results of our Principal Person Appearance based Re-ID method (PPA) and other competing approaches. Deep learning based methods (RCNN [24], CNN+KISSME [41], CNN+XQDA [41]) have considerably pushed the state-of-the-art in recent years on PRID. At the same time, performance of AvgTAPR [11] demonstrates that efficient temporal pooling of hand-crafted features can still yield competitive results. When using same ingredients as [11], *i.e.* LOMO features and XQDA metric learning, our method aims to learn stable person appearance instead of focusing on walk-cycles. As a consequence, we significantly outperform [11] and all other competing methods on PRID, including recent deep learning based approaches.

In comparison to PRID, iLIDS-VID dataset has significant occlusions, as it was collected at an airport. It consists of 300 persons observed from two cameras. We again follow evaluation protocol of [33] and report average performance over 10 trials of random train-test splits based on person IDs. As PPA explicitly addresses recovery of principal appearance in presence of transient occlusions, it achieves best performance on this dataset as well (Table 2). Even though our method considerably improves state-of-the-art, due to challenging nature of iLIDS-VID, the rank-1 rate of all methods is considerably lower than on PRID. However, our method is the only one that achieves above 90% recognition rate at rank-5.

No. of Groups	iLIDS-VID Cam1	iLIDS-VID Cam2	PRID2011 Cam1	PRID2011 Cam2
Mean	3.5	3.0	3.2	3.3
StdDev	1.0	0.8	0.6	0.8
Median	3	3	3	3
Maximum	10	7	7	8

Table 1: Statistics of optimal number of principal groups found per person. Results are split based on Camera ID for each dataset.

	PRID				iLIDS-VID			
Method	**r=1**	**r=5**	**r=10**	**r=20**	**r=1**	**r=5**	**r=10**	**r=20**
HOG3D+RankSVM [6]	19.4	44.9	59.3	77.2	12.1	29.3	41.5	56.3
Color+RankSVM [6]	29.7	49.4	59.3	71.1	16.4	37.3	48.5	62.6
DVR [6]	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.6
ColorLBP [6]+RankSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
DVDL [6]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
Color+LFDA [6]	43.0	73.1	82.9	90.3	28.0	55.3	70.6	88.0
AFDA [6]	43.0	72.7	84.6	91.9	37.5	62.7	73.0	81.8
DSVR [6]	40.0	71.1	84.5	92.2	39.5	61.1	71.7	81.0
MTL-LORAE [6]	-	-	-	-	43.0	60.1	70.3	85.3
STFV3D+KISSME [6]	64.1	87.3	89.9	92.0	43.8	69.3	80.0	90.0
CNN+KISSME [6]	69.9	90.6	-	98.2	48.8	75.6	-	92.6
RFA-Net+RankSVM [6]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
CNN+XQDA [6]	77.3	93.5	-	99.3	53.0	81.4	-	95.1
AvgTAPR+XQDA [6]	68.6	94.6	97.4	98.9	55.0	87.5	93.8	97.2
TDL [6]	56.7	80.0	87.6	93.6	56.3	87.6	95.6	98.3
RCNN [6]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
PPA+Euclidean	66.6	90.1	93.5	96.7	29.6	55.7	67.6	79.7
PPA+KISSME	85.7	98.9	**99.9**	**100.0**	65.7	92.3	96.8	99.1
PPA+XQDA	**87.6**	**99.2**	99.6	99.9	**66.8**	**93.9**	**97.8**	**99.8**

Table 2: Comparison of recognition rates (%) at different ranks of various Re-ID methods on PRID and iLIDS-VID. Best results are highlighted in bold.

7 Conclusion

Due to articulated motion, irrelevant pixels and detection/tracking imperfection, input (image sequences) to a re-identification system displays high intra-sample variance. This also leads to high inter-sample variance and adversely affects multi-shot Re-ID method's performance. This paper advocates for explicitly addressing high intra-sample variance at an early stage of Re-ID pipeline to learn robust appearance models that lead to better Re-ID performance. Towards this end, this paper introduces a principal appearance discovery method using low-rank recovery formulation. A track is divided into multiple groups based on recovered principal appearance before being represented as a set of a small number of representative signatures. These signatures are then matched with a conventional adaptive ranking function to achieve state-of-the-art results on publicly available datasets.

References

[1] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013.

[2] Samy Bengio, Fernando Pereira, Yoram Singer, and Dennis Strelow. Group sparse coding. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 82–89, 2009.

[3] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[5] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[6] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[7] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1268–1277, 2016.

[8] Yeong-Jun Cho and Kuk-Jin Yoon. Improving person re-identification via pose-aware multi-shot matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1362, 2016.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[10] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[11] C. Gao, J. Wang, L. Liu, and N. Sang J-G. Yu. Temporally aligned pooling representation for video-based person re-identification. In *IEEE International Conference on Image Processing (ICIP)*, pages 4284–4288, 2016.

[12] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102, 2011.

[13] Martin Hirzer, Peter Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 780–793, 2012.

[14] Kui Jia, Tsung-Han Chan, Zinan Zeng, Shenghua Gao, Gang Wang, Tianzhu Zhang, and Yi Ma. Roml: A robust feature correspondence approach for matching objects in a set of images. *arXiv preprint arXiv:1403.7877*, 2014.

[15] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4516–4524, 2015.

[16] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2295, 2012.

[17] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *British Machine Vision Conference (BMVC)*, page 8, 2012.

[18] Y. Li, Z. Wu, S. Karanam, and R.J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *British Machine Vision Conference (BMVC)*, page 2, 2015.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, 2015.

[20] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[21] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision (ECCV)*, pages 391–401, 2012.

[22] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3810–3818, 2015.

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[24] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334, 2016.

[25] Yadong Mu, Jian Dong, Xiaotong Yuan, and Shuicheng Yan. Accelerated low-rank visual recovery by random projection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2609–2616, 2011.

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

[27] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1306–1315, 2016.

[28] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3200–3208, 2015.

[29] Xianbiao Shu, Fatih Porikli, and Narendra Ahuja. Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3874–3881, 2014.

[30] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3739–3747, 2015.

[31] Yao Sui, Yafei Tang, and Li Zhang. Discriminative low-rank tracking. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3002–3010, 2015.

[32] Ming-Chia Tsai, Chia-Po Wei, and Yu-Chiang Frank Wang. Graph regularized low-rank matrix recovery for robust person re-identification. In *IEEE International Conference on Image Processing (ICIP)*, pages 4654–4658, 2015.

[33] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision (ECCV)*, pages 688–703, 2014.

[34] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, 2016.

[35] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision (ECCV)*, pages 701–716, 2016.

[36] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1345–1353, 2016.

[37] Zinan Zeng, Tsung-Han Chan, Kui Jia, and Dong Xu. Finding correspondence from multiple images via sparse and low-rank decomposition. In *European Conference on Computer Vision (ECCV)*, pages 325–339, 2012.

[38] Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. Learning by associating ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 708–715, 2013.

[39] Tianzhu Zhang, Kui Jia, Changsheng Xu, Yi Ma, and Narendra Ahuja. Partial occlusion handling for visual tracking via robust part matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1258–1265, 2014.

[40] Tianzhu Zhang, Si Liu, Narendra Ahuja, Ming-Hsuan Yang, and Bernard Ghanem. Robust visual tracking via consistent low-rank sparse learning. *International Journal of Computer Vision*, 111(2):171–190, 2015.

[41] 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 *European Conference on Computer Vision (ECCV)*, pages 868–884, 2016.

[42] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4678–4686, 2015.