moM: Mean of Moments Feature for Person Re-identification*

Mengran Gou, Octavia Camps, Mario Sznaier
Electrical and Computer Engineering
Northeastern University, Boston, MA 02115, US
{mengran,camps,msznaier}@coe.neu.edu

Abstract

Person re-identification (re-id) has drawn significant attention in the recent decade. The design of view-invariant feature descriptors is one of the most crucial problems for this task. Covariance descriptors have often been used in person re-id because of their invariance properties. More recently, a new state-of-the-art performance was achieved by also including first-order moment and two-level Gaussian descriptors. However, using second-order or lower moments information might not be enough when the feature distribution is not Gaussian. In this paper, we address this limitation, by using the empirical (symmetric positive definite) moment matrix to incorporate higher order moments and by applying the on-manifold mean to pool the features along horizontal strips. The new descriptor, based on the on-manifold mean of a moment matrix (moM), can be used to approximate more complex, non-Gaussian, distributions of the pixel features within a mid-sized local patch. We have evaluated the proposed feature on five widely used re-id datasets. The experiments show that the moM and hierarchical Gaussian descriptor (GOG) [30] features complement each other and that using a combination of both features achieves a comparable performance with the state-of-the-art methods.

1. Introduction

Person re-identification (re-id) is the problem of matching images of a pedestrian across cameras with no overlapping fields of view. It is one of the key tasks in surveillance video processing. Due to the extremely large inter-class variances across different cameras (e.g., poses, illumination, viewpoints), the performance of the state-of-the-art person re-id algorithms is still far from ideal [18, 50]. Most of the existing re-id literature focuses on two aspects of the problem: 1) designing viewpoint invariant feature descriptors [3, 12, 14, 21, 27, 28, 29, 30, 44, 48] and/or 2) learning a supervised classifier to alleviate the effect of the variances across the cameras [19, 21, 22, 33, 35, 37, 43, 45, 47, 24]. Recently, deep neural networks have been adopted to learn both the descriptor and classifier simultaneously [1, 6, 20, 40]. For more details, we refer the reader to [11, 18, 42, 50].

Figure 1. Sample images where up to second-order moments are not enough to distinguish targets. Each column shows one pair of samples from VIPeR, CUHK01, PRID450s and GRID, respectively. The images on the first row are from the "probe" view and the second row are from the "gallery" view. The blue rectangle indicates a 16×16 patch. The third row shows ranking results using the proposed moM feature and the GOG [30] feature (lower is better). In these examples, with the help of higher order moments, moM is more discriminative when the person has fine-detailed appearance, e.g., the checkered pattern in column 1 and 2, the salient white collar in column 3 and the flower pattern in column 4.

*[1]This work was supported in part by NSF grants IIS1318145 and ECCS1404163 and AFOSR grant FA9550-15-1-0392. This material is based upon work supported by the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Grant Award 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

This paper focuses on the first aspect. In the past decade, different types of descriptors have been proposed and tested on the re-id problem. Two recently proposed techniques led to a significant improvement in the quality of these descriptors [22, 30].

The first technique replaces the simple computation of histograms with more advanced feature-encoding methods. Along this line, covariance matrices are used to encapsulate the second-order moment information in a local patch [3, 14]. By recognizing the importance of also including the first-order moment in the feature representation, [27, 30] achieved the state-of-the-art performance using a symmetric positive definite (SPD) embedded Gaussian descriptor. However, a limitation of this descriptor is the implicit assumption that the underlying distribution is a Gaussian. When the assumption does not hold, (see Figure 1), up to second-order moment information is not sufficient to completely represent relatively complex local regions. Though Fisher Vector encoding features can mimic a non-Gaussian distribution with a Gaussian mixture model (GMM) and achieve decent results on re-id [12, 29], it assumes that the variables at the pixel-level feature are independent from each other. Moreover, the GMM needs a training set to learn its parameters. In contrast, here we propose to take into account higher (greater than two) order moment information by using the empirical moment matrix to approximate arbitrary non-Gaussian distributions in the local region without requiring learning parameters.

The second technique applies a strip level pooling step to further improve cross-view invariance. As identities are roughly aligned along the vertical direction (Figure 1), different viewpoints would mainly affect the appearance distribution in the horizontal direction. Based on this assumption, Liao *et al.* [21] applies maximum pooling along the same height and Matsukawa *et al.* [30] uses another Gaussian model to approximate the distribution of the dense patches descriptors. In this paper, we also use horizontal mean pooling to improve the feature viewpoint invariance. Furthermore, since moment matrices are on a SPD manifold, we also propose to use the on-manifold mean and flattening on its tangent space.

Experiments on five public benchmark datasets illustrate the benefits of encapsulating higher order moments information. The combination of proposed mean of moment (moM) features with GOG [30] achieves comparable or better state-of-the-art performance on all the tested datasets.

2. Related Work

Person re-id specific hand-crafted features mainly focus on the invariance across different cameras. In [10], based on the symmetric axis of each body part, a carefully designed body configuration was modeled. Then, the weighted color histogram was computed, depending on the distance between the pixel and the axis. The final representation was also combined with maximally stable color regions (MSCR) and recurrent high-structured patches (RHSP). Ma *et al.* [28] used the biological inspired feature (BIF) as the raw feature and compressed it using the similarity between the covariance matrices of small patches. Since then, following the development of metric learning methods, researchers tend to use native but redundant features to feed into the supervised learned metric. Gary and Tao [13] used 8 color channels (RGB, HS, and YUV) and 21 texture filters. In [31], responses of texture filters were substituted by LBP features. Instead of color histogram, in [15], the local mean of each patch was adopted. Pedagadi *et al.* [35] added the first three moments to the color histogram to represent a small patch. More recently, Zhao *et al.* [48] combined the LAB histogram with dense SIFT descriptors on a densely sampled grid. To obtain a stable representation, color names have been applied recently. In [44], salience color name distributions were computed over different color models to remedy the illumination variance. Zheng *et al.* [49] encoded the local color name descriptors through Bag-of-Words. Liao *et al.* [21] proposed maximum-pooling the color and SILTP [23] histogram along the same horizontal strip to achieve better viewpoint invariance.

Covariance and Gaussian descriptors have been applied in person re-id, to compress more information than histogram and local mean. In [3], pixel level color intensity and gradient in a local patch are compressed into a covariance matrix. Ma *et al.* [27] modeled the low level feature with a Gaussian distribution and compare the Gaussian with the product on Lie group. In [29], GMM is used to model the pixel feature by assuming the variants are independent of each other. Inspired by LOMO [21], a hierarchical Gaussian feature (GOG) was proposed in [30]. Similar to previous work, pixel features in a small patch are modeled by a Gaussian distribution, which is embedded in an SPD manifold. Then, the second level models the distribution of the first level descriptors within a strip around the same height.

Because the covariance matrix lies on a Riemannian manifold, several on-manifold metric based methods have been proposed for different computer vision applications. In [39], the covariance matrix and on-manifold classification were applied for pedestrian detection. Huang *et al.* [16] proposed on-manifold metric learning for image set classification. By generalizing VLAD [17] to Riemannian manifold, Faraki *et al.* [9] showed the effectiveness of on-manifold VLAD in different applications. Zhang *et al.* [46] compared different on-manifold metrics and applied them for skeleton activity classification.

Our proposed feature moM generalizes the naive Gaussian distribution model or independent multi-variant GMM with the empirical moment matrix. Using higher order moments, it can approximate arbitrary non-Gaussian distribu-

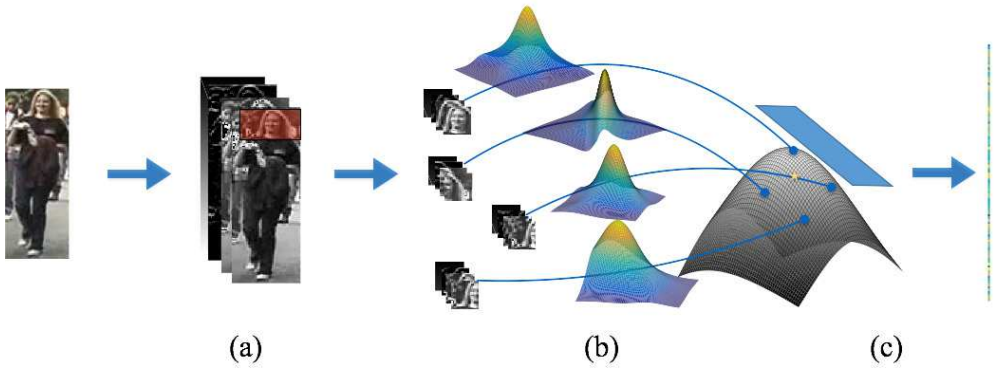


Figure 2. moM feature extraction: Starting with a pedestrian image, (a) pixel features are computed to extract the color and gradient information and (b) each patch is modeled with a moment matrix, which lies on a SPD manifold. (c) On-manifold mean is applied to pool the information along horizontal strips and then the mean matrix is flattened to its tangent space and vectorized to form the final descriptor of the strip.

tions. Furthermore, an SPD embedded Gaussian matrix is a special case of the empirical moment matrix when the order is 1 (please see Sec. 4.2 for the proof).

3. Notation

For ease of reference, in this section we summarize the notation used in this paper.

$\mathbb{R}, \mathbb{N}$	set of real numbers, set of nonnegative integers
$x, \mathbf{x}, \mathbf{X}$	scalar, a column vector in $\mathbb{R}^n$, a matrix in $\mathbb{R}^{m\times n}$
$\mathbf{x}(i)$	the i-th entry of $\mathbf{x}$
$\mathbf{X}(i,j)$	the (i,j)-th entry of $\mathbf{X}$
$\|\mathbf{X}\|_F$	Frobenius norm of the matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ $\|\mathbf{X}\|_F \doteq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{X}(i,j)^2}$
$\|\mathbf{x}\|_2$	ℓ_2-norm of the vector $\mathbf{x} \in \mathbb{R}^n$ $\|\mathbf{x}\|_2 \doteq \sqrt{\sum_{i=1}^{n}\mathbf{x}(i)^2}$
$\|\mathbf{x}\|_1$	ℓ_1-norm of the vector $\mathbf{x} \in \mathbb{R}^n$ $\|\mathbf{x}\|_1 \doteq \sum_i \|x_i\|$
$s_{m,D}$	$\binom{m+D}{m}$

4. Mean of Moment (moM) Feature

Next, we describe the moM features. Figure 2 illustrates the pipeline to extract them, and the corresponding step-by-step procedure is summarized in Alg. 1.

4.1. Base pixel features

Following the work [30], we also use the following pixel level features to represent local appearance information:

$$\mathbf{x}_k = [y, A_{0^\circ}, A_{90^\circ}, A_{180^\circ}, A_{270^\circ}, C_a, C_b, C_c]^T \quad (1)$$

where y is the y coordinate of pixel z_k, $A_{\theta\in\{0^\circ,90^\circ,180^\circ,270^\circ\}}$ are the magnitudes of gradient along four directions, and $C_{\{a,b,c\}}$ are intensity values in the corresponding color channel. All dimensions are normalized to the range $[0,1]$. In this paper, we will use RGB, HSV, LAB or normalized RG as the color channel.

4.2. Empirical moment matrix

Given a dataset consisting of N samples $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^N$, where $\mathbf{x}_k = [x_{k1}, x_{k2}, ...x_{km}] \in \mathbb{R}^m$, the collection of all monomials of $\mathbf{x}_k \in \mathbb{R}^m$ up to order D is defined as

$$\mathbf{v}_k \in \mathbb{R}^{s_{m,D}}, \text{ with } \mathbf{v}_k(i) = x_{k1}^{d_{i1}} x_{k2}^{d_{i2}} \cdots x_{km}^{d_{im}}, \forall_{i=1}^{s_{m,D}} \quad (2)$$

where the tuple $\mathbf{d}_i \doteq (d_{i1}, d_{i2}, \ldots, d_{im}) \in \mathbb{N}^m$ denotes the exponents of $x_{k1}, x_{k2}, \ldots, x_{km}$ in the term $\mathbf{v}_k(i)$, satisfying $0 \le \|\mathbf{d}_i\|_1 \le D$. The D-th[1] order empirical moment matrix is defined as:

$$\begin{aligned}\mathbf{M} \doteq& \mathcal{E}\{\mathbf{v}\mathbf{v}^T\} \in \mathbb{R}^{s_{m,D}\times s_{m,D}}, \text{ with}\\ \mathbf{M}(i,j) \doteq& \mathcal{E}\{\mathbf{v}(i)\mathbf{v}(j)\}\\ =& \frac{1}{N}\sum_{k=1}^{N}\mathbf{v}_k(i)\mathbf{v}_k(j), \forall i,j = 1,\ldots,s_{m,D}\end{aligned} \quad (3)$$

When $D = 1$, the moment matrix is given by:

$$\begin{bmatrix} 1 & \mathcal{E}(\mathbf{x}) \\ \mathcal{E}(\mathbf{x})^T & \mathcal{E}\{\mathbf{x}\mathbf{x}^T\} \end{bmatrix} \quad (4)$$

[1] Please note the D-th order $\mathbf{M}$ has moments up to order $2D$.

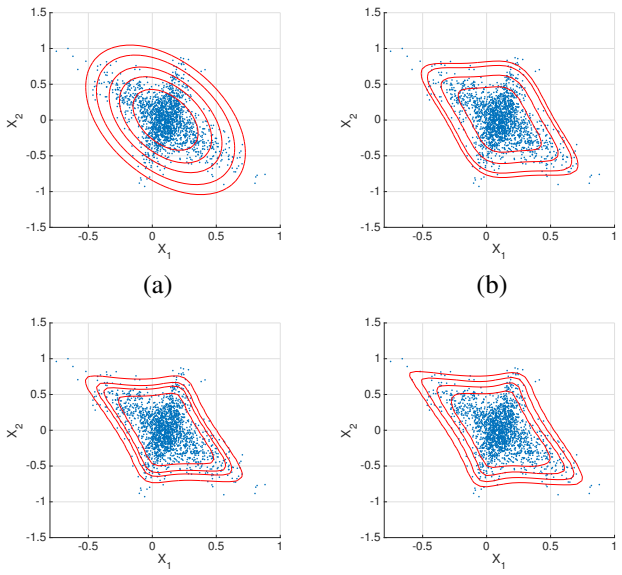


Figure 3. Level sets of (5) with different Ds and Ts. (a) $D = 1$; (b) $D = 2$; (c) $D = 3$; (d) $D = 4$

which is the same as the transformation from a Gaussian distribution to the SPD manifold, except for the normalization term [25].

In comparison to commonly used properties such as the mean and covariance, a moment matrix of higher order ($D \geq 2$) contains richer statistical information. Pauwels and Lasserre [34] noted that the level set of polynomial

$$\mathbf{v}^T \times \mathbf{M}^{-1} \times \mathbf{v} = T \tag{5}$$

(with large enough D) can be used to represent the shape of an arbitrary distribution. Figure 3 illustrates the merit of this representation, where we plot the (red) level sets described as (5) for different values of T and the moment matrix computed from the (blue) given samples of a cross-shaped distribution, computed for increasing values of $D = 1, \ldots, 4$. As observed in the figure, using a moment matrix with higher D captures the shape of the cloud of samples more accurately.

In the sequel, within a mid-sized patch p, we will use the moment matrix $\mathbf{M}_p$ defined as (3) to model the local appearance feature distribution.

4.3. On-manifold mean pooling

As shown in Figure 1, pedestrians inside the bounding boxes are roughly aligned along the vertical direction. The current state-of-the-art re-id features, GOG [30] and LOMO [21], take advantage of this assumption and apply information pooling along horizontal strips. In this work, we also use mean pooling to represent the patches at the same height. However, since $\mathbf{M}$ is an SPD matrix, all $\mathbf{M}$s lie on an SPD manifold. Then, on-manifold distance should be applied to compute the mean matrix. Here, we adopt the Log-Euclidean Riemannian Metric (LE) [2] as in (6) to calculate the distances between two SPD matrices:

$$\sigma_{LE}(\mathbf{M}_{p1}, \mathbf{M}_{p2}) = \| \log(\mathbf{M}_{p1}) - \log(\mathbf{M}_{p2}) \| \tag{6}$$

and the associated on-manifold mean for strip s is:

$$\bar{\mathbf{M}}_s = \exp\left(\frac{1}{Q} \sum_{p=1}^{Q} \log(\mathbf{M}_p)\right) \tag{7}$$

where Q is the number of patches in strip s and $\exp(\cdot)$ denotes the matrix exponential operator.

The benefits of using LE as the on-manifold metric are two-fold: 1) it has a closed form solution and can be computed very efficiently; 2) to feed the feature to off-shelf metric learning methods, one can transfer the SPD matrices into Euclidean space by taking the logarithm, which will cancel the $\exp(\cdot)$.

The vectorized moM feature $\mathbf{g}_s$ for strip s is obtained by equation (8)

$$\begin{aligned} \mathbf{\Gamma}_s &= \log(\bar{\mathbf{M}}_s) \\ \mathbf{g}_s &= \text{vec}(\Gamma_s) \\ &= \left[\mathbf{\Gamma}(1,1), \sqrt{2}\mathbf{\Gamma}(1,2), \ldots, \mathbf{\Gamma}(2,2), \sqrt{2}\mathbf{\Gamma}(2,3), \ldots\right] \end{aligned} \tag{8}$$

where $\log(\cdot)$ denotes the matrix logarithm operator and $\sqrt{2}$ applies on off-diagonal elements to keep the condition $\|\mathbf{\Gamma}_s\|_F = \|\mathbf{g}_s\|_2$ holding. To reduce numerical problems caused by the logarithm of small eigenvalues of the moment matrix, all $\mathbf{M}_p$ are normalized to $\det(\mathbf{M}_p) = 1$.

Finally, the global feature vector $\mathbf{f}$ is defined as the concatenation of all $\mathbf{g}_s$ in all strips. Following the setting in [30, 36], we also apply mean removal and power normalization. Thus, the moM descriptor is normalized by (9)

$$\mathbf{f}_{norm} = \text{sign}(\mathbf{f} - \boldsymbol{\mu}_f)|\mathbf{f} - \boldsymbol{\mu}_f|^{0.5} \tag{9}$$

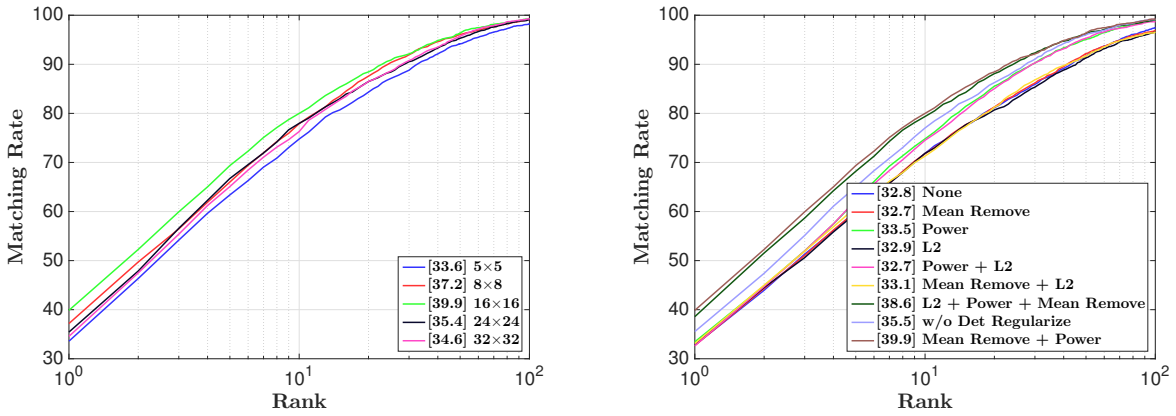
where $|\cdot|$ is the absolute value and $\boldsymbol{\mu}_f$ is the mean of all moM features in the training set.

5. Experiments

5.1. Datasets

We evaluate the proposed moM feature using four widely used hand labeled re-id benchmark datasets and one large-scale automatic detected dataset.

VIPeR [13] contains 632 persons. Each person has two images taken from different viewpoints. All identities are separated into training and testing sets equally. One view is fixed as the probe view. This procedure is repeated 10 times and the average performance is reported.



(a) (b)

Figure 4. Performance analysis on VIPeR dataset.

Table 1. Comparing with different Ds and on-manifold means. The best results in each dataset are marked in red.

Dataset	VIPeR				CUHK01				PRID450s				GRID			
Methods	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
$\text{moM}_{\text{rgb}}^{\text{Jeff}}$ (D=1)	31.4	59.2	71.8	82.7	38.2	57.9	66.5	75.2	38.4	63.9	75.2	84.6	12.2	27.8	35.0	45.8
$\text{moM}_{\text{rgb}}^{\text{JBLD}}$ (D=1)	33.0	61.0	73.1	82.8	41.8	62.4	70.7	79.2	46.4	71.0	80.2	87.9	15.4	30.0	37.8	48.5
$\text{moM}_{\text{rgb}}^{\text{LE}}$ (D=1)	33.1	59.4	72.1	82.7	42.9	62.7	70.8	79.3	47.9	70.0	80.3	88.9	15.5	30.0	38.5	48.6
$\text{moM}_{\text{rgb}}^{\text{Jeff}}$	40.8	71.0	82.1	90.8	48.6	70.9	78.9	86.4	59.6	82.1	89.2	94.5	17.8	39.4	49.8	62.2
$\text{moM}_{\text{rgb}}^{\text{JBLD}}$	39.0	71.1	81.1	89.6	51.7	73.4	81.1	87.8	59.5	82.6	89.4	95.0	20.4	40.1	51.2	62.7
$\text{moM}_{\text{rgb}}^{\text{LE}}$	39.9	69.4	80.0	88.4	52.1	73.4	80.9	87.6	62.5	83.7	90.7	96.5	21.4	42.0	51.9	62.6
GOG_{rgb}	41.4	74.7	85.4	92.6	53.7	76.0	83.6	89.8	62.9	84.6	92.0	96.1	20.2	38.7	49.2	59.8
$\text{moM}_{\text{rgb}}^{\text{JBLD}}$+$\text{GOG}_{\text{rgb}}$	46.0	77.3	86.7	93.8	62.3	83.2	89.3	93.5	67.6	87.6	93.8	97.4	22.2	44.2	55.7	66.1
$\text{moM}_{\text{rgb}}^{\text{LE}}$+$\text{GOG}_{\text{rgb}}$	46.9	77.4	87.2	93.0	62.4	83.0	88.9	93.3	68.6	89.2	94.8	97.4	23.1	44.5	56.2	66.7
$\text{moM}_{\text{f}}^{\text{JBLD}}$	48.0	77.9	86.6	92.2	57.7	78.5	85.3	90.8	66.0	85.9	92.6	97.1	22.6	44.6	54.9	64.8
$\text{moM}_{\text{f}}^{\text{LE}}$	48.0	76.8	85.4	92.1	57.3	78.1	85.1	90.7	65.9	87.2	93.1	97.2	23.4	44.6	54.8	65.4
GOG_{f}	48.8	79.6	88.8	94.6	57.3	79.9	87.0	92.5	68.4	88.5	94.2	97.2	21.8	43.3	52.7	63.5
$\text{moM}_{\text{f}}^{\text{JBLD}}$+$\text{GOG}_{\text{f}}$	52.1	82.1	89.2	94.5	64.3	**85.1**	**90.7**	**94.9**	**71.1**	91.2	**95.4**	97.8	23.6	**46.3**	**57.4**	**67.4**
$\text{moM}_{\text{f}}^{\text{LE}}$+$\text{GOG}_{\text{f}}$	**53.3**	**82.3**	**89.5**	**94.8**	**64.6**	84.9	90.6	94.8	**71.1**	**91.3**	**95.4**	**97.9**	**24.5**	46.1	56.8	66.9

Algorithm 1 moM feature extraction

Require: Image I, number of horizontal strips S, number of patches per strip Q, moment matrix order D.

1: Compute pixel features in (1)
2: **for** strip $s = 1$ to S **do**
3: **for** patch $p = 1$ to Q **do**
4: Compute moment matrix $\mathbf{M}_p$ based on (3)
5: **end for**
6: Compute on-manifold mean $\bar{\mathbf{M}}_s$ based on (7)
7: Compute the feature of $\mathbf{g}_s$ based on (8)
8: **end for**
9: Concatenate $\mathbf{g}_{1,2,\ldots,S}$ to form the final moM feature $\mathbf{f}$

CUHK01 [51] contains 971 persons from two views and each person has 2 images per view. One camera is set as probe with equally separated train and test sets. Average performance of ten randomly trails is reported.

QMUL underGround Re-IDentification (GRID) [26] dataset has 250 paired pedestrians and 775 un-paired distractions captured in a subway station. The large size of the gallery set and relatively low image quality make it one of the most challenging re-id datasets. We use the provided partition configuration.

PRID450s [38] is a subset of PRID2011 [14] with 450 persons and 2 cameras. Each person has one image per camera. Similar to VIPeR dataset, the train and test sets are equally separated and one camera is fixed as the probe one. Ten repeated evaluations are performed and the average result is reported.

Market1501 [49] dataset is a recently proposed large scale re-id dataset. It contains 1,501 identities from 6 cameras. All bounding boxes are automatically detected with the DPM [10] algorithm and manually annotated. In total, it contains 32,668 bounding boxes including 2,793 false alarms from the person detector. We adopt the provided single query train/test partition to evaluate our feature.

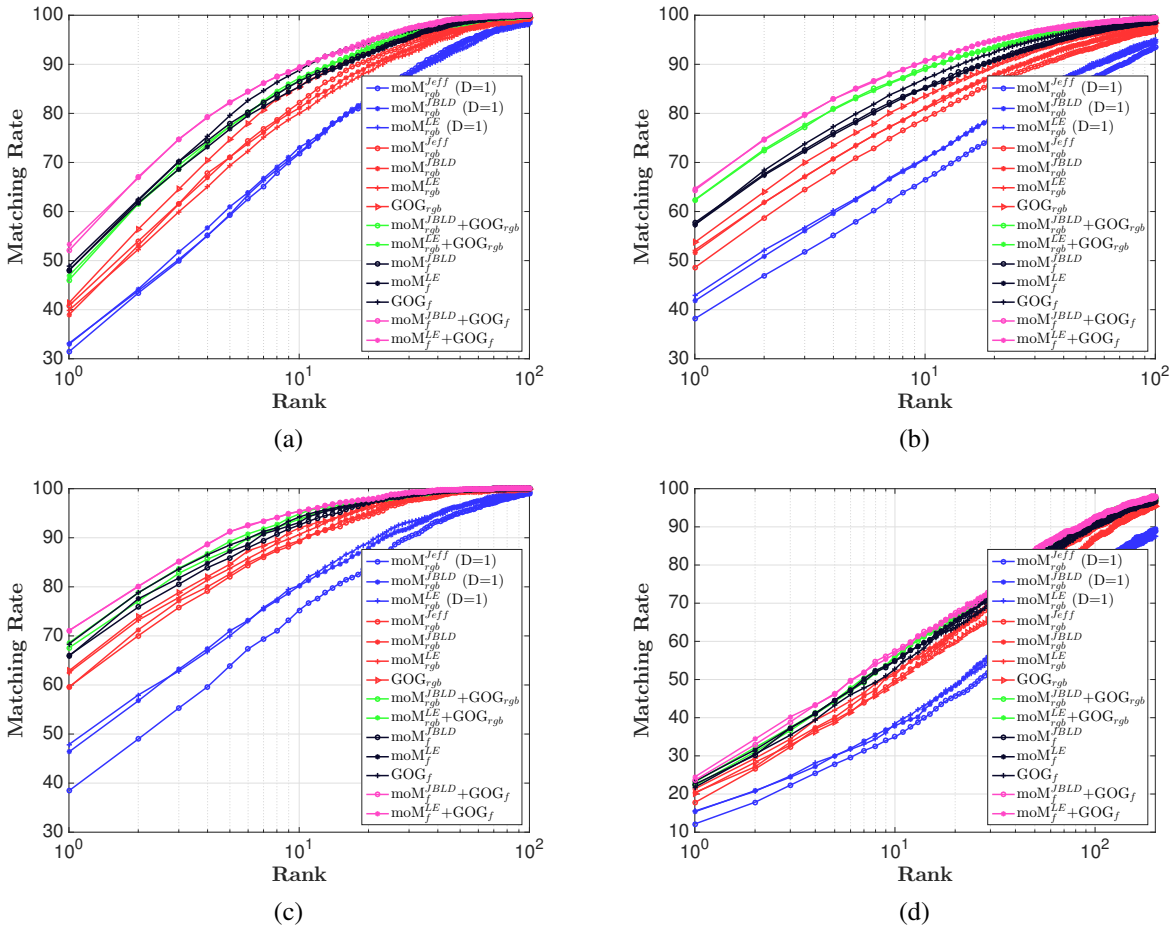


Figure 5. CMC curves for (a)VIPeR, (b)CUHK01, (c)PRID450s and (d)GRID datasets.

5.2. Implementation details

First of all, we reshape all images to the size of 128×48, and the patch size is set to 16×16 with 50% overlap which will generate 15 horizontal strips. The order D is set to 2, so the moment matrix size is 45×45. Therefore, the final dimension of the feature with RGB is $(45 \times 46/2) \times 15 = 15,525$. Following the setting in [30], we weight the patches according to their position on x axis as $w_p = e^{(-(x_p - x_c)^2/2\sigma^2)}$, where $x_c = W/2$ and $\sigma = W/4$. x_p is the x coordinate of the center point of patch p and W is the width of the image. We also fuse moM from different color channels to boost the performance. Results with fused feature are noted with subscripts "f". Because of the relatively high dimensionality of the feature space, we adopt kLFDA with linear kernel [43] as the metric learning algorithm for all experiments.

5.3. Method analysis

Patch size: We investigated the effect of the size of the patch on the performance. Figure 4(a) shows the results for different values. For a fair comparison, we keep the adjacent patches with 50% overlapping. We note that the performance decreases when the patch size is either too small or too large. On the one hand, there is not enough number of pixels within small patches to estimate the higher order moment matrix. Moreover, small patches tend to be less discriminant because they only model local information. As shown in the results, the rank 1 performance downgrades 6.3% when the patch size shrinks to 5×5. On the other hand, although large patches provide enough samples to estimate complex distributions, they encode specific pose and lose multi-view invariance. Therefore, for the remainder of the experiments we use a patch size of 16×16 as a compromise between the discriminating and invariant properties.

Normalization: Figure 4(b) illustrates the effects of applying different normalizations. By forcing the product of eigenvalues to be 1, a determinant normalization improves the result by 4.4%. Because most of the elements of higher order moments are small numbers, their logarithms are large negative values, which overwhelm the variance on that di-

Table 2. Comparing with state-of-the-art methods. The best results in each dataset are marked in red and the second best in blue.

Methods	Reference	VIPeR				CUHK01				PRID450s				GRID			
		r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
moM$_f^{LE}$+GOG$_f$+kLFDA[3]	**Ours**	53.3	82.3	89.5	94.8	**64.6**	**84.9**	90.6	94.8	**71.1**	**91.3**	**95.4**	**97.9**	24.5	46.1	56.8	66.9
HIPHOP + LOMO	PAMI17[5]	**54.2**	**82.4**	**91.5**	**96.9**	**78.8**	**92.6**	**95.3**	**97.8**	-	-	-	-	**26.0**	**50.6**	**62.5**	**73.3**
SSDAL + XQDA	ECCV16[8]	43.5	71.8	81.5	89.0	-	-	-	-	-	-	-	-	22.4	39.2	48.0	58.4
SCSP	CVPR16[4]	**53.5**	**82.6**	**91.5**	**96.7**	-	-	-	-	-	-	-	-	24.2	44.6	54.1	65.2
GOG$_f$ + XQDA	CVPR16[30]	49.7	79.7	88.7	94.5	57.9	79.2	86.2	92.1	**68.0**	**88.7**	**94.4**	**97.6**	**24.8**	**47.0**	**58.4**	**68.9**
TCP	CVPR16[6]	47.8	74.7	84.8	91.1	53.7	84.3	**91.0**	**96.3**	-	-	-	-	-	-	-	-
SS-SVM	CVPR16[47]	42.7	-	84.3	91.9	-	-	-	-	60.5	-	88.6	93.6	22.4	-	51.3	61.2
MLAPG	ICCV15[22]	40.7	-	82.3	92.4	-	-	-	-	-	-	-	-	16.6	-	41.2	53.0
Metric Ensumble	CVPR15[33]	45.9	77.5	88.9	**95.8**	53.4	76.4	84.4	90.5	-	-	-	-	-	-	-	-
LOMO+XQDA	CVPR15[21]	40.0	-	80.5	91.1	49.2	75.5	84.2	90.8	62.6	85.6	92.0	96.6	16.6	-	41.8	52.4
SCNCD	ECCV14[44]	37.8	68.5	81.2	90.4	-	-	-	-	41.6	68.9	79.4	87.8	-	-	-	-

Table 3. Comparing with state-of-the-art on Market1501 dataset.

Method	Reference	r=1	mAP
moM$_f^{LE}$+GOG$_f$	**Ours**	**71.6**	**43.5**
moM$_f^{LE}$	**Ours**	61.0	30.3
HIPHOP+LOMO	PAMI17[5]	**71.8**	**45.5**
GOG$_f$	CVPR16 [30] [3]	66.7	38.5
Gated S-CNN	ECCV16[40]	65.9	39.6
S-LST	ECCV16[41]	61.6	35.3
SSDAL+XQDA	ECCV16[8]	39.4	19.6
SCSP	CVPR16[4]	51.9	26.4
DNS	CVPR16[45]	55.4	29.9
BoW+KISSME	ICCV15[49]	44.4	20.8

mension. The mean removal step centers all dimensions while keeping the variance at the same time. The power normalization reduces the "spike" situation further more. Combining these two steps improves the rank 1 accuracy by 7.1%, but adding ℓ_2 normalization decreases it by 1.3%.

Moment matrix order: In Table 1, the first three rows show the results for $D = 1$. Comparing to the results from the following three rows for $D = 2$, the average rank1 performance on different on-manifold means increases by 7.4%, 9.8%, 16.3% and 5.5% on VIPeR, CUHK01, PRID450s and GRID, respectively. One can also observe a distinct margin between blue curves and other curves in Figure 5. This shows that the higher order moments are informative to boost the performance of the descriptor.

On-manifold metric: Besides the LE metric, there are several other metrics for the SPD manifold. Here, we also compare the performance using Jeffery Divergence (Jeff) [32] and Jensen-Bregman Log-det Divergence (JBLD) [7] on four datasets. Experimental results are shown in Table 1 with different superscripts.

$$\bar{\mathbf{M}}_{JBLD}^{(t+1)} = \left[\frac{1}{Q}\sum_{p=1}^{Q}\left(\frac{\mathbf{M}_p + \bar{\mathbf{M}}_{JBLD}^{(t)}}{2}\right)^{-1}\right]^{-1} \tag{10}$$

$$\bar{\mathbf{M}}_{Jeff} = \mathbf{P}^{-1/2}(\mathbf{P}^{1/2}\mathbf{Q}\mathbf{P}^{1/2})^{1/2}\mathbf{P}^{-1/2}, \text{with}$$
$$\mathbf{P} = \sum_{p=1}^{Q}\mathbf{M}_p^{-1}, \mathbf{Q} = \sum_{p=1}^{Q}\mathbf{M}_p. \tag{11}$$

When only using RGB as the color channel, although moM$_{rgb}^{Jeff}$ outperforms moM$_{rgb}^{LE}$ by 0.9% on VIPeR rank1 result, moM$_{rgb}^{LE}$ beats the others on CUHK01, PRID450s and GRID datasets. On average, moM$_{rgb}^{LE}$ achieves 2.3% and 1.3% higher rank1 performance along the four datasets compared with moM$_{rgb}^{Jeff}$ and moM$_{rgb}^{JBLD}$, respectively. When fusing with other color channels and GOG feature, moM with LE performs slightly better than JBLD.

5.4. Comparison with GOG descriptor

The results in Table 1 and Figure 5 compare the performances of the moM features, the GOG features and the combination of both of them. We ran the code provided by the authors of [30] and set the patch size to 15×15[2] and number of strips to 15. Among all four datasets, moM$_{\text{rgb}}^{\text{LE}}$ obtains slightly worse results in VIPeR and CUHK01, comparable result in PRID450s and better result in GRID. When fusing with all different color channels, moM$_{\text{f}}^{\text{LE}}$ performs worse in VIPeR and PRID450s, comparable in CUHK01 and better in GRID. However, by simply concatenating moM and GOG, a consistent outperformance can be achieved. In Figure 5, one can observe a clear margin between green and red curves and pink and black curves. Specifically, with the RGB color channel, adding moM$_{\text{rgb}}^{\text{LE}}$ to GOG$_{\text{rgb}}$ improves the rank 1 performance by 5.5%, 8.7%, 5.7% and 2.9%, respectively. After fusing with all different color channels, adding moM$_{\text{rgb}}^{\text{LE}}$ to GOG$_{\text{rgb}}$ further improve the rank 1 performance by 4.5%, 7.3%, 2.7% and 2.7%, respectively. The result implies that moM and GOG features encapsulate complementary appearance informations.

This observation can be explained by noting that the GOG feature has up to 2nd order information of the distribution representing the patches. However, it contains no information about the higher order (greater than 2) moments of these patches. On the other hand, the moM feature has information about the mean value (across patches) of the higher order moments, but not about their variance. Thus, one can think of the combination of GOG and moM as a

[2]The code provided can only accept an odd number as the patch size

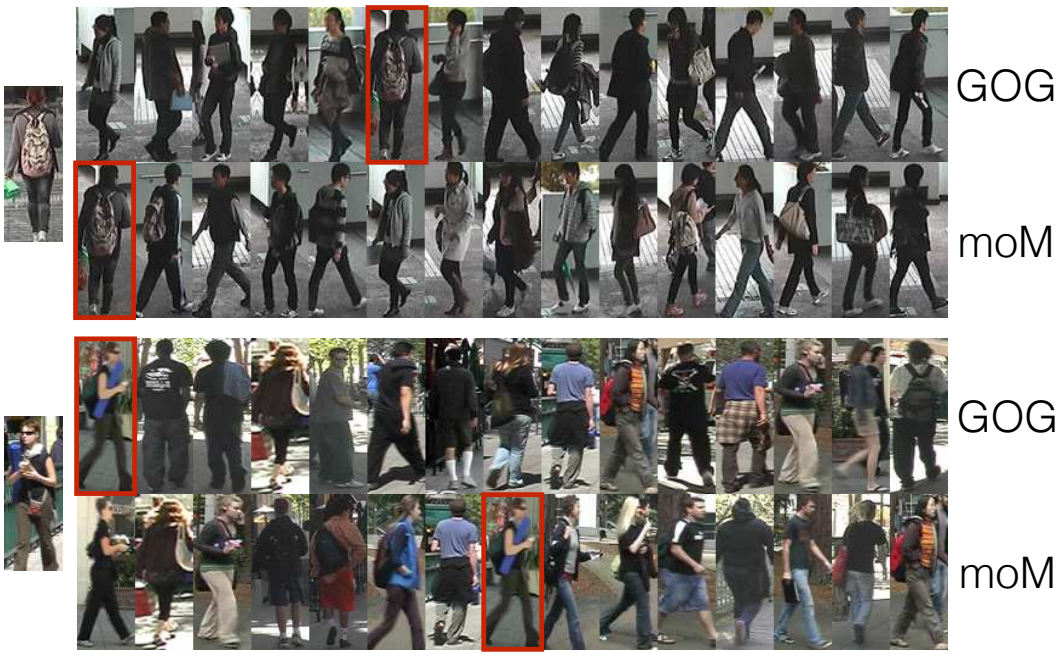


Figure 6. Examples for moM and GOG features. The very left image is the probe image and the first row on the right hand side is the result from GOG_{rgb} and the second row is from $\text{moM}_{\text{rgb}}{}^{\text{LE}}$. The correct match is labeled with a red box. The first example shows the situation moM feature is preferred while the second one shows the case GOG feature is better. Please see the text for more analysis.

tractable approximation to a "Moments of Moments" feature, where GOG provides information about the variance of 1st and 2nd order moments while moM provides information about the mean value of all moments (up to 4th order). For $\mathbf{x} \in \mathbb{R}^8$ this leads to a feature vector with $O(10^3)$ elements, as opposed to a true Moment of Moments feature ($D = 2$) that would have $O(10^6)$ elements.

Figure 6 gives two qualitative analysis examples. When the identity has fine-detailed appearance patterns, moM feature preserves those patterns better than the GOG feature. In the first example, moM feature captures the backpack with rich texture in the probe image and retrieves the gallery images with similar pattern to top two and finds the correct matching at rank 1. On the other hand, when the identity has homogeneous local texture but relatively complex patterns along the strip, GOG feature is preferred. In the second example, the strip level second-order moment helps to preserve the blue/black/skin color pattern along the upper body part.

5.5. Comparing with state-of-the-art methods

In Table 2, we compare the combination of $\text{moM}_{\text{f}}{}^{\text{LE}}$ and GOG_{f}[3] with recently published re-id methods. We achieve comparable performance on all four datasets and set the new state-of-the-art in PRID450s dataset. To show the generalization of moM on a large scale, automatic detected dataset, we compare with state-of-the-art works on the Market1501 dataset in Table 3. To be consistent with previous experiments, we report the result of GOG_f with the same setting in Table 1. By combining with our proposed moM feature, the complementary information brings a 4.9% improvement on rank 1 performance and increases by 5% on mAP comparing with GOG_f only.

[3]Please note that the GOG feature we used has a different setting from [30]

6. Conclusion

We proposed a novel mean of moment (moM) feature for the person re-id problem. The proposed feature generalizes the Gaussian assumption used in previous work, by using the empirical moment matrix and adopting the on-manifold mean to alleviate the cross-view variance. Extensive experimental results on five datasets illustrate that the moM and GOG features complement each other. The combination of both features achieves comparable or better performance on five public datasets.

In the future, instead of using an on-manifold mean, we plan to apply a more sophisticated second level pooling schema to model the global distribution of the descriptors of patches.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. *Differences*, 5:25, 2015.

[2] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 179–184. IEEE, 2011.

[4] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature aug- mentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (DOI: 10.1109/TPAMI.2017.2666805)*.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2161–2174, 2013.

[8] J. X. W. G. Chi Su, Shiliang Zhang and Q. Tian. Deep attributes driven person re-identification. In *European Conference on Computer Vision*. Springer, 2016.

[9] M. Faraki, M. T. Harandi, and F. Porikli. More about vlad: A leap from euclidean to riemannian manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4951–4960, 2015.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*, volume 1. Springer, 2014.

[12] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, O. Camps, and M. Sznaier. Person re-identification in appearance impaired scenarios. In *Proceedings of the British Machine Vision Conference 2016*. BMVA Press, 2016.

[13] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[14] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[15] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Computer Vision–ECCV 2012*, pages 780–793. Springer, 2012.

[16] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 720–729, 2015.

[17] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

[18] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2016.

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

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 152–159. IEEE, 2014.

[21] 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.

[22] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

[23] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1301–1306. IEEE, 2010.

[24] 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.

[25] M. Lovric, M. Min-Oo, and E. A. Ruh. Multivariate normal distributions parametrized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 74(1):36–48, 2000.

[26] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE, 2009.

[27] B. Ma, Q. Li, and H. Chang. Gaussian descriptor based on local features for person re-identification. In *Asian Conference on Computer Vision*, pages 505–518. Springer, 2014.

[28] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *British Machive Vision Conference*, pages 11–pages, 2012.

[29] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 413–422. Springer, 2012.

[30] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672. IEEE, 2012.

[32] M. Moakher and P. G. Batchelor. Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*, pages 285–298. Springer, 2006.

[33] 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.

[34] E. Pauwels and J. B. Lasserre. Sorting out typicality with the inverse moment matrix sos polynomial. In *Advances in Neural Information Processing Systems*, pages 190–198, 2016.

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

[36] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[37] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[38] 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.

[39] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1713–1727, 2008.

[40] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[41] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

[42] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.

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

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

[45] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[46] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[47] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[48] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2528–2535. IEEE, 2013.

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[50] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[51] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 649–656. IEEE, 2011.