Re-identification by Covariance Descriptors

Slawomir Bak, François Bremond

Chapter 3
Re-identification by Covariance Descriptors

Sławomir Bąk and François Brémond

Abstract This chapter addresses the problem of appearance matching, while employing the covariance descriptor. We tackle the extremely challenging case in which the same non-rigid object has to be matched across disjoint camera views. Covariance statistics averaged over a Riemannian manifold are fundamental for designing appearance models invariant to camera changes. We discuss different ways of extracting an object appearance by incorporating various training strategies. Appearance matching is enhanced either by discriminative analysis using images from a single camera or by selecting distinctive features in a covariance metric space employing data from two cameras. By selecting only essential features for a specific class of objects (*e.g.* humans) without defining *a priori* feature vector for extracting covariance, we remove redundancy from the covariance descriptor and ensure low computational cost. Using a feature selection technique instead of learning on a manifold, we avoid the over-fitting problem. The proposed models have been successfully applied to the person re-identification task in which a human appearance has to be matched across non-overlapping cameras. We carry out detailed experiments of the suggested strategies, demonstrating their pros and cons *w.r.t.* recognition rate and suitability to video analytics systems.

3.1 Introduction

The present work addresses the re-identification problem that consists in appearance matching of the same subject registered by non-overlapping cameras. This task is particularly hard due to camera variations, different lighting conditions, different color responses and different camera viewpoints. Moreover, we focus on non-rigid

Sławomir Bąk
INRIA, Sophia Antipolis, France. e-mail: slawomir.bak@inria.fr

François Brémond
INRIA, Sophia Antipolis, France. e-mail: francois.bremond@inria.fr

objects (*i.e.* humans) that change their pose and orientation contributing to the complexity of the problem.

In this work we design two methods for appearance matching across non-overlapping cameras. One particular aspect is a choice of an image descriptor. A good descriptor should capture the most distinguishing characteristics of an appearance, while being invariant to camera changes. We offer to describe an object appearance by using the *covariance descriptor* [26] as its performance is found to be superior to other methods (section 3.3).

By averaging descriptors on a Riemannian manifold, we incorporate information from multiple images. This produces *mean Riemannian covariance* (section 3.3.2) that yields a compact and robust representation.

Having an effective descriptor, we design efficient strategies for appearance matching. The first method assumes a predefined appearance model (section 3.4.2), introducing discriminative analysis, which can be performed online. On the other hand, the second technique learns an appearance representation during an offline stage, guided by an entropy-driven criterion (section 3.4.3). This removes redundancy from the descriptor and ensures low computational cost.

We carry out detailed experiments of proposed methods (section 13.5), while investigating their pros and cons *w.r.t.* recognition rate and suitability to video analytics systems.

3.2 Related work

Recent studies have focused on the appearance matching problem in the context of pedestrian recognition. Person re-identification approaches concentrate either on *metric learning* regardless of the representation choice, or on *feature modeling*, while producing a distinctive and invariant representation for appearance matching.

Metric learning approaches use training data to search for strategies that combine given features maximizing inter-class variation whilst minimizing intra-class variation. These approaches do not pay too much attention to a feature representation. In the result, metric learning techniques use very simple features such as color histograms or common image filters [10, 21, 30]. Moreover, for producing robust metrics, these approaches usually require hundreds of training samples (image pairs with the same person/object registered by different cameras). It raises numerous questions about practicability of these approaches in a large camera network.

Instead, feature-oriented approaches concentrate on an invariant representation, which should handle view point and camera changes. However, these approaches usually do not take into account discriminative analysis [6, 5, 14]. In fact, learning using a sophisticated feature representation is very hard or even unattainable due to a complex feature space.

It is relevant to mention that both approaches proposed in this work belong more to feature-oriented approaches as they employ the covariance descriptor [26]. The covariance matrix can be seen as a meta descriptor that can fuse efficiently different

types of features and their modalities. This descriptor has been extensively used in the literature for different computer vision tasks.

In [27] covariance matrix is used for designing a robust human detection algorithm. Human appearance is modeled by a dense set of covariance features extracted inside a detection window. Covariance descriptor is computed from sub-windows with different sizes sampled from different locations. Then, a boosting mechanism selects the best regions characterizing a human silhouette.

Unfortunately, using covariance matrices, we also influence significantly computational complexity. This issue has been addressed in [28]. The covariance matrices of feature subsets rather than the full feature vector, provide similar performance while significantly reducing the computation load.

Covariance matrix has also been successfully applied to tracking. In [23] object deformations and appearance changes were handled by a model update algorithm using the Lie group structure of the positive definite matrices.

The first approach which employs the covariance descriptor for appearance matching across non-overlapping cameras is [2]. In this work an HOG-based detector establishes the correspondence between body parts, which are matched using a spatial pyramid of covariance descriptors.

In [22] we can find biologically inspired features combined with the similarity measure of covariance descriptors. The new descriptor is not represented by the covariance matrix but by a distance vector computed using the similarity measure between covariances extracted at different resolution bands. This method shows promising results not only for person re-identification but also for face verification.

Matching groups of people by covariance descriptor is the main topic of [7]. It is shown that contextual cues coming from group of people around a person of interest can significantly improve the re-identification performance. This contextual information is also kept by the covariance matrix.

In [4] the authors use *one-against-all* learning scheme to enhance distinctive characteristic of covariances for a specific individual. As covariances do not live on Euclidean space, binary classification is performed on a Riemannian manifold. Tangent planes extracted from positive training data points are used as a classification space for a boosting algorithm. Similarly, in [19] discriminative models are learned by a boosting scheme. However, covariance matrices are transformed into *Sigma Points* to avoid learning on a manifold, which often produces a over-fitted classifier.

Although discriminative approaches show promising results, they are usually computationally intensive, which is unfavorable in practice. In general, discriminative methods are also accused of non-scalability. It may be noted that an extensive learning phase is necessary to extract discriminative signatures at every instant when a new person is added to the set of existing signatures. This updating step makes these approaches very difficult to apply in a real world scenario.

In this work we overcome the mentioned issues twofold: (1) by offering an efficient discriminative analysis, which can be performed online even in a large camera network or (2) by an offline learning stage, which learns a general model for ap-

pearance matching. Using a feature selection technique instead of learning on a manifold, we avoid the over-fitting problem.

3.3 Covariance descriptor

In [26] covariance of d-features has been proposed to characterize an image region. The descriptor encodes information on feature variances inside the region, their correlations with each other and their spatial layout. It can fuse different types of features, while producing a compact representation. The performance of the covariance descriptor is found to be superior to other methods, as rotation and illumination changes are absorbed by the covariance matrix.

Covariance matrix can be computed from any type of image such as a one dimensional intensity image, three channel color image or even other types of images, *e.g.* infrared.

Let I be an image and F be a d-dimensional feature image extracted from I

$$F(x,y) = \phi(I,x,y), \tag{3.1}$$

where function ϕ can be any mapping, such as color, intensity, gradients, filter responses, *etc*. For a given rectangular region $Reg \subset F$, let $\{f_k\}_{k=1\ldots n}$ be the d-dimensional feature points inside Reg (n is the number of feature points, *e.g.* the number of pixels). We represent region Reg by the $d \times d$ covariance matrix of the feature points

$$C_{Reg} = \frac{1}{n-1}\sum_{k=1}^{n}(f_k - \mu)(f_k - \mu)^T, \tag{3.2}$$

where μ is the mean of the points.

Such a defined positive definite and symmetric matrix can be seen as a tensor. The main problem is that such defined tensor space is a manifold that is not a vector space with the usual additive structure (does not lie on Euclidean space). Hence, many usual operations, such as *mean* or *distance*, need a special treatment. Therefore, the covariance manifold is often specified as Riemannian to determine a powerful framework using tools from differential geometry [24].

3.3.1 Riemannian geometry

A manifold is a topological space which is locally similar to a Euclidean space. It means that every point on the m-dimensional manifold has a neighborhood homeo-

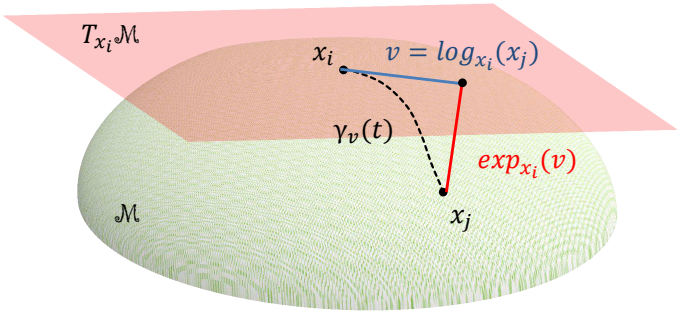


Fig. 3.1: An example of a two-dimensional manifold. We show the tangent plane at x_i, together with the exponential and logarithm mappings related to x_i and x_j [16].

morphic to an open subset of the m-dimensional space $\Re^m$. Performing operations on the manifold involves choosing a metric.

Specifying manifold as Riemannian gives us Riemannian metric. This automatically determines a powerful framework to work on the manifold by using tools from differential geometry [24]. Riemannian manifold $\mathcal{M}$ is a differentiable manifold in which each tangent space has an inner product which varies smoothly from point to point. Since covariance matrices can be represented as a connected Riemannian manifold, we apply operations such as *distance* and *mean* computation using this differential geometry.

Figure 3.1 shows an example of a two-dimensional manifold, a smooth surface living in $\Re^3$. Tangent space $T_x\mathcal{M}$ at x is the vector space that contains the tangent vectors to all 1-D curves on $\mathcal{M}$ passing through x. Riemannian metric on manifold $\mathcal{M}$ associates to each point $x \in \mathcal{M}$, a differentiable varying inner product $\langle \cdot, \cdot \rangle_x$ on tangent space $T_x\mathcal{M}$ at x. This induces a norm of tangent vector $v \in T_x\mathcal{M}$ such that $\|v\|_x^2 = \langle v, v \rangle_x$. The minimum length curve over all possible smooth curves $\gamma_v(t)$ on the manifold between x_i and x_j is called *geodesic*, and the length of this curve stands for geodesic distance $\rho(x_i, x_j)$.

Before defining geodesic distance, let us introduce the exponential and the logarithm functions, which take as an argument a square matrix. The exponential of matrix W can be defined as the series

$$\exp(W) = \sum_{k=0}^{\infty} \frac{W^k}{k!}. \tag{3.3}$$

In the case of symmetric matrices, we can apply some simplifications. Let $W = U\ D\ U^T$ be a diagonalization, where U is an orthogonal matrix, and $D = DIAG(d_i)$ is the diagonal matrix of the eigenvalues. We can write any power of W in the same way $W^k = U\ D^k\ U^T$. Thus

$$\exp(W) = U\ DIAG(\exp(d_i))\ U^T, \tag{3.4}$$

and similarly the logarithm is given by

$$\log(W) = U\ DIAG(\log(d_i))\ U^T. \tag{3.5}$$

According to a general property of Riemannian manifolds, geodesics realize a local diffeomorphism from the tangent space at a given point of the manifold to the manifold. It means that there is the mapping which associates to each tangent vector $v \in T_x\mathcal{M}$ a point of the manifold. This mapping is called the exponential map, because it corresponds to the usual exponential in some matrix groups.

The exponential and logarithmical mappings have the following expressions [24]:

$$\exp_\Sigma(W) = \Sigma^{\frac{1}{2}} \exp(\Sigma^{-\frac{1}{2}} W \Sigma^{-\frac{1}{2}}) \Sigma^{\frac{1}{2}}, \tag{3.6}$$

$$\log_\Sigma(W) = \Sigma^{\frac{1}{2}} \log(\Sigma^{-\frac{1}{2}} W \Sigma^{-\frac{1}{2}}) \Sigma^{\frac{1}{2}}, \tag{3.7}$$

where

$$\Sigma^{\frac{1}{2}} = \exp\left(\frac{1}{2}(\log(\Sigma))\right). \tag{3.8}$$

Given tangent vector $v \in T_{x_i}\mathcal{M}$, there exists a unique geodesic $\gamma_v(t)$ starting at x_i (see figure 3.1). The exponential map $\exp_{x_i} : T_{x_i}\mathcal{M} \to \mathcal{M}$ maps tangent vector v to the point on the manifold that is reached by this geodesic. The inverse mapping is given by logarithm map denoted by $\log_{x_i} : \mathcal{M} \to T_{x_i}\mathcal{M}$. For two points x_i and x_j on manifold $\mathcal{M}$, the tangent vector to the geodesic curve from x_i to x_j is defined as $v = \overrightarrow{x_i x_j} = \log_{x_i}(x_j)$, where the exponential map takes v to the point $x_j = \exp_{x_i}(\log_{x_i}(x_j))$. The Riemannian distance between x_i and x_j is defined as $\rho(x_i, x_j) = \|\log_{x_i}(x_j)\|_{x_i}$. It is relevant to note that an equivalent form of geodesic distance can be given in terms of generalized eigenvalues [13].

The distance between two symmetric positive definite matrices C_i and C_j can be expressed as

$$\rho(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)}, \tag{3.9}$$

where $\lambda_k(C_i, C_j)_{k=1\ldots d}$ are the generalized eigenvalues of C_i and C_j, determined by

$$\lambda_k C_i x_k - C_j x_k = 0, \quad k = 1 \ldots d \tag{3.10}$$

and $x_k \neq 0$ are the generalized eigenvectors.

We have already mentioned that we are more interested in extracting covariance statistics from several images rather than from a single image. Then, having a suitable metric, we can define *mean Riemannian covariance*.

3.3.2 Mean Riemannian covariance (MRC)

Let $C_1,\ldots,C_N$ be a set of covariance matrices. The Karcher or Fréchet mean is the set of tensors minimizing the sum of squared distances. In the case of tensors, the manifold has a non-positive curvature, so there is a unique mean value μ:

$$\mu = arg\min_{C\in\mathcal{M}} \sum_{i=1}^{N} \rho^2(C,C_i). \tag{3.11}$$

As the mean is defined through a minimization procedure, we approximate it by the intrinsic Newton gradient descent algorithm. The following mean value at estimation step $t+1$ is given by:

$$\mu_{t+1} = exp_{\mu_t}\left[\frac{1}{N}\sum_{i=1}^{N} log_{\mu_t}(C_i)\right], \tag{3.12}$$

where exp_{μ_t} and log_{μ_t} are mapping functions (see equations 3.6 and 3.7). This iterative gradient descent algorithm usually converges very fast (in experiments 5 iterations were sufficient, which is similar to [24]). This mean value is referred to as *mean Riemannian covariance* (MRC).

MRC *vs* volume covariance Covariance matrix could be directly computed from a video by merging feature vectors from many frames into a single content (similarly to 3D descriptors, *i.e.* 3D HOG). Then, this covariance could be seen as *mean covariance*, describing characteristics of the video. Unfortunately, such solution disturbs time dependencies (time order of features is lost). Further, context of the features might be lost and at the same time some features will not appear in the covariance.

Figure 3.2 illustrates the case, in which edge features are lost during computation of the volume covariance. This is a consequence of loosing information that the feature appeared in specific time. Computing volume covariance, order of the feature appearances and their spatial correlations can be lost by merging feature distribution in time. This clearly shows that MRC holds much more information than covariance computed directly from the volume.

3.4 Efficient models for human re-identification

In this section we focus on designing efficient models for appearance matching. These models are less computationally expensive than boosting approaches [4, 19], while enhancing distinctive and descriptive characteristics of an object appearance. We propose two strategies for appearance extraction: (1) by using a hand-designed

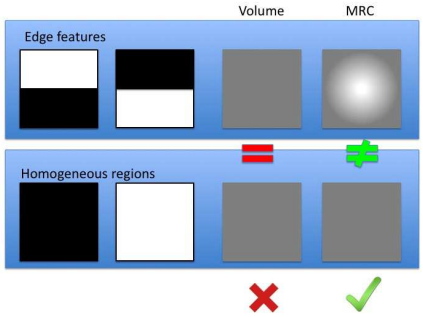


Fig. 3.2: Difference between covariance computed directly from the video content (volume covariance) and MRC. Volume covariance looses information on edge features and can not distinguish two given cases - two edge features (first row) from two homogeneous regions (second row). MRC holds information on the edges, being able to differentiate both cases.

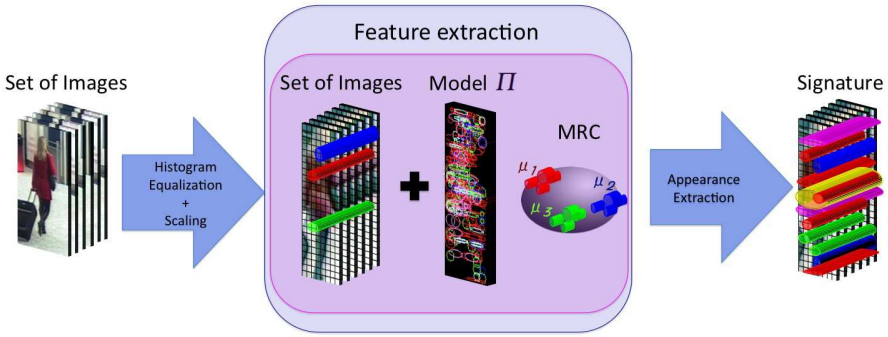


Fig. 3.3: Appearance extraction: features are determined using model Π for computing *mean Riemannian covariances* (MRC), which stand for the final appearance representation - *signature*.

model which is enhanced by a fast discriminative analysis (section 3.4.2) and (2) by employing machine learning technique that selects the most accurate features for appearance matching (section 3.4.3).

3.4.1 General scheme for appearance extraction

The input of the appearance extraction algorithm is a set of cropped images obtained by human detection and tracking results corresponding to a given person of interest (see figure 3.3). Color dissimilarities caused by variations in lighting conditions are

minimized by applying *histogram equalization* [20]. This technique maximizes the entropy in each color channel (RGB) producing more camera-independent color representation. Then, the normalized image is scaled into a fixed size $W \times H$.

From such scaled and normalized images, we extract covariance descriptors from image sub-regions and we compute MRC-s (section 3.3.2). Every image sub-region (its size and position) as well as features from which covariance is extracted is determined by a model. The final appearance representation is referred to as a *signature*.

3.4.2 MRCG model

Mean Riemannian covariance grid (MRCG) [3] has been designed to deal with low resolution images and a crowded environment where more specialized techniques (e.g. based on background subtraction) might fail. It combines a dense descriptor philosophy [9] with the effectiveness of MRC descriptor.

MRCG is represented by a dense grid structure with overlapping spatial square sub-regions (*cells*). First, such dense representation makes the signature robust to partial occlusions. Second, as the grid structure, it contains relevant information on spatial correlations between MRC *cells*, which is essential to carry out discriminative power of the signature. MRC *cell* describes statistics of an image square subregion corresponding to the specific position in the grid structure. In case of MRCG, we assume a fixed size of *cells* and a fixed feature vector for extracting covariances. Let Π be a model which is actually represented by a set of MRC cells. This model is enhanced by using our discriminative analysis, which weights each cell depending on its distinctive characteristics. These weighs are referred to as MRC *discriminants*.

3.4.2.1 MRC discriminants

The goal of using discriminants is to identify the relevance of MRC *cells*. We present an efficient way to enhance discriminative features, improving matching accuracy. By employing *one-against-all* learning schema, we highlight distinctive features for a particular individual. The main advantage of this method is its efficiency. Unlike [4], by using simple statistics on Riemannian manifold we are able to enhance features, without applying any time consuming training process.

Let $\mathfrak{S}^c = \{\mathfrak{s}_i^c\}_{i=1}^p$ be a set of signatures, where $\mathfrak{s}_i^c$ is signature i from camera c and p is the total number of pedestrians recorded in camera c. Each signature is extracted using model Π: $\mathfrak{s}_i^c = \{\mu_{i,1}^c, \mu_{i,2}^c, \ldots, \mu_{i,|\Pi|}^c\}$, where $\mu_{i,j}^c$ stands for MRC *cell*. For each $\mu_{i,j}^c$ we compute the variance between the human signatures from camera c defined as

$$\sigma^c_{i,j} = \frac{1}{p-1} \sum_{k=1;k\neq i}^{p} \rho^2(\mu^c_{i,j}, \mu^c_{k,j}). \tag{3.13}$$

In the result, for each human signature $\mathfrak{s}^c_i$ we obtain the vector of discriminants related to our MRC *cells*, $d^c_i = \{\sigma^c_{i,1}, \sigma^c_{i,2}, \ldots, \sigma^c_{i,|\Pi|}\}$. This idea is similar to methods derived from text retrieval where a frequency of *terms* is used to weight relevance of a *word*. As we do not want to quantize covariance space, we use $\sigma^c_{i,j}$ of MRC *cell* to extract its relevance. The MRC is assumed to be more significant when its variance is larger in the class of humans. Here, it is a kind of *"killing two birds with one stone"*: (1) it is obvious that the most common patterns belong to the background (the variance is small); (2) the patterns which are far from the rest are at the same time the most discriminative (the variance is large).

We thought about normalizing the $\sigma^c_{i,j}$ by the variance *within the class* (similarly to Fisher's linear discriminants, we could compute the variance of covariances related to a given cell). However, the results have shown that such normalization does not improve matching accuracy. We believe that it is due to the fact that a given number of images per individual is not sufficient for obtaining the reliable variance of MRC *within the class*.

Scalability Discriminative approaches are often accused of non-scalability. It is true that in the most of these approaches an extensive learning phase is necessary to extract discriminative signatures. This makes these approaches very difficult to apply in a real world scenario where in every minute new people appear.

Fortunately, proposing MRC discriminants, we employ a very simple discriminative method which is able to perform in a real world system. It is true that every time when a new signature is created we have to update all signatures in the database. However, for $10,000$ signatures, the update takes less than 30 seconds. Moreover, we do not expect more than such a number of signatures in the database as the re-identification approaches are constraint to *one day period* (the strong assumption about the same clothes). Further, one alternative solution might be to use a fixed *reference dataset*, which can be used as training data for discriminating new signatures.

3.4.3 COSMATI model

In the previously presented model, we assumed *a priori* the size of MRC cells, the grid layout and the feature vector, from which covariance is extracted. However, depending on image resolution and image characteristics (object class), we could use different feature vectors extracted from different image regions. Moreover, it may happen that different regions of the object appearance ought to be matched

using various feature vectors to obtain a distinctive representation. Then, we actually can formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects. This approach is referred to as *COrrelation-based Selection of covariance MATrIces* (COSMATI) [1].

In contrast to the previous model and to the most of state of the art approaches [4, 19, 26], we do not limit our covariance descriptor to a single feature vector. Instead of defining *a priori* feature vector, we use a machine learning technique to select features that provide the most descriptive apperance representation. The following sections describe our feature space and the learning, by which the appearance model for matching is generated.

3.4.3.1 Feature Space

Let $L = \{R, G, B, I, \nabla_I, \theta_I, \dots\}$ be a set of feature layers, in which each layer is a mapping such as color, intensity, gradients and filter responses (texture filters, *i.e.* Gabor, Laplacian or Gaussian). Instead of using covariance between all of these layers, which would be computationally expensive, we compute covariance matrices of a few relevant feature layers. These relevant layers are selected depending on the region of an object (see Section 3.4.3.2). In addition, let layer $\mathfrak{D}$ be a distance between the center of an object and the current location. Covariance of distance layer $\mathfrak{D}$ and three other layers l ($l \in L$) form our descriptor, which is represented by a 4×4 covariance matrix. By using distance $\mathfrak{D}$ in every covariance, we keep a spatial layout of feature variances, which is rotation invariant. State of the art techniques very often use pixel location (x, y) instead of distance $\mathfrak{D}$, yielding better description of an image region. Conversely, among our detail experimentation, using $\mathfrak{D}$ rather than (x, y), we did not decrease the recognition accuracy in the general case, while decreasing the number of features in the covariance matrix. This discrepancy may be due to the fact that we hold spatial information twofold, (1) by location of a rectangular sub-region from which the covariance is extracted and (2) by $\mathfrak{D}$ in covariance matrix. We constraint our covariances to combination of 4 features, ensuring computational efficiency. Also, bigger covariance matrices tend to include superfluous features which can clutter the appearance matching. 4×4 matrices provide sufficiently descriptive correlations while keeping low computational time needed for calculating generalized eigenvalues during distance computation.

Different combinations of three feature layers produce different kinds of covariance descriptor. By using different covariance descriptors, assigned to different locations in an object, we are able to select the most discriminative covariances according to their positions. The idea is to characterize different regions of an object by extracting different kinds of features (*e.g.* when comparing human appearances, edges coming from shapes of arms and legs are not discriminative enough in most cases as every instance possess similar features). Taking into account this phenomenon, we minimize redundancy in an appearance representation by an entropy-driven selection method.

Fig. 3.4: A meta covariance feature space. Example of three different covariance features. Every covariance is extracted from a region (P), *distance* layer ($\mathfrak{D}$) and three channel functions (*e.g.* bottom covariance feature is extracted from region P_3 using layers: $\mathfrak{D}$, I-intensity, ∇_I-gradient magnitude and θ_I-gradient orientation).

Let us define index space $\mathbb{Z} = \{(P, l_i, l_j, l_k) : P \in \mathbf{P}; l_i, l_j, l_k \in L\}$, of our meta covariance feature space $\mathfrak{C}$, where $\mathbf{P}$ is a set of rectangular sub-regions of the object; and l_i, l_j, l_k are color/intensity or filter layers. Meta covariance feature space $\mathfrak{C}$ is obtained by mapping $\mathbb{Z} \rightarrow \mathfrak{C} : cov_P(\mathfrak{D}, l_i, l_j, l_k)$, where $cov_P(\phi)$ is the covariance descriptor [26] of features ϕ: $cov_P(\phi) = \frac{1}{|P|-1} \sum_{k \in P} (\phi_k - \mu)(\phi_k - \mu)^T$. Fig. 3.4 shows different feature layers as well as examples of three different types of covariance descriptor. The dimension $n = |\mathbb{Z}| = |\mathfrak{C}|$ of our meta covariance feature space is the product of the number of possible rectangular regions by the number of different combinations of feature layers.

3.4.3.2 Learning in a Covariance Metric Space

Let $\mathbf{a}_i^c = \{\mathbf{a}_{i,1}^c, \mathbf{a}_{i,2}^c, \ldots \mathbf{a}_{i,m}^c\}$ be a set of relevant observations of an object i in camera c, where $\mathbf{a}_{ij}^c$ is a n-dimensional vector composed of all possible covariance features extracted from image j of object i in the n-dimensional meta covariance feature space $\mathfrak{C}$. We define the distance vector between two samples $\mathbf{a}_{i,j}^c$ and $\mathbf{a}_{k,l}^{c'}$ as follows

$$\delta(\mathbf{a}_{i,j}^c, \mathbf{a}_{k,l}^{c'}) = \left[\rho(\mathbf{a}_{i,j}^c[z], \mathbf{a}_{k,l}^{c'}[z])\right]_{z \in \mathbb{Z}}^T, \tag{3.14}$$

where ρ is the geodesic distance between covariance matrices [13], and $\mathbf{a}_{i,j}^c[z]$, $\mathbf{a}_{k,l}^{c'}[z]$ are the corresponding covariance matrices (the same region P and the same combination of layers). The index z is an iterator of $\mathfrak{C}$.

We cast the appearance matching problem into the following *distance learning* problem. Let δ^+ be distance vectors computed using pairs of relevant samples (of the same people captured in different cameras, $i = k$, $c \neq c'$) and let δ^- be distance vectors computed between pairs of related irrelevant samples ($i \neq k$, $c \neq c'$). Pairwise elements δ^+ and δ^- are distance vectors, which stand for positive and negative samples, respectively. These distance vectors define a *covariance metric*

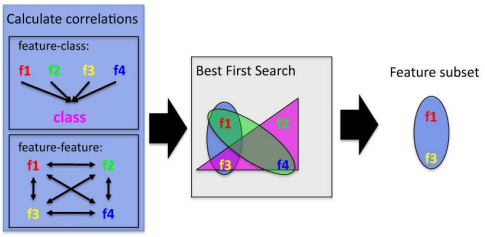


Fig. 3.5: Correlation-based feature selection. *Best first search* evaluates different feature subsets using feature-class and feature-feature correlations (equation 3.15). The best feature subset stands for model Π that is used for appearance extraction and matching.

space. Given δ^+ and δ^- as training data, our task is to find a general model of appearance to maximize matching accuracy by selecting relevant covariances and thus defining a distance.

Learning on a manifold This is a difficult and unsolved challenge. Methods [4, 27] perform classification by regression over the mappings from the training data to a suitable tangent plane. By defining tangent plane over the Karcher mean of the positive training data points, we can preserve a local structure of the points. Unfortunately, models extracted using means of the positive training data points tend to over-fit. These models concentrate on tangent planes obtained from training data and do not have generalization properties. We overcome this issue by employing a feature selection technique for identifying the most salient features. Based on the hypothesis: *"A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other"* [18], we build our appearance model using covariance features chosen by a *correlation-based feature selection*.

Correlation-based Feature Selection (CFS) [18] is a filter algorithm that ranks feature subsets according to a correlation-based evaluation function. This evaluation function favors feature subsets which contain features highly correlated with the class and uncorrelated with each other. In the *metric learning* problem, we define positive and negative class by δ^+ and δ^-, as relevant and irrelevant pairs of samples (see Section 3.4.3.2).

Further, let feature $f_z = \delta[z]$ be characterized by a distribution of the zth elements in distance vectors δ^+ and δ^-. The feature-class correlation and the feature-feature inter-correlation are measured using a symmetrical uncertainty model [18]. As this model requires nominal valued features, we discretize f_z using the method of Fayyad and Irani [11]. Let X be a nominal valued feature obtained by discretization of f_z (discretization of distances).

We assume that a probabilistic model of X can be formed by estimating the probabilities of the values $x \in X$ from the training data. The information content can be measured by entropy $H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$. A relationship between fea-

tures X and Y can be given by $H(X|Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y) \log_2 p(x|y)$. The amount by which the entropy of X decreases reflects additional information on X provided by Y and is called the *information gain* (*mutual information*) defined as $Gain = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.

Even if the *information gain* is a symmetrical measure, it is biased in favor of features with more discrete values. Thus, the symmetrical uncertainty r_{XY} is used to overcome this problem $r_{XY} = 2 \times \left[Gain/\big(H(X) + H(Y)\big)\right]$.

Having the correlation measure, a subset of features $\mathfrak{S}$ is evaluated using function $\mathfrak{M}(\mathfrak{S})$ defined as

$$\mathfrak{M}(\mathfrak{S}) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k+1)\overline{r_{ff}}}}, \tag{3.15}$$

where k is the number of features in subset $\mathfrak{S}$, $\overline{r_{cf}}$ is the average feature-class correlation and $\overline{r_{ff}}$ is the average feature-feature inter-correlation

$$\overline{r_{cf}} = \frac{1}{k} \sum_{f_z \in \mathfrak{S}} r_{cf_z}, \quad \overline{r_{ff}} = \frac{2}{k(k-1)} \sum_{\substack{f_i, f_j \in \mathfrak{S} \\ i<j}} r_{f_i f_j}, \tag{3.16}$$

where c is the class, or relevance feature, which is $+1$ on δ^+ and -1 on δ^-. The numerator in Eq. 3.15 indicates predictive ability of subset $\mathfrak{S}$ and the denominator stands for redundancy among the features (for details of $\mathfrak{M}(\mathfrak{S})$, the interested reader is pointed to [18]).

Equation 3.15 is the core of CFS, which ranks feature subsets in the search space of all possible feature subsets. Since exhaustive enumeration of all possible feature subsets is prohibitive in most cases, a heuristic search strategy has to be applied. We have investigated different search strategies, among which *best first search* [25] performs the best.

Best first search is an *Artificial Intelligence* (AI) search strategy that allows backtracking along the search path. Our *best first* starts with no feature and progresses forward through the search space adding single features. The search terminates if T consecutive subsets show no improvement over the current best subset (we set $T = 5$ in experiments). By using this stopping criterion we prevent the best first search from exploring the entire feature subset search space. Fig. 3.5 illustrates CFS method. Let Π be the output of CFS that is the feature subset of $\mathfrak{C}$. This feature subset Π forms a model that is used for appearance extraction and matching.

3.4.4 Appearance matching

Let $\mathfrak{s}_a^c$ and $\mathfrak{s}_b^{c'}$ be the object signatures. The signature consists of *mean Riemannian covariance* matrices extracted using set Π. The similarity between two signatures $\mathfrak{s}_a^c$ and $\mathfrak{s}_b^{c'}$ is defined as

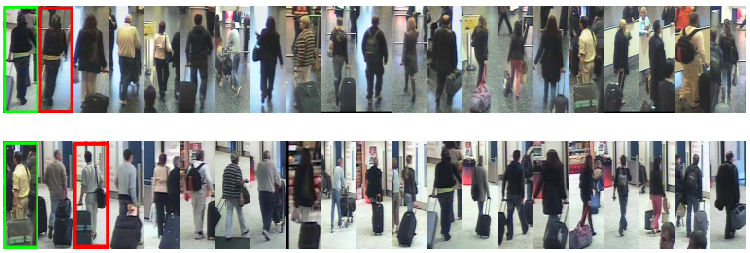


Fig. 3.6: Example of the person re-identification on i-LIDS-MA. The left-most image is the probe image. The remaining images are the top 20 matched gallery images. The red boxes highlight the correct matches.

$$S(\mathfrak{s}_a^c, \mathfrak{s}_b^{c'}) = \frac{1}{|\Pi|} \sum_{i \in \Pi} \frac{\sigma_{a,i}^c + \sigma_{b,i}^{c'}}{\max(\rho(\mu_{a,i}^c, \mu_{b,i}^{c'}), \varepsilon)}, \tag{3.17}$$

where ρ is a geodesic distance, $\mu_{a,i}^c$ and $\mu_{b,i}^{c'}$ are mean covariance matrices extracted using covariance feature $i \in \Pi$ and $\varepsilon = 0.1$ is introduced to avoid the denominator approaching to zero. $\sigma_{a,i}^c$ and $\sigma_{b,i}^{c'}$ are discriminants of the corresponding MRC-s (see section 3.4.2.1). If discriminants have not been computed then the nominator is set to 1 ($\sigma_{a,i}^c + \sigma_{b,i}^{c'} = 1$). Using the average of similarities computed on feature set Π the appearance matching becomes robust to noise.

3.5 Experiments

In this section we mostly focus on comparing MRCG with COSMATI model. We carry out experiments on 3 i-LIDS datasets[1]: i-LIDS-MA [4], i-LIDS-AA [4] and i-LIDS-119 [29]. These datasets have been extracted from the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset for evaluating the re-identification task. The results are analyzed in terms of recognition rate, using the *cumulative matching characteristic* (CMC) [17] curve. The CMC curve represents the expectation of finding the correct match in the top n matches (see figure 3.6). We also employ a quantitative scalar appraisal of CMC curve by computing the *normalized area under curve* (nAUC) value.

[1] The Image Library for Intelligent Detection Systems (i-LIDS) is the UK government's benchmark for Video Analytics (VA) systems

3.5.1 Experimental setup

Comparing the proposed models we keep the experimental settings presented in [1, 3]. *Model_Name*$^+$ means that signatures were enhanced by using discriminative analysis (section 3.4.2.1). It should be noted that this discriminative analysis can be applied to MRCG as well as to COSMATI model.

3.5.1.1 MRCG model

Every human image is scaled into a fixed size of 64×192 pixels (size of the grid). We extract the MRC *cells* of 16×16 pixels, on a fixed grid of 8 pixels step (it gives in total 161 *cells*). The feature vector consists of 11 features:

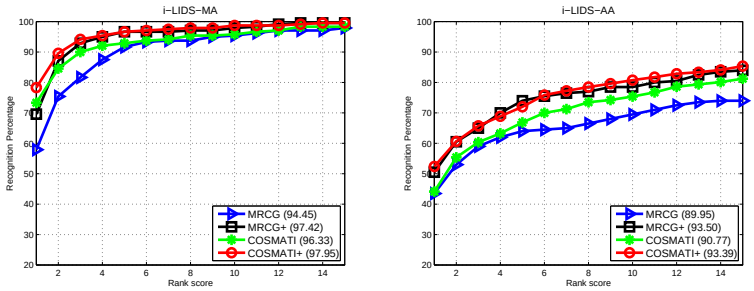
$$\left[x, y, R_{xy}, G_{xy}, B_{xy}, \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right], \tag{3.18}$$

where x and y are pixel location, R_{xy}, G_{xy}, B_{xy} are RGB channel values and ∇ and θ correspond to gradient magnitude and orientation in each channel, respectively.

3.5.1.2 COSMATI model

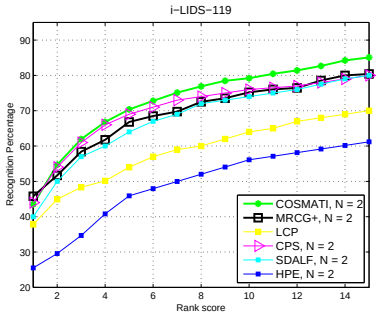
Feature space We scale every human image into a fixed size window of 64×192 pixels. The set of rectangular sub-regions $\mathbf{P}$ is produced by shifting 32×8 and 16×16 pixel regions with 8 pixels step (up and down). It gives $|\mathbf{P}| = 281$ overlapping rectangular sub-regions. We set $L = \{(I, \nabla_I, \theta_I)_{I=I,R,G,B}, G_{i=1\ldots4}, \mathcal{N}, \mathfrak{L}\}$, where I, R, G, B refer to intensity, red, green and blue channel, respectively; ∇ is the gradient magnitude; θ corresponds to the gradient orientation; G_i are Gabor's filters with parameters $\gamma, \theta, \lambda, \sigma^2$ set to $(0.4, 0, 8, 2)$, $(0.4, \frac{\pi}{2}, 8, 2)$, $(0.8, \frac{\pi}{4}, 8, 2)$ and $(0.8, \frac{3\pi}{2}, 8, 2)$, respectively; $\mathcal{N}$ is a gaussian and $\mathfrak{L}$ is a laplacian filter. A learning process involving all possible combinations of three layers would not be computationally tractable (229296 covariances to consider in section 3.4.3.2). Thus instead, we experimented with different subsets of combinations and selected a reasonably efficient one. Among all possible combinations of the three layers, we choose 10 combinations ($C_{i=1\ldots10}$) to separate color and texture features, while ensuring inexpensive computation. We set C_i to (R, G, B), $(\nabla_R, \nabla_G, \nabla_B)$, $(\theta_R, \theta_G, \theta_B)$, (I, ∇_I, θ_I), (I, G_3, G_4), $(I, G_2, \mathfrak{L})$, $(I, G_2, \mathcal{N})$, $(I, G_1, \mathcal{N})$, $(I, G_1, \mathfrak{L})$, (I, G_1, G_2), respectively. Note that we add to every combination C_i layer $\mathfrak{D}$, thus generating our final 4×4 covariance descriptors. The dimension of our meta covariance feature space is $n = |\mathfrak{C}| = 10 \times |\mathbf{P}| = 2810$.

Learning and testing Let us assume that we have $(p+q)$ individuals seen from two different cameras. For every individual, m images from each camera are given. We take q individuals to learn our model, while p individuals are used to set up the



(a) p = 30 (b) p = 100

Fig. 3.7: Performance comparison. Evaluation of COSMATI is performed using the models learned on i-LIDS-MA. nAUC values are presented within parentheses.



(a) p = 119

Fig. 3.8: Comparison with state of the art approaches on i-LIDS-119 dataset: LCP [4], CPS [8], SDALF [5] and HPE [6].

gallery set. We generate positive training examples by comparing m images of the same individual from one camera with m images from the second camera. Thus, we produce $|\delta^+| = q \times m^2$ positive samples. Pairs of images coming from different individuals stand for negative training data, thus producing $|\delta^-| = q \times (q-1) \times m^2$ negative samples.

3.5.2 Results

i-LIDS-MA [4] This dataset consists of 40 individuals extracted from two non-overlapping camera views. For each individual a set of 46 images is given. The dataset contains in total $40 \times 2 \times 46 = 3680$ images. For each individual we randomly select $m = 10$ images. Then, we randomly select $q = 10$ individuals to learn

COSMATI models. The evaluation is performed on the remaining $p = 30$ individuals. We evaluate MRCG and COSMATI on the same sets of people. Every signature is used as a query to the gallery set of signatures from the other camera. This procedure has been repeated 10 times to obtain averaged CMC curves.

We compare COSMATI with MRCG in figure 3.7(a). The best performance is achieved by COSMATI$^+$. It appears that discriminative analysis has more significant impact on MRCG than on COSMATI model. This result may be due to the fact that COSMATI already selects distinctive representation for appearance matching. We can also note that MRCG$^+$ achieves similar recognition rate as COSMATI$^+$. However, it is relevant to mention that COSMATI is significantly faster than MRCG, as it uses small covariance matrices (4×4). The experiment bears out that by designing the efficient feature space (section 3.4.3.1) and employing the effective selection method (section 3.4.3.2), we are able to produce efficient and effective models without loosing the recognition performance.

The disadvantage of COSMATI is the offline learning phase. The approaches which are based on training data requiring positive pairs (two images with the same person registered in different cameras) may have difficulties while employed in video analytics systems. Annotations of training data from c cameras and training $\binom{c}{2} = \frac{c!}{2!(c-2)!}$ models, can be unaffordable in practice in case of large c. However, we have to stress that unlike regular metric learning approaches [10, 21, 30], COSMATI does not need a lot of training samples. Most of metric learning techniques produce matching strategies by using 100-300 subjects, while our method uses only 10 persons for obtaining an effective and efficient model.

i-LIDS-AA [4] This dataset contains 100 individuals automatically detected and tracked in two cameras. Cropped images are noisy, which makes the dataset more challenging (*e.g.* detected bounding boxes are not accurately centered around the people, only part of the people is detected due to occlusion). For minimizing misalignment issues, we employ *discriminatively trained deformable part models* [12, 15], which slightly improve detection accuracy. COSMATI is evaluated using the models learned on i-LIDS-MA. Figure 3.7(b) illustrates the results. Although, data are noisy, we can observe the same trends as in the case of evaluating on i-LIDS-MA data.

i-LIDS-119 [29] For comparing MRCG and COSMATI models with state of the art techniques, we select i-LIDS-119 data. This dataset is extensively used in the literature for testing the person re-identification approaches. It consists of 119 individuals with 476 images. The dataset is very challenging since there are many occlusions and often only the top part of the person is visible. As only few images are given per individual, we extract signatures using maximally $N = 2$ images.

In figure 3.8 we compare MRCG$^+$ and COSMATI with LCP [4], CPS [8], SDALF [5] and HPE [6]. In case of COSMATI, we have used models learned on i-LIDS-MA to evaluate our approach on the full dataset of 119 individuals.

COSMATI performs the best among all considered methods. We believe that it is due to the informative appearance representation obtained by CFS technique (sec-

tion 3.4.3.2). It clearly shows that a combination of the strong covariance descriptor with the efficient selection method produces distinctive models for the appearance matching problem. For more extensive evaluation and competitive results of COSMATI and of MRCG, the interested reader is pointed to [1] and [3], respectively.

Computational speed The level of performance achieved by COSMATI comes with a significant computational gain *w.r.t.* MRCG. In our experiments, for $q = 10$ and $m = 10$ we generate $|\delta^+| = 1000$ and $|\delta^-| = 9000$ training samples. Learning on 10.000 samples takes around 20 minutes on Intel quad-core 2.4GHz. COSMATI model is composed of 150 covariance features in average, which is similar to MRCG (161 cells). Comparing the time calculation of generalized eigenvalues (distance computation) of 4×4 covariance with 11×11 covariance, we always reach $10 - 15$ speedup depending on the hardware architecture. In the result, we can expect the same speedup while retrieving signatures in video analytics systems.

3.6 Conclusion

This chapter presented two strategies for appearance matching, while employing covariance statistics averaged over a Riemannian manifold. We discussed different ways of extracting an object appearance by incorporating various training strategies. We showed that by applying efficient discriminative analysis, we are able to improve re-identification accuracy. Further, we demonstrated that by introducing an offline learning stage, we can characterize an object appearance in a more efficient and distinctive way. In the future, we plan to integrate the notion of motion in our recognition framework. This would allow to distinguish individuals using their shape characteristics and to extract only the features which surely belong to foreground region.

Acknowledgements This work has been supported by VANAHEIM, ViCoMo and PANORAMA European projects.

References

1. Bak, S., Charpiat, G., Corvee, E., Bremond, F., Thonnat, M.: Learning to match appearances by correlations in a covariance metric space. In: Proceedings of the 12th European Conference on Computer Vision,, ECCV. IEEE Computer Society (2012)
2. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. IEEE Computer Society (2010)
3. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: Proceedings of the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. IEEE Computer Society (2011)

4. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Boosted human re-identification using riemannian manifolds. Image and Vision Computing **30**(6-7), 443 – 452 (2012)
5. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding **117**(2), 130–144 (2013)
6. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters **33**(7), 898–903 (2012). Special Issue on Awards from ICPR 2010
7. Cai, Y., Takala, V., Pietikainen, M.: Matching groups of people by covariance descriptor. In: Proceedings of the 20th International Conference on Pattern Recognition, ICPR, pp. 2744–2747. IEEE Computer Society (2010)
8. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the 22nd British Machine Vision Conference, BMVC, pp. 68.1–68.11. BMVA Press (2011)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 18th Conference on Computer Vision and Pattern Recognition, CVPR, pp. 886–893. IEEE Computer Society (2005)
10. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Proceedings of the 10th Asian Conference on Computer Vision, ACCV, pp. 501–512. IEEE Computer Society (2010)
11. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the International Joint Conference on Uncertainty in AI, IJCAI, pp. 1022–1027 (1993)
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (2010)
13. Förstner, W., Moonen, B.: A metric for covariance matrices. In: Quo vadis geodesia ...?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday, TR Dept. of Geodesy and Geoinformatics, Stuttgart University (1999)
14. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proceedings of the 19th Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1528–1535. IEEE Computer Society (2006)
15. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/
16. Goh, A., Vidal, R.: Unsupervised riemannian clustering of probability density functions. In: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I, ECML PKDD, pp. 377–392. Springer-Verlag, Berlin, Heidelberg (2008)
17. Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In: Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, PETS. IEEE Computer Society (2007)
18. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, Department of Computer Science, University of Waikato (1999)
19. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Proceedings of the 17th Scandinavian Conference on Image Analysis, SCIA, pp. 91–102. Springer-Verlag, Berlin, Heidelberg (2011)
20. Hordley, S.D., Finlayson, G.D., Schaefer, G., Tian, G.Y.: Illuminant and device invariant colour using histogram equalisation. Pattern Recognition (2005)
21. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Proceedings of the 25th Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2288–2295 (2012)
22. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: Proceedings of the 23rd British Machine Vision Conference, BMVC (2012)

23. Oncel, F.P., Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based on lie algebra. In: Proceedings of the 19th Conference on Computer Vision and Pattern Recognition, CVPR (2006)
24. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. International Journal on Computer Vision **66**(1), 41–66 (2006)
25. Rich, E., Knight, K.: Artificial Intelligence. McGraw-Hill Higher Education (1991)
26. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: Proceedings of the 9th European Conference on Computer Vision, ECCV, pp. 589–600. Springer-Verlag (2006)
27. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1713–1727 (2008)
28. Yao, J., Odobez, J.M.: Fast human detection from joint appearance and foreground feature subset covariances. Computer Vision and Image Understanding **115**(3), 1414–1426 (2011)
29. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: Proceedings of the 20th British Machine Vision Conference, BMVC. BMVC Press (2009)
30. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Proceedings of the 24th Conference on Computer Vision and Pattern Recognition, CVPR, pp. 649 –656. IEEE Computer Society (2011)