Exploiting Feature Correlations by Brownian Statistics for People Detection and Recognition

Slawomir Bak, Marco San Biagio, Ratnesh Kumar, Vittorio Murino, François Bremond

Exploiting Feature Correlations by Brownian Statistics for People Detection and Recognition

Sławomir Bąk[1], Marco San Biagio[2], Ratnesh Kumar[1], Vittorio Murino[2] and François Brémond[1]
[1]STARS Lab, INRIA Sophia Antipolis Méditerranée, Sophia Antipolis, 06902 Valbonne, France
[2]Pattern Analysis and Computer Vision (PAVIS), IIT IStituto Italiano di Tecnologia, 16163 Genova, Italy

Characterizing an image region by its feature inter-correlations is a modern trend in computer vision. In this paper, we introduce a new image descriptor that can be seen as a natural extension of a covariance descriptor with the advantage of capturing nonlinear and non-monotone dependencies. Inspired from the recent advances in mathematical statistics of Brownian motion, we can express highly complex structural information in a compact and computationally efficient manner. We show that our Brownian covariance descriptor can capture richer image characteristics than the covariance descriptor. Additionally, a detailed analysis of the Brownian manifold reveals that in opposite to the classical covariance descriptor, the proposed descriptor lies in a relatively flat manifold, which can be treated as a Euclidean. This brings significant boost in the efficiency of the descriptor. The effectiveness and the generality of our approach is validated on two challenging vision tasks, pedestrian classification and person re-identification. The experiments are carried out on multiple datasets achieving promising results.

***Index Terms*—brownian descriptor, covariance descriptor, pedestrian detection, re-identification.**

I. Introduction

DESIGNING proper image descriptors is a crucial step in computer vision applications, including scene detection, target tracking and object recognition. A good descriptor should be invariant to illumination, scale and viewpoint changes. This usually involves a high-dimensional floating-point vector encoding a robust representation of an image region [1], [2]. Typically, descriptors employ simple statistics (*i.e.* histograms) of features extracted by different kinds of image filters (gradients [3], binary patterns [4]). In recent studies, a trend has emerged that consists in discarding the intrinsic value of the features, encoding instead their inter-correlations. The most well-known image descriptor following this idea is the covariance descriptor [5]. This descriptor encodes information on feature covariances inside an image region, their inter-feature linear correlations and their spatial layout. The correlation-based descriptors show a consistent invariance to many aspects (scale, illumination, rotation), making them a good choice for representing object classes with high intra-class variability (*e.g.* pedestrians [6]). Moreover, correlation-based descriptors are superior to other methods for absorbing inter-camera changes (*e.g.* for matching objects registered by different cameras [7], [8]).

In this paper, we focus on *correlation-based* descriptors, revisiting fundamentals of covariance. We highlight that the covariance descriptor measures only linear dependence between features, which might not be enough to capture the complex structure of many objects. As an example see Fig. 1 which illustrates the correlation between two features extracted from the patch of a pedestrian image. Intensity values and the corresponding gradient magnitudes are plotted together to show the dependency. Most pixels of the patch have high intensity and low gradient (homogeneous regions). This produces the dense distribution in the lower-right corner of the plot. The most informative pixels are captured by the strap

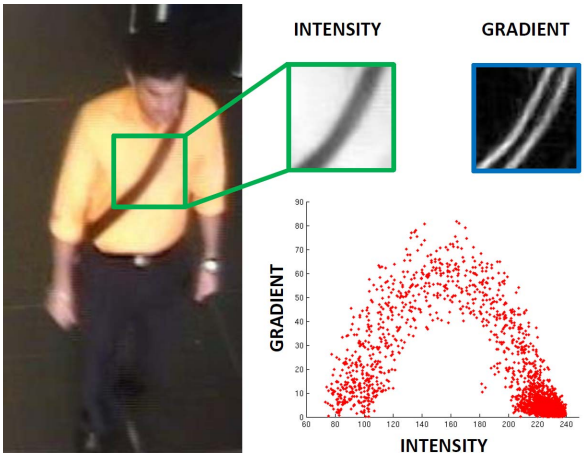


Fig. 1. Nonlinear dependency between two features extracted from a patch. Intensity values and gradient magnitudes are plotted together to illustrate the relation.

structure that show a non-monotone (nonlinear) dependency. Interestingly, the classical covariance will not capture this information as it measures only the linear correlation between features. In the result the covariance descriptor may produce a diagonal matrix, which is not a sufficient condition for statistical independence; actually, a non-monotone relation exists. This indicates information loss when using the covariance descriptor.

We overcome this issue by employing a novel descriptor based on Brownian covariance [9], [10]. The classical covariance measures the degree of linear relationship between features, whereas Brownian covariance measures the degree of *all kinds of possible relationships* between features. We show that our novel descriptor is computationally efficient and more effective than the covariance descriptor (Sec. II).

This paper makes the following contributions:

- We discuss the covariance descriptor and highlight its

constraints and limitations as a dependency measure (Sec. III-A).
- We propose a new image region descriptor that is a natural extension of covariance (Sec. III-B): the proposed descriptor is referred to as *Brownian descriptor* due to its analogy to the Brownian covariance.
- We illustrate advantages of the new descriptor over the classical covariance descriptor using synthetic data and theoretical analysis (Sec. III-D) and we provide an efficient algorithm for extracting the descriptor employing *integral images* (Sec. III-C).
- We show the generality of the descriptor, validating it on different vision tasks (Sec. IV). We show that this descriptor can handle both inter- and intra-class variations, *e.g.* pedestrian classification and person re-identification. The results bear out that this descriptor reaches sufficient trade off between discriminative power and invariance. Finally, we demonstrate that the Brownian descriptor outperforms the classical covariance descriptor in terms of both efficiency and accuracy.

The paper draws conclusions in Sec. V by discussing future perspectives.

II. Related work

One of the most common problems in object detection and recognition is to find a suitable object representation. For historical and computational reasons, vector descriptors that encode the local appearance of the data have been broadly applied. In this sense, many different techniques have been developed in the literature. As shown in [11], many of these techniques follow two complementary paradigms: "feature-based" and "relation-based". The former takes into account measurable intrinsic characteristics of an object, such as color or shape information. Most of well-known descriptors included in this sub-group are: *Scale-Invariant Feature Transform* (SIFT) [12], *Histogram of Oriented Gradients* (HOG) [3], *Local Binary Pattern* (LBP) [13], [4]. The latter paradigm consists of considering the intrinsic value of these cues, encoding their inter-relations: the most known descriptor following this line is the covariance of features (COV) [5], in which linear relations between features are exploited as elementary patterns.

A. Feature-based descriptors

SIFT descriptor, originally proposed in [12], is used for a large number of purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. SIFT descriptor is invariant to translations, rotations and scaling transformations in an image domain and robust to moderate perspective transformations and illumination variations. Experimentally, SIFT descriptor has been proven to be very useful for image matching and object recognition under real-world conditions [14], [15], [16], [17].

However image descriptors must not only be accurate but also highly efficient. SIFT unfortunately is represented by high-dimensional floating point vector bringing significant computational burden, while employed to tasks that require real-time performance. Consequently, HOG descriptor has been revealed. This descriptor is of particular interest in object detection and recognition as it is fast to compute and provides high performance [3], [18], [19], [20], [21], [22]. HOG is considered as the most popular feature used for pedestrian detection.

In [23], we can find PHOG, an extension of classical HOG descriptor for pedestrian detection. The authors showed that PHOG can yield better classification accuracy than the conventional HOG and it is much computationally lighter while having smaller dimension. However, these HOG-like features that capture edge and local shape information might perform poorly when the background is cluttered with noisy edges [4].

Originally proposed by [13], Local Binary Pattern (LBP) is a simple but very efficient texture operator which labels pixels of an image according to the differences between values of the pixel itself and the surrounding ones. It has been widely used in various applications and has achieved very good accuracy in face recognition [24]. LBP is highly discriminative and its key advantages, namely its invariance to monotonic gray level changes and computational efficiency, make it suitable for demanding image analysis tasks such as human detection [25].

B. Relation-based descriptors

Recently, in contrast to the classical feature descriptors discussed above, a novel trend has emerged that consists of considering the intrinsic value of image features, encoding their inter-relations. The most popular descriptors exploiting feature correlations in images is the covariance descriptor [5]. This descriptor represents an image patch by the covariance of its features such as spatial location, intensity, higher order derivatives, *etc.*

Covariance descriptor was first introduced for object matching and texture classification [5]. Since then it has also been intensively employed in many other computer vision applications, such as pedestrian detection [6], [26], [27], person re-identification [7], [28], [29], [30], [31], object tracking [32], action recognition [33] and head orientation classification [34].

As covariance matrices do not lie on a Euclidean space, each of these studies addresses the problem of using the covariance descriptor in a non-trivial machine learning framework. Several optimization algorithms on manifolds have been proposed for the space of positive semi-definite matrices Sym_d^+ [5], [6], [27], [34], [35]. The most common approach consists in mapping covariance matrices to the tangent space that can be treated as an approximation of a Euclidean space [6], [27]. Performing mapping operations involves choosing the tangent point on the manifold, which usually is determined either by the mean of the training data points - Karcher mean [36], or by the identity matrix [34]. The logarithmic and exponential maps are iteratively used to map points from the manifold to the tangent space, and vice-versa. Unfortunately, the resulting algorithms suffer from two drawbacks: first, the iterative use of the logarithmic and exponential maps makes them computationally expensive, and, second, they only approximate

true distances on the manifold by Euclidean distances on the tangent space. Another possibility is to compute a metric directly on Sym_d^+, which estimates the geodesic distance [5], [34]. Using this approach, we preserve real distances between each pair of samples. Unfortunately, both solutions involve a high computational cost.

For the above mentioned reasons, in contrast to the previous approaches, we present new insights into the covariance descriptor, raising fundamental limitations of covariance as a dependence measure. In this paper, we design a novel descriptor driven by recent achievements in mathematical statistics related to Brownian motion [37], [10]. The new descriptor can be treated as a point in a Euclidean space, making the descriptor computationally efficient and useful for real-time applications. This novel descriptor not only brings tremendous matching speed-up in comparison to the classical covariance, but also keeps more information on feature correlations inside an image region.

In the following sections we raise fundamental constraints of covariance as a dependency measure and we define the *Brownian descriptor*.

III. Brownian descriptor

This section introduces the Brownian descriptor, discussing its advantages over the classical covariance. Before elaborating the Brownian descriptor, we discuss the classical covariance descriptor proposed in [5], highlighting its limitations.

A. Limitations of the classical covariance

Image feature inter-relation are often captured by the *covariance matrix*. This descriptor encodes information on feature variances inside the image region, their covariance with each other and their spatial layout. It enables to fuse different types of features, while producing a compact representation.

Let $\mathcal{I}$ be an image and F be a n-dimensional feature image extracted from $\mathcal{I}$

$$F = \phi(\mathcal{I}), \tag{1}$$

where function ϕ can be any mapping. Usually, the most applied mappings contain intensity values, color, gradients, filter responses, *etc.* . Recently, we can also find other types of mappings, *e.g.* based on infrared images, depth or motion flow. In the result, each pixel can be expressed by a n-dimensional feature point determined by mapping ϕ.

For a given region $R \subset F$ containing Z pixels, let $\{f_z\}_{z=1...Z}$ be the n-dimensional feature points inside R. Each feature point f_z is characterized by function ϕ. We represent region R by the $n \times n$ covariance matrix (C_R) with (i,j)-th element expressed as

$$C_R(k,l) = \frac{1}{Z-1}\sum_{z=1}^{Z}\Big(f_z(k)-\mu(k)\Big)\Big(f_z(l)-\mu(l)\Big), \tag{2}$$

where μ is the mean of f_z points. The diagonal entries are variances of each feature, whereas the off-diagonal entries are the covariances between pairs of features.

Standardization. Covariance values are very often normalized by the product of corresponding standard deviations

$$\rho_{kl} = \frac{C_R(k,l)}{\sqrt{C_R(k,k)C_R(l,l)}}, \tag{3}$$

and are referred to as *the Pearson Product-Moment Correlation Coefficients*.

1) Limitations due to linear dependency measure

ρ measures a *linear* correlation between two variables (the strength of the linear dependence). However, as it is computed with respect to the mean of the feature (note μ components in Eq. (2)), it is not able to measure nonlinear or non-monotone dependence (see Sec. III-D and Fig. 2 for elaboration).

2) Limitations due to choice of metric

As we have already mentioned in Sec. II, covariance matrices do not lie on a Euclidean space. Computing distance between two covariance descriptors, we need to either assume a Riemannian manifold employing *geodesic distance* or map covariance to a tangent space approximating distances. Both solutions are computationally intensive and unfavorable in practice. Moreover, well known machine learning techniques are not adequate for learning on complex manifolds, often producing over-fitted classifiers.

B. Brownian covariance

Brownian descriptor inherits the theory from recent advances in mathematical statistics related to Brownian covariance [10]. In particular, it is based on the *sample distance covariance* statistics that measures dependence between random vectors in arbitrary dimension. In the following sections we introduce *distance covariance* $\mathcal{V}^2$, *sample distance covariance* $\mathcal{V}_n^2$, and their relations to Brownian covariance $\mathcal{W}$. The mathematical notations and formulas are in accordance with [10].

1) Distance covariance $\mathcal{V}^2$

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors, where p and q are positive integers. f_X and f_Y denote the characteristic functions of X and Y, respectively, and their joint characteristic function is denoted $f_{X,Y}$. In terms of characteristic functions, X and Y are independent if and only if $f_{X,Y} = f_X f_Y$. A natural way of measuring the dependence between X and Y is to find a suitable norm to measure the distance between $f_{X,Y}$ and $f_X f_Y$.

The *Distance covariance* $\mathcal{V}^2$ [37] is a new measure of dependence between random vectors and can be defined by

$$\mathcal{V}^2(X,Y) = ||f_{X,Y}(t,s) - f_X(t)f_Y(s)||^2 \tag{4}$$

$$= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s)-f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}}dtds, \tag{5}$$

where c_p and c_q are constants determining norm function in $\mathbb{R}^p \times \mathbb{R}^q$, $t \in X$, $s \in Y$. This measure is analogous to classical covariance, but with the important property that $\mathcal{V}^2(X,Y) = 0$ if and only if X and Y are independent. In [37] *distance covariance* is seen as a natural extension and a generalization

of the classical covariance measure. It extends the ability to measure linear association to all types of dependence relations. Further, distance covariance can be computed between any random vectors in arbitrary dimension. For more theoretical and practical advantages of this new dependency measure the interested reader is refereed to [37].

2) Sample distance covariance $\mathcal{V}_n^2$

Designing a new image descriptor, we are interested in finding relations between finite distributions (limited amount of pixels). Thus, we can employ a sample counterpart of distance covariance [10]. The sample distance covariance $\mathcal{V}_n^2$ between random vectors X and Y is defined as

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}, \tag{6}$$

where A_{kl} and B_{kl} are simple linear functions of the pairwise distances between n sample elements. These functions are defined in the following.

For a random sample $(\mathbf{X},\mathbf{Y}) = \{(X_k, Y_k) : k = 1 \dots n\}$ of n i.i.d random vectors (X, Y) from their joint distribution, compute the Euclidean distance matrices $(a_{kl}) = (|X_k - X_l|_p)$ and $(b_{kl}) = (|Y_k - Y_l|_q)$. Define

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}, \quad k,l = 1,\dots,n, \tag{7}$$

where

$$\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}. \tag{8}$$

Similarly, we define $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$.

Standardization. Similarly to covariance which has its standardized counterpart ρ, $\mathcal{V}_n^2$ has its standardized version referred to as *distance correlation* $\mathcal{R}_n^2$ and defined by

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases} \tag{9}$$

where

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X,X) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2. \tag{10}$$

3) Brownian covariance $\mathcal{W}$

Brownian motion is a stochastic process invented for modeling random movements of particles suspended in a fluid. It describes their trajectories and interactions. These interactions can be expressed by Brownian covariance. Let $\mathcal{W}$ be a Brownian covariance. According to [10] and [37], $\mathcal{W}$ measures *all kinds of possible relationships* between random particles (variables). This means that $\mathcal{W}(X,Y) = 0$ if and only if X and Y are independent.

The surprising coincidence is that for arbitrary $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ with finite second moments

$$\mathcal{W}(X,Y) = \mathcal{V}(X,Y). \tag{11}$$

For the proof, the interested reader is pointed out to THEOREM 8 in [10]. Further, we see THEOREM 2 from [10] that says:

If $E|X|_p^\alpha < \infty$ and $E|Y|_q^\alpha < \infty$, then almost surely

$$\lim_{n\to\infty} \mathcal{V}_n(X,Y) = \mathcal{V}(X,Y), \tag{12}$$

where α is a positive exponent on Euclidean distance. This equality holds only if the α moments are finite and $0 < \alpha < 2$. Although $\mathcal{V}$ can be defined for $\alpha = 2$, it does not characterize independence. Indeed, the case $\alpha = 2$ (squared Euclidean distance) leads to the classical covariance measure. In the results in algorithm 1, we assume $\alpha = 1$ that leads to employing ℓ_1 metric, while computing distance matrix (a_{kl}).

From equations (11) and (12) we can see that

$$\mathcal{W}(X,Y) = \lim_{n\to\infty} \mathcal{V}_n(X,Y) \propto \mathcal{R}_n^2(X,Y). \tag{13}$$

In the result if $\mathcal{R}_n^2(X,Y) = 0$, we expect no dependence between variables. *This is the main advantages of $\mathcal{R}_n^2(X,Y)$ over ρ. $\rho = 0$ means that there is no linear correlation between variables, while nonlinear or non-monotone dependence may exist.* Although $\mathcal{R}_n^2$ is just a sample counterpart of $\mathcal{R}$, we believe that $\mathcal{R}_n^2$ keeps more information than ρ while characterizing an image region, which is clarified in the subsequent sections.

C. Efficient algorithm for computing Brownian descriptor

Let $\mathcal{I}$ be an image and $\mathcal{L} = \{L_1, L_2, \dots, L_n\}$ be a set of feature layers defined by mapping ϕ. In other words, after applying mapping ϕ on $\mathcal{I}$, each pixel z of the image can be expressed as the following feature vector:

$$\phi(\mathcal{I}, z) = [L_1(z), L_2(z), \dots, L_n(z)]. \tag{14}$$

The task is to provide a discriminative representation of a given image region R containing Z pixels.

We propose to treat each layer L_i as a point in a Z-dimensional space and to express the Brownian descriptor as $\mathcal{R}_n^2(\mathcal{L},\mathcal{L})$. We design the Brownian descriptor defining an algorithm (See Algorithm 1), in which we employ the computing formula for the *sample distance covariance* $\mathcal{V}_n^2$ (see Sec. III-B2). The final descriptor is expressed by the standardized version $\mathcal{R}_n^2(\mathcal{L},\mathcal{L})$. Note that $\mathcal{R}_n^2(\mathcal{L},\mathcal{L})$ is actually a scalar value - Eq. (9). Rather than representing an image region by a scalar value, we keep distance coefficients in the form of matrix $(\mathcal{R}_{kl}^2)$. We believe that this provides finer and more distinctive representation.

Similarly to the classical covariance matrix, the Brownian descriptor is represented by a positive definite symmetric matrix and it provides a natural way of fusing multiple features. This descriptor does not contain any information regarding the order and the number of pixels. This implies a certain scale and rotation invariance over the image regions in different images as long as layers L_i are invariant (similarly to the classical covariance descriptor [5]).

Intuitively, the difference between the classical covariance descriptor and our Brownian is that covariance computes correlation with respect to μ of each feature layer (see Eq. (2)), while Brownian statistics are based on distances between all feature layers (a_{kl}).

1) Extraction complexity

The computation time and memory complexity for both Brownian descriptor and the classical covariance matrix [5] is the same; the computation complexity for both descriptors is $O(n^2 Z)$, where n is the number of feature layers and Z is the

Algorithm 1: Brownian descriptor algorithm

Data: Layers $\mathcal{L} = \{L_1, L_2, \ldots, L_n\}$, $L_i \in \mathbb{R}^Z$
Result: Brownian descriptor $\mathcal{R}^2_{kl}$
begin
Compute the Euclidean distance matrix (a_{kl})
$(a_{kl}) = (|L_k - L_l|)$
Let $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$, where
$\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}$ *(mean of the k-th row)*
$\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$ *(mean of the l-th column)*
$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$ *(mean of distances)*
Let $\mathcal{V}^2_n(\mathcal{L}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A^2_{kl}$
$\mathcal{R}^2_{kl} = \frac{A^2_{kl}}{\mathcal{V}^2_n(\mathcal{L})}$
end

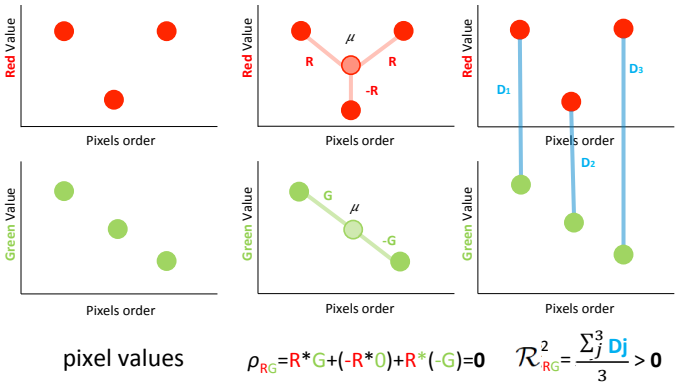


Fig. 2. Comparison of $\mathcal{R}^2_{kl}$ *vs.* ρ_{kl}. For the sake of simplicity we consider only 3 pixel values of two feature layers - Red and Green channel. $\rho_{kl} = 0$ and $\mathcal{R}^2_{kl} > 0$, while actually layers Red and Green are in non-monotone correlation.

number of pixels. For fast descriptor computation, similarly to [5], we can construct *integral images* that need to be extracted for each $|L_k - L_l|$ and for each $\sum$ in Algorithm 1. After computing integral images, the descriptor can be computed in constant time $O(1)$.

2) Matching complexity

Instead of using *geodesic distance* or tangent plane projections at the identity matrix, we can directly employ an Euclidean metric for expressing distance between two Brownian descriptors (see Sec. IV-A3 for elaboration). This makes our descriptor computationally efficient in opposite to the classical covariance descriptors. The descriptor performance with respect to several metrics is evaluated in Sec. IV-A. Its efficiency is discussed in Sec. IV-C.

D. $\mathcal{R}^2_{kl}$ vs. ρ_{kl}

In the Brownian descriptor ρ_{kl} is replaced by coefficients of $\mathcal{R}^2_{kl}$ for measuring dependence between image features. We claim that $\mathcal{R}^2_{kl}$ coefficients keep more information on dependence between features included in the mapping ϕ.

Fig. 2 illustrates a comparison between $\mathcal{R}^2_{kl}$ and ρ_{kl}, while handling non-monotone dependency between two feature layers (red and green channels). We can notice that ρ_{kl} ignores non-monotone correlation due to mean-dependent computation (see Eq. (2)). It results in $\rho_{kl} = 0$. This is the fundamental problem of covariance, in which ρ_{kl} *may go very close to zero even if the two variables are highly correlated*. In contrary, $\mathcal{R}^2_{kl}$ keeps information on the dependence between features even when they exhibit non-monotone or nonlinear correlation. Fig. 1 illustrates a real case where we observe non-monotone correlation between features.

IV. Experimental results

This section focuses on evaluating the Brownian descriptor on two vision tasks: *pedestrian detection* in Sec. IV-A and *person re-identification* in Sec. IV-B. We concentrate on a comparison of the Brownian descriptor with the classical covariance descriptor. Additionally, we carry out an analysis of feature-based descriptors, *e.g.* HOG, LBP, illustrating superiority of relation-based descriptors. Sec. IV-C discusses the efficiency of the proposed descriptor.

A. Pedestrian detection

Pedestrian detection is an important and complex task in Computer Vision [38], representing one of the most basic operations in many significant applications such as car assistance[39], video-surveillance, robotics and content-based image/video retrieval. The articulated structure and variable appearance of the human body, combined with illumination and pose variations, different point of views and low image resolution contribute the complexity of the problem in real-world applications. Furthermore, in case of a moving camera in a dynamic environment, changing backgrounds and partial occlusions may cause additional problems.

In this section, we explore the Brownian descriptor and employ it for detecting pedestrians. We carry out our experiments on two challenging data, Daimler Multi-Cue Occluded Pedestrian Classification Benchmark Dataset [25] and INRIA Pedestrian Dataset [3] that provide different low-level features (*e.g.* depth, motion). Figure 3 shows some examples from these datasets.

We evaluate five cases: (1) BROWNIAN; (2) BROWNIAN$_{Proj.}$; (3) COVARIANCE; (4) COVARIANCE$_{Proj.}$ and (5) HOG. By label $Proj.$, we indicate that a descriptor is assumed to be an element of a Riemannian manifold. In this case, we project the descriptor on the tangent plane at the identity matrix [34]. In all cases, we employ a linear SVM [40] for classification.

1) Daimler Multi-Cue Pedestrian Dataset [25]

This dataset contains 77720 unoccluded positive and 48700 negative samples. These are split into 52122 positive and 32465 negative samples for training and 25608 positive and 16235 negative samples for testing. Each image, of size 96×48, is composed of three image modalities: standard visible gray scale image $V(x, y)$, depth $D(x, y)$, and motion flow $M(x, y)$. For each image, we have the following dense feature map $\mathbf{F}(x, y)$:

$$\mathbf{F}(x, y) = \left[\mathbf{F}^V(x, y), \mathbf{F}^D(x, y), \mathbf{F}^M(x, y), x, y\right], \qquad (15)$$

where each 96×48 map $\mathbf{F}^V(x, y)$, $\mathbf{F}^D(x, y)$ and $\mathbf{F}^M(x, y)$ represents low-level features extracted from the visible, depth,

(a) Daimler Multi-Cue Dataset (b) INRIA Dataset

Fig. 3. Positive and negative examples extracted from the two pedestrian datasets.

and motion flow modality, respectively; x and y are the horizontal and vertical pixel coordinates. These two last features are particularly interesting, since they allow to instantiate relations that hold between particular cues and their spatial position. In particular, on each modality we extract the following low-level features:

$$\mathbf{F}^i_{(x,y)} = \left[I^i, \left|I^i_x\right|, \left|I^i_y\right|, \left|I^i_{xx}\right|, \left|I^i_{yy}\right|, \sqrt{{I^i}^2_x + {I^i}^2_y}, LBP(I^i) \right], \tag{16}$$

where I^i, I^i_x, I^i_y, I^i_{xx} and I^i_{yy}, are the intensity, first- and second-order derivatives of the three image modalities, and the last term represents the LBP [41][1]. For the depth and motion flow modalities, the depth value and the module of the motion flow are considered as image intensities. Therefore, the resulting number of feature layers is $n = 23$.

In the first pedestrian detection experiment, we decided to use a simple object model and a simple classifier to demonstrate descriptors' performance (BROWNIAN, BROWNIAN$_{Proj.}$, COVARIANCE, COVARIANCE$_{Proj.}$, HOG).

For each image, BROWNIAN and COVARIANCE descriptors are extracted on a set of patches of size 12×12, fusing together the different modalities, resulting in 13×5 matrices. The global feature vector fed to a linear SVM classifier is given by $(n + n^2)/2$ (276) elements of the vectorized descriptor multiplied by the total number of patches (65). HOG descriptor is extracted in the same way, following the procedure of [42].

We show the classification performance in figure 4(a) using the Detection Trade-Off (DET) curve that expresses the Miss Rate $\left(\frac{\#FalseBackground}{\#TotalPedestrians}\right)$ against the False Positives Rate (FPRate) $\left(\frac{\#FalsePedestrians}{\#TotalBackgrounds}\right)$ on a log scale.

One can notice two important results: first, the performances of BROWNIAN and BROWNIAN$_{Proj.}$ are almost similar. This may be due to the fact that a Brownian manifold is sufficiently flat and no projection is needed, saving computational cost time (see section IV-A3 for elaboration). Second, the performance of BROWNIAN and COVARIANCE$_{Proj.}$ are even comparable, demonstrating the quality of our descriptor. In fact, considering a false positives rate of 10^{-2}, the miss rate is equal to 0.054 for COVARIANCE$_{Proj.}$ and 0.055 for BROWNIAN, with no statistical differences between the two descriptors. Furthermore, both descriptors are better than HOG, which has a miss rate greater than BROWNIAN and COVARIANCE, equal to 0.093. The HOG descriptor was computed following the same protocol of [42]. As expected, the performance of COVARIANCE without projection is lower than the others descriptors with a miss rate value over 0.1. Similar result was reported in [5]. This confirms that manifold projection is crucial for achieving good performance by the covariance descriptor.

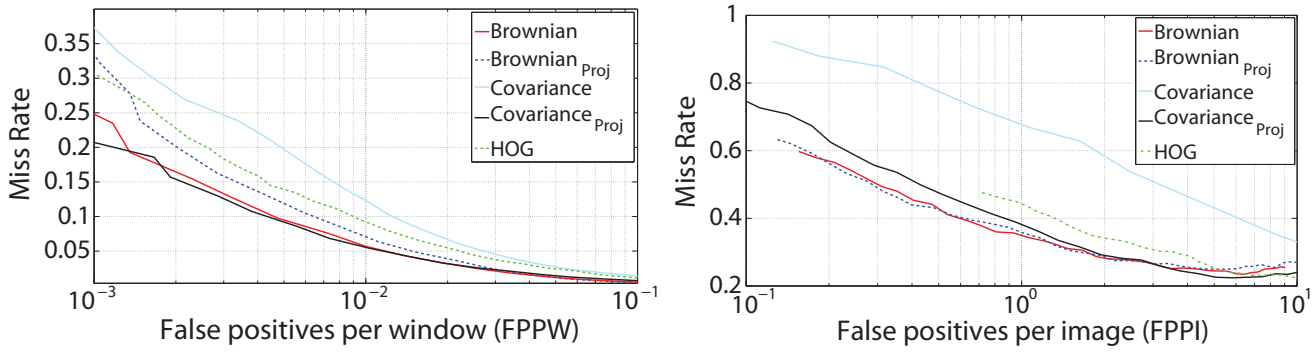
[1]Note that here LBP is employed as a low-level feature which provides just a single value, from 0 to 255, for each image pixel.

2) INRIA Pedestrian Dataset [3]

This dataset contains $1,774$ pedestrian annotations ($3,548$ with reflections) and $1,671$ person-free images. The pedestrian annotations are scaled into fixed size window of 64×128 pixels (with a margin of 16 pixels around the pedestrians). We divide the data into two: $2,416$ pedestrian annotations and $1,218$ person-free images for training, and $1,126$ pedestrian annotations and 453 person-free images for testing. Detection on the INRIA pedestrian dataset is challenging since it includes subjects with variations in pose, clothing, illumination, background, and partial occlusions. The framework used to evaluate our descriptor is the same of [3] and [6]. We detect pedestrians on each test image (positives and negatives) in all positions and scale, computing the descriptors on a patch size of 16×16 pixels with a step size of 8×8 and a scale factor of 1.2. Multi-scale and nearby detections are merged using non maximal suppression and a list of detected bounding boxes are given out. Evaluation on the list of detected bounding boxes is done using the PASCAL criterion which counts a detection to be correct if the overlap of the detected bounding box and ground truth bounding box is greater than 0.5. For each sliding window, we have the following dense feature map $\mathbf{F}(x, y)$:

$$[x, y, \Phi_R, \Phi_G, \Phi_B, M_R, M_G, M_B, R, G, B, LBP(I)] \tag{17}$$

where x and y represent horizontal and vertical pixel coordinates, Φ represents the first-order gradient vector, M its magnitude, R, G B are the channels values and LBP is computed on intensity. As in the previous experiment, the



(a) Daimler Multi-Cue dataset (b) INRIA dataset

Fig. 4. Pedestrian detection, DET curves for comparison between BROWNIAN, BROWNIAN$_{Proj.}$, COVARIANCE, COVARIANCE$_{Proj.}$ and HOG [3]: (a) Daimler Multi-Cue dataset; (b) INRIA dataset.

DESCRIPTOR	Miss Rate
BROWNIAN	**0.343**
BROWNIAN$_{Proj.}$	0.355
COVARIANCE	0.673
COVARIANCE$_{Proj.}$	0.379
HOG	0.448

TABLE I
PEDESTRIAN DETECTION. MISS RATE VALUES ON INRIA DATASET FOR A FALSE POSITIVE RATE EQUAL TO 10^0:

performances are evaluated by adopting the Detection Error Trade-Off (DET) curve.

Putting a threshold on the SVM scores, between -5 and 5, we obtain the DET curves (figure 4(b)). In this dataset, we compared our BROWNIAN descriptor with COVARIANCE and HOG. Covariance descriptor was already compared to HOG in [6], showing that it outperforms HOG [3]. However, due to time complexity and sophisticated manifold learning required for the covariance, HOG has been applied more often to pedestrian detection task.

The results clearly show that BROWNIAN descriptor outperforms COVARIANCE and HOG (experiments for HOG and COVARIANCE are done following the same framework of [6] and [3] and results are reported in Fig 4(b)) and Table I. Considering a false positives rate per image equal to 10^0, the equivalent miss rate values for BROWNIAN, COVARIANCE$_{Proj.}$ and HOG are 0.34, 0.38 and 0.45, respectively. This result illustrates that with the same number of false positive per image, BROWNIAN is able to detect 4% more pedestrians with respect to COVARIANCE, and 11% more with respect to HOG. The good performance of Brownian demonstrates good encoding of nonlinear relations that covariance fails to capture.

3) Manifold curvature analysis

The previous result surprisingly develops that Brownian descriptor, an element of Sym_d^+, performs relatively good in a Euclidean space. To further investigate this phenomenon, we employ a quantitative measure of nonflatness of the manifold, that is *the sectional curvature* κ_p [34].

Given a Riemannian manifold $(\mathcal{M}, \langle , \rangle)$, its *sectional curvature* $\kappa_p(X_p, Y_p)$ at $p \in \mathcal{M}$, if X_p and Y_p are linearly independent tangent vectors at p, is given by:

$$\kappa_p(X_p, Y_p) = \frac{\langle R(X_p, Y_p)X_p, Y_p\rangle_p}{\langle X_p, X_p\rangle_p \langle Y_p, Y_p\rangle_p - \langle X_p, Y_p\rangle_p^2} \qquad (18)$$

MANIFOLD	mean	standard deviation
BROWNIAN	-0.1000	± 0.0811
COVARIANCE	-0.2066	± 0.1398

TABLE II
CURVATURE ANALYSIS OF Sym_d^+.

where R is denoting the Riemann curvature operator.

If we use the identity matrix as a projection point $p = I_d$, we can re-write the formula 18 as:

$$\begin{aligned} \kappa_{I_d}(X, Y) &= \frac{\langle R(X, Y)X, Y\rangle}{\|X\|^2\|Y\|^2 - \langle X, Y\rangle^2} \\ &= 2\frac{Tr((XY)^2 - X^2Y^2)}{Tr(X^2)Tr(Y^2) - (Tr(XY))^2}, \end{aligned} \qquad (19)$$

where Tr is the trace operator. The sectional curvature for Sym_d^+ is nonpositive at any point. The lower κ_{I_d}, the stronger a Riemannian differs from a flat one (*i.e.* Euclidean).

The numerical evaluation of the curvature κ_{I_d} in correspondence to training samples of a particular descriptor allows us to understand concavity of the related manifolds. Having extracted BROWNIAN and COVARIANCE descriptors for INRIA dataset, we compute the mean value and the standard deviation of κ_{I_d} for both descriptors (see Table II). The mean value obtained for Brownian manifold is twice larger than for the Covariance manifold with also smaller standard deviation. That confirms our hypothesis that the manifold of the Brownian is flatter than the one of the Covariance. This might suggest that the Brownian manifold is sufficiently flat and no projection is needed for achieving good performance.

B. Person re-identification

Person re-identification is a visual search of the same person across a network of non-overlapping cameras. This task requires models dealing with significant appearance changes caused by variations in lighting conditions, pose changes and sensor scarce resolution. It is crucial that these models are based on visual features, which show a good trade-off between their discriminative power and invariance to camera changes. This trade-off can be learned [43] but it requires significant amount of labeled data which might be unattainable in a large camera network. Alternatively, it has been shown that relation-based descriptors perform relatively well in the

re-identification scenario [7], [28], [29], [30], [31] but they usually involve a high computational cost. In this section, we show that our descriptor captures distinctive information while showing practical invariance to appearance changes and keeping computational efficiency. We carry out experiments on four various re-identification datasets evaluating descriptor performance on different challenges: significant variations in illumination - PRID2011 [7]; low resolution images - CAVIAR4REID [44]; cluttered environments with occlusions - i-LIDS [45]; and serious perspective and pose changes - SAIVT-SOFTBIO database [46].

1) Experimental setup

In the past few years the re-identification problem has been the focus of intense research bringing proper metrics and datasets for evaluation. Re-identification performance is analyzed in terms of recognition rate, using the averaged *cumulative matching characteristic* (CMC) curve [47]. The CMC represents the expectation of finding the correct match in the top matches. The nAUC is a scalar obtained by normalizing the area under the CMC curve.

Every human annotation is scaled into a fixed size window of 64×192 pixels. The set of rectangular sub-regions is produced by shifting 32×32 regions with a step size of 16 pixels in either direction. This results in 33 overlapping rectangular sub-regions. From each sub-region, we extract 5 descriptors; three histogram-based descriptors: (1) COLOR$_{RGB}$ histogram, (2) LBP histogram and (3) HOG histogram, and two correlation-based descriptors: (4) COVARIANCE$_{Proj}$ and (5) BROWNIAN. We employ 11-dimensional feature map from [28]:

$$\left[x, y, R_{xy}, G_{xy}, B_{xy}, \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right] \quad (20)$$

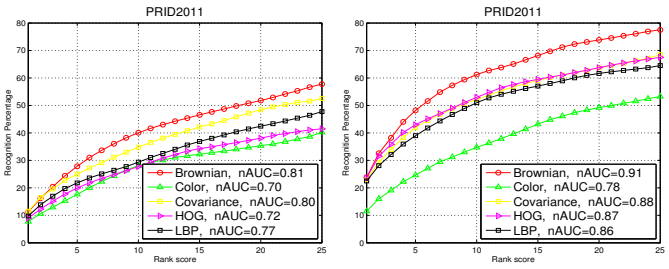
where x and y are pixel location, R_{xy}, G_{xy}, B_{xy} are RGB channel values and ∇ and θ corresponds to gradient magnitude and orientation in each channel, respectively. Motivated by the curvature analysis of Sym^+_d (section IV-A3) in these experiments we assume BROWNIAN to be an element of a Euclidean space for avoiding expensive projections on the tangent plane.

For each subject we compute signatures using randomly selected K consecutive images. We evaluated both single-shot ($K = 1$) and multiple-shot ($K > 1$) scenario. In multiple-shot case, descriptor values are simply averaged to encode a set of K images depicting the same subject. Every signature is used as a query to the gallery set of signatures from different cameras. The procedure is repeated 10 times to produce average CMC curves and nAUC values.

2) PRID2011 dataset [7]

The PRID2011 dataset consists of person images recorded from two different static surveillance cameras. Images are extracted from trajectories providing roughly 50 to 100 images per subject and camera view. Characteristic challenges of this dataset are significant differences in illumination, viewpoint and pose changes (see Fig. 5). Although, one camera view contains up to 749 subjects, only 200 person appear in both cameras. In our evaluation we used only these 200 subjects. We selected $K = 1$ and $K = 20$.

Fig. 5. The sample images from PRID2011 dataset. Top and bottom lines correspond to images from different cameras. We illustrate the first 10 subjects from the dataset.



(a) K=1 (b) K=20

Fig. 6. Comparison of different descriptors using CMC curves on PRID2011 dataset: signatures have been computed using (a) $K = 1$ and (b) $K = 20$ images.

Fig. 6 illustrates a comparison between different descriptors. We can notice that COLOR histograms are less discriminative than other descriptors. In particular, significant difference is found for multi-image signatures ($K = 20$). This effect can be explained by strong illumination changes (compare rows in Fig. 5) and reasonable image quality. Although all images are re-scaled to a uniform size (64×192), the original person images were typically 100 to 200 pixels high. This yields better quality images containing edge and texture information. The best performance among all descriptors is obtained by the BROWNIAN descriptor. The recognition accuracy per rank is given in Table III.

3) CAVIAR4REID dataset [44]

This dataset comes from the CAVIAR project[2]. Video clips are recorded from two different points of view in a shopping center in Lisbon. The resolution is half-resolution PAL standard (384×288 pixels, 25 frames per second). Small

[2]CAVIAR webpage: *http://homepages.inf.ed.ac.uk/rbf/CAVIAR*

K=1	DESCRIPTOR	$r = 1$	$r = 5$	$r = 10$	$r = 25$
	BROWNIAN	**11.43%**	**28.14%**	**40.17%**	**58.93%**
	COLOR	8.16%	17.92%	27.12%	**40.09%**
	LBP	9.98%	23.11%	29.22%	48.41%
	HOG	9.62%	20.83%	28.14%	42.79%
	COVARIANCE	11.41%	25.55%	35.39%	52.13%
K=20		$r = 1$	$r = 5$	$r = 10$	$r = 25$
	BROWNIAN	**24.31%**	**49.13%**	**62.22%**	**78.11%**
	COLOR	11.32%	26.33%	34.55%	**86.32%**
	LBP	23.51%	39.64%	51.51%	64.91%
	HOG	24.02%	44.81%	54.92%	68.61%
	COVARIANCE	23.73%	42.55%	52.19%	69.43%

TABLE III
DESCRIPTOR PERFORMANCE COMPARISON ON PRID2011 DATASET AT DIFFERENT RANKS r.

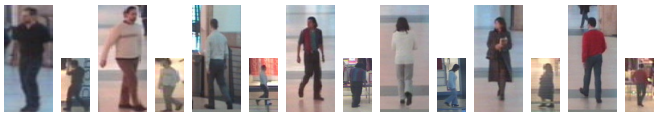
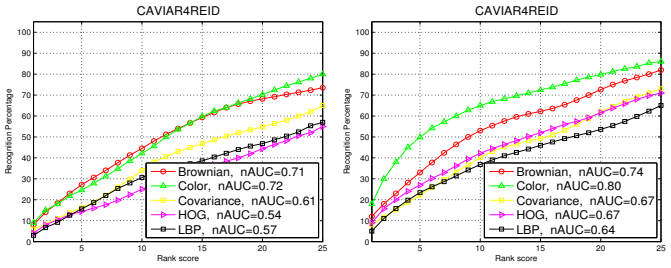


Fig. 7. The sample images from CAVIAR dataset. Pairs of images highlight appearance changes due to different camera resolution and illumination conditions.



(a) K=1 (b) K=5

Fig. 8. Comparison of different descriptors using CMC curves on CAVIAR4REID dataset: signatures have been computed using (a) $K = 1$ and (b) $K = 5$ images.

appearances and a very low resolution in one of the two cameras make the appearance matching very challenging (see Fig. 7). Of the 72 different individuals identified (with images varying from 17×39 to 72×144), 50 are captured by both views and 22 from only one camera. In our evaluation we only used these 50 people leaving out the remaining 22. The dataset provides up to 10 images per subject and camera, thus we assumed $K = 1$ and $K = 5$ for evaluation. The average CMC curves are displayed in Fig. 8.

Interestingly, neither BROWNIAN nor COVARIANCE obtained the best performance. The best accuracy is achieved by simple COLOR histogram. We believe that this is due to the significant changes in camera resolution and illumination conditions. In Fig. 7, we can notice that low resolution images barely contain edge and texture information. Note that BROWNIAN and COVARIANCE maps (Eq. (20)) mainly focus on extracting correlations between gradient-based layers in opposite to quantitative descriptors like color histograms. This experiment illustrates that both, BROWNIAN and COVARIANCE are quite sensitive to low resolution images (some persons in this dataset are less than 40 pixels height). The low resolution explains as well low performance of remaining descriptors, which mostly exploit texture information that in fact is not present in these images. However, although BROWNIAN has slightly worse performance than COLOR histogram, it outperforms all remaining descriptors for both $K = 1$ and $K = 5$ modalities. Table IV provides detailed results with respect to the considered rank and the number of K images.

4) SAIVT-SOFTBIO database [46]

This dataset consists of 152 people moving through a network of 8 cameras. Subjects travel in uncontrolled manner thus most of subjects appear only in a subset of the camera network. This provides a highly unconstrained environment reflecting a real-world scenario. In average, each subject is registered by 400 frames spanning up to 8 camera views in challenging

K=1	DESCRIPTOR	$r = 1$	$r = 5$	$r = 10$	$r = 25$
	BROWNIAN	8.61%	**27.34%**	**46.57%**	74.23%
	COLOR	**9.22%**	24.89%	42.82%	**79.80%**
	LBP	3.12%	15.12%	32.82%	57.11%
	HOG	4.02%	13.73%	26.14%	55.49%
	COVARIANCE	6.93%	16.35%	35.39%	64.13%
K=5		$r = 1$	$r = 5$	$r = 10$	$r = 25$
	BROWNIAN	12.93%	33.23%	53.52%	82.09%
	COLOR	**18.32%**	**51.23%**	**65.45%**	**86.32%**
	LBP	5.21%	24.34%	38.91%	65.31%
	HOG	9.02%	28.13%	44.49%	73.19%
	COVARIANCE	8.13%	23.65%	45.09%	74.13%

TABLE IV
DESCRIPTOR PERFORMANCE COMPARISON ON CAVIAR4REID DATASET AT DIFFERENT RANKS r.

DESCRIPTOR	$r = 1$	$r = 5$	$r = 10$	$r = 25$
BROWNIAN	**16.06%**	**37.53%**	**49.37%**	**72.09%**
COLOR	8.12%	22.09%	32.79%	54.60%
LBP	11.30%	25.62%	38.92%	59.91%
HOG	12.02%	26.73%	40.64%	62.89%
COVARIANCE	15.83%	33.65%	45.09%	66.13%

TABLE V
DESCRIPTOR PERFORMANCE COMPARISON ON SAIVT-SOFTBIO DATASET. VALUES CORRESPOND TO THE RECOGNITION ACCURACY AVERAGED AMONG ALL 56 PAIRS OF CAMERAS AT DIFFERENT RANKS r.

surveillance conditions (significant illumination, pose and view point changes, see Fig. 9(a)). Each camera captures data at 25 frames per second at resolution of 704×576 pixels. Although some cameras have overlap, we do not use this information while testing re-identification algorithms. Authors [46] provide XML files with annotations given by coarse bounding boxes indicating the location of the subjects. For each subject we randomly select the first frame in such way that we can create the signature from the next $K = 75$ frames. Every signature is used as a query to the gallery set of signatures from the other cameras. This procedure has been repeated 10 times to obtain averaged CMC results.

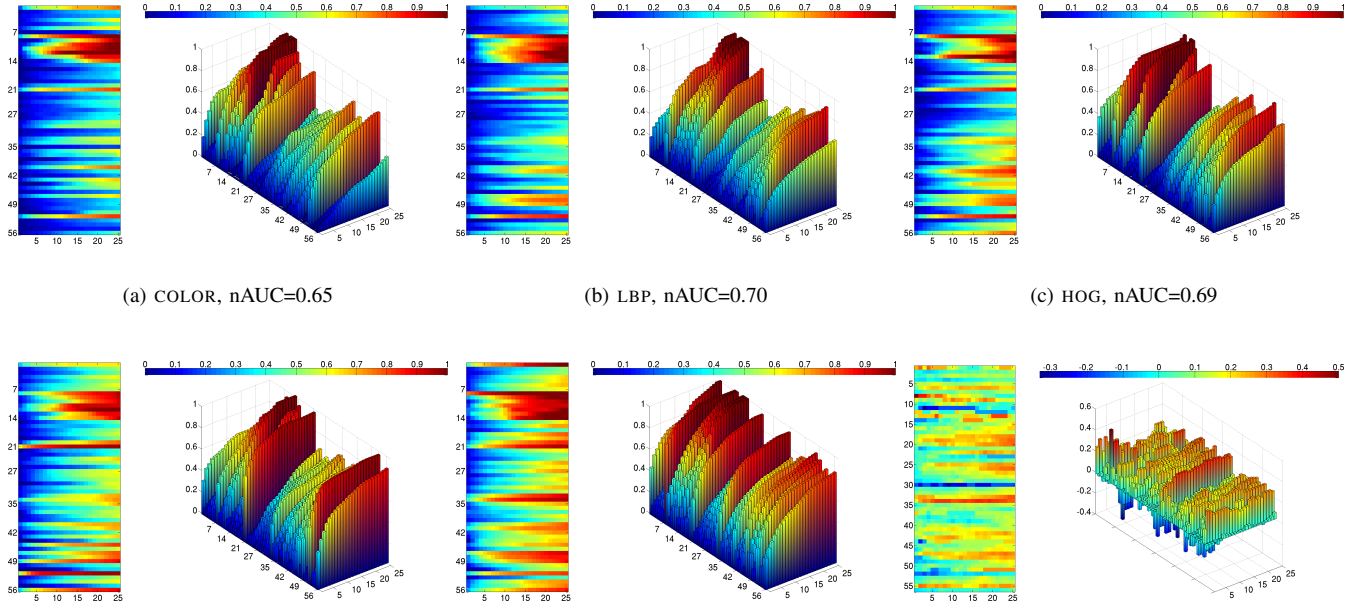
As SAIVT-SOFTBIO consists of several cameras, we display the CMC results using 3D bar-charts (see Fig. 10). The horizontal axis corresponds to recognition accuracy, while on the vertical axis the first 25 ranks are presented for each camera pair (*i.e.* having 8 cameras we actually can produce 56 CMC bar series that present recognition accuracy for each camera pair). We also color the CMC bars with respect to recognition accuracy and display it as a top-view image of 3D bar. In the result we can see that re-identification accuracy might be strongly associated with a particular pair of cameras (similar/non-similar camera view, resolution, the number of registered subjects). Fig. 10(a-e) illustrates the retrieval results for each descriptor. From the results it is apparent that Brownian descriptor outperforms the rest of descriptors. Table V shows the averaged (among all 56 camera pairs) recognition accuracy with respect to the rank and Fig. 11(a) illustrates averaged CMC curves. We can see that the Brownian descriptor consistently achieves the best performance for all ranks.

5) i-LIDS dataset [45]

Since the achieved performance showed the advantage of the Brownian descriptor, we have decided to employ MRCG model [28] that consists of a dense grid of covariances, and replace the classical covariance with BROWNIAN (without pro-

(a) SAIVT-SOFTBIO dataset (b) i-LIDS dataset

Fig. 9. Sample images from: (a) SAIVT-SOFTBIO dataset. Rows correspond to 2 first subjects from the database, while columns show the appearance from different camera views (notice significant difference in appearance); (b) i-LIDS dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.



(a) COLOR, nAUC=0.65 (b) LBP, nAUC=0.70 (c) HOG, nAUC=0.69

(d) COV_{Proj}, nAUC=0.73 (e) BROW, nAUC=0.79 (f) (e)-(d), ΔnAUC=0.06

Fig. 10. Descriptor performances as CMC bars for 56 camera pairs (a-e) of SAIVT-SOFTBIO dataset. nAUC is a weighted (by gallery size) average of nAUC obtained by each pair of cameras. For each descriptor the top view and 3D chart is presented. Red color indicates high recognition accuracy. For each descriptor we can notice the red region on the top view (see rows $7-14$). This is the retrieval result for the second camera in which only few subjects were registered (29 out of 152). The rest of cameras is more balanced (about 100 subjects per camera). (f) illustrates the difference between BROWNIAN and COVARIANCE_{Proj}. We can notice that BROWNIAN performed better for most of camera pairs (bluish color correspond to opposite case).

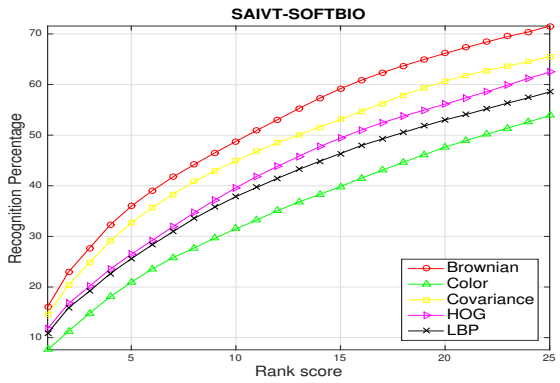
jection). This allows us to compare the descriptor with state-of-the-art approaches on i-LIDS data. This dataset contains 476 images with 119 individuals registered by two different cameras. It is very challenging dataset since there are many occlusions and often only the top part of the person is visible (see Fig. 9(b)). We reproduce the same experimental settings as [28]. Signatures are generated using $N = 2$ images. Fig. 11(b) illustrates a comparison with re-identification state-of-the-art approaches. We see that combination of MRCG technique with the Brownian descriptor outperforms state-of-the-art performance. Table VI provides the recognition accuracy with respect to the considered rank.

METHOD	$r = 1$	$r = 5$	$r = 10$	$r = 25$
OUR	**48.75%**	**75.09%**	**83.75%**	**96.70%**
MRCG[28]	45.79%	66.80%	75.21%	85.71%
CPS[44]	44.02%	69.30%	76.31%	86.09%
SDALF[48]	38.93%	64.61%	74.19%	85.13%
GROUP[49]	25.98%	45.03%	55.07%	70.01%

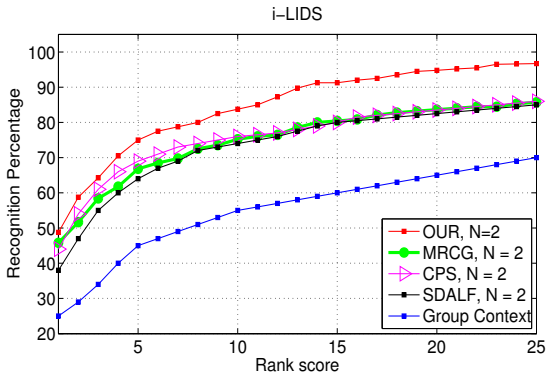
TABLE VI
PERFORMANCE COMPARISON ON I-LIDS DATASET WITH STATE-OF-THE-ART APPROACHES AT DIFFERENT RANKS r.

C. Descriptor effectiveness and efficiency

The significant improvement can be noticed on person re-identification. This confirms that the Brownian descriptor is less dependent on camera parameters than the covariance. We believe that it is due to the descriptor design based on



(a) SAIVT-SOFTBIO



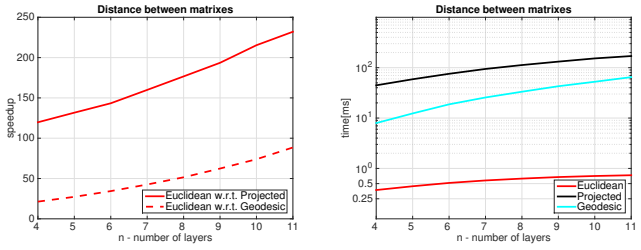
(b) i-LIDS

Fig. 11. Performance comparison using CMC curves: (a) averaged values among all 56 pairs of cameras; (b) our *vs*. MRCG [28], CPS [44], SDALF [48], Group Context [49].

statistics computed on distances between all feature layers. It bears out that this descriptor reaches sufficient trade off between discriminative power and camera invariance. Thus, we recommend Brownian as a valuable descriptor for vision tasks that require camera independence.

Moreover, other main benefits of using Brownian come from the significant speed-up in matching/classification without losing descriptive properties. Instead of projecting the descriptor on the tangent plane or using the geodesic distance, we can directly use a Euclidean metric (see Sec. IV-A3). In the results, we speedup 4 times the whole pedestrian detection framework in comparison with the classical covariance (feature extraction & classification).

For the same set of features, the proposed Brownian descriptor achieves similar or better accuracy than the classical covariance based descriptor. Importantly, the matching accuracy by Brownian is achieved at n^2 times faster speed than classical covariance (see Fig. 12(a)). Note that the larger number of feature layers n in a descriptor, the bigger speedup is achieved (theoretically speedup is lower bounded by $o(n^2)$ due to SVD computation in geodesic distance for covariance). This has a tremendous impact on the re-identification task, where speedup in matching has a direct effect on the whole retrieval framework.



(a) speedup (b) time (ms)

Fig. 12. Comparison of time complexity with respect to a chosen metric: a Euclidean metric, a tangent projection at the identity matrix and a geodesic distance.

Fig. 12 illustrates comparison of distance computation while using a Euclidean distance, distance with a projection on a tangent plane at the identity matrix and a geodesic distance. Time complexity in Fig. 12(b) was computed on Intel quad-core 2.4GHz without applying any hardware-dependent optimization routines (*e.g.* no block operations optimized for architecture). We can notice that matching is significantly faster applying a Euclidean metric.

V. Conclusions

In this paper, we introduced a novel descriptor based on mathematical statistics related to Brownian covariance. This new descriptor can be seen as a natural extension of the classical covariance descriptor. While the classical covariance measure is limited to model only linear dependencies between features, the Brownian descriptor is capable to measure all kinds of possible relationships between low-level features of visual entities.

The advantages of the proposed descriptor were presented by the theoretical analysis and the experimental evaluation on different vision tasks. We extensively evaluated the proposed approach, outperforming covariance on INRIA pedestrian detection dataset and bringing novel state-of-the-art performance in person re-identification. The significant improvement on person re-identification task *w.r.t.* the classical covariance suggests that Brownian descriptor indeed helps in correlating non linearly related features and hence can be applied to many vision tasks requiring camera invariant descriptors.

We shown not only the effectiveness of the Brownian descriptor, but also elaborate its efficiency. For computing the distance between two Brownian descriptors, we can use a Euclidean metric that is $o(n^2)$ times faster than the classical geodesic distance, where n is the number of feature layers.

We believe that this descriptor is valuable beyond the scope of the presented applications and can be used in many diverse scenarios. In future, we plan to integrate Brownian descriptors with alternative classifiers (*e.g.* boosting, decision trees) and apply it to other vision tasks. Furthermore, detailed analysis of layer-significance *w.r.t.* the vision task will be investigated.

Acknowledgment

This work has also been supported by PANORAMA European project.

REFERENCES

[1] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding (CVIU) **110**(3) (2008) 346–359

[2] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) (2004)

[3] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition (CVPR). Volume 1. (2005) 886–893

[4] Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: International Conference on Computer Vision (ICCV). (2009) 32–39

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

[6] Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **30** (2008) 1713–1727

[7] Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis (SCIA). (2011) 91–102

[8] Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: British Machine Vision Conference (BMVC). (2012) 1–11

[9] Bak, S., Kumar, R., Bremond, F.: Brownian descriptor: a rich meta-feature for appearance matching. In: Winter Applications in Computer Vision (WACV). (2014) 363–370

[10] Székely, G.J., Rizzo, M.L.: Brownian distance covariance. The Annals of Applied Statistics **3**(4) (2009) 1236–1265

[11] San Biagio, M., Crocco, M., Cristani, M., Martelli, S., Murino, V.: Heterogeneous auto-similarities of characteristics (hasc): Exploiting relational information for classification. In: International Conference on Computer Vision (ICCV). (2013) 809–816

[12] Lowe, D.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision (ICCV). Volume 2. (1999) 1150–1157

[13] Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1) (1998) 51–59

[14] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: International Conference on Computer Vision (ICCV). Volume 2. (2003) 1470–1477

[15] Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition (CVPR). CVPR 2005, Washington, DC, USA, IEEE Computer Society (2005) 524–531

[16] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, IEEE Computer Society (2006) 2169–2178

[17] Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: European Conference on Computer Vision (ECCV), Springer (2006) 490–503

[18] Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In Leonardis, A., Bischof, H., Pinz, A., eds.: European Conference on Computer Vision (ECCV). Volume 3952 of Lecture Notes in Computer Science., Graz, Autriche (2006) 428–441

[19] Zhu, Q., Zhu, Q., Avidan, S., Avidan, S., chen Yeh, M., chen Yeh, M., ting Cheng, K., ting Cheng, K.: Fast human detection using a cascade of histograms of oriented gradients. In: Computer Vision and Pattern Recognition (CVPR). (2006) 1491–1498

[20] Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: Computer Vision and Pattern Recognition (CVPR). (2007) 1–8

[21] Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition (CVPR). Volume 32. (2008) 1627–1645

[22] Weinrich, C., Volkhardt, M., Gross, H.M.: Appearance-based 3d upper-body pose estimation and person re-identification on mobile robots. In: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. (Oct 2013) 4384–4390

[23] Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. Computer Vision and Pattern Recognition (CVPR) **1** (2008) 1–8

[24] Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) **28**(12) (December 2006) 2037–2041

[25] Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M.: Multi-cue pedestrian classification with partial occlusion handling. In: Computer Vision and Pattern Recognition (CVPR). (2010) 990–997

[26] Jayasumana, S., Hartley, R., Salzmann, M., Hongdong, L., Harandi, M.: Kernel methods on the riemannian manifold of symmetric positive definite matrices. In: Computer Vision and Pattern Recognition (CVPR). (2013) 73–80

[27] Yao, J., Odobez, J.M.: Fast human detection from joint appearance and foreground feature subset covariances. Computer Vision and Image Understanding (CVIU) **115**(3) (2011) 1414–1426

[28] Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: Advanced Video and Signal-Based Surveillance (AVSS). (2011) 179–184

[29] Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Boosted human re-identification using riemannian manifolds. Image and Vision Computing **6-7** (2011) 443–452

[30] Bak, S., Charpiat, G., Corvee, E., Bremond, F., Thonnat, M.: Learning to match appearances by correlations in a covariance metric space. In: European Conference on Computer Vision (ECCV). Volume 7574. (2012) 806–820

[31] Cai, Y., Takala, V., Pietikainen, M.: Matching groups of people by covariance descriptor. In: International Conference on Pattern Recognition (ICPR). (2010) 2744–2747

[32] Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based on lie algebra. In: Computer Vision and Pattern Recognition (CVPR). Number 728-735 in CVPR (2006)

[33] Harandi, M.T., Sanderson, C., Sanin, A., Lovell, B.C.: Spatio-temporal covariance descriptors for action and gesture recognition. In: Winter Applications in Computer Vision (WACV). (2013) 103–110

[34] Tosato, D., Spera, M., Cristani, M., Murino, V.: Characterizing humans on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **35** (2013) 1972–1984

[35] Goh, A., Vidal, R.: Clustering and dimensionality reduction on riemannian manifolds. In: Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society (2008) 1–7

[36] Karcher, H.: Riemannian center of mass and mollifier smoothing. Communications on Pure and Applied Mathematics **30**(5) (1977) 509–541

[37] Bakirov, N., Székely, G.: Brownian covariance and central limit theorem for stationary sequences. Theory Probab. Appl **55**(3) (2011) 371 – 394

[38] Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **34**(4) (2012) 743–761

[39] Xu, Y., Xu, D., Lin, S., Han, T.X., Cao, X., Li, X.: Detection of sudden pedestrian crossings for driving assistance systems. IEEE Transactions on Systems, Man, and Cybernetics, Part B **42**(3) (2012) 729–739

[40] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research (JMLR) **9** (2008) 1871–1874

[41] Zheng, Y., Shen, C., Hartley, R.I., Huang, X.: Effective pedestrian detection using center-symmetric local binary/trinary patterns. CoRR (2010)

[42] Enzweiler, M., Gavrila, D.M.: A multilevel mixture-of-experts framework for pedestrian classification. Transaction on Image Processing (TIP) **20** (October 2011) 2967–2979

[43] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Computer Vision and Pattern Recognition (CVPR). (2012)

[44] Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference (BMVC). Volume 68. (2011) 1–11

[45] Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference (BMVC). Volume 23. (2009) 1–11

[46] Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P.: A database for person re-identification in multi-camera surveillance networks. In: Digital Image Computing Techniques and Applications (DICTA). (2012) 1–8

[47] Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. PETS (2007)

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

[49] Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: Conference on Graphics, Patterns and Images (SIBGRAPI). (2009) 322–329

Sławomir Bąk is a Research Engineer at STARS team, INRIA Sophia Antipolis. He received his PhD degree from INRIA, University of Nice in 2012 for a thesis on *person re-identification*. He obtained his Master degree in 2008 at Poznan University of Technology in GRID computing. In 2007 he was a member of the Automated Scheduling Optimization and Planning (ASAP) research group at University of Nottingham. He obtained his Bachelor's Degree in 2007 at Poznan University of Technology, Faculty of Computing Science. Since 2008, he has conducted research in video surveillance as a joint research scientist between INRIA and PSNC (Poznan Supercomputing and Networking Center). His research interest are in computer vision, machine learning and optimization techniques.

Marco San Biagio Marco San Biagio received the M.Sc. degree cum laude in Informatics Engineering from the University of Palermo, Italy, in 2010, and the Ph.D. in computer engineering from University of Genoa and Istituto Italiano di Tecnologia (IIT), Italy, in 2014, under the supervision of Prof. Vittorio Murino and Prof. Marco Cristani working on "Data Fusion in Video Surveillance". Currently, he is a post-doc at the Pattern Analysis and Computer Vision department (PAVIS) in IIT, Genoa, Italy. His main research interests include statistical pattern recognition and data fusion techniques for object detection and classification.

Ratnesh Kumar has obtained his Master's degree from University of Florida, Gainesville in Fall 2010, and Bachelors in Engineering from Manipal Institute of Technology, Manipal at India. Starting 2011, he is pursuing his PhD in the area of Video Segmentation and Multiple Object Tracking at STARS Team, INRIA, Sophia Antipolis France.

Vittorio Murino is full professor and head of the Pattern Analysis and Computer Vision (PAVIS) department at the Istituto Italiano di Tecnologia (IIT), Genoa, Italy. He received the Ph.D. in Electronic Engineering and Computer Science in 1993 at the University of Genoa, Italy. Then, he was first at the University of Udine and, since 1998, at the University of Verona, where he was chairman of the Department of Computer Science from 2001 to 2007. His research interests are in computer vision and machine learning, in particular, probabilistic techniques for image and video processing, with applications on video surveillance, biomedical image analysis and bio-informatics. He is also member of the editorial board of Pattern Recognition, Pattern Analysis and Applications, and Machine Vision & Applications journals, as well as of the IEEE Transactions on Systems Man, and Cybernetics. Finally, he is senior member of the IEEE and Fellow of the IAPR.

François Brémond is a Research Director at INRIA Sophia Antipolis. He created the STARS team on the 1st of January 2012 and was previously the head of the PULSAR INRIA team in September 2009. He obtained his Master degree in 1992 from ENS Lyon. He has conducted research works in video understanding since 1993 both at Sophia-Antipolis and at USC (University of Southern California), LA. In 1997 he obtained his PhD degree from INRIA in video understanding and pursued his research work as a post doctorate at USC on the interpretation of videos taken from UAV (Unmanned Airborne Vehicle) in DARPA project VSAM (Visual Surveillance and Activity Monitoring). In 2007 he obtained his HDR degree from Nice University on Scene Understanding: perception, multi-sensor fusion, spatio-temporal reasoning and activity recognition.