Chapter 2
Discriminative Image Descriptors for Person Re-identification

Bingpeng Ma, Yu Su and Frédéric Jurie

Abstract This chapter looks at *person re-identification* from a *computer vision* point of view, by proposing two new image descriptors designed for matching people bounding boxes in images. Indeed, one key issue of person re-identification is the ability to measure the similarity between two person-centered image regions, allowing to predict if these regions represent the same person despite changes in illumination, viewpoint, background clutter, occlusion, and image quality/resolution. They hence heavily rely on the signatures or descriptors used for representing and comparing the regions. The first proposed descriptor is a combination of Biologically Inspired Features (BIF) and covariance descriptors, while the second builds on the recent advances of Fisher Vectors. These two image descriptors are validated through experiments on two different person re-identification benchmarks (VIPeR and ETHZ), achieving state-of-the-art performance on both datasets.

2.1 Introduction

In recent years, person re-identification in unconstrained videos (i.e. without subjects' knowledge and in uncontrolled scenarios) has attracted more and more research interest. Generally speaking, person re-identification consists of recognizing an individual through different images (e.g., coming from cameras in a distributed network

B. Ma (✉)
School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China
e-mail: bpma@ucas.ac.cn

Y. Su · F. Jurie
GREYC—CNRS UMR 6072, University of Caen Basse-Normandie, Caen, France
e-mail: yu.su@unicaen.fr

F. Jurie
e-mail: frederic.jurie@unicaen.fr

or from the same camera at different time). It is done by measuring the similarity between two person-centered bounding boxes and predicting—based on this similarity—if they represent the same person. This is challenging in unconstrained scenarios because of illumination, viewpoint, and background changes, as well as occlusions or low resolution.

In order to tackle this problem, researchers have concentrated their effort on either (1) the design of visual features to describe individual images or (2) the use of adapted distance measures (e.g., obtained by metric learning). This chapter focuses on the former by proposing two novel image representations. The proposed image representations can be used to measure effectively the similarity between two persons, without requiring any preprocessing step (e.g., background subtraction or body part segmentation).

The first representation is based on Biologically Inspired Features (BIF) [30] extracted through the use of Gabor filters (S1 layer) and MAX operator (C1 layer). They are encoded by the covariance descriptor of [37], used to compute the similarity of BIF features at neighboring scales. The Gabor filters and the covariance descriptor improve the robustness to the illumination variation, while the MAX operator increases the tolerance to scale changes and image shifts. Furthermore, we argue that measuring the similarity of neighboring scales limits the influence of the background (see Sect. 2.3.3 for details). By overcoming illumination, scale, and background changes, the performance of person re-identification is widely improved.

The second one builds on the recently proposed Fisher Vectors for image classification [26] which encodes higher order statistics of local features, and gives excellent performance for several object recognition and image retrieval tasks [27, 28]. Motivated by the success of Fisher Vector, we combine Fisher Vectors with a novel and very simple seven-dimensional local descriptor adapted to the representation of person images, and use the resultant representation (*Local Descriptors encoded by Fisher Vector* or LDFV) as a person descriptor.

These two representations have been experimentally validated on two person re-identification databases (namely the VIPeR and ETHZ datasets), which are challenging since they contain pose changes, viewpoint and lighting variations, and occlusions. Furthermore, as they are commonly used in the recent literature, they allow comparisons with state-of-the-art approaches.

The remainder of this chapter is organized as follows: Sect. 2.2 reviews the related works on image representation for person re-identification in videos. Section 2.3 describes the first proposed descriptor in detail, analyzes its advantages, and then shows its effectiveness on the VIPeR and ETHZ datasets. The second person descriptor and its experimental validation are given Sect. 2.4. Finally, Sect. 2.5 concludes the chapter.

2.2 Related Work

Person re-identification in the literature has been considered either as a *on the fly* [21] or as an *offline* [33] problem. More formally, person re-identification can be defined as finding the correspondences between the images of a *probe set* representing a single person and the corresponding images in a *gallery set*. Depending on the number of available images per individual (i.e., the size of the probe set), different scenarios have been addressed: (a) Single versus Single (S vs. S) if only one exemplar per individual is available both in probe and in gallery sets [17]; (b) Multiple versus Single (M vs. S) if multiple exemplars per individual are available in the gallery set [12]; (c) Multiple versus Multiple (M vs. M) if multiple exemplars per individual are available both in the probe and gallery sets [33].

As explained before, the image descriptors used for comparing persons are important as they strongly impact the overall performance. The recent literature abounds with such image descriptors. They can be based on (1) color—widely used since the color of clothing constitutes a simple but efficient visual signature—usually encoded within histograms of RGB or HSV values [6], (2) shape, e.g., HOG-based signatures [25, 33], (3) texture, often represented by Gabor filters [18, 29, 40], differential filters [18, 29], Haar-like representations [4] and Co-occurrence Matrices [33], (4) interest points, e.g., SURF [15] and SIFT [21, 41] and (5) image regions [6, 25].

Region-based methods usually split the human body into different parts and extract features for each part. In [6, 9], Maximally Stable Color Regions (MSCR) are extracted, by grouping pixels of similar color into small stable clusters. Then, the regions are described by their area, centroid, second moment matrix, and average color. The Region Covariance Descriptor (RCD) [1, 5, 40] has also been widely used for representing regions. In RCD, the pixels of a region are first represented by a feature vector which captures their intensity, texture, and shape statistics. The so-obtained feature vectors are then encoded by a covariance matrix.

Besides these generic representations, there are some more specialized representations. For example, Epitomic Analysis [7], Spin Images [2, 3], Bag-of-Words based descriptors [41], Implicit Shape Models (ISM) [21], or Panoramic Maps [14] have also been applied to person re-identification.

Since the elementary features (color, shape, texture, etc.) capture different aspects of the information contained in images, they are often combined to give a richer signature. For example, [29] combined 8 color features with 21 texture filters (Gabor and differential filters). Bazzani et al. [6] and Cheng et al. [9] combined MSCR descriptors with weighted Color Histograms, achieving state-of-the-art results on several widely used person re-identification datasets. Interestingly, RCD can be generalized to any type of images such as one-dimensional intensity images, three channel color images, or even other types of images (e.g., infrared). For example, in [40], Gabor features and Local Binary Patterns (LBP) are combined to form a Covariance descriptor which handles the difficulties of varying illumination, viewpoint changes, and nonrigid body deformations.

Different representations need different similarity functions. For example, representations based on histograms can be compared with Bhattacharyya distance [6, 7, 9] or Earth Mover's Distance (EMD) [2, 3]. When the dimensionalities of the representations to be compared are different, EMD can also be used as it allows many-to-many association [25]. Feature selection has been used to improve the discriminative power of the distance function, *e.g.* with boosting. In [18], the authors select the most relevant features (color and texture) by a weighted ensemble of likelihood ratio tests, obtained with AdaBoost. Similarly, in [4] Haar-like features are extracted from the whole body and the most discriminative ones are selected by AdaBoost.

Metric learning has also been used to provide a metric adapted to person re-identification (e.g. [17, 29, 41]). Most distance metric learning approaches learn a Mahalanobis-like distance such as Large Margin Nearest Neighbors (LMNN) [38], Information Theoretic Metric Learning (ITML) [10], Logistic Discriminant Metric Learning (LDML) [19], or PCCA [23]. LMNN minimizes the distance between each training point and its K nearest similarly labeled neighbors, while maximizing the distance between all differently labeled points which are closer than the aforementioned neighbors' distances plus a constant margin. In [11], the authors improved the LMNN with rejection and successfully applied their method to person re-identification. Besides Adaboost and metric learning, RVM [29], Partial Least Squares (PLS) and multiple instance learning [31, 32] have also been applied to person re-identification, with the same idea of improving the performance.

Our approach builds on these recent works, and shows that carefully designed visual features can provide us with state-of-the art results, without the need for any complex distance functions.

2.3 Bio-inspired Covariance Descriptor for Person Re-identification

Our first descriptor is a covariance descriptor using bio-inspired features, BiCov for short. It is a two-stage representation (see Fig. 2.1) in which biologically inspired features are encoded by computing the difference of covariance descriptors at different scales. In the following, the two stages are presented and motivated.

2.3.1 Low-Level Biologically Inspired Features (BIF)

Based on the study of the human visual system, bio-inspired features [30] have obtained excellent performances on several computer vision tasks such as object category recognition [34], face recognition [22], age estimation [20], and scene classification [36].
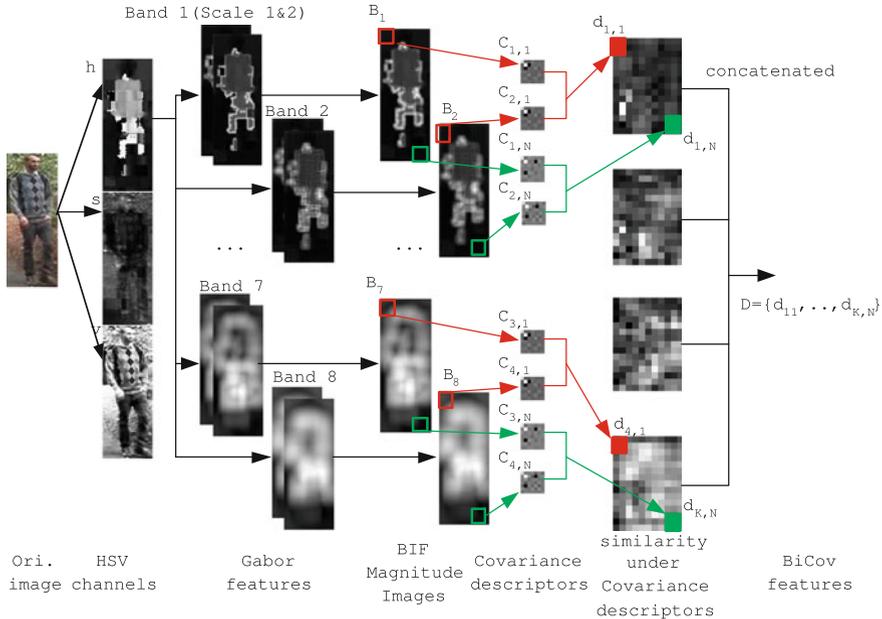
Fig. 2.1 Flowchart of the proposed approach: (1) color images are split into three color channels (HSV), (2) for each channel, Gabor filters are computed at different scales, (3) pairs of neighboring scales are grouped to form one band, (4) magnitude images are produced by applying the MAX operator within the same band, (5) magnitude images are divided into small bins and each bin is represented by a covariance descriptor, and (6) the difference of covariance descriptors between two consecutive bands is computed for each bin and concatenated to form the image representation

Considering the great success of these BIFs, the first step consists of extracting such features to model the low-level properties of images. For an image $I(x, y)$, we compute its convolution with Gabor filters according to the following equations [39]:

$$G(\mu, \nu) = I(x, y) * \psi_{\mu,\nu}(z) \tag{2.1}$$

with:

$$\psi_{\mu,\nu}(z) = \frac{\left\| k_{\mu,\nu} \right\|^2}{\sigma^2} e^{\left(\frac{-\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2} \right)} \left[e^{i k_{\mu,\nu} z} - e^{\frac{-\sigma^2}{2}} \right] \tag{2.2}$$

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu}, k_\nu = 2^{-\frac{\nu+2}{2}} \pi, \phi_\mu = \mu \frac{\pi}{8} \tag{2.3}$$

where μ and ν are scale and orientation parameters, respectively. In our work, μ is quantized into 16 scales while the ν is quantized into eight orientations.

In practice, we have observed that for person re-identification, the image representations $G(\mu, \nu)$ for different orientations can be averaged without significant loss of performance. Thus, in this case, we replace $\psi_{\mu,\nu}(z)$ in Eq. 2.1 by

Table 2.1 Scales of Gabor filters in different bands

Band	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
Filter sizes	11×11	15×15	19×19	23×23	27×27	31×31	35×35	39×39
Filter sizes	13×13	17×17	21×21	25×25	29×29	33×33	37×37	41×41

Fig. 2.2 A pair of images and their BIF Magnitude Images. From *left* to *right* the original image, its three HSV channels, six BIF Magnitude Images for different bands

$\psi_\mu(z) = \frac{1}{8} \sum_{v=1}^{8} \psi_{\mu,v}(z)$. This simplification makes the computations of $G(\mu)$—which is the average of $G(\mu, v)$ over all orientations—more efficient.

In all our experiments, the number of scales is fixed to 16 and two neighborhood scales are grouped into one band (we therefore have eight different bands). The scales of Gabor filters in different bands are shown in Table 2.1. We then apply the *MAX pooling* over two consecutive scales (within the same orientation if the orientations are not merged):

$$B_i = \max(G(2i-1), G(2i)) \tag{2.4}$$

The MAX pooling operation increases the tolerance to small-scale changes which often occur, even for the same person, since images are only roughly aligned. We refer to B_i $i \in [1, \ldots, 8]$ as the *BIF Magnitude Images*. Figure 2.2 shows a pair of images of one person and its respective BIF Magnitude Images. The image in the first column is the input image, while the ones in the second column are three HSV channels. The images from the third to the eigth column are the BIF Magnitude Images for six different bands.

2.3.2 BiCov Descriptor

In the second stage, BIF Magnitude Images are divided into small overlapping rectangular regions, allowing the preservation of some spatial information. Then, each

region is represented by a covariance descriptor [37]. Covariance descriptors can capture shape, location, and color information, and their performances have been shown to be better than other methods in many situations, as rotation and illumination changes are absorbed, to some extent, by the covariance matrix [37].

In order to do this, each pixel of the BIF Magnitude Image B_i is encoded into a seven-dimensional feature vector which captures the intensity, texture, and shape statistics:

$$f_i(x, y) = [x, y, B_i(x, y), B_{i_x}(x, y), B_{i_y}(x, y), B_{i_{xx}}(x, y), B_{i_{yy}}(x, y)] \quad (2.5)$$

where x and y are the pixel coordinates, $B_i(x, y)$ is the raw pixel intensity at position (x, y), $B_{i_x}(x, y)$ and $B_{i_y}(x, y)$ are the derivatives of image B_i with respect to x and y, and $B_{i_{xx}}(x, y)$ and $B_{i_{yy}}(x, y)$ are the second-order derivatives.

Finally, the covariance descriptor is computed for each region of the image:

$$C_{i, r} = \frac{1}{n - 1} \sum_{(x, y) \in region\ r} (f_i(x, y) - \bar{f}_i)(f_i(x, y) - \bar{f}_i)^T \quad (2.6)$$

where $\bar{f}_i$ is the mean of $f_i(x, y)$ over the region r and n is the size of region r (in pixels).

Usually, the covariance matrices computed by Eq. 2.6 are considered as the image representation. Covariance matrices are positive definite symmetric matrices lying on a manifold of the Euclidean space. Hence, many usual operations (like the l2 distance) cannot be used directly.

In this chapter, differently from past approaches using covariance descriptors, we compute (for each region separately) the difference of covariance descriptors between two consecutive bands:

$$d_{i, b} = d(C_{2i-1, r}, C_{2i, r}) = \sqrt{\sum_{p=1}^{P} \ln^2 \lambda_p(C_{2i-1, r}, C_{2i, r})} \quad (2.7)$$

where $\lambda_p(C_{2i-1, r}, C_{2i, r})$ is the p-th generalized eigenvalues of $C_{2i-1, r}$ and $C_{2i, r}$, $i = 1, 2, 3, 4$. Finally, the differences are concatenated to form the image representation:

$$D = (d_{1,1}, \cdots, d_{1,R}, \cdots, d_{K,1}, \cdots, d_{K,R}) \quad (2.8)$$

where R is the number of regions and K is the number of band pairs (four in our case). The distance between two images I_i and I_j is obtained by computing the Euclidian distance between their representations D_i and D_j:

$$d(I_i, I_j) = ||D_i - D_j|| \quad (2.9)$$

It is worth pointing out that color images are processed by splitting the image into three color channels (HSV), extracting the proposed descriptor on each channel separately, and finally concatenating the three descriptors into a single signature.

As mentioned in Sect. 2.2, it is usually better to combine several image descriptors. In this chapter, we combine the BiCov descriptor with two other ones, namely the (a) Weighted Color Histogram (wHSV) and (b) the MSCR, such as that defined in [6]. For simplicity, we denote this combination as eBiCov (enriched BiCov). The difference between two eBicov signatures $D_1 = (HA_1, MSCR_1, BiCov_1)$ and $D_2 = (HA_2, MSCR_2, BiCov_2)$ is computed as:

$$d_{eBiCov}(D_1, D_2) = \frac{1}{3}d_{wHSV}(HA_1, HA_2) + \frac{1}{3}d_{MSCR}(MSCR_1,$$
$$MSCR_2) + \frac{1}{3}d(BiCov_1, BiCov_2) \qquad (2.10)$$

Obviously, further improvements could be obtained by optimizing the weights (i.e., *using a supervised approach*), but as we are looking for an unsupervised method, we fix them once for all. Regarding the definition of d_{wHSV} and d_{MSCR}, we use the ones given in [6].

2.3.3 BiCoV Analysis

By combing Gabor filters and covariance descriptors—which are both known to be tolerant to illuminations changes [37]—the BiCov representation is robust to illumination variations.

In addition, BiCov is also robust to background variations. Roughly speaking, background regions are not as contrasted as foreground ones, making their Gabor features (and therefore their covariance descriptors) at different neighboring scales very similar. Since the BiCov descriptor is based on the difference of covariance descriptors, background regions are, to some extent, filtered out.

Finally, it is worth pointing out that our approach makes a very different use of the covariance descriptor. In the literature, covariance-based similarity is defined by the difference between covariance descriptors computed on two different images. Knowing how time-consuming it is to compute eigenvalues, the standard approach which requires to evaluate Eq. 2.7 for computing the distance between the query and each image of the gallery can hardly be used with large galleries. In contrast, BiCov computes the similarity of covariance descriptors *within* the same image, between two consecutive scales, once for all. These similarities are then concatenated to obtain the image signature, and the difference of probe and gallery images is obtained by simply computing the $l2$ distance between their signatures.

Fig. 2.3 VIPeR dataset: Sample images showing same subjects from different viewpoints

2.3.4 Experiments

The proposed representation has been experimentally validated on two datasets for person re-identification (VIPeR [17] and ETHZ [12]).

Person Re-identification on the VIPeR Dataset

VIPeR is specifically made for viewpoint-invariant pedestrian re-identification. It contains 1,264 images of 632 pedestrians. There are exactly two views per pedestrian, taken from two nonoverlapping viewpoints. All images are normalized to 128×48 pixels. The VIPeR dataset contains a high degree of viewpoint and illumination variations: most of the examples contain a viewpoint change of 90 degrees, as can be seen in Fig. 2.3. This dataset has been widely used and is considered to be one of the benchmarks of reference for person re-identification. All the experiments on this dataset address the unsupervised setting, i.e., without using training data, and therefore not involving any metric leaning.

We use the Cumulative Matching Characteristic (CMC) curve [24] and Synthetic Reacquisition Rate (SRR) curve [17], which are the two standard performance measurements for this task. CMC measures the expectation of the correct match at rank r while SRR measures the probability that any of the m best matches is correct.

Figure 2.4 shows the performance of the eBicov representation, and gives comparisons with SDALF [6] which is the state-of-the-art approach for this dataset. We follow the same experimental protocol as [6] and report the average performance over 10 different random sets of 316 pedestrians. We can see that eBiCov
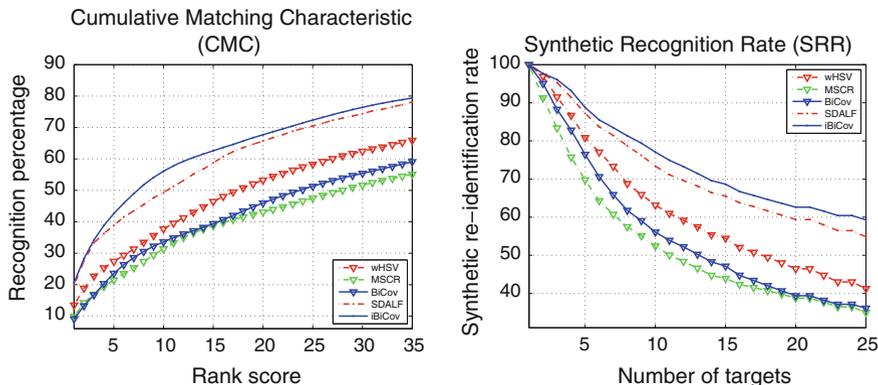
Fig. 2.4 VIPeR dataset: CMC and SRR curves

consistently outperforms SDALF: matching rate at rank 1 for eBiCov is 20.66 %
while that of SDALF is 19.84 %. The matching rate at rank 10 for eBiCov is 56.18
while that of SDALF is 49.37. This improvement can be explained in two ways: on
one hand, most of the false positives are due to severe lighting changes, which the
combination of Gabor filters and covariance descriptors can handle efficiently. On
the other hand, since many people tend to dress in very similar ways, it is important
to capture as fine image details as possible. This is what BIF does. In addition, it is
worth noting that for these experiments the orientation of Gabor filters is not used,
allowing to reduce the computational cost. We have indeed experimentally observed
that the performance is almost as good as that with orientations.

Finally, Fig. 2.4 also reports the performance of the three components of the
eBicov components (i.e., BiCov, wHSV, and MSCR) when used alone.

Person Re-identification on the ETHZ Dataset

The ETHZ dataset contains three video sequences of crowded street scenes captured
by two moving cameras mounted on a chariot. SEQ. #1 includes 4,857 images of 83
pedestrians, SEQ. #2 1,961 images of 35 pedestrians, and SEQ. #3 1,762 images of
28 pedestrians. The most challenging aspects of ETHZ are illumination changes and
occlusions. We follow the evaluation framework proposed by [6] to perform these
experiments.

Figure 2.5 shows the CMC curves for the three different sequences, for both single
($N = 1$) and multiple shots ($N = 2, 5, 10$) cases. In the single-shot case, we can see
that the performance of BiCov alone is already much better than that of SDALF, on
all of the three sequences. The performance of eBiCov[1] is greatly improved on SEQ.
1 and 2. In particular, on SEQ. 1, eBiCov is 7 % better than SDALF at ranks between

---

[1] Remember that eBiCov is the combination of BiCov, MSCR, and wHSV.
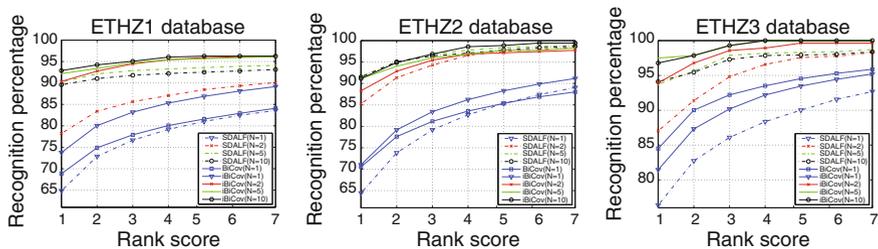
Fig. 2.5 The CMC *curves* on the ETHZ dataset

1 and 7. In SEQ. 2, matching rate at rank 1 is around 71 % for eBiCov and 64 % for SDALF. Compared with the improvements observed on VIPeR, improvements on ETHZ are even more obvious. As the images come from a few video sequences, they are rather similar and the performance is more heavily dependent on the quality of the descriptor.

Besides the single-shot setting, we also tested our method in the multishot case. As in [6], N is set to 2, 5, 10. The results are given in Fig. 2.5. It can be seen that on SEQ. 1 and 3, the proposed eBiCoV gives much better results than SDALF. It is even more obvious on SEQ. 3 for which our method's CMC is equal to 100 % for $N = 5, 10$, which experimentally validates our descriptor.

2.4 Fisher Vector Encoded Local Descriptors for Person Re-identification

This section presents our second descriptor and experimentally demonstrates its effectiveness on the two previously mentioned benchmarks.

As explained in the Introduction, this descriptor is based on local features embedding. The most common approach for combining local features into a global signature is the Bag-of-Words (BoW) model [35], in which local features extracted from an image are mapped to a set of pre-learned visual words, the image being represented as a histogram of visual word occurrences. The BoW model has been used for person re-identification in [41], where the authors built groups of descriptors by embedding the visual words into concentric spatial structures and by enriching the BoW description of a person by the contextual information coming from the surrounding people. Recently, the BoW model has been greatly enhanced by the Fisher Vector [26] which encodes higher order statistics of local features. Compared with BoW, Fisher Vectors encode how the parameters of the model should be changed to optimally represent the image, rather than only the number of visual words occurrences. It has been shown that the resultant Fisher Vector gives excellent performance for several challenging object recognition and image retrieval tasks [27, 28]. Motivated by these recent advances, we propose to combine Fisher Vectors with a novel and very simple seven-dimensional local descriptor adapted to the representation of persons

images, and to use the resultant representation (*Local Descriptors encoded by Fisher Vector* or LDFV) to describe persons. Specifically, in LDFV, each pixel of an image is converted into a seven-dimensional local feature, which contains the coordinates, the intensity, the first-order and second-order derivative of this pixel. Then, the local features are encoded and aggregated into a global Fisher Vector, i.e., the LDFV representation. In addition, metric learning can be used to further improve the performance by providing a metric adapted to the task (e.g. [17, 29, 41]). We used in this section the Pairwise Constrained Component Analysis (PCCA) proposed by [23].

2.4.1 Local Image Descriptor

In order to capture the local properties of images, we have designed a very simple seven-dimensional descriptor inspired by [37] as well as by the method proposed in the first section of this chapter:

$$f(x, y, I) = (x, y, I(x, y), I_x(x, y), I_y(x, y), I_{xx}(x, y), I_{yy}(x, y)) \qquad (2.11)$$

where x and y are the pixel coordinates, $I(x, y)$ is the raw pixel intensity at position (x, y), I_x and I_y are the first-order derivatives of image I with respect to x and y, and I_{xx} and I_{yy} are the second-order derivatives.

Let $M = \{m_t, t = 1, \ldots, T\}$ be the set of the T local descriptors extracted from an image. The key idea of Fisher Vectors [26] is to model the data with a generative model and compute the gradient of the likelihood of the data with respect to the parameters of the model, i.e., $\nabla_\lambda \log p(M|\lambda)$. We model M with a Gaussian mixture model (GMM) using Maximum Likelihood (ML) estimation. Let $\hat{u}_\lambda$ be the GMM model: $\hat{u}_\lambda(m) = \sum_{i=1}^{K} w_i u_i(\mu_i, \sigma_i)$, where K is the number of Gaussian components. The parameters of the models are $\lambda = \{w_i, \mu_i, \sigma_i, i = 1, \ldots, K\}$, where w_i denotes the weight of the i-th component, while μ_i and σ_i are its mean and its standard deviations. We assume the covariance matrices are diagonal and σ_i represents the vector of standard deviations of the i-th component of the model. It is worth pointing out that, considering the computational efficiency for each image in the training set, only a randomly selected subset of local features is used to train the GMM model.

After getting the GMM, image representations are computed using Fisher Vector, which is a powerful method for aggregating local descriptors and has been demonstrated to outperform the BoW model by a large margin [8].

Let $\gamma_t(i)$ be the soft assignment of the descriptor m_t to the component i:

$$\gamma_t(i) = \frac{w_i u_i(m_t)}{\sum_{j=1}^{K} w_j u_j(m_t)} \qquad (2.12)$$

$G_{\mu,i}^M$ and $G_{\sigma,i}^M$ are the 7-dimensional gradients with respect to μ_i and σ_i of the component i. They can be computed using the following derivations:

$$G_{\mu,i}^M = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{m_t - \mu_i}{\sigma_i} \right) \tag{2.13}$$

$$G_{\sigma,i}^M = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(m_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \tag{2.14}$$

where the division between vectors is performed as a term-by-term operation. The final gradient vector G is the concatenation of the $G_{\mu,i}^M$ and $G_{\sigma,i}^M$ vectors for $i = 1, \dots, K$ and is therefore $2 \times 7 \times K$-dimensional.

LDFV on color images. Previous works have shown that using color is a useful cue for person re-identification. We use the color information by splitting the image into three color channels (HSV), extract the proposed descriptor on each channel separately, and finally concatenate the three descriptors into a single signature.

Similarity between LDFV representations. Finally, the distance between two images I_i and I_j can be obtained by computing the Euclidean distance between their representations :

$$d(I_i, I_j) = ||LDFV_i - LDFV_j||. \tag{2.15}$$

2.4.2 Extending the Descriptor

Adding spatial Information. To provide a rough approximation of the spatial information, we divide the image into many rectangular bins and compute one LDFV descriptor per bin. Please note that for doing this we compute one GMM per bin. Then, the descriptors of the different bins are concatenated to form the final representation. It is denoted by bLDFV, for bin-based LDFV.

It must be pointed out that our method does not use any body part segmentation. However, adapting the bins to body parts would be possible and could make the results even better.

Combining LDFV with other features. As mentioned in the Introduction, combining different types of image descriptors is generally useful. In this chapter, we combine our bLDFV descriptor with two other descriptors: the Weighted Color Histogram (wHSV) and the MSCR, shown to be efficient for this task [6]. We denote this combination as eLDFV (enriched LDFV). In eLDFV, the difference between two image signatures $eD_1 = (HA_1, MSCR_1, bLDFV_1)$ and $eD_2 = (HA_2, MSCR_2, bLDFV_2)$ is computed as:

$$d_{eLDFV}(eD_1, eD_2) = \frac{1}{6}d_{wHSV}(HA_1, HA_2) + \frac{1}{6}d_{MSCR}(MSCR_1,$$
$$MSCR_2) + \frac{2}{3}d_{bLDFV}(bLDFV_1, bLDFV_2). \quad (2.16)$$

Regarding the definition of d_{wHSV} and d_{MSCR}, we use those given in [6]. For simplicity reasons and because it is not the central part of the chapter, we have set the mixing weights by hand, giving more importance to the proposed descriptor. Learning them could certainly improve the results further.

Using metric learning. In addition to the unsupervised similarity function (Eq. 2.15), we have also evaluated a supervised similarity function in which we use PCCA [23] to learn the metric. This variant is denoted sLDFV for supervised bLDFV. Any metric learning could have been done but we chose PCCA because of its success in person re-identification [23]. PCCA learns a projection into a low-dimensional space where the distance between pairs of data points respects the desired constraints, exhibiting good generalization properties in the presence of high-dimensional data. Please note that the bLDFV descriptors are preprocessed by applying a whitened PCA before PCCA, to make the computation faster. In sLDFV, PCCA is used with a linear kernel.

2.4.3 Experiments

The proposed approach has been experimentally validated on the two previously introduced person re-identification datasets (VIPeR [17] and ETHZ [12, 33]). We present in this section several experiments showing the efficiency of our simple LDFV descriptor and its extensions.

Evaluation of the Image Descriptor

In this section, our motivation is to evaluate the intrinsic properties of the descriptor. For this reason we do not use any metric learning but simply measure the similarity between two persons using the Euclidean distance between their representations.

Evaluation of the simple feature vector. The core of our descriptor is the seven-dimensional simple feature vector given by Eq. 2.11. This first set of experiments aims at validating this feature vector by comparing it with several alternatives, the rest of the framework being exactly the same. We performed experiments with (1) SIFT features (reduced to 64 dimensions by PCA) and (2) Gabor features [13] (with eight scales and eight orientations). For these experiments, we divide the bounding box into 12 bins (3×4) and the number of GMM components is set to 16. For each bin and each one of the three color channels (HSV), we compute the FV model and concatenate the 12 descriptors for obtaining the final representation. The size of the final descriptor is therefore $7 \times 16 \times 12 \times 2 \times 3$ for our 7-d descriptor, $64 \times 16 \times 12 \times 2 \times 3$ for both the SIFT and Gabor descriptor based FV. We then
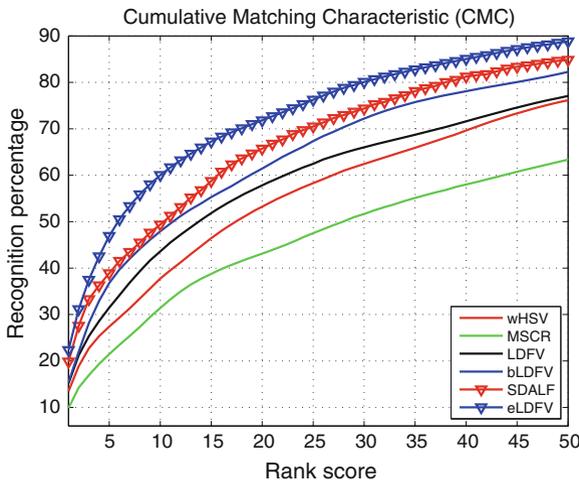
Fig. 2.6 VIPeR dataset: CMC *curves* obtained with LDFV, bLDFV, eLDFV and SDALF

compute CMC normalized Area under Curve (nAUC) on VIPeR and get 83.17, 86.37, and 91.60 %, respectively, for SIFT, Gabor and bLDFV using our seven-dimensional feature vector. Consequently, the proposed descriptor, in addition to being compact and very simple to compute, gives much better results than SIFT and Gabor filters for this task.

We have evaluated the performance of our descriptor for different number of GMM components (16, 32, 50, and 100), and have observed that the performance is not very sensitive to this parameter. Consequently, we use 16 components in all of our experiments, which is a good tradeoff between performance and efficiency.

A set of representative images is required to learn the GMM. We conducted a set of experiments in order to evaluate how critical the choice for these images is. Our experiments have shown that using the whole dataset or only a smaller training set independent from the test set makes almost no difference, showing that, in practice, a small set of representative images is more than enough for learning the GMM.

Single-shot experiments. Single-shot means that a single image is used as the query. We first present some experiments on the VIPeR dataset, showing the relative importance of the different components of our descriptor. The full descriptor (eLDFV) is based on a basic Fisher encoding of the simple seven-dimensional feature vector (LFDV) computed on the three color channels (HSV). The two extensions are (1) bLFDV which embeds spatial encoding and (2) the combination with two other features (namely wHSV and MSCR).

Figure 2.6 shows the performance of eLDFV as well as the performance of wHSV, MSCR, and bLDFV alone. We follow the same experimental protocol as that of [6], and report the average performance over 10 random splits of 316 persons. The figure also gives the performance of the state-of-the-art SDALF [6]. We can draw several conclusions: (1) LDFV alone performs much better than MSCR and wHSV (2) using
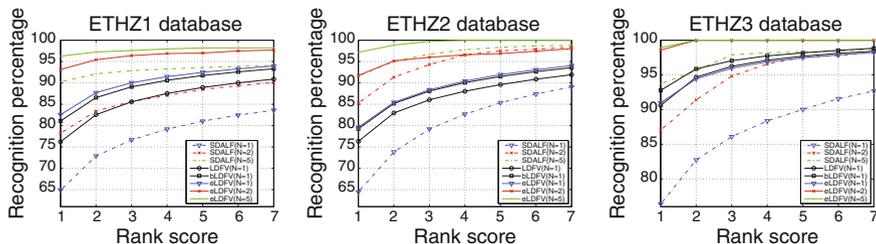
Fig. 2.7 CMC *curves* obtained on the ETHZ dataset

spatial information (bLFDV) improves the performance of LDFV (3) combining the three components (eLDFV) gives a significant improvement over bLDFV and any of the individual components (4) the proposed approach outperforms SDALF by a large margin. For example, the CMC score at rank 1, 10, and 50 for eLDFV are 22.34, 60.04, and 88.82 %, respectively, while those of SDALF are 19.84, 49.37, and 84.84 %.

We have also tested the proposed descriptor on the ETHZ database, in the single-shot scenario ($N = 1$). Here again we follow the evaluation protocol proposed by [6]. Figure 2.7 shows the CMC curves for the three different sequences. In the figure, dashed results come from [6]. The solid line is given by the proposed method. We can see that the performances of LDFV, bLDFV, and eLDFV are all much better than that of SDALF, on all the three sequences, and improvements are even more visible than on VIPeR. Especially, on SEQ. 1 and 3, the performances of eLDFV are much worse than those of bLDFV though eLDFV is the combination of bLDFV, wHSV, and MSCR. We attribute this to the low accuracy of wHSV and MSCR. In particular, on SEQ. 1, the minimum and maximum of the matching rate between the eLDFV and SDALF is about 10 and 18 %, respectively. In SEQ. 2, the matching rate at rank 1 is around 80 % for eLDFV and 64 % for SDALF. The average difference of the matching rate between eLDFV and SDALF, at rank 7, is about 10 % in SEQ. 3.

Multishot experiments on ETHZ. Besides the single-shot case, we also test our descriptors in the multishot case. In this case $N \geq 2$ images are used as queries. We again follow the evaluation framework proposed by [6], the number of query images N being set to 2 and 5. Results are also shown in Fig. 2.7. We can see that on SEQ. 1 and 3, eLDFV gives almost perfect results. Especially, on SEQ. 3, the performance of eLDFV is 100 % with $N \geq 2$, for ranks greater than 2.

Comparison with Recent Approaches

In this section we compare our framework with recent approaches. For making comparison fair, we use here the metric learning algorithm described in Sect. 2.4.2.

We first present some experiments done on the VIPeR dataset. Following the standard protocol for this dataset, the dataset is split into a train and a test set by randomly selecting 316 persons out of the 632 for the test set, the remaining persons
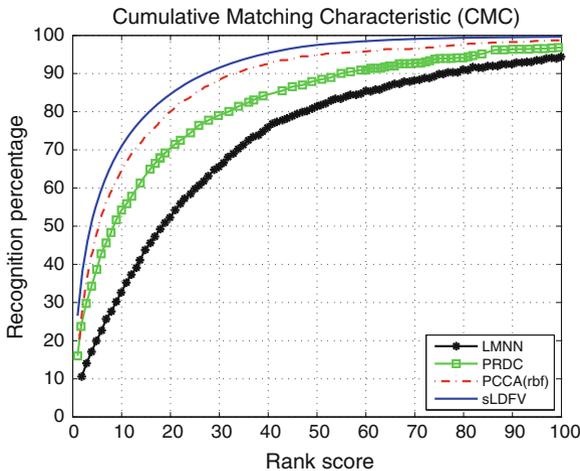
Fig. 2.8 VIPeR dataset: CMC *curves* with 316 persons

Table 2.2 VIPeR dataset: matching rates (%) at rank *r* with 316 persons

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
PRDC [42]	15.66	38.42	53.86	70.09
MCC [42]	15.19	41.77	57.59	73.39
ITML [42]	11.61	31.39	45.76	63.86
LMNN [42]	6.23	19.65	32.63	52.25
CPS [9]	21.00	45.00	57.00	71.00
PRSVM [29]	13.00	37.00	51.00	68.00
ELF [18]	12.00	31.00	41.00	58.00
PCCA-sqrt $n^- = 10$ [23]	17.28	42.41	56.68	74.53
PCCA-rbf $n^- = 10$ [23]	19.27	48.89	64.91	80.28
sLDFV $n^- = 10$	**26.53**	**56.38**	**70.88**	**84.63**

The values of bold are the best performance of different methods at the specific rank.

being in the training set. As in [23], one negative pair is produced for each person, by randomly selecting one image of another person. We produce 10 times more negative pairs than positive ones. The process is repeated 100 times and the results are reported as the mean/std values over the 100 runs.

Figure 2.8 and Table 2.2 compare our approach (sLDFV) with three different approaches using metric learning: PRDC [42], LMNN [38] and PCCA [23]. The results of PRDC and LMNN are taken from [42] while the ones of PCCA come from [23]. For PRDC and LMNN, the image representation is the combination of RGB, YCbCr, and HSV color features and two texture features extracted by local derivatives and Gabor filters on six horizontal strips. For PCCA, the feature descriptor is a 16 bin color histogram in three color spaces (RGB, HSV, and YCrCb) as well as texture histograms based on Local Binary Patterns (LBP) computed on six nonoverlapping horizontal strips. PCCA [23] reports state-of-the-art results for per-

son re-identification, improving over Maximally Collapsing Classes [16], ITML [10] or LMNN-R [11].

Figure 2.8 and Table 2.2 show that the proposed approach (sLDFV) performs much better than any previous approaches. For example, if we compare sLDFV with PCCA, we can see that matching rates at rank 1, 10, and 20 are 26.53, 70.88, and 84.63 % for sLDFV, while those of PCCA are only 19.27, 64.91, and 80.28 %. It must be pointed out that sLDFV is not using any nonlinear kernel, from which we can expect further improvements.

2.5 Conclusions

This chapter proposes two novel image representations for person re-identification, with the objective of being as robust as possible to background, occlusions, illumination, or viewpoint changes. The first representation—so-called BiCov—combines Biologically Inspired Features (BIF) and covariance descriptors. BiCov is more robust to illumination, scale, and background variations than competing approaches which makes it suitable for person re-identification. The second representation—namely LDFV—is based on a simple seven-dimensional feature representation encoded by Fisher Vectors. We have validated these two descriptors on two challenging public datasets (VIPeR and ETHZ) for which they outperformed all current state-of-the-art methods.

Though both the two proposed representations outperform state-of-the-art approaches, they have their own characteristics. While BiCov is usually not as well performing as LDVF, it is worth pointing out that it does not need any training images, which is a huge advantage for real applications. In addition, it is very fast as the most computational demanding step is to extract the low-level features. On the other hand, LDFV requires to build a GMM model during the training stage, which is time-consuming. However, after getting the GMM model, the computation of the representation of the testing sample is very fast, which makes it usable in online systems.

References

1. Ayedi, W., Snoussi, H., Abid, M.: A fast multi-scale covariance descriptor for object re-identification. Pattern Recogn. Lett. (2011)
2. Aziz, K., Merad, D., Fertil, B.: People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation. In: Proceedings

of International Conference on Advanced Video and Signal-Based Surveillance, pp. 303–308 (2011)

3. Aziz, K., Merad, D., Fertil, B.: Person re-identification using appearance classification. In: International Conference on Image Analysis and Recognition, Burnaby (2011)

4. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using haar-based and DCD-based signature. In: Proceedings of International Workshop on Activity Monitoring by Multi-camera Surveillance Systems (2010)

5. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean Riemannian covariance grid. In: Proceedings of International Conference on Advanced Video and Signal-Based Surveillance (2011)

6. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Comput. Vis. Image Underst. **117**(2), 130–144 (2013)

7. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recogn. Lett. **33**(7), 898–903 (2012). (Special Issue on Awards from ICPR 2010)

8. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proceedings of British Machine Vision Conference (2011)

9. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of British Machine Vision Conference (2011)

10. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of International Conference on Machine Learning, pp. 209–216 (2007)

11. Dikmen, M., Akbas, E., Huang, T., Ahuja, N.: Pedestrian recognition with a learned metric. Proc. Asian Conf. Comput. Vis. **4**, 501–512 (2010)

12. Ess, A., Leibe, B., Schindler, K., van Gool, L.: A mobile vision system for robust multi-person tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)

13. Fisher, R.A.: The use of multiple measures in taxonomic problems. Ann. Eugenics **7**, 179–188 (1936)

14. Gandhi, T., Trivedi, M.: Person tracking and re-identification: introducing panoramic appearance map (PAM) for feature representation. Mach. Vis. Appl. **18**(3–4), 207–220 (2007)

15. Gheissari, N., Sebastian, T., Tu, P., Rittscher, J., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1528–1535 (2006)

16. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Advances in Neural Information Processing Systems (2006)

17. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2007)

18. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer Vision, pp. 262–275 (2008)

19. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: Proceedings of the IEEE International Conference on Computer Vision(2009)

20. Guo, G., Mu, G., Fu, Y., T.S. Huang: Human age estimation using bio-inspired features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 112–119 (2009)

21. Kai, J., Bodensteiner, C., Arens, M.: Person re-identification in multi-camera networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio Workshops, pp. 55–61 (2011)

22. Meyers, E., Wolf, L.: Using biologically inspired features for face processing. Int. J. Comput. Vis. **76**(1), 93–104 (2008)

23. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2012)

24. Moon, H., Phillips, P.: Computational and performance aspects of PCA-based face-recognition algorithms. Perception **30**(3), 303–321 (2001)
25. Oreifej, O., Mehran, R., Shah, M.: Human identity recognition in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)
26. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
27. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: Proceedigs of the IEEE Conference on Computer Vision and Pattern Recognition (2010)
28. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: Proceedings of the European Conference on Computer Vision, pp. 143–156 (2010)
29. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference (2010)
30. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nat. Neurosci. **2**(11), 1019–1025 (1999)
31. Satta, R., Fumera, G., Roli, F.: Exploiting dissimilarity representations for person re-identification. In: Proceedings of the International Workshop on Similarity-Based Pattern Analysis and Recognition (2011)
32. Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V.: A multiple component matching framework for person re-identification. In: International Conference on Image Analysis and Processing (2011)
33. Schwartz, W., Davis, L.: Learning discriminative appearance based models using partial least squares. In: Brazilian Symposium on Computer Graphics and Image Processing (2009)
34. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 994–1000 (2005)
35. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proceedings of IEEE International Conference on Computer Vision (2003)
36. Song, D., Tao, D.: Biologically inspired feature manifold for scene classification. IEEE Trans. Image Process. **19**, 174–184 (2010)
37. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. **30**(10), 1713–1727 (2008)
38. Weinberger, K., Saul, L.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
39. Wiskott, L., Fellous, J.M., Krüger, N., Malsburg, C.V.D.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 775–779 (1997)
40. Zhang, Y., Li, S.: Gabor-LBP based region covariance descriptor for person re-identification. In: International Conference on Image and Graphics, pp. 368–371 (2011)
41. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: Proceedings of British Machine Vision Conference (2009)
42. Zheng, W., Gong, S., Xiang, T.: Re-identification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 653–668 (2013)