Hierarchical Gaussian Descriptor for Person Re-Identification

Tetsu Matsukawa[1], Takahiro Okabe[2], Einoshin Suzuki[1], Yoichi Sato[3]
[1] Kyushu University [2] Kyushu Institute of Technology [3] The University of Tokyo
{matsukawa, suzuki}@kyushu-u.ac.jp, okabe@ai.kyutech.ac.jp, ysato@iis.u-tokyo.ac.jp

Abstract

Describing the color and textural information of a person image is one of the most crucial aspects of person re-identification. In this paper, we present a novel descriptor based on a hierarchical distribution of pixel features. A hierarchical covariance descriptor has been successfully applied for image classification. However, the mean information of pixel features, which is absent in covariance, tends to be major discriminative information of person images. To solve this problem, we describe a local region in an image via hierarchical Gaussian distribution in which both means and covariances are included in their parameters. More specifically, we model the region as a set of multiple Gaussian distributions in which each Gaussian represents the appearance of a local patch. The characteristics of the set of Gaussians are again described by another Gaussian distribution. In both steps, unlike the hierarchical covariance descriptor, the proposed descriptor can model both the mean and the covariance information of pixel features properly. The results of experiments conducted on five databases indicate that the proposed descriptor exhibits remarkably high performance which outperforms the state-of-the-art descriptors for person re-identification.

1. Introduction

Appearance matching of person images observed in disjoint camera views, referred to as person re-identification, is receiving increasing attention, mainly because of its broad range of applications [11]. In this task, the person images are captured from various viewpoints and under different illuminations, resolutions, human poses, and background environments. These large intra-personal variations in person images cause serious difficulties. In addition, similar clothes among different persons add further challenges.

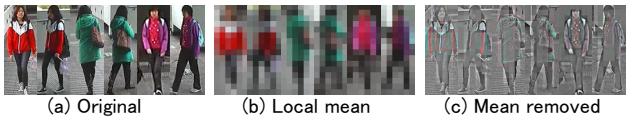
To address these difficulties, researchers are actively working on appearance descriptors [27, 28, 29, 43, 46, 47] and methods for matching them [22, 32, 33, 42, 48]. Descriptors characterize the appearance (color and textural) information of human clothes. A good descriptor should be robust against intra-personal variations and at the same time have high discriminative power to distinguish different persons.

Figure 1. Importance of hierarchal distribution: (a) Regions that have the same distribution (mean/covariance) of pixel features (each color indicates the same feature vector). (b) Local patches inside the regions which have different pixel feature distribution. (c) Regions can be distinguished via distributions of patch level distributions.

Person images are low in resolution and have large pose variations; consequently, it has been proved that the most important cue for person re-identification is color information such as color histograms and color name descriptors [43]. Because they cannot sufficiently differentiate different persons of similar color, textural descriptors such as Local Binary Pattern (LBP) and the responses of filter banks are often combined with color descriptors [33, 42, 48].

A covariance descriptor [40] describes a region of interest as a covariance of pixel features. It provides a natural way to fuse different modalities, *e.g.*, color and texture, of pixel features into a single meta-descriptor. Since the covariance descriptor is obtained by averaging features inside the region, it remedies the effects of noise and spatial misalignments. Consequently, it has been successfully applied to person re-identification [4, 5, 44].

In this paper, we propose a novel region descriptor based on hierarchical Gaussian distribution of pixel features for person re-identification. More specifically, we densely extract local patches inside a region and regard the region as a set of local patches. We firstly model the region as a set of multiple Gaussian distributions, each of which represents the appearance of one local patch. We refer to such a Gaussian distribution representing each local patch as a *patch Gaussian*. The characteristics of the set of patch Gaussians are again described by another Gaussian distribution. We refer to this Gaussian distribution as a *region Gaussian*. The parameters of the region Gaussian are then used as feature vector to represent the region.



(a) Original (b) Local mean (c) Mean removed

Figure 2. Importance of mean: (a) Original images. (b) Images that show mean RGB values of 10×10 pixel patches of (a). (c) Mean removed images (each RGB value is scaled over the range [0,255] for visualization). It is easy to determine the same persons from (b), whereas it is hard from (c).

Our motivation of the use of a hierarchical model stems from the appearance structure of person images. The persons' clothes consist of local parts, each of which has local color/texture structures. The spatial arrangement of these parts determines the global appearance structure. However, most of the existing meta descriptors [9, 10, 27, 30, 37, 40] are based on a global distribution of pixel features inside a region. Thereby, the local structure of the person image is lost. In contrast, our proposed descriptor describes the global distribution using the local distribution of the pixel features. Indeed, it can distinguish the textures which have the same global distribution but different local structures, as shown in Fig. 1.

We use the Gaussian distribution as a base component of the hierarchy. The motivation of the use of the distribution comes from the importance of the mean color of local parts. Although the hierarchical representation of covariance descriptors has been proposed [18, 36], the mean information is not included in each hierarchy. The loss of mean information is a crucial problem when they are applied to person re-identification. This is because the clothes a person wears tend to consist of a small number of colors in each local part, and therefore the mean color in the local parts tends to be the major discriminative information of the persons. As shown in Fig. 2, the mean images of local color contain highly distinguished information of different persons.

We name the proposed hierarchical method Gaussian Of Gaussian (GOG) descriptor. The GOG descriptor provides a conceptually simple and consistent way to generate discriminative and robust features that describe color and textural information simultaneously. The results of extensive experiments conducted on five public datasets reveal that, despite its simplicity, our proposed descriptor can achieve surprisingly high performance on person re-identification.

2. Related Work

Feature design and distance metric learning are two key components for person re-identification. In the feature design, several works have been conducted by focusing on the characteristic properties of person images. Symmetry-Driven Accumulation of Local Features (SDALF) [6] exploits the symmetric property of a person through obtaining head, torso, and leg positions to handle view variations. Unsupervised salience learning [46] estimates rare patches among different images, to perform matching of rare-appearances such as rare-colored coats, baggages and folders. Attribute based descriptors obtain lingual description of person images [17]. These works have been mainly conducted on unsupervised settings.

In the recent half decade, a supervised approach, *i.e.*, metric learning, has shown more impressive results in terms of accuracies [31, 32, 33, 23, 42, 48]. The features used for metric learning are rather simple compared to the features for the unsupervised settings. For metric learning, the features need not necessarily be robust or discriminative when unsupervised matching is performed, however, it requires to contain enough information within them. For example, high dimensional features composed of densely sampled color histograms, LBPs and SIFTs are often used [32, 42]. The design of features would largely affect the matching accuracy of metric learning methods. Nevertheless, most of the previous works focused on algorithm of metric learning [22, 24, 33, 48], and only few works focused on the feature design [23, 28, 43].

Our use of two-level (patch/region) statistics for person re-identification is motivated by the recently proposed Local Maximal Occurrence (LOMO) [23], which is a high dimensional representation of color and Scale Invariant Local Ternary Pattern (SILTP) histograms. This method locally constructs a histogram of pixel features, and then takes its maximum values within horizontal strips to overcome viewpoint variations while maintaining local discrimination. Indeed, LOMO describes only mean information of pixel features. Covariance-of-Covariance feature [36], where region covariance is estimated over local patch covariances of pixel features, motivated us to add covariance information in each hierarchy.

Making use of mean information to enhance the covariance descriptor is motivated by several works, such as Shape of Gaussians [10], Global Gaussian [30] and Gaussians of Local Descriptors (GOLD) [37]. By benefitting from the recent advances on Riemannian geometry, we treat a Gaussian distribution on a point of Symmetric Positive Definite (SPD) matrix manifold with the same manner as [13, 19] and apply log Euclidean metric and half-vectorization to flatten the manifold-valued data as in the works [9, 37]. Though such a Gaussian coding of low level pixel features is introduced into person re-identification [27], previous Gaussian descriptors are not constructed on a hierarchal manner.

Convolutional Neural Network (CNN) is one of the state-of-the art recognition algorithms that leverage hierarchal structure [16] and CNN has been recently adopted for person re-identification [1, 21]. However, their accuracies are not high compared to metric learning approaches, especially in small sampled datasets such as VIPeR [12]. This is because the CNN requires a large number of labeled train-

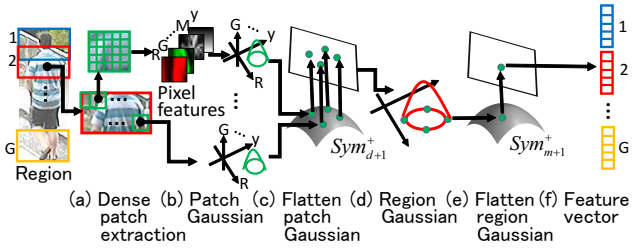


Figure 3. **GOG descriptor:** (a) For each region, we extract local patches densely. (b) We then describe each of local patches via a Gaussian distribution of pixel features, we refer to these Gaussians as patch Gaussians. (c) Each of patch Gaussians is flattened and vectorized by considering the underlining geometry of Gaussians. (d) Then the patch Gaussians inside a region are summarized into a region Gaussian. (e) We further flatten the region Gaussian and create a feature vector. (f) Finally, the feature vectors extracted from all regions are concatenated into one vector.

ing samples to obtain good performance. Although several descriptors are proposed by focusing on CNN like hierarchy [8, 39], they require learning processes for feature extraction in each hierarchy. In contrast, our descriptor requires no learning process because in each hierarchy, our descriptor describes regions via mean and covariance estimations which are not involved in learning.

3. Hierarchical Gaussian Descriptor

We outline the proposed hierarchal Gaussian descriptor named GOG in Fig. 3. To achieve the feature representation of a person image, we adopt a part-based model [35]. We assume that G regions of a person image are given in advance, which are typically horizontal stripes of the image. The proposed descriptor returns a feature vector of the regions. The rest of this section describes the details of the descriptor.

3.1. Pixel features

Let us focus on one of the G regions of a person image. To describe the local structure of the region, we densely extract squared (k $\times$ k pixels) patches with the p pixel intervals (Fig 3 (a)). In order to characterize each pixel in the patch, we extract d dimensional feature vector $\boldsymbol{f}_i$ for every pixel i. The feature vector can be any type of features, such as color, intensity, gradient orientation and filter response.

Since the number of pixels in each patch is small, the dimension d is preferable to be low for robustly estimating the covariance matrices of patch Gaussians in the next step. In this work, we extract 8-dimensional pixel features defined as:

$$\boldsymbol{f}_i = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, R, G, B]^T, \quad (1)$$

where y is the pixel location in the vertical direction, $M_{\theta \in \{0^\circ,\ldots,270^\circ\}}$ are the magnitudes of pixel intensity gradient along four orientations, and R, G, B are color channel values. Each dimension of $\boldsymbol{f}_i$ is linearly stretched to the range [0, 1] for equalizing the scales of the different feature values.

The pixel location is introduced to leverage spatial information within each region. The use of only vertical image location comes from the analysis in [27]; the person images tend to be well aligned in vertical direction while pose/viewpoint change causes a large misalignment in the horizontal direction. Note that one would like to set y_i from the top (or center) of the current region as in [9]. However, each pixel belongs to multiple regions and such a setting increases computational complexity. Since person images are coarsely aligned, we directly set y_i from the top of the image.

The gradient information is introduced to describe textural information of clothes. Gradient orientation $O = \arctan(\mathrm{I_y}/\mathrm{I_x})$ is calculated from x and y derivatives I_x, I_y of intensity I. We quantize the orientation into four bins; $O_{\theta \in \{0^\circ,90^\circ,180^\circ,270^\circ\}}$. To complement the loss of information by the quantization, we use soft voting into nearby two orientation bins. The voting weights are linearly determined from the distances from the quantized orientations as in the GO vector in [15]. To focus on high gradient edges, we multiply the gradient magnitude $M = \sqrt{I_y^2 + I_y^2}$ to the quantized orientation O_θ and obtain the oriented gradient magnitude; $M_\theta = MO_\theta$.

Color information is the most important cue for person re-identificaiton. We use the color channel values of the most basic color space: RGB. Other color spaces, *e.g.*, Lab, HSV and YCbCr, might be used. In fact, we will extend our pixel features in different color spaces (Sec. 3.5).

3.2. Patch Gaussians

After we extract the pixel features inside a patch, we then summarize them via the most classical parametric distribution which has mean and covariance as parameters: Gaussian distribution (Fig. 3 (b)). For every patch s, we model feature vectors as the patch Gaussian $\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_s \boldsymbol{\Sigma}_s)$ defined as,

$$\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1} (\boldsymbol{f} - \boldsymbol{\mu}_s)\right)}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_s|}, \quad (2)$$

where $|\cdot|$ is the determinant of a matrix, $\boldsymbol{\mu}_s$ is the mean vector and $\boldsymbol{\Sigma}_s$ is the covariance matrix of the sampled patch s. The mean vector and the covariance matrix are respectively estimated by: $\boldsymbol{\mu}_s = \frac{1}{n_s} \sum_{i \in \mathcal{L}_s} \boldsymbol{f}_i$ and $\boldsymbol{\Sigma}_s = \frac{1}{n_s - 1} \sum_{i \in \mathcal{L}_s} (\boldsymbol{f}_i - \boldsymbol{\mu}_s)(\boldsymbol{f}_i - \boldsymbol{\mu}_s)^T$, where $\mathcal{L}_s$ is the area of the sampled patch s and n_s denotes the number of pixels in $\mathcal{L}_s$.

Note that the densely sampled mean vectors and covariance matrices can be efficiently calculated through integral images [41]. Since regions can be overlapped, we construct

the integral images of pixel features for an overall person image rather than creating them for each region.

For a more precise description of distributions, Gaussian Mixture Model (GMM) might be used. Since a local patch is expected to consist of a small number of colors/textures, we assume that the unimodal Gaussian is sufficient for describing the distribution of its pixel features.

3.3. Tangent space mapping and half vectorization

As we will explain in the next subsection, our descriptor is a summarized representation of patch Gaussians in a region. For this summarization, mathematical operations such as mean or covariance of the Gaussian are required.

From the viewpoint of information geometry, the space of probability distribution is considered as a Riemannian manifold where the Euclidean operation cannot be applied directly [2]. A Riemannian manifold can be locally flatten into a Euclidean space by projecting it into a tangent space endowed with Riemannian metric. The space of the Symmetric Positive Definite (SPD) matrix is also considered as a Riemannian manifold and this space is recently well understood. The log Euclidean metric [3] for SPD matrix provides a solid way to map a point on the manifold to a Euclidean tangent space via a matrix logarithm.

To leverage the benefit of the log Euclidean metric, we embed the patch Gaussians in the SPD matrix in the similar manner to the work [19]. From an analysis in the information geometry literature [25], the space of d-dimensional multivariate Gaussians can be embedded into the $d+1$ dimensional SPD matrices space denoted by Sym^+_{d+1}. We represent the d-dimensional patch Gaussian $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ into Sym^+_{d+1} as $\boldsymbol{P}_s$:

$$\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \sim \boldsymbol{P}_s = |\boldsymbol{\Sigma}_s|^{-\frac{1}{d+1}} \begin{bmatrix} \boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T & \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_s^T & 1 \end{bmatrix}. \tag{3}$$

For more detailed theory of this embedding, one may refer the literature [25].

The covariance matrix of the local patch often becomes singular due to the lack of sufficient number of pixels within the patch. We avoid this problem by adding the identity matrix $\boldsymbol{I}_d$ to $\boldsymbol{\Sigma}_s$ with a small positive constant value, ϵ_s, as $\boldsymbol{\Sigma}_s \leftarrow \boldsymbol{\Sigma}_s + \epsilon_s \boldsymbol{I}_d$.

In order to describe the region distribution in a Euclidean operation, we then map each of patch Gaussians $\boldsymbol{P}_s$ into a tangent space via a matrix logarithm (Fig. 3 (c)).

We then store the upper triangular part of the mapped matrix as a vector since the matrix is symmetric. By considering the off-diagonal entries as being counted twice during the norm computation [41], the matrix of patch Gaussian $\boldsymbol{P}_s$ becomes $m = (d^2+3d)/2+1$ dimensional vector $\boldsymbol{g}_s$, defined as, $\boldsymbol{g}_s = \mathrm{vec}(\log(\boldsymbol{P}_s)) = [b_{s(1,1)}, \sqrt{2}b_{s(1,2)}, \cdots, \sqrt{2}b_{s(1,d+1)}, b_{s(2,2)}, \sqrt{2}b_{s(2,3)}, \cdots, b_{s(d+1,d+1)}]^T$, where $\log(\cdot)$ is the matrix logarithm operator and $b_{s(i,j)}$ is the (i,j) element of $\boldsymbol{B}_s = \log(\boldsymbol{P}_s)$.

3.4. Region Gaussian on tangent space

Due to the pose variation of person images, the positions of local parts vary in different observations. Thus we summarize the local patches into an orderless representation of them. More specifically, we summarize the flattened patch Gaussians in the previous section into a region distribution (Fig.3 (d)). For this summarization, we also use a Gaussian distribution that can describe not only covariance but also mean. Again, GMM might be used to describe more precise distributions. However, matching among GMMs is not a trivial problem [19] and will cause complexity to match among region descriptors. The summarization with a Gaussian distribution is performed by considering a spatial property of patches as follows.

A person image often contains background regions which significantly differ in places. To suppress the effect of background regions, we introduce a weight for each patch in a similar manner as for the weighted color histograms [6]. In most cases, the person is centered in each image; thus a higher value is assigned to the patches which are closer to the center y axis of an image: $w_s = \exp(-(x_s - x_c)^2/2\sigma^2)$ where $x_c = W/2$, $\sigma = W/4$. Here x_s denotes the x coordinate of the center pixel of patch s and W is the image width. Then we define the weighted mean vector and covariance matrix as

$$\boldsymbol{\mu}^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s \boldsymbol{g}_s, \tag{4}$$

$$\boldsymbol{\Sigma}^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s (\boldsymbol{g}_s - \boldsymbol{\mu}^{\mathcal{G}})(\boldsymbol{g}_s - \boldsymbol{\mu}^{\mathcal{G}})^T, \tag{5}$$

where $\mathcal{G}$ is the region in which the patch Gaussians are summarized. Using the mean vector and covariance matrix, we represent the region as the region Gaussian $\mathcal{N}(\boldsymbol{g}; \boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$.

For matching among region descriptors, it is convenient to flat the region Gaussian in the Euclidean space since most of the matching methods such as metric learning are designed on a Euclidean space. For this purpose, we embed m dimensional region Gaussian into $m+1$ dimensional SPD matrices in the same manner as Eq.(3): $\mathcal{N}(\boldsymbol{g}; \boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}) \sim \boldsymbol{Q}$ where $\boldsymbol{Q}$ is a $(m+1)\times(m+1)$ SPD matrix. Here the covariance matrix $\boldsymbol{\Sigma}^{\mathcal{G}}$ is regularized as $\boldsymbol{\Sigma}^{\mathcal{G}} \leftarrow \boldsymbol{\Sigma}^{\mathcal{G}} + \epsilon^{\mathcal{G}} \boldsymbol{I}_m$, We then map $\boldsymbol{Q}$ into the tangent space of Sym^+_{m+1} by using matrix logarithm and half-vectorize it to form a $(m^2+3m)/2+1$ dimensional feature vector, which we denote $\boldsymbol{z}$ (Fig.3 (e)).

By extracting the region Gaussian for each of G regions, we obtain feature vectors $\{\boldsymbol{z}_g\}_{g=1}^G$. In order to maintain the spatial location of these vectors, we concatenate them and form a feature vector (Fig.3(f)). Then the feature representation of a person image becomes $\boldsymbol{z} = [\boldsymbol{z}_1^T, .., \boldsymbol{z}_G^T]^T$.

3.5. Fusion descriptor of different color spaces

It has been proved that descriptors extracted from different color spaces have complementary properties, and their fusion improves re-identification accuracies [43].

To extract more color information in GOG descriptors, we replace the RGB channel values in the pixel feature in Eq.(1) with three alternative color channels values {Lab, HSV, nRGB} and fuse their GOG descriptors. Here the nRGB is the normalized color space (*e.g.*, nR = R/(R+G+B)). Since there is a redundancy in this space, we only use {nR, nG} in this color space. Thus, the pixel feature dimension of each {RGB, Lab, HSV, nRnG} color space is $d = \{8, 8, 8, 7\}$ and therefore the dimension of patch Gaussian vector is $m = \{45, 45, 45, 36\}$.

We denote the GOG descriptor $\boldsymbol{z}$ extracted from the Eq.(1) as $\mathrm{GOG}_{\mathrm{RGB}}$, and the descriptors extracted from the alternative color channels as $\mathrm{GOG}_{\mathrm{Lab}}$, $\mathrm{GOG}_{\mathrm{HSV}}$ and $\mathrm{GOG}_{\mathrm{nRnG}}$, respectively. The fusion is simply performed by concatenating GOG descriptors on different pixel features as $\mathrm{GOG}_{\mathrm{Fusison}}$ = [$\mathrm{GOG}^T_{\mathrm{RGB}}$, $\mathrm{GOG}^T_{\mathrm{Lab}}$, $\mathrm{GOG}^T_{\mathrm{HSV}}$, $\mathrm{GOG}^T_{\mathrm{nRnG}}$]^T. Therefore, the dimensionality of the fusion descriptor is 3 (color spaces) $\times$ 1081 (= $(45^2 + 3 \times 45)/2 + 1) \times G$ (regions) + 1 (color space) $\times$ 703 (= $(36^2 + 3 \times 36)/2 + 1) \times G$ (regions).

3.6. Normalization of GOG

For high dimensional features, normalization is an important factor to improve their performance [34]. Since the GOG descriptor is high dimensional, we normalize the descriptor by using the L2 norm normalization, which is the most widely adopted normalization.

We observed that there exist dimensions which have commonly high/small values among different images within the GOG descriptor. This is because we use pixel features which has different properties of its distributions, *e.g.*, gradient magnitude distributes sparsely in images, and color intensity distributes more uniformly. In such a case, the cosine distance, *i.e.*, the Euclidean distance after the normalization, would be dominated by the biased dimensions.

To remedy such biased dimensions, we remove the mean vector of training samples before normalizing the feature vector. The normalization of GOG becomes as follows:

$$\boldsymbol{z} = (\boldsymbol{z} - \overline{\boldsymbol{z}})/\|\boldsymbol{z} - \overline{\boldsymbol{z}}\|_2, \tag{6}$$

where $\overline{\boldsymbol{z}}$ is the sample mean of the GOG descriptors. For the fusion descriptor, we normalize each of the GOG descriptors extracted on four color spaces before concatenating them.

For the Bag-of-Words representation, similar normalization is proposed to reflect co-missing words for cosine similarity [14]. In contrast, we employ it to remedy the effect of biased dimensions.

4. Experiments

4.1. Setup

We evaluate the proposed descriptor on five benchmark datasets: VIPeR [12], CUHK01 [20], GRID [26], PRID450S [33] and CUHK03 [21]. We resize each image in the dataset to 128×48 pixels to facilitate the evaluation with the common parameters of the descriptor.

We extract the GOG descriptor from seven overlapping horizontal strips ($G = 7$). Each of the strips consists of 32×48 pixels. By considering the trade-off between the computational time and the predictive accuracy, we extract local patches at two-pixel intervals ($p = 2$) in each region. We set the local patch size to 5×5 pixels ($k = 5$). Following the setting of [13], we set the regularization parameter for region Gaussian as $\epsilon^{\mathcal{G}} = \epsilon_0 \mathrm{Tr}(\Sigma^{\mathcal{G}})$. Here $\mathrm{Tr}(\cdot)$ is the trace norm of the matrix. In several patch Gaussians, the trace norm of covariance matrix becomes nearly zero when the patch contains only nearly equal pixel values. Thus, we set a small constant to ϵ_s for patch Gaussian, then we have $\epsilon_s = \epsilon_0 \max(\mathrm{Tr}(\Sigma_s), 10^{-2})$. We set $\epsilon_0 = 10^{-3}$ for both patch and region Gaussians.

We evaluate the proposed descriptor with a distance metric learning, Cross-view Quadratic Discriminant Analysis (XQDA) [23]. The KISS Metric learning (KISSME) [33] is commonly used in person re-identification. However, it is more sensitive to dimensionality of subspace where the distance metric is learned. The XQDA learns a discriminative subspace and a distance metric simultaneously, and is able to select the optimal dimensionality automatically.

4.2. Performance analysis on VIPeR

In this section, we compare the performance within our approach using VIPeR dataset [12]. The VIPeR is a challenging dataset containing 632 person image pairs from two camera views. The testing protocol is to split the number of the person into half, 316 for training 316 for testing. We conduct the evaluation procedure for 10 splits and report the average Cumulative Matching Characteristic (CMC) curves.

As a default, we use RGB color space for pixel features ($\mathrm{GOG}_{\mathrm{RGB}}$) and the normalization in Sec. 3.6.

Distribution modeling: We compare other distribution models to GOG in Fig. 4(a). The Mean, Cov and Gauss are global distribution descriptors of pixel features within each region. The Cov-of-Cov, Cov-of-Gauss and GOG are hierarchical distribution descriptors. The tangent space mapping using log-Euclidean and half vectorization are applied for all descriptors except Mean. The regularization parameter of the covariance matrix is set as the same manner as GOG. The concatenated feature vector of the 7 regions is used for all descriptors. For a fair comparison, we adopted the weighted pooling for all descriptors.

First, we compare the global distribution descriptors. The rank-1 rates of Mean and Cov are 11.6% and 23.6%, respectively. By adding the mean and the covariance information, Gauss performs 7.7% better than Cov in rank-1 rate. This result confirms the importance of the use of both mean and covariance information of pixel features.

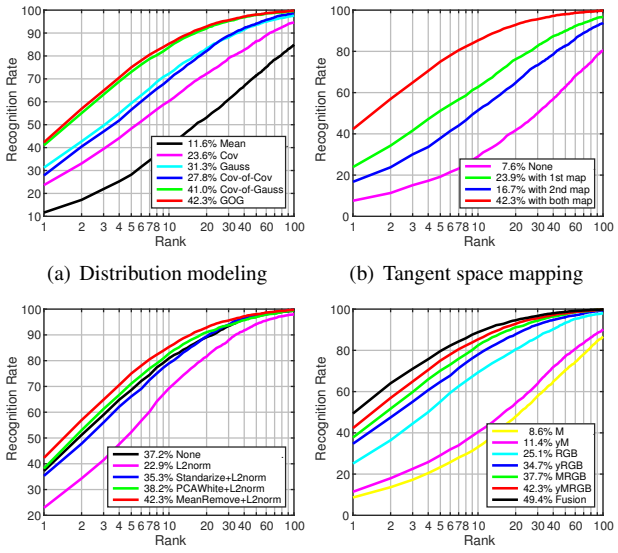
We then compare the hierarchical distribution descriptors. The Cov-of-Cov uses covariance matrix in both patch and region modeling, which is similar to [18, 36]. It performs 4.2% better than Cov in rank-1 rate. The Cov-of-Gauss uses Gaussian for patch and covariance matrix for region modeling. It improves the performance of Gauss by 9.7% in rank-1 rate. These results confirm the importance of covariance information of patch Gaussians. By adding mean information in region modeling, GOG improves the performance of Cov-of-Gauss by 1.3% in rank-1 rate.

Tangent space mapping: We compare the effect of flattening the manifold in Fig. 4 (b). The None shows the results when the tangent space mapping is not applied for constructing the vector of both region and patch Gaussians. The 1st and 2nd map respectively shows the results when the mapping is applied to one of the patch or region Gaussians. When either the 1st or 2nd mapping is applied, rank-1 rates increase by 16.3% and 9.1%, respectively. By applying the both mappings, the rank-1 rate increase by 34.7%. From these results, we can see that the consideration of the underlying geometry of Gaussian is necessity.

Normalization: We compare normalization in Fig. 4 (c). Due to the dimensions which have commonly high/small values among samples, the standard L2 norm degrades performance largely, 14.3% in rank-1 rate. We also compare the standardization and PCA whitening. For PCA whitening, we varied the dimension of PCA and the best results are reported. After applying the standardization or the whitening, L2 norms are normalized. The standardization drops rank-1 rate by 1.9% and the improvement by PCA whiting is small, 1.0% in rank-1 rate. We suspect these methods magnify the noise of dimensions where the standard deviations are small. We can set that the proposed normalization is most effective; it increases rank-1 rate by 5.1%.

Pixel features: We compare the components within pixel features in Fig. 4 (d). The color channel information, RGB, is more effective than the gradient magnitude information, M_θ, when comparing only these two components. By combining these two components, $M_\theta RGB$ achieves 12.6% better rank-1 rate than RGB alone. The use of vertical pixel location y also improves the performance, *e.g.*, rank-1 rate of $yM_\theta RGB$ is better than that of $M_\theta RGB$ by 4.6%. The $Fusion$ of GOG descriptors extracted from four color spaces is 7.1% better than $yM_\theta RGB$ in rank-1 rate [1].

[1]We also compared the GOG descriptor with pixel features used in other articles; 11-d [5] and 7-d [27] pixel features. In these cases, rank-1 identification rates were 35.4% and 39.7%, respectively.



(a) Distribution modeling (b) Tangent space mapping

(c) Normalization (d) Pixel features

Figure 4. Performance analysis of the GOG descriptor.

4.3. Performance comparison

We compare the GOG descriptor with other descriptors using the four datasets: VIPeR [12], CUHK01 [20], GRID [26] and PRID450S [33], respectively contains 632, 971, 250 and 450 images of individuals captured in two disjoint camera views.

The VIPeR, GRID and PRID450S datasets contain one image for each person in one camera view and CUHK01 contains two images. We conduct experiments with the *single shot* setting. Following the conventions [23, 31, 43], we randomly divide each dataset into training and test sets containing half of the available individuals. The number of probe images is equal to the gallery images in all datasets except GRID. For GRID, we add additional 775 images that do not belong to the person of 250 image pairs into the gallery set. We repeat the above evaluation procedure 10 times and obtain the average rank scores and also report the Proportion of Uncertainty Removed (PUR) [32] which is the measure to evaluate the whole ranks of CMC curve.

Other meta descriptors: We compare the GOG descriptor with other meta-descriptors: Heterogeneous Auto-Similarities of Characteristics (HASC) [7], Local Descriptors encoded by Fisher Vector (LDFV) [28], Second-order Average Pooling (2AvgP) [9] and GOLD [37].

The HASC is composed of the covariance descriptor and the Entropy and Mutual Information (EMI) descriptor. The EMI descriptor captures the non-linear dependency within pixel features and it has equal dimensionality to the covariance descriptor. The GOLD describes an image region by mean vector and covariance matrix. The covariance matrix is flattened by log-Euclidean and half-vectorization is applied. The vectors of mean and covariance are concatenated into a feature vector. The 2AvgP describes an image

Table 1. Comparison with XQDA metric learning (CMC@rank-r and PUR). (a) GOG descriptor. (b) Other meta descriptors. (c) Other descriptors for person re-identification. The best scores on (a) are shown in red, (b) and (c) are shown in blue.

	Methods	Pixel feature	# of Region	Dim.	Patch Weight	VIPeR r=1	VIPeR r=10	VIPeR r=20	VIPeR PUR	CUHK01 r=1	CUHK01 r=10	CUHK01 r=20	CUHK01 PUR	PRID450S r=1	PRID450S r=10	PRID450S r=20	PRID450S PUR	GRID r=1	GRID r=10	GRID r=20	GRID PUR
(a)	GOG_{Fusion}	Fusion	7	27,622	Y	**49.7**	**88.7**	**94.5**	**63.6**	**57.8**	**86.2**	**92.1**	**66.8**	**68.4**	**94.5**	**97.8**	**73.9**	**24.7**	**58.4**	**69.0**	**45.9**
	GOG_{RGB}	$yM_\theta RGB$	7	7,567	Y	42.3	85.3	92.8	58.8	55.8	85.5	91.3	65.8	63.6	91.5	96.2	69.8	22.8	52.3	64.1	43.1
	GOG_{Fusion}	Fusion	7	27,622	N	47.0	89.2	94.8	62.6	54.4	83.2	89.7	63.3	61.6	91.1	96.5	68.6	24.6	53.8	63.8	44.1
(b)	Cov-of-Cov [36]	Fusion	7	16,828	N	33.9	76.6	87.7	50.9	40.9	72.5	81.1	52.1	47.0	83.4	91.6	56.8	16.6	45.0	55.2	36.2
	GOLD [37]	Fusion	7	1,169	N	27.1	66.5	77.7	41.9	35.3	65.2	74.2	44.5	40.5	73.8	82.2	46.7	10.9	29.2	37.4	25.9
	2AvgP [9]	Fusion	7	952	N	28.8	68.5	79.2	43.3	36.1	68.1	76.3	46.2	44.7	75.8	83.8	49.8	12.9	36.7	47.4	30.3
	HASC [7]	Fusion	7	1,904	N	30.9	70.6	81.8	46.0	38.6	68.7	77.1	48.4	41.8	76.3	85.2	49.5	12.9	35.6	47.3	31.2
	LDFV [28]	Fusion	7	6,944	N	25.3	66.8	79.4	42.4	36.4	71.0	80.3	49.1	32.1	66.9	77.6	40.3	16.2	41.9	53.1	35.1
	Cov [40]	Fusion	7	952	N	26.9	65.8	77.1	41.2	34.5	64.5	73.6	43.8	40.4	73.4	82.1	46.4	10.6	29.0	36.7	25.5
(c)	LOMO [23]	CH+SILTP	40	26,960	N	41.1	82.2	91.1	56.8	49.2	84.2	90.8	62.4	62.6	92.0	96.6	69.4	17.9	46.3	56.2	36.9
	CH+LBP [42]	CH+LBP	75	32,250	N	27.7	69.3	82.4	45.0	31.3	70.4	81.5	48.8	21.5	60.8	74.4	35.0	16.2	45.0	57.1	36.7
	gBiCov [29]	BIF	-	5940	N	22.8	64.0	77.8	40.4	24.1	55.6	67.2	37.6	27.9	67.2	76.8	38.6	10.6	30.4	41.4	28.2

Table 2. Comparison of state-of-the-art results (CMC@rank-r). The best and second best scores are respectively shown in red and blue.

Methods	Reference	VIPeR r=1	VIPeR r=5	VIPeR r=10	VIPeR r=20	CUHK01 (M=1) r=1	CUHK01 (M=1) r=5	CUHK01 (M=1) r=10	CUHK01 (M=1) r=20	CUHK01 (M=2) r=1	CUHK01 (M=2) r=5	CUHK01 (M=2) r=10	CUHK01 (M=2) r=20	PRID450S r=1	PRID450S r=5	PRID450S r=10	PRID450S r=20	GRID r=1	GRID r=5	GRID r=10	GRID r=20
GOG_{Fusion}+XQDA	**Ours**	**49.7**	**79.7**	88.7	94.5	**57.8**	**79.1**	**86.2**	**92.1**	**67.3**	**86.9**	**91.8**	**95.9**	**68.4**	**88.8**	**94.5**	**97.8**	**24.7**	**47.0**	**58.4**	**69.0**
MetricEnsemble	CVPR2015 [31]	45.9	77.5	**88.9**	**95.8**	53.4	76.4	84.4	90.5	-	-	-	-	-	-	-	-	-	-	-	-
LOMO+XQDA	CVPR2015 [23]	40.0	-	80.5	91.1	49.2	75.7	84.2	90.8	63.2	-	90.8	94.9	62.6	85.6	92.0	96.6	16.6	-	41.8	52.4
SCNCD	ECCV2014 [43]	37.8	68.5	81.2	90.4	-	-	-	-	-	-	-	-	41.6	68.9	79.4	87.8	-	-	-	-
Semantic	CVPR2015 [38]	31.1	68.6	82.8	94.9	32.7	51.2	64.4	76.3	-	-	-	-	43.1	70.5	78.2	86.2	-	-	-	-
SalMatch	ICCV2013 [45]	30.2	52	65	-	28.5	45	55	-	-	-	-	-	-	-	-	-	-	-	-	-
MLFL	CVPR2014 [47]	29.1	-	65.9	70.9	34.3	55	65	75	-	-	-	-	-	-	-	-	-	-	-	-

region by the zero-mean covariance matrix, and applies log-Euclidean and half-vectorization to obtain a feature vector. The LDFV encodes pixel features using Fisher Vector coding, which encodes difference of pixel features from pre-trained GMM means. By following the recommended setting [28], we set the number of GMM components to 16 [2].

We focus on the encoding process of pixel features only, and discard other options on the above descriptors, such as the spatial pyramid in GOLD. We extract each of the meta-descriptors from the same horizontal strips as GOG. The descriptor extracted from the 7 regions are concatenated. As well as GOG, we use the fusion approach that concatenates the meta descriptors extracted from 4 pixel feature vectors into one vector. For normalization, the mean removal and the L2 normalization are applied to each descriptor since we found it generally improves their performances.

We list the performance of GOG and the compared meta descriptors in Table 1 (a) and (b). All the descriptors in (b) except Cov-of-Cov are not hierarchical descriptors, which discard local structures of regions. The descriptors which use single layered distribution (Cov, HASC, LDFV, 2AvgP and GOLD) have similar performances. On the other hand, Cov-of-Cov clearly outperforms them. These results confirm the effectiveness of the hierarchical distribution. The GOG outperforms Cov-of-Cov, since it also contains the mean information, which is absent in covariance.

[2] It is reported that their 7-d pixel feature produces much better results than FV on the SIFT descriptors which is widely used in image classification [34]. When their pixel features were adopted to the LDFV in our settings, the rank1 recognition rates were 24.8%, 28.2%, 20.2% and 10.5% in VIPeR, CUHK01, PRID450s and GRID dataset, respectively.

Descriptors for metric learning: We compare the GOG descriptor with other descriptors used in metric learning for person re-identification: LOMO [23], Color Histogram (CH)+LBP [42] and gBiCov [29]. For these descriptors, we use the source codes provided by the authors. The default parameters of the codes are used for LOMO and gBiCov. Xiong et al. [42] conducted experiments using different region numbers to extract 28 bin color histogram and 2 uniform LBPs. Among them, we use 75 regions that was the best setting. For a fair comparison, the same normalization as GOG and XQDA metric learning are commonly applied.

The experimental results are shown in the Table 1 (c). It can be shown that GOG_{Fusion} clearly outperforms LOMO with nearly equal dimensionality. The rank-1 identification rates of GOG_{Fusion} are 8.6%, 8.6%, 5.8% and 8.1% better in VIPeR, CUHK01, PRID450s and GRID datasets, respectively. Although, LOMO and CH+LBP use more spatial regions and high dimensional pixel features, the GOG descriptor outperforms these descriptors by a large margin. When the patch weights are used, GOG_{RGB}, which is extracted from pixel features with only RGB color information, outperforms LOMO with a much smaller dimensionality. The superiority of the GOG descriptor comes from its hierarchal use of the mean and covariance information of pixel features, whereas LOMO uses only the mean information.

State-of-the-arts: In Table 2, we compare the performance of the reported results on the state-of-the-art methods, including MidLevel Filter Learning (MLFL) [47], Salience Matching (SalMatch) [45] , SCNCD [43], Semantic attribute representation [38], Metric Ensemble [31]

Table 3. State-of-the-art results on CUHK03 (CMC@rank-r).

Methods	Reference	Labeled r=1	Labeled r=5	Labeled r=10	Detected r=1	Detected r=5	Detected r=10
GOG_{Fusion}+XQDA	**Ours**	**67.3**	**91.0**	**96.0**	**65.5**	**88.4**	**93.7**
MetricEmsemble	CVPR2015 [31]	62.1	89.1	94.3	-	-	-
LOMO+MLAPG	ICCV2015 [24]	58.0	-	-	51.2	-	-
LOMO+XQDA	CVPR2015 [23]	52.2	-	-	46.3	-	-
ImprovedDeep	CVPR2015 [1]	54.7	88.3	93.3	45.0	75.7	83.0
DeepReID	CVPR2014 [21]	20.7	51.7	68.3	19.9	49.0	64.3

(a) Labeled (b) Detected

Figure 5. Example images from CUHK03 dataset [21]. Images in the same column represent the same person.

and LOMO [23] [3]. It can be observed that the GOG descriptor achieves the new state-of-the-art results, 49.7%, 57.8%, 67.3%, 68.4% and 24.7% of rank-1 rate on VIPeR, CUHK01 (M=1), CUHK01 (M=2), PRID450S and GRID dataset, respectively. Since GOG and LOMO adopt the common metric learning, it is clear that the success of our approach comes from our design of a better feature descriptor. The metric ensemble [31] uses four base metrics, each of them is learned on SIFT, color histogram + LBP, covariance descriptor and CNN. Our descriptor also outperforms such an ensemble of different descriptors.

4.4. Comparison on automatic detected dataset

To show the generality of the GOG descriptor on a large and automatic detected dataset, we compare the performances on CUHK03 dataset [21]. The CUHK03 dataset includes 13,164 images of 1,360 persons, captured by disjoint camera views. Each person in the dataset has in average 4.8 images in each view. In addition to manually cropped person images, the dataset contains images detected by the state-of-the-art person detector. Therefore, realistic variations such as misalignment, occlusions and missing body parts are contained in person images. Fig. 5 shows some example images of the dataset.

We evaluate the GOG descriptor with the common setting to previous works [1, 21, 23, 24, 31]. Namely, we divide the images of the dataset into 1,160 persons for the train set and 100 persons for the test set. The random division are repeated 20 times and we report the average results.

Table 3 lists the performance comparison with the state-of-the-art results. The GOG_{Fusion} achieves 67.3% and 65.5% rank-1 identification rates with the labeled and the automatically detected bounding boxes, respectively, which clearly outperforms the state-of-the-art LOMO features [23, 24] and deep learning methods [1, 21] by a large margin. The performance decrease in rank-1 rate between labeled and detected data is 1.8% in the case of GOG, which is more than three times smaller than 5.95% of LOMO+XQDA. This might be because the LOMO feature is extracted from narrower horizontal stripes than regions of GOG. High dimensionality of LOMO is partially due to such a large number of narrow horizontal strips. In contrast, the high dimensionality of GOG is due to the Gaussian matrix, which is composed of the mean vector and the covariance matrix. Such a dimension enhancement of pixel features does not decrease the robustness to misalignment, and thus the GOG descriptor is more preferable in realistic situations with misalignments of person images.

[3] The experimental setting of CUHK01 (M=1) is the single shot setting, which is common to [31] and CUHK01 (M=2) is the multi shot setting in [23]. The results of the CUHK01(M=1) and PRID450S datasets of LOMO are obtained from the code provided by the author.

Table 4. Time of feature extraction (seconds/image).

LOMO [23]	Cov_{RGB}	GOG_{RGB}	GOG_{Fusion}	gBiCov [29]
0.016	0.021	0.34	1.34	7.8

4.5. Running time

The GOG descriptor is implemented in Matlab [4] with MEX function for calculation of covariance matrices, and run on a PC equipped with Intel Xeon E5-2687W @3.1GHz CPU. The running times of the descriptors are shown in Table 4. The listed times are the average of all images of the VIPeR dataset. The matching cost of GOG_{Fusion} is nearly equal to LOMO since their dimensionalities are almost the same. The GOG descriptor is about 16 times slower than the covariance descriptor when the same pixel feature is used, and GOG_{Fusion} is about 84 times slower than LOMO. However, it is 5.8 times faster than gBiCov. Considering other methods which require more computational cost [6, 45, 47], the running time of the GOG descriptor is still appealing.

5. Conclusions

We have proposed a novel hierarchical Gaussian descriptor for person re-identification. The proposed descriptor models both mean and covariance information of pixel features in each of the patch and region hierarchies. The results of our extensive experiments revealed that the proposed descriptor can achieve surprisingly high performance which improves the state-of-the-art performances on five public datasets.

In our future work, we plan to investigate the deep hierarchy of Gaussian descriptors to describe more in-depth the hierarchical structure of person appearances. In addition, we would like to test ensembles of the GOG descriptors extracted from different kinds of pixel features for further improvements of identification accuracies.

[4] http://www.i.kyushu-u.ac.jp/~matsukawa/ReID/

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3908–3916, 2015. 2, 8

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*. volume 191 of Translations of mathematical monographs. American Mathematical Society, 2001. 4

[3] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Analysis Applications*, 29(1):328–347, 2006. 4

[4] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *European Conference on Computer Vision (ECCV)*, volume 7574, pages 806–820, 2012. 1

[5] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Boosted human re-identification using Riemannian manifolds. *Image Vision Computing*, 30(6-7):443–452, 2012. 1, 6

[6] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013. 2, 4, 8

[7] M. S. Biagio, M. Crocco, M. Cristani, S. Martelli, and V. Murino. Heterogeneous auto-similarities of characteristics (HASC): exploiting relational information for classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 809–816, 2013. 6, 7

[8] L. Bo, K. Lai, X. Ren, and D. Fox. Object recognition with hierarchical kernel descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1729–1736, 2011. 3

[9] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Free-form region description with second-order pooling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1177–1189, 2015. 2, 3, 6, 7

[10] L. Gong, T. Wang, and F. Liu. Shape of Gaussians as feature descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2366–2371, 2009. 2

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition, Springer, 2014. 1

[12] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision (ECCV)*, pages 262–275, 2008. 2, 5, 6

[13] 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 *International Conference on Machine Learning (ICML)*, pages 720–729, 2015. 2, 5

[14] H. Jégou and O. Chum. Negative evidences and cooccurences in image retrieval: The benefit of PCA and whitening. In *European Conference on Computer Vision (ECCV)*, pages 774–787, 2012. 5

[15] T. Kobayashi and N. Otsu. Image feature extraction using gradient local auto-correlations. In *European Conference on Computer Vision (ECCV)*, pages 346–358, 2008. 3

[16] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 2

[17] R. Layne, T. M. Hospedales, and S. Gong. Person re-identification by attributes. In *British Machine Vision Conference, (BMVC)*, pages 1–11, 2012. 2

[18] P. Li and Q. Wang. Local log-Euclidean covariance matrix (L2ECM) for image representation and its applications. In *European Conference on Computer Vision (ECCV)*, pages 469–482, 2012. 2, 6

[19] P. Li, Q. Wang, and L. Zhang. A novel earth mover's distance methodology for image matching with Gaussian mixture models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1689–1696, 2013. 2, 4

[20] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision (ACCV)*, pages 31–44, 2012. 5, 6

[21] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159, 2014. 2, 5, 8

[22] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3610–3617, 2013. 1, 2

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, 2015. 2, 5, 6, 7, 8

[24] S. Liao and S. Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *The IEEE Conference on Computer Vision (ICCV)*, pages 3685–3693, 2015. 2, 8

[25] M. Lovrić, 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. 4

[26] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010. 5, 6

[27] B. Ma, Q. Li, and H. Chang. Gaussian descriptor based on local features for person re-identification. In *Asian Conference on Computer Vision (ACCV) Workshop*, pages 505–518, 2014. 1, 2, 3, 6

[28] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by Fisher vectors for person re-identification. In *European Conference on Computer Vision (ECCV) Workshop*, pages 413–422, 2012. 1, 2, 6, 7

[29] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face

verification. *Image and Vision Computing*, 32(6):379–390, 2014. 1, 7, 8

[30] H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian approach for scene categorization using information geometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2336–2343, 2010. 2

[31] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1846–1855, 2015. 2, 6, 7, 8

[32] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian. Local Fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, 2013. 1, 2, 6

[33] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267, 2014. 1, 2, 5, 6

[34] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 5, 7

[35] R. Satta. Appearance descriptors for person re-identification: a comprehensive review. *CoRR*, abs/1307.5748, 2013. 3

[36] G. Serra, C. Grana, M. Manfredi, and R. Cucchiara. Covariance of covariance features for image classification. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, page 411, 2014. 2, 6, 7

[37] G. Serra, C. Grana, M. Manfredi, and R. Cucchiara. GOLD: Gaussians of local descriptors for image representation. *Computer Vision and Image Understanding*, 134:22–32, 2015. 2, 6, 7

[38] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4184–4193, 2015. 7

[39] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. In *Neural Information Processing Systems (NIPS)*, pages 163–171, 2013. 3

[40] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision (ECCV)*, pages 589–600, 2006. 1, 2, 7

[41] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(10):1713–1727, 2008. 3, 4

[42] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2014. 1, 2, 7

[43] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 536–551, 2014. 1, 2, 5, 6, 7

[44] M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu. Efficient person re-identification by hybrid spatiogram and covariance descriptor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 48–56, 2015. 1

[45] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2528–2535, 2013. 7, 8

[46] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3593, 2013. 1, 2

[47] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 144–151, 2014. 1, 7, 8

[48] W. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013. 1, 2