Covariance Descriptor based on Bio-inspired Features for Person re-Identification and Face Verification

Bingpeng Ma[a], Yu Su[b] and Frédéric Jurie[b]

[a] *School of Computer and Control Engineering, University of China Academy Science, Beijing, China.*
Email: bpma@ucas.ac.cn. Tel: +86-10-6967-1794. Fax: +86-10-6967-1794.
[b] *GREYC — CNRS UMR 6072, University of Caen Basse-Normandie, Caen, France.*
Email: nanosuyu@gmail.com, frederic.jurie@unicaen.fr

Abstract

Avoiding the use of complicated pre-processing steps such as accurate face and body part segmentation or image normalization, this paper proposes a novel face/person image representation which can properly handle background and illumination variations. Denoted as gBiCov, this representation relies on the combination of Biologically Inspired Features (BIF) and Covariance descriptors [1]. More precisely, gBiCov is obtained by computing and encoding the difference between BIF features at different scales. The distance between two persons can then be efficiently measured by computing the Euclidean distance of their signatures, avoiding some time consuming operations in Riemannian manifold required by the use of Covariance descriptors. In addition, the recently proposed KISSME framework [2] is adopted to learn a metric adapted to the representation. To show the effectiveness of gBiCov, experiments are conducted on three person re-identification tasks (VIPeR, i-LIDS and ETHZ) and one face verification task (LFW), on which competitive results are obtained. As an example, the matching rate at rank 1 on the VIPeR dataset is of 31.11%, improving the best previously published result by more than 10%.

Keywords: image representation, person re-identification, face verification, biologically inspired features, covariance descriptor.

1. Introduction

The task of person re-identification consists in recognizing an individual through different cameras in a distributed network or through the same camera capturing images at different time. This is a challenging problem that has attracted a lot of attention in recent years. The key issue of such systems lies in their ability to measure the similarity between two person-centered bounding boxes, i.e. to predict if they represent to the same person, despite changes in illumination, pose, viewpoint, background, partial occlusions and low resolution. In order to tackle this problem, the dominant strategy is to combine feature sets into templates, used as person descriptors, and to measure the similarity between templates to predict persons' identities. Descriptors adapted to the re-identification of faces are usually different than those for person re-identification. Indeed, face verification required to be able to capture smaller details of the input image, as intra-class and inter-class variations is smaller that for person re-identification. It is challenging for a descriptor to handle both tasks at the same time. Finally, even if such person or face descriptors have received a lot of attention during the last decade, they still need some improvement before they can be used for real applications. This is the motivation for the presented work.

Extending the work presented in [3], this paper presents a novel image representation for person re-identification and face verification. Specifically, the proposed image representation allows to measure efficiently the similarity between two persons/faces without any pre-processing step (e.g., precise background subtraction or body part segmentation). This paper mainly focuses on person re-identification which has received less attention than face verification, however we experimentally demonstrate that the proposed representation also works well for face verification. In both scenarios, we assume that pedestrians/faces have been previously detected and cropped.

The proposed method, denoted as gBiCov, includes three steps. In the first step, Biologically Inspired Features (BIF) [4] are extracted. BIFs are based on the study of human visual system and have shown excellent performances on several computer vision tasks [5], [6], [7]. In particular, we use the S1 layer (Gabor filters) and C1 layer (MAX operator) of BIF. While the Gabor filters can improve robustness to illumination variations, the MAX operator increases the tolerance to scale changes and image shifts. In the second step, a Covariance descriptor is used to compute the similarity of BIF features taken at neighboring scales. Covariance descriptors can capture shape, location and color information, and their performance have been shown to be better than other methods in many situations, as rotations and illumination changes are absorbed by the covariance matrix [1]. Furthermore, we argue that measuring the similarity of BIF at neighboring scales decreases the influence of the background (see Section 3.5 for details). In the third step, BIF and covariance descriptors are combined into a single representation. Finally, we show that the performance of the proposed representation can be further enhanced by the use of metric leaning

(we use the KISSME framework of [2]). Since the resulting representation is robust to illumination, scale and background changes, the performance for person re-identification and face verification can be significantly improved.

In addition to presenting an approach performing well on real datasets, one interesting contribution of the paper lies in the use of Covariance descriptors in a novel way. In traditional covariance-based method, the similarity of two images can be obtained by comparing their covariance descriptors [8, 9, 10], which is a time-consuming process. In contrast, in the proposed approach, the similarities of Covariance descriptors between consecutive bands of BIF features in the same image are measured. These similarities are then concatenated to produce image signatures, and the similarity between probe and gallery images is obtained by simply computing the $L2$ distance between their signatures. It avoids the expensive computation of the similarity between Covariance descriptors of probe image and each gallery image, which can be extremely time-consuming when the gallery is large.

The proposed method is experimentally validated on three public datasets for person re-identification: VIPeR, i-LIDS and ETHZ. They are among the most challenging ones, since all the above-described issues such as pose changes, viewpoint and lighting variations, and occlusions, are present. As an illustration of the performance, the matching rate at rank 1 (*i.e.*, considering only the most similar image of the gallery) is of 31.11% on the VIPeR dataset (10% better than best previously published result). Knowing that the matching rate at low ranks is the most important criterion for real-life applications, this improvement is very significant. The proposed method is also validated on a face verification dataset, the Labeled Faces in the Wild (LFW) dataset, and compared to recently published state-of-the-art approaches.

The remaining of this paper is organized as follows: Section 2 reviews the related works on person's re-identification and face verification. Section 3 describes the proposed method in details and discusses its advantages. Experimental validations are given in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

Person/face re-identification – which is the task of associating the same person through different cameras or at different time – is a challenging problem as the association has to be done despite view point, illumination and pose changes. It has received a lot of attention in the recent literature, reflecting the interest for the important applications that can be addressed with these technologies.

More formally, the task of person re-identification can be defined as finding the correspondences between the images of a *probe set* representing a single person and the images of a *gallery set*. Depending on the number of available images per individual (size of the probe set), the scenarios can be defined as: (a) single-shot [11, 12], if only one frame per individual is available both in the probe and gallery sets; and (b) multiple-shot [11, 12], if multiple frames per individual are available both in the probe and gallery sets.

One of the key ingredient of face/person re-identification approaches lies in the encoding of images into visual signatures that can be compared more efficiently than raw pixel intensities. The recent literature abounds with such image descriptors for person re-identification. They can be based on (i) color, widely used since the color of clothing constitutes simple but efficient visual signatures, usually encoded within histograms of RGB or HSV values [11], (ii) shape, *e.g.* using HOG based signature [13, 14], (iii) texture, often represented by Gabor filters [15, 10, 16], differential filters [15, 16], Haar-like representations [17] and Co-occurrence Matrices [14], (iv) interest points, *e.g.* SURF [18] and SIFT [19, 20] and (v) image regions [13, 11, 12]. Besides these generic representations, there are some more specialized representations, *e.g.* Epitomic Analysis [21], Spin Images [22, 23], Bag-of-Word-based description [20], Implicit Shape Model (ISM) [19] and Panoramic Map [24]. Since different elementary features capture different and complementary aspects of the image, better performance is obtained by combining several signatures. We point this out in the following section.

Among these methods, those based on representing humans by a collection of parts have attracted more and more attention. Part-based methods split the human body into different parts and encode each part separately. In [11, 12], the authors use Maximally Stable Color Regions (MSCR) to build a representation of human body. MSCR consists in grouping pixels having similar colors into maximally stable regions during a clustering process. The regions are subsequently described by their area, centroid, second moment matrices and average colors. Interestingly, covariance descriptors have also been widely used for representing regions [8, 9, 10]; the pixels within a region are represented by a feature vector consisting of intensity, texture and shape statistics, while the regions are represented by the covariance matrix of these feature vectors.

As mentioned above, since different elementary features (color, shape, texture, etc.) capture different and complementary characteristics of the image, they are often combined to give a richer signature. For example, [15] combines 8 color features with 21 texture filters (Gabor and differential filters). [11] and [12] combine MSCR descriptors with weighted color histograms, achieving state-of-the-art results on several wildly-used person re-identification datasets. The Covariance descriptor can be generalized to any type of images (three channel color images, infrared images, *etc.*), and can be used to combined different descriptors. For example, in [10], Gabor features and Local Binary Patterns (LBP) are combined with a covariance descriptor which handles, to some extent, illumination and viewpoint changes as well as non-rigid deformations.

Different representations usually require different similarity measures. For example, representations based on histograms can be compared with the Bhattacharyya distance [11, 21, 12] or the Earth Movers Distance (EMD). When the representation includes two or more different features/channels, the similarity is usually computed by combining their respective similarities (late fusion) *e.g.* using a linear combination [21, 12, 11]. Re-

garding the methods based on the covariance descriptor, even if the similarity of two regions is computed by estimating the distance between two covariance matrices in a pairwise manner, the similarity of human body described themselves by a set of covariance matrices has to combine several region similarities. This combination can be based, for example, on the mean covariance distance between corresponding regions [25] or by the minimum difference between corresponding body regions [10]. To capture the correlation between body parts, [17] uses spatial pyramid matching and designs a new similarity measure between human signatures. In [9], the authors argue that the covariance matrices lie in a Riemannian manifold, and combine the efficiency of the mean Riemannian covariance descriptor with the spatial information carried out by a dense grid structure. In [8], the authors propose a multi-scale covariance descriptor which describes an image quadrant through its corresponding sub-tree.

In order to improve the performance of these representations in the context of person re-identification, several papers have proposed to use discriminative classifiers on top of them: these classifiers can be based on Adaboost [16, 17], Rank SVM [15], Partial Least Squares (PLS), multi-feature learning [26] or multiple instance learning [27, 28].

Different from these classifiers, metric learning can provide a way to adapt a similarity function to the given task. Simple but efficient are the metric learning methods based on Mahalanobis-like distance functions. Approaches such as Large Margin Nearest Neighbors (LMNN) [29], Information Theoretic Metric Learning (ITML) [30], Logistic Discriminant Metric Learning (LDML) [31], Pairwise Constrained Component Analysis (PCCA) [32] and Keep It Simple and Straightforward Metric Learning (KISSME) [2] have been used successfully in the context of face verification and person re-identification. From the statistical inference perspective, KISSME [2] computes the covariance matrix of similar and dissimilar pairs respectively, and uses the difference of the inverse covariance matrix as a projection matrix. It is very simple and efficient, since it does not involve any iterative optimization procedure. In this paper, KISSME is used to define a metric between descriptor pairs.

Face verification is also a challenging topic which has been studied for several decades. Image representation is, here also, one of the key steps. Compared with person images, the intra-class variations of face images are much smaller, explaining why face verification relies more on smaller details of the input images. Most of the recently proposed face descriptors are built on local descriptors since they allow to capture image details and are robust to variations like expression, illumination, aging, etc. The most widely used face descriptors in the literature include Gabor wavelets, LBP [33] or SIFT/HOG [34] and their variants. In addition to these low-level descriptors, the feature pooling methods allowing to produce a global representation/signature from local features were also intensively investigated (such as the Bag-of-Words [35], Fisher Vectors [36] and Sparse coding [37]). Sparse coding based methods have achieved great success in face representation and recognition. A set of over-complete dictionary is first learned from image patches, allowing to represent images as weighted sums of a small number of code words. This mechanism is, in some sense, similar to the human vision system since in the visual cortex only a small number of neurons are activated at the same time. In parallel, some researchers also tried to build computational models which directly simulate the human vision process [4]. On the classifier side, the nearest neighbor classifier, SVM, Neural Networks are widely used. Recently, the traditional Bayesian model was also revisited [38] and impressive results have been obtained on the challenging LFW dataset.

This paper extends [3] in two directions: (i) in contrast with [3], the proposed image representation combines covariance similarity with biologically inspired features. It results in a significant performance improvement, as the match rate at rank 1 of this new representation is significantly improved (+3%). The mean accuracy on LFW is also improved from 74.03% to 84.48%. (ii) A metric learning stage (relying on the KISSME algorithm) is used to learn a similarity function adapted to the gBiCov representation, giving better results (on VIPeR the matching rate at rank 1 is improved up to more than 31%).

3. Covariance Descriptor based on Bio-inspired Features

This section presents the proposed novel image representation: the Covariance descriptor based on Bio-inspired Features (gBiCov for short). There are three steps in this representation: (i) Biologically Inspired Features (BIF) are first extracted, (ii) BIF are then encoded by comparing their Covariance descriptors at different scales, and, (iii) BIF are combined with Covariance descriptors. The flowchart of the two first steps of gBiCov is given in Fig. 1. In the following, we first present these three steps then introduce some extensions, and, finally outline the advantages of this representation.

3.1. Low-Level Biologically Inspired Features (BIF)

BIF [4], based on the study of the human visual system, have been proposed to address several computer vision tasks such as object category recognition [5], face recognition [6], age estimation [7] and scene classification [39], on which they have obtained excellent performance.

Our representation builds on these prior works. More specifically, for an image $I(z)$ where $z = (x, y)$, we compute its convolution with the Gabor filter $\psi(z)$ accordingly to the following equation [40]:

$$G(\mu, \nu) = I(z) * \psi_{\mu,\nu}(z) \tag{1}$$

where:

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

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu}, k_\nu = 2^{-\frac{\nu+2}{2}}\pi, \phi_\mu = \mu\frac{\pi}{8}, \tag{3}$$

and where μ and ν are scale and orientation parameters respectively. In our work, μ is quantized into 24 scales while ν is quantized into 8 orientations. Gabor filters are inspired by the human visual system and their kernels are very similar to the

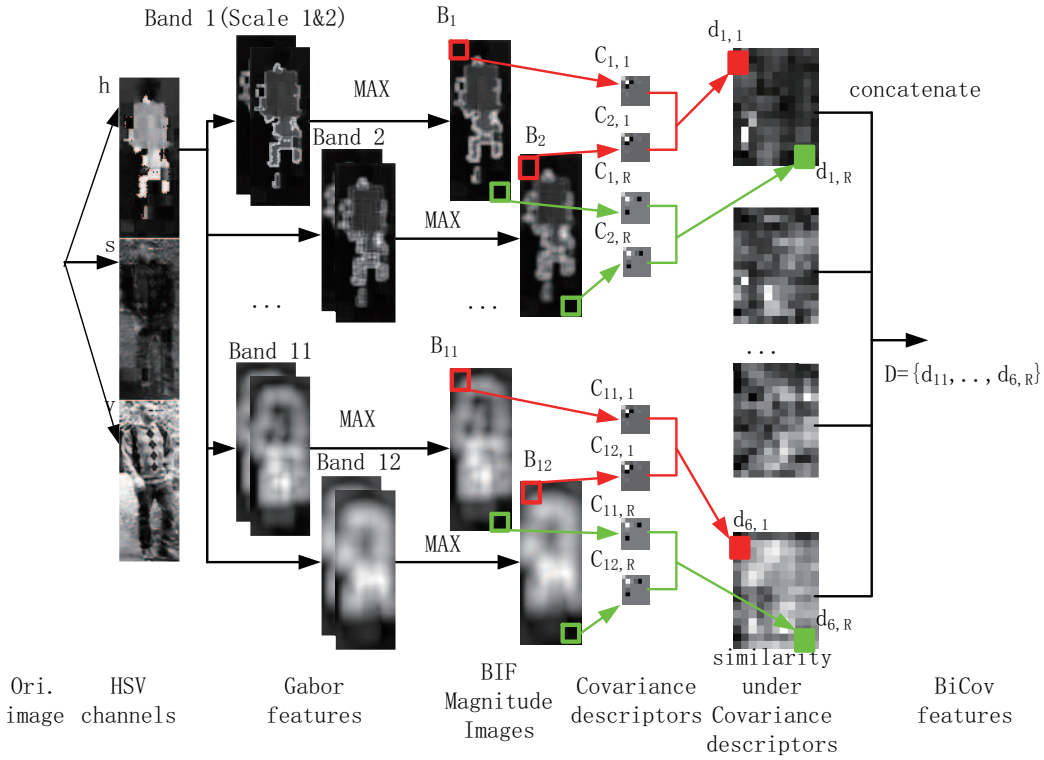


Figure 1: Flowchart of the gBiCov representation. Color images are first split into 3 color channels (HSV). These input images are convolved with Gabor filters at different scales, and the neighboring scales are grouped into bands. BIF Magnitude Images are obtained by using the max operator within the same band of the Gabor features. BIF Magnitude Images are then divided into small regions, represented by covariance descriptors. We compute the difference of covariance descriptors between the corresponding regions of the different bands. These differences are then concatenated to form the final representation of the image.

2-D receptive field profiles of the mammalian cortical simple cells.

In practice, we have observed that for the person re-identification task, the image representations $G(\mu, \nu)$ for different orientations can be averaged without significant loss of performance. Thus, we replace $\psi_{\mu,\nu}(z)$ in Eq. 1 by $\psi_{\mu}(z)$:

$$\psi_{\mu}(z) = \frac{1}{8} \sum_{\nu=1}^{8} \psi_{\mu,\nu}(z) \tag{4}$$

This simplification makes the computations of $G(\mu)$ – which is the average of $G(\mu, \nu)$ over all orientations – more efficient.

In practice, the number of scales is fixed to 24 and two consecutive scales are grouped into one band (we therefore have 12 different bands). The size of the Gabor filters for the different bands are shown in Tab. 1. We then apply a max-pooling over two consecutive scales (within the same orientation if the orientations are not merged):

$$B_i = \max(G(2i-1), G(2i)) \tag{5}$$

Max-pooling increases the tolerance to small scale changes which often appears in person and face images since they are only roughly aligned. We call B_i $i \in [1, \ldots, 12]$ as *BIF magnitude images*. Fig. 2 shows a pair of images of the same person and its respective BIF magnitude images. The image in the first column is the input image while the second column shows its three HSV channels. The images from the 3rd column to the 8th column are BIF magnitude images corresponding to the 6 different bands.

3.2. Covariance descriptors in the gBiCov Descriptor

During this step, each BIF magnitude image is divided into small overlapping regions. In this way, the spatial information of the images is kept. Then, each region is represented by a covariance descriptor [1]. Covariance descriptors can capture shape, location and color information, and their performance have been shown to be better than other methods in many situations, as rotations and illuminations changes are partly absorbed by the covariance matrix [1].

For each pixel of the BIF magnitude image B_i, a 7-dimensional feature vector is computed to capture the intensity, texture and shape statistics:

$$f_i(x, y) = [x, y, B_i, B_{i_x}, B_{i_y}, B_{i_{xx}}, B_{i_{yy}}] \tag{6}$$

where x and y are the pixel coordinates, B_i is the raw pixel intensity at position (x, y), B_{i_x} and B_{i_y} are the derivatives of image B_i with respect to x and y, $B_{i_{xx}}$ and $B_{i_{yy}}$ are the second order derivatives of image B_i with respect to x and y. The input image region is mapped to the covariance region represented by a 7×7 matrix.

After that, we compute the covariance descriptor for each one of the small overlapping regions previously introduced:

$$C_{i,r} = \frac{1}{n-1} \sum_{(x,y)\in r} (f_i(x, y) - \bar{f}_i)(f_i(x, y) - \bar{f}_i)^T \tag{7}$$

Table 1: Scales of Gabor filters for the different bands.

Band	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}
filter sizes	3×3	7×7	11×11	15×15	19×19	23×23	27×27	31×31	35×35	39×39	43×43	47×47
filter sizes	5×5	9×9	13×13	17×17	21×21	25×25	29×29	33×33	37×37	41×41	45×45	49×49

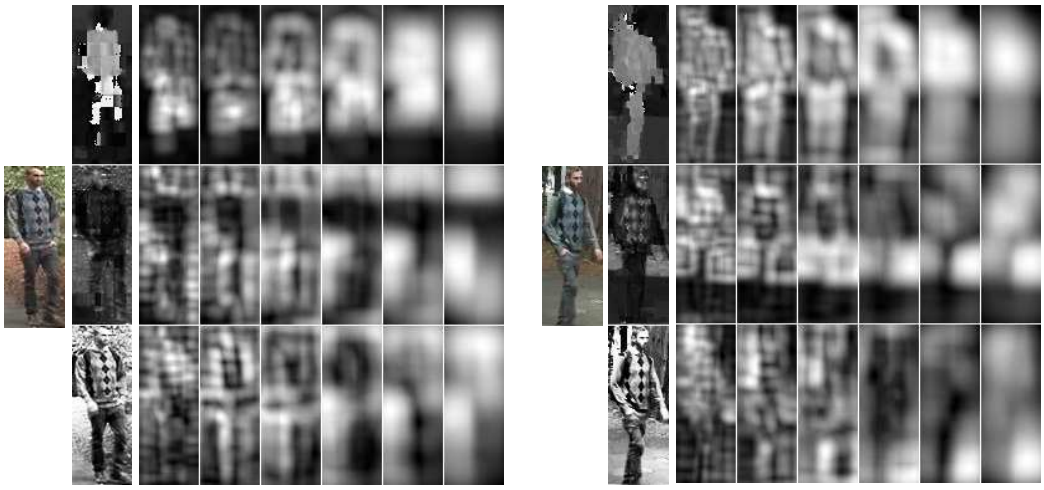


Figure 2: A pair of images representing the same person, and their BIF magnitude images. The images in the first and second column are the input images and their three channels in H, S and V channel, respectively. The images from the 3th column to the 8th column are the BIF Magnitude Images for the different bands.

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).

In traditional covariance-based methods, covariance matrices computed by Eq. 7 are considered as the image representation. Differently, in this paper, we compute for each region the difference of covariance descriptors between two consecutive bands:

$$d_{i,r} = d(C_{2i-1,r}, C_{2i,r}) = \sqrt{\sum_{p=1}^{P} \ln^2 \lambda_p(C_{2i-1,r}, C_{2i,r})} \quad (8)$$

where $\lambda_p(C_{2i-1,r}, C_{2i,r})$ is the $p-$th generalized eigenvalues of $C_{2i-1,r}$ and $C_{2i,r}$, $i = 1, \ldots, 6$.

3.3. BIF in the gBiCov Descriptor

Though we can take $d_{i,r}$ as the representation of gBiCov directly, considering the success of BIF magnitude features in many areas, we also combine the BIF magnitude features in the gBiCov descriptor. BIF magnitude features can be seen as appearance-based features while the covariance matrices can be seen as a description of feature properties. To a certain extent, BIF magnitude features and covariance matrices are two different levels of the entire representation.

By denoting $\bar{B}_{2i-1,r}$ and $\bar{B}_{2i,r}$ the mean of BIF magnitude features of region r under band $2i-1$ and $2i$, respectively, we compute $b_{i,r}$ as the average of the BIF magnitude features of these two bands:

$$b_{i,r} = \frac{\bar{B}_{2i-1,r} + \bar{B}_{2i,r}}{2} \quad (9)$$

We simply concatenate BIF features $b_{i,r}$ with covariance feature $d_{i,r}$, after normalizing them:

$$\hat{d}_{i,r} = \frac{\sqrt{|d_{i,r}|}}{\sqrt{\sum_{i=0}^{M}\sum_{r=0}^{R} d_{i,r}^2}} \quad (10)$$

$$\hat{b}_{i,r} = \frac{\sqrt{|b_{i,r}|}}{\sqrt{\sum_{i=0}^{M}\sum_{r=0}^{R} b_{i,r}^2}} \quad (11)$$

where R and M are the number of regions and bands, respectively.

Finally, they are concatenated to form the image representation $\mathbf{D}$:

$$\mathbf{D} = (\hat{d}_{1,1}, \cdots, \hat{d}_{M,R}, \hat{b}_{1,1}, \cdots, \hat{b}_{M,R}) \quad (12)$$

It is worth pointing out that color images are processed by splitting the image into 3 color channels (HSV), extracting gBiCov descriptors on each channel separately and finally concatenating the 3 descriptors into a single signature.

The resulting feature $\mathbf{D}$ lies in a high dimensional space. Here we show that simple dimensionality reduction method such as Principal Component Analysis (PCA) [41] is a good option for compressing the features. PCA is a linear transform technique, which reduces the dimensionality of features while preserving most of their variance. The projection matrix $\mathbf{W}_{pca}$ is made of the orthogonal eigenvectors of the covariance matrix. The experiment section shows that the drop in performance is small, even when the dimensionality reduction is significant. In some cases, the low-dimensional features even perform better, which can be interpreted over-fitting reduction.

Finally, the distance between two images I_i and I_j is obtained by computing the Euclidean distance between their low-dimensional representations $\mathbf{D}_i$ and $\mathbf{D}_j$:

$$d(I_i, I_j) = \|\mathbf{W}_{pca} \times \mathbf{D}_i - \mathbf{W}_{pca} \times \mathbf{D}_j\| \tag{13}$$

3.4. gBiCov Extensions

3.4.1. eBiCov: combining gBiCov with additional image features

As mentioned in Section 2, especially in the context of person re-identification, the performance is usually improved by combining different types of image descriptors. In this paper, we follow the same methodology and combine the gBiCov descriptor with other two representations: (i) Weighted Color Histograms (wHSV) and (ii) MSCR as defined in [11]. For notational simplicity, we denote this combination as eBiCov (for enriched gBiCov). While SDALF, which is the current state-of-the-art approach for unsupervised person re-identification, uses a combination of wHSV, MSCR and Recurrent High-Structured Patches (RHSP), [12] has observed that RHSP can be removed without significant loss in performance. Consequently, eBiCov can be seen as the combination of SDALF and gBiCov. In eBiCov, the difference between two image signatures $\mathbf{D}_1^{'} = (HA_1, MSCR_1, gBiCov_1)$ and $\mathbf{D}_2^{'} = (HA_2, MSCR_2, gBiCov_2)$ is computed as:

$$\begin{aligned} d_{eBiCov}(\mathbf{D}_1^{'}, \mathbf{D}_2^{'}) = \frac{1}{3} d_{wHSV}(HA_1, HA_2) + \frac{1}{3} d_{MSCR} \\ (MSCR_1, MSCR_2) + \frac{1}{3} d_{BiCov}(gBiCov_1, gBiCov_2) \end{aligned} \tag{14}$$

Improvements might be obtained by optimizing the weights based on additional information, *e.g.* class labels, other priors and cross validation. However, to show the intrinsic quality of the descriptor, we have simply used this simple fixed-weights combination. Regarding the definition of d_{wHSV} and d_{MSCR}, we use the ones given in [11].

3.4.2. kBiCov: comparing gBiCov signatures using learnt metrics

In addition to the simple Euclidean distance (Eq. 13), we have also investigated how a learnt metric could improve performance, assuming a training set is available.

More precisely, we focused our investigations on the class of Mahalanobis-like distance functions, which has gained considerable interest in the recent computer vision literature (see Section 2). In this paper, considering its great success in face recognition and person re-identification, we build on the KISSME framework [2] as a general metric learning approach.

As stated by the authors of [2], the main advantage of KISSME is the simplicity and efficiency of the learning process, as it only requires the computation of two small-sized covariance matrices, one for the positive class (pairs of vectors of the same class) and the other for the negative class (pairs of vectors from different classes). The similarity is based on a likelihood-ratio test applied to the difference of the two vectors to be compared, computing plausibility that the difference belongs to either the positive or the negative class.

More precisely, the matrix $\mathbf{M}$ is computed by the following equations:

$$\mathbf{M} = \mathbf{\Sigma}_{y_{ij}=1}^{-1} - \mathbf{\Sigma}_{y_{ij}=0}^{-1} \tag{15}$$

where

$$\mathbf{\Sigma}_{y_{ij}=1} = \sum_{y_{ij}=1} (x_i - x_j)(x_i - x_j)^T \tag{16}$$

$$\mathbf{\Sigma}_{y_{ij}=0} = \sum_{y_{ij}=0} (x_i - x_j)(x_i - x_j)^T \tag{17}$$

where y_i is the label of sample x_i. $y_{ij} = 1$ means similar pairs, *i.e.,* if the samples share the same class label ($y_i = y_j$) and $y_{ij} = 0$ otherwise.

In practice, using a projection matrix is more convenient than using $\mathbf{M}$ directly. We therefore compute the corresponding projection metric using Cholesky factorization:

$$\mathbf{M} = \mathbf{W}_{kiss}^T \times \mathbf{W}_{kiss} \tag{18}$$

At this stage, no dimensionality reduction is performed. However, for the reasons given in the previous section, reducing the dimensionality is generally useful. We here again use PCA, combined with the KISSME metric, giving the following projection matrix:

$$\mathbf{D}^K = \mathbf{W}_{kiss} \times \mathbf{W}_{pca} \times \mathbf{D} \tag{19}$$

Finally, the similarity between two vectors is computed by projecting them using $\mathbf{W}_{kiss} \times \mathbf{W}_{pca}$ (the projection matrix combining PCA and KISSME) and by computing the Euclidean distance between their projected vectors.

This variant is denoted as kissme-gBiCov (kBiCov for short) in the experiments.

3.5. Advantages of gBiCoV

First, combining Gabor filters with covariance descriptors makes gBiCov very robust to illumination variations. On one hand, Gabor filters are known to be robust to illumination changes; on the other hand the covariance descriptor also absorbs illuminations changes [1]. As being the combination of Gabor filters and covariance descriptor, gBiCov can be shown to be even more robust to illumination variations.

Second, gBiCov is also more robust to background variations, *i.e.,* it can achieve good performance without any accurate foreground/background segmentation or body parts detection, which are often even more difficult tasks. Roughly speaking, background regions are usually less textured, which makes their Gabor features (and hence their covariance descriptors) at different neighboring scales very similar. Since the gBiCov descriptor is based on the difference of covariance descriptors at different scales, the gBiCov descriptors extracted from background regions are small and do not impact the similarity between descriptors a lot.

Third, the way of using the covariance descriptor in this paper is very different than what is usually done. Indeed, to measure the distance between two images, the traditional way is



Figure 3: VIPeR dataset: sample images showing the same subjects from different viewpoints.

to compute the difference between their covariance descriptors. Since finding the eigenvalues (required for comparing the covariance descriptors) is very time-consuming, it is computationally prohibitive when the gallery set is large. In contrast, gBiCov computes the similarity of covariance descriptors of consecutive scales, and these similarities are concatenated to obtain the image signature. In other words, covariance descriptors are used to capture self-similarities, and not exploited to perform matching between different signatures. Therefore, the time needed to calculate distances between covariances are solely used during the building of the signature. The matching holds in a Euclidean space, which makes it very fast.

4. Experiments

This section presents the experimental validation of the proposed gBiCov representation. The validation is done on three datasets for person re-identification (namely VIPeR [42], i-LIDS [20] and ETHZ [43, 14]) and one for face verification (namely LFW [34]).

4.1. Pedestrian re-identification on the VIPeR Dataset

The *Viewpoint Invariant Pedestrian Recognition* (VIPeR) dataset – as indicated by its name – has been designed for viewpoint-invariant pedestrian re-identification. It contains 1264 images of 632 pedestrians. There are exactly two views per pedestrian, taken from two different viewpoints. All images are resized to 128×48 pixels. Most of the examples contain a viewpoint change of 90 degrees and strong illumination variation, as it can be seen in Fig. 3. This dataset has been widely used and is considered to be one of the benchmarks of reference for pedestrian re-identification.

Measuring the performance of person re-identification is usually done with the Cumulative Matching Characteristic (CMC) curve [44] and the normalized Area Under Curve (nAUC). CMC curves treat re-identification as a ranking problem by representing the probability of finding the correct match over the first k ranks. In other words, $CMC(k)$ can be seen as the recall at k. In contrast, the *Synthetic Reacquisition Rate* (SRR) curve [42] measures the probability that any of the k best matches is correct. The nAUC is the area under the CMC curve, which is the scalar appraisal of CMC curves and can be used to summarize the overall performance. The higher the nAUC is, the better the performance is.

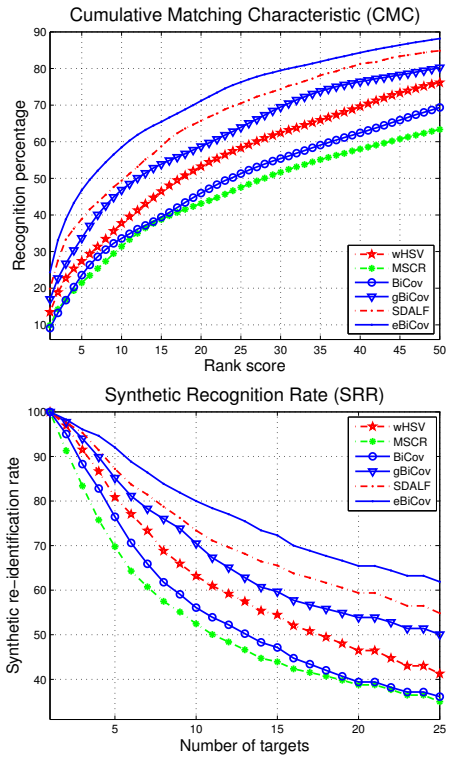


Figure 4: VIPeR dataset: CMC and SRR curves.

4.1.1. Comparing gBicov with other methods

Fig. 4 shows the CMC and SSR curves obtained with the eBicov representation (extended gBicov), as well as the one given by the state-of-the-art person re-identification algorithm, namely SDALF [11]. Since the matching rates at small ranks are very important in real-life applications, Tab. 2 also shows the matching rates at ranks 1, 5, 10, 20 and 50. We follow the same experimental protocol as [11] and report the average performance over 10 different random sets of 316 pedestrians. To show the performance of the descriptor alone, we also report the performance of the 3 components of the eBicov individually (*i.e.,* gBiCov, wHSV and MSCR, as defined in Section 3). Since the performance of the third component in SDALF is much worse than those of wHSV and MSCR, the combination of wHSV and MSCR can be seen as a good approximation of SDALF. To emphasize the improvement of gBiCov over BiCov [3], we report both in the aforementioned table and figure.

From the abovementioned figure and table, we can see that eBiCov consistently outperforms SDALF. For example, the matching rate at rank 1 of eBiCov is 24.34% while the one of SDALF is 19.84%. The good performance of eBiCov is explained by the good performance of gBiCov: its matching rate at ranks 1, 10 and 50 are of 17.01%, 46.84% and 80.24% respectively, while those of wHSV are of 13.49%, 37.36% and 76.17% respectively. When comparing the performance of one single component (MSCR, wHSV) with gBiCov, the advan-

Table 2: VIPeR dataset: Top ranked matching rates (%).

Method	r=1	r=5	r=10	r=20	r=50
wHSV	13.49	27.41	37.36	53.24	76.17
MSCR	9.88	21.46	31.41	43.13	63.36
BiCov	9.01	23.59	33.59	45.95	69.37
gBiCov	17.01	33.67	46.84	58.72	80.24
SDALF	19.87	38.89	49.37	65.73	83.07
eBiCov	**24.34**	**46.75**	**58.48**	**71.17**	**88.18**

tage of the proposed gBiCov descriptor is even more obvious. For example, the matching rate at rank 1 is only of 13.49% and 9.88% for wHSV and MSCR, while that of gBiCov is of 17.01%. In addition, the performance of gBiCov is much better than that of BiCov. This demonstrates the advantage of combining BIF features with covariance descriptors. This improvement can be attributed to the different complementary levels that BIF features and covariance descriptor brings to the image representation.

Compared with wHSV and MSCR, the advantage of gBiCov comes from two factors: on one hand, most of the false positives are due to severe lighting changes. In gBiCov, the combination of Gabor filters and Covariance descriptors strongly alleviates this effect. 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 to overcome the ambiguity introduced by similar clothing. This is where BIF does well. In addition, it is worth noting that for these experiments the orientation of Gabor filters is not used (see Sec. 3), allowing reducing the computational cost. We have experimentally observed that the performance is almost as good as when including orientations.

4.1.2. Analysis of the parameters (region size and overlap)

In gBiCov, there are two important parameters: the size of the regions and their overlap. To show the influence of these parameters, we experimented with different region sizes and with different overlaps. The width of the region is ranged from 4 to 12 pixels while the height is from 8 to 24 pixels. The overlap is set to 25%, 50% or 100% of the region size. The nAUCs of gBiCov with different region and overlapping sizes are shown in Tab. 3. In the table, we also show the dimensionality of the whole representation since it varies greatly under different region and overlapping sizes. The main conclusion is that the performance is not influenced a lot by any of the two parameters. However, one can also see that the performance is as better as the overlap is important and, in general, better for larger regions. However, when the overlap is important, more regions are necessary to cover the image, increasing the dimensionality of the representation. A tradeoff between the performance and the computational cost has therefore to be made. In practice, we set the region size to 16×16 and the overlap to 4×4 in all of our experiments.

4.2. Pedestrian re-identification on the ETHZ dataset

In addition to the previous experiments, we have also experimented the gBiCov representation on the ETHZ database.

The ETHZ dataset contains three video sequences of crowded street scenes captured by two moving cameras mounted on a chariot. The three sequences have: $4,857$ images of 83 pedestrians for SEQ. #1, $1,961$ images of 35 pedestrians for SEQ. #2, and $1,762$ images of 28 pedestrians for SEQ. #3. The most challenging aspects of ETHZ are illumination changes and occlusions. We follow the evaluation framework proposed by [11] to perform the experiments. Besides the single-shot case, we also tested gBiCov in the multi-shot case.

Fig. 5 shows the CMC curves for the three sequences, for both single ($N = 1$) and multiple shots ($N = 2, 5, 10$). In case of single shot, we can see that the performance of gBiCov alone is already much better than that of SDALF, for the three sequences. After adding MSCR and wHSV to gBicov (giving the so called eBiCov representation), the performance is greatly improved. In particular, on SEQ. #1, eBiCov is 9% better than SDALF for ranks between 1 and 7. On SEQ. #2, the matching rate at rank 1 around 76% for eBiCov and 64% for SDALF. Compared with the improvements observed on the VIPeR dataset, improvements on the ETHZ dataset are even more obvious. The reason seems to be that the images of the same person come from video sequences, which makes the task of person re-identification much easier for all the methods.

In case of multi-shots, as in [11], N is set to 2, 5 and 10. From Fig. 5, it can be seen that on SEQs. #1 and #3, the proposed eBiCoV obtains much better results than SDALF. It is even more obvious on SEQ. #3 for which our method's CMC is 100% for $N = 5, 10$, which experimentally validates the effectiveness of our descriptor for person re-identification.

4.3. Person re-identification on the i-LIDS dataset

The i-LIDS MCTS dataset has been captured by multiple non-overlapping cameras at a busy airport arrival hall. There are 119 pedestrians with total 476 images. All the images are normalized to the size of 128×64 pixels. Many of these images undergo quite large illumination changes and occlusions (see Fig. 6).

We tested the proposed descriptors in the single-shot scenario. We follow the same experimental settings of [11, 12]. Considering there are 4 images on average for each pedestrian, we randomly select one image for each pedestrian to build the gallery set, while the rest (357 images) form the probe set. We repeat this procedure 10 times and compute the average CMC and nAUC. On the i-LIDS dataset, the best single-shot published performance is obtained by a covariance-based technique (SCR) [45]. Fig. 7 shows the CMC curves given by gBiCov, SCR [45], Custom Pictorial Structures (PS) [12] and SDALF [11].

Fig. 7 shows that gBiCov outperforms SDALF on this dataset, obtaining results which are comparable to the PS and SCR approaches. However, contrarily to PS and SCR, gBiCov does not need any body detection stage nor any background elimination pre-processing algorithm. This is significant advantage, knowing that body segmentation is still an open problem under real conditions.

Table 3: VIPeR dataset: nAUCs of gBiCov with different region size and different overlaps.

Region Size	4×8	4×8	8×8	8×8	8×8	8×8	8×16	8×16	8×16
Overlapping Size	2×4	4×8	2×4	4×4	4×8	8×8	2×4	4×8	8×8
Dim	25668	6912	23436	12276	6336	3456	21924	5960	3240
nAUC	89.32	88.90	89.87	89.72	89.54	89.04	90.37	90.11	89.68
Region Size	8×16	16×16	16×16	16×16	16×16	12×24	12×24	12×24	12×24
Overlapping Size	8×16	2×4	4×4	4×8	8×8	2×4	4×8	8×8	6×12
Dim	1728	17748	9396	4860	2700	18468	5040	2520	2268
nAUC	89.15	90.68	90.67	90.59	90.38	90.59	90.45	90.36	90.00

4.4. Person re-identification using metric learning

In this section we experimentally validate the combination of the proposed descriptor with the metric learning approach described in section 3.4.2. We compare our kBiCov (kBiCov = gBiCov + metric learning) approach with recent approaches based on metric learning, on the VIPeR and i-LIDS datasets.

4.4.1. kBiCov on VIPeR

To make comparisons fair, we follow the standard protocol for this dataset. We randomly take 316 persons out of the 632 for the test set, the remaining persons being in the train set. Like in [32], 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.

To face the increase of computational complexity due to the metric learning stage, MSCR is discarded and wHSV is replaced by a simple histogram. We use color histograms extracted from 8×24 rectangular regions to represent images. The rectangular regions are densely collected from a regular grid with 4 pixel spacing in vertical direction and 12 pixel spacing in horizontal direction. This step size is equal to half the width and length of the rectangles.

We compare kBiCov with four different approaches using metric learning: PRDC [46], LMNN [46], PCCA [32] and KISSME [2]. 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 6 horizontal strips. For PCCA, the feature descriptor is a 16-bin color histograms in 3 different color spaces (RGB, HSV and YCrCb) as well as texture histograms based on Local Binary Patterns (LBP) computed on 6 non-overlapping horizontal strips. PCCA [32] reports the state-of-the-art results for person re-identification, improving over Maximally Collapsing Classes [47], ITML [30] or LMNN-R [48]. For KISSME, the representation includes two components: HSV and Lab histograms on overlapping blocks of size 8×16 and stride of 8×8, and texture information captured by LBPs. The concatenated descriptors are projected into a 34 dimensional subspace by PCA.

Fig. 8 shows CMC curves of the different methods while Tab. 4 shows the nAUCs at ranks 1, 5, 10 and 20. The results of PRDC, LMNN and PCCA are taken from their original papers. From the figure and the table, we can see that kBiCov performs much better than any of the other approaches. For example, the matching rates for ranks 1, 10 and 20 are of 31.11%, 70.71% and 82.44% for kBiCov while those of PCCA are of 19.27%, 64.91% and 80.28%. Compared with the results of KISSME [2], the 1-rank matching rate is improved from 19.60% to 31.11%, validating the proposed representation. Indeed, the only difference between kBiCov and [2] is the image representation.

Table 4: VIPeR dataset: Top ranked matching rates (%) with 316 persons.

Method	r=1	r=5	r=10	r=20
PRDC [46]	15.66	38.42	53.86	70.09
MCC[46]	15.19	41.77	57.59	73.39
ITML[46]	11.61	31.39	45.76	63.86
LMNN[46]	6.23	19.65	32.63	52.25
CPS [12]	21.00	45.00	57.00	71.00
PRSVM [15]	13.00	37.00	51.00	68.00
ELF [16]	12.00	31.00	41.00	58.00
PCCA-sqrt [32]	17.28	42.41	56.68	74.53
PCCA-rbf [32]	19.27	48.89	64.91	80.28
KISSME [2]	19.60	-	62.60	-
kBiCov	**31.11**	**58.33**	**70.71**	**82.44**

Interestingly, the advantage of kBiCov over other approaches is obvious at low ranks. For example, the best matching rate at rank 1 among state of the art methods is of 21.00% while for kBiCov the matching rate is of 31.11%, which means that the improvement is nearly of 50%. This improvement is very significant for real world applications where it can effectively decrease the need of human intervention and make the search of a specific person easier.

In kBiCov, one important parameter is the dimensionality of the projected space (after PCA). The nAUCs for different dimensionalities are given Tab. 5. It can be seen from this table that the nAUCs are almost the same, until it reaches 80. For higher dimensionalities, the performance drops, probably because of over-fitting. In practice and in all of our experiments, the size of the low-dimensional space is set to 60, which is a good tradeoff between accuracy and efficiency.

4.4.2. kBiCov on i-LIDS

We also experimented with the supervised setting of the i-LIDS dataset. For making the comparison fair, we follow the experimental setting of [49] by randomly selecting the images

Table 5: VIPeR dataset: nAUCs with different dimensions.

Dimension	10	20	30	40	50	60	70	80	90	100
nAUC	92.40	94.89	95.82	96.25	96.42	96.48	96.50	96.57	96.46	96.39

Table 6: i-LIDS dataset: Top ranked matching rates (%) with 30 persons in the gallery set.

Method	r =1	r =5	r =10	r =20
PRDC [49]	44.05	72.74	84.69	96.29
Adaboost [49]	35.58	66.43	79.88	93.22
LMNN [49]	33.68	63.88	78.17	92.64
ITM [49]	36.37	67.99	83.11	95.55
MCC [49]	40.24	73.64	85.87	96.65
Xing's [49]	31.8	62.62	77.29	90.63
PLS [49]	25.76	57.36	73.57	90.31
L1-norm [49]	35.31	64.62	77.37	91.35
Bhat. [49]	31.77	61.43	74.19	89.53
kBiCov	39.17	68.19	82.10	95.26

Table 7: Mean classification accuracy (%) and standard deviation on the LFW dataset, unrestricted setting.

Method	m $\pm\sigma$
SD-MATCHES, 125x12512 [51], aligned	64.10±0.62
H-XS-40, 81x15012 [51], aligned	69.45±0.48
GJD-BC-100, 122x22512 [51], aligned	68.47±0.65
LARK unsupervised20 [52], aligned	72.23±0.49
POEM [53], aligned	82.71±0.59
G-LQP [50], aligned	82.10±0.26
I-LQP [50], aligned	86.20±0.46
gBiCov, aligned	84.48±0.70

of 30 persons for the test set while the remaining ones are attributed to the train set. In the test set, there is one image of each person which is randomly selected as a gallery image while other images constitute the probe set. The training set has 10 times more negative pairs than positive pairs.

Tab. 6 shows the matching rates of the different methods investigated. From the table, we can see that the matching rates of kBiCov are significantly better than those of other methods, except PRDC and MCC.

4.5. Face verification in uncontrolled environments

Besides person re-identification, we also experimented gBiCov for face verification on the LFW dataset [34]. LFW consists of 13,233 images of 5,749 people which are originally gathered from news articles on the web. Face verification on the LFW dataset is a challenging problem due to the variations in facial poses, illumination or expressions. Fig. 9 shows typical images of the LFW dataset.

We tested the proposed descriptor on the View 2 of the LFW, following the protocol described in [34]. In View 2, the dataset is split into 10 disjoint folds. Each fold contains 600 pairs of images: 300 positive pairs (*i.e.,* two images of the same person) and 300 negative ones (pairs of different persons). The task is to verify if a test pair represents the same individual or not. In detail, two images are predicted to be the same person if the distance between face signatures is smaller than a threshold. Otherwise the pair is supposed to contain different persons. The verification performance is reported as the mean recognition rate and the corresponding standard deviation over 10-folds. The training and testing splits are defined on the LFW website[1], from which we also obtained the aligned version of the face images (80×150 images).

[1]http://vis-www.cs.umass.edu/lfw/index.html

Experiments on face verification are different from the ones on person re-identification in several ways. First, for person re-identification, the orientation information of Gabor filters is discarded for improving the computational efficiency, without significant loss in accuracy. However, in face verification, orientations should be taken into account to preserve fine details. Here, we compute one BIF image per orientation. As we have 8 different orientations, the size of the descriptor is 8 times bigger than the descriptors used for person re-identification.

Second, in this set of experiments, we have evaluated the performance of gBiCov and kBiCov, but not the one of eBiCov. Indeed, the information needed for person re-identification and face verification are quite different. For example, wHSV and MSCR perform well on person re-identification, but they are not suitable for face verification. In fact, it is one of the advantages of the proposed gBiCov that it can be applied on both person re-identification and face verification.

Finally, the image representation $\mathbf{D}$ is projected into a lower space using *whiten* PCA. Whitening the data essentially means rotating them into a space of principal components, dividing each dimension by square root of variance in that direction, and rotating back to pixel space. The dimension of the whiten PCA space is set to 60. This normalization has been reported to be very useful in this context [50].

Tab. 7 reports the performance of gBiCov, as well the performance of state-of-the-art methods. These results are taken from the LFW website which keeps track of any published results on this dataset. By giving a mean classification accuracy of 84.48%, the performance of the proposed gBiCov descriptor is comparable to that of the state-of-the-art such as I-LQP. However, compared with I-LQP, gBiCov does not need any training images, which is a big advantage for real world applications. In addition, gBiCov is more computationally efficient than I-LQP, since I-LQP needs to learn the codebook of $3^{16} = 43$ million distinct codes.

Besides the unrestricted setting, we also tested kBiCov under the image restricted training setting. In this setting, we just

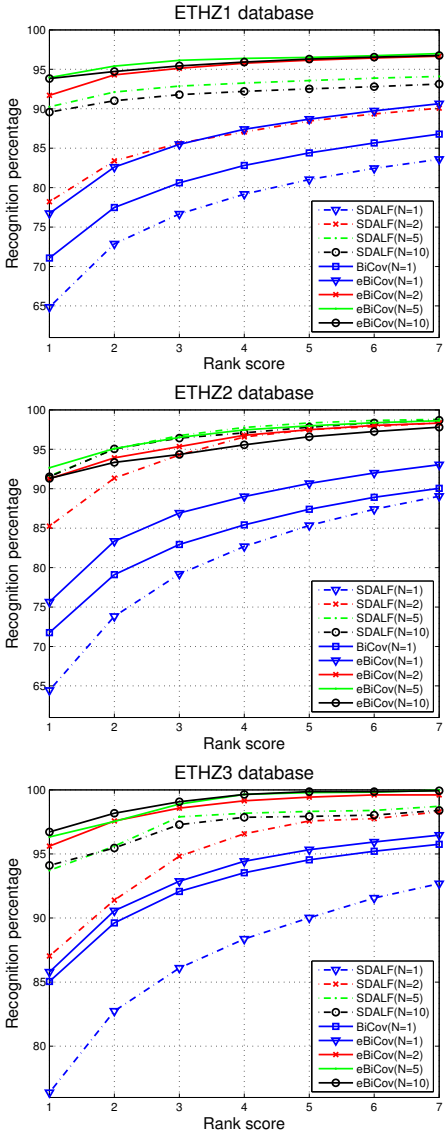


Figure 5: ETHZ dataset: CMC curves.

know that a training pair is either positive or negative pair; we do not know the identity of the persons. In kBiCov, the feature of gBiCov are reduced to 60 dimensions by PCA first, and then we use KISSME to learn the projection. State-of-the-art under this setting can be also found on the LFW website. Tab. 8 shows the accuracy of kBiCov and compare it to the state-of-the-art results.

Tab. 8 shows that Fisher vectors perform best on LFW, with a mean accuracy of 87.47%. However, the mean accuracy obtained by kBiCov is of 86.80%, which is comparable to the Fisher vectors and much better than those of any other methods. It must be pointed out that the Fisher vectors method requires a huge amount of time to learn the GMM model in the feature extraction stage, while kBiCov does not need any feature learning stage. On the whole, the results obtained with both gBiCov and kBiCov show the good performance of the proposed image representations on the face verification task.



Figure 6: Some images in the i-LIDS MCTS dataset. The images in the same column are belonging to the same person.

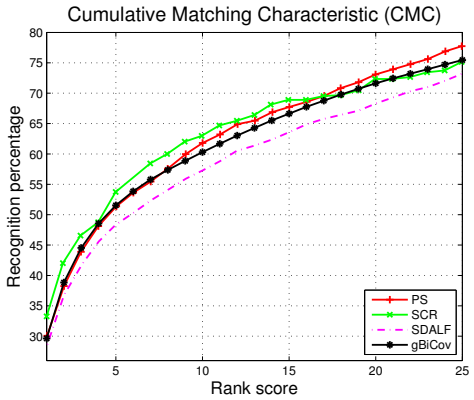


Figure 7: i-LIDS dataset: CMC curves of the different methods in the single shot scenario.

5. Conclusions

This paper proposes a novel image representation – referred as the gBiCov representation – which combines Biologically Inspired Features (BIF) and the Covariance descriptor. gBiCov is robust to illumination, scale and background variations, which makes it suitable for both person re-identification and face verification tasks. Furthermore, the paper shows that the discriminative ability of gBiCov can be improved by the use of metric learning. Experiments on three pedestrian datasets (VIPeR, i-LIDS and ETHZ) and one face dataset (LFW) show that the proposed gBiCoV achieves the state-of-the-art performances in both unsupervised setting and supervised setting, while being at the same time efficient and robust, in the sense that it is fast to compute and quite insensitive to parameter tuning.

Acknowledgment

This work was partly realized as part of the Quaero Program funded by the OSEO, French State agency for innovation and by the ANR, grant reference ANR-08-SECU-008-01/SCARFACE. The first author is partially supported by National Natural Science Foundation of China under contract Nos. 61003103 and 61173065.

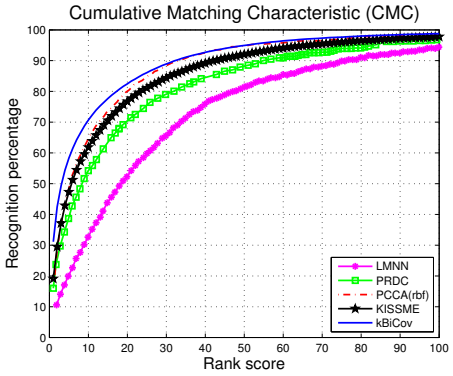


Figure 8: VIPeR dataset: CMC curves with 316 persons.



Figure 9: Example images of LFW dataset. The two images on the same column belong to the same subject.

Table 8: Mean classification accuracy (%) and standard deviation on the LFW dataset, restricted setting.

Method	m±σ
Eigenfaces, original	60.02 ± 0.79
Nowak, original	72.45 ± 0.40
Nowak[2], funneled	73.93 ± 0.49
Hybrid descriptor-based, funneled	78.47 ± 0.51
3x3 Multi-Region Histograms (1024)	72.95 ± 0.55
Pixels/MKL, funneled	68.22 ± 0.41
V1-like/MKL, funneled	79.35 ± 0.55
APEM (fusion)	84.08 ± 1.20
MRF-MLBP	79.08 ± 0.14
Fisher vector faces	87.47 ± 1.49
kBiCov	86.80 ± 0.79

References

[1] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[2] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295, 2012.

[3] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," *Proc. British Machine Vision Conference*, 2012.

[4] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[5] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 994–1000, 2005.

[6] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 93–104, 2008.

[7] G. Guo, G. Mu, Y. Fu, and T. Huang, "Human age estimation using bio-inspired features," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 112–119, 2009.

[8] W. Ayedi, H. Snoussi, and M. Abid, "A fast multi-scale covariance descriptor for object re-identification," *Pattern Recognition Letters*, pp. 1902–1907, 2011.

[9] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean Riemannian covariance grid," *Proc. International Conference on Advanced Video and Signal-Based Surveillance*, 2011.

[10] Y. Zhang and S. Li, "Gabor-LBP based region covariance descriptor for person re-identification," *International Conference on Image and Graphics*, pp. 368–371, 2011.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[12] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," *Proc. British Machine Vision Conference*, 2011.

[13] O. Oreifej, R. Mehran, and M. Shah, "Human identity recognition in aerial images," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 709–716, 2010.

[14] W. Schwartz and L. Davis, "Learning discriminative appearance based models using partial least squares," *Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[15] B. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," *Proc. British Machine Vision Conference*, 2010.

[16] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Proc. European Conference on Computer Vision*, pp. 262–275, 2008.

[17] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using haar-based and DCD-based signature," *Proc. International Workshop on Activity Monitoring by Multi-camera Surveillance Systems*, 2010.

[18] N. Gheissari, T. Sebastian, P. Tu, J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1528–1535, 2006.

[19] J. Kai, C. Bodensteiner, and M. Arens, "Person re-identification in multi-camera networks," *Proc. IEEE Conference on Computer Vision and Pattern Recognitio Workshops*, pp. 55–61, 2011.

[20] W. Zheng, S. Gong, and T. Xiang, "Associating groups of people," *Proc. British Machine Vision Conference*, 2009.

[21] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by HPE signature," *Proc. IEEE International Conference on Pattern Recognition*, pp. 1413–1416, 2010.

[22] K. Aziz, D. Merad, and B. Fertil, "People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation," *Proc. International Conference on Advanced Video and Signal-Based Surveillance*, pp. 303–308, 2011.

[23] K. Aziz, D. Merad, and B. Fertil, "Person re-identification using appearance classification," *Image Analysis and Recognition*, pp. 170–179, 2011.

[24] T. Gandhi and M. Trivedi, "Person tracking and re-identification: introducing panoramic appearance map (PAM) for feature representation," *Machine Vision and Applications*, vol. 18, no. 3-4, pp. 207–220, 2007.

[25] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," *Image Analysis*, pp. 91–102, 2011.

[26] D. Figueira, L. Bazzani, Ha Quang Minh, M. Cristani, A. Bernardino, and V. Murino, "Semi-supervised multi-feature learning for person re-identification," *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 111–116, 2013.

[27] R. Satta, G. Fumera, F. Roli, M. Cristani, and V. Murino, "A multiple component matching framework for person re-identification," *Int. Conf. on Image Analysis and Processing*, pp. 2360–2367, 2011.

[28] R. Satta, G. Fumera, and F. Roli, "Exploiting dissimilarity representations for person re-identification," *International Workshop on Similarity-Based Pattern Analysis and Recognition*, 2011.

[29] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[30] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon, "Information-theoretic metric learning," *Proc. International Conference on Machine Learning*, pp. 209–216, 2007.

[31] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," *Proc. IEEE International Conference on Computer Vision*, 2009.

[32] A. Mignon and F. Jurie, "PCCA: a new approach for distance learning from sparse pairwise constraints," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2666–2672, 2012.

[33] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary pattern," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[34] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[35] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," *Proc. of IEEE Computer Vision and Pattern Recognition*, p. 524C531, 2005.

[36] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3384–3391, 2010.

[37] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210C227, 2009.

[38] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun, "Bayesian face revisited: A joint formulation," *European Conference on Computer Vision*, 2012.

[39] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Transactions on Image Processing*, vol. 19, pp. 174–184, 2010.

[40] L. Wiskott, J. Fellous, N. Krüger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[41] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[42] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2007.

[43] A. Ess, B. Leibe, K. Schindler, and L. Gool, "A mobile vision system for robust multi-person tracking," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[44] H. Moon and P. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[45] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," *Proc. International Conference on Advanced Video and Signal-Based Surveillance*, 2010.

[46] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[47] A. Globerson and S. Roweis, "Metric learning by collapsing classes," *Advances in Neural Information Processing Systems*, 2006.

[48] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," *Proc. Asian Conference on Computer Vision*, vol. 4, pp. 501–512, 2010.

[49] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[50] S. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," *British Machine Vision Conference*, 2012.

[51] J. Ruiz-del-Solar, R. Verschae, and M. Correa, "Recognition of faces in unconstrained environments: a comparative study," *EURASIP Journal on Advances in Signal Processing*, 2009.

[52] H. Seo and P. Milanfar, "Face verification using the LARK face representation," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1275–1286, 2011.

[53] N. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1352–1365, march 2012.