Boosting Fisher Vector based Scoring Functions for Person Re-Identification

Stefano Messelodi, Carla Maria Modena

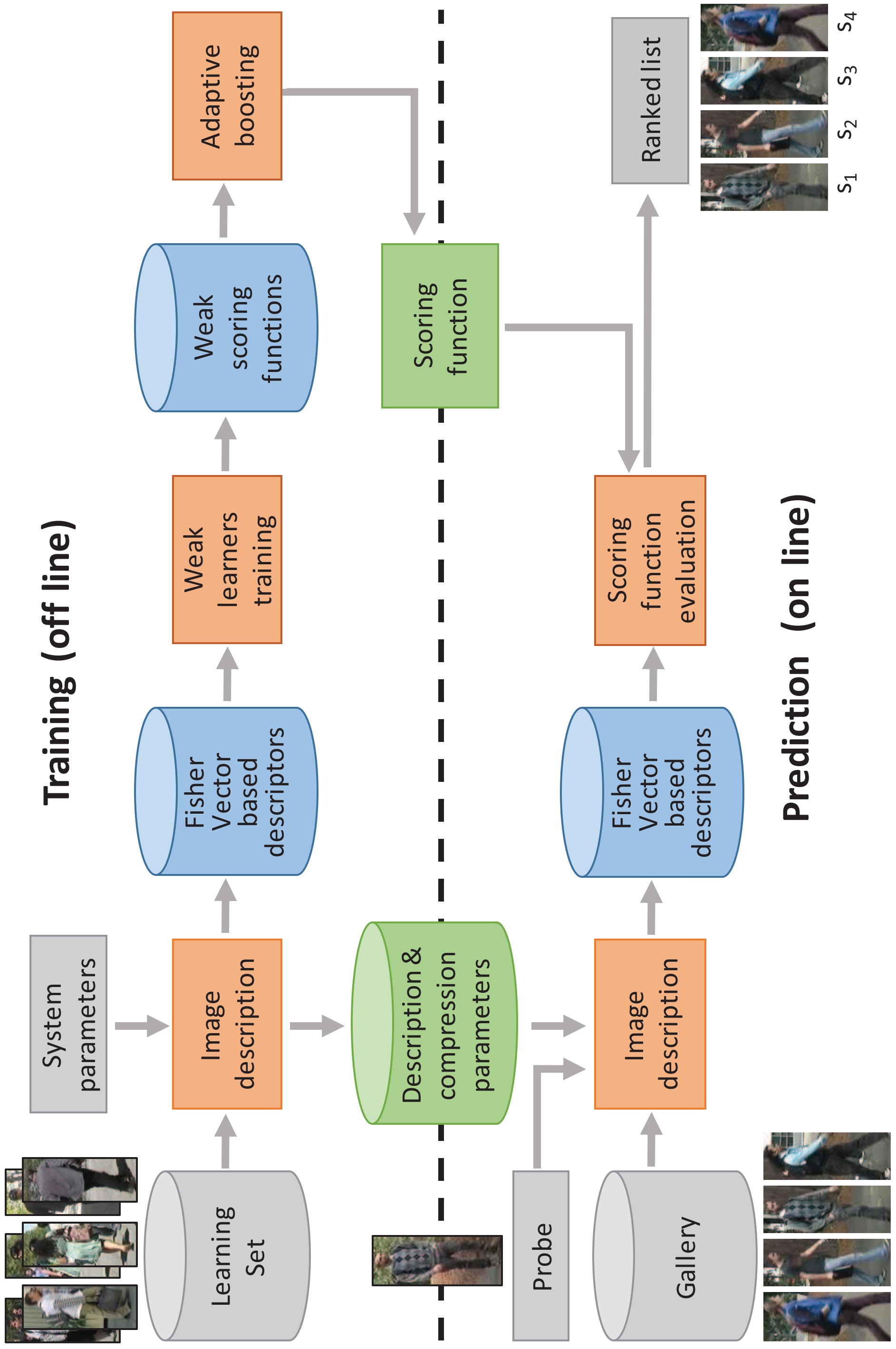
PII: S0262-8856(15)00114-6
DOI: doi: 10.1016/j.imavis.2015.09.008
Reference: IMAVIS 3437

To appear in: *Image and Vision Computing*

Received date: 30 August 2014
Revised date: 17 August 2015
Accepted date: 16 September 2015

Please cite this article as: Stefano Messelodi, Carla Maria Modena, Boosting Fisher Vector based Scoring Functions for Person Re-Identification, *Image and Vision Computing* (2015), doi: 10.1016/j.imavis.2015.09.008

Graphical Abstract



Boosting Fisher Vector based Scoring Functions for Person Re-Identification

Stefano Messelodi, Carla Maria Modena[1]

Fondazione Bruno Kessler, Via Sommarive 18, I-38123 Trento, Italy
e-mail: messelod@fbk.eu, modena@fbk.eu

Abstract

In recent years, much effort has been put into the development of novel algorithms to solve the person re-identification problem. The goal is to match a given person's image against a gallery of people. In this paper, we propose a single-shot supervised method to compute a scoring function that, when applied to a pair of images, provides a score expressing the likelihood that they depict the same individual. The method is characterized by: (i) the usage of a set of local image descriptors based on Fisher Vectors, (ii) the training of a pool of scoring functions based on the local descriptors, and (iii) the construction of a strong scoring function by means of an adaptive boosting procedure. The method has been tested on four data-sets and results have been compared with state-of-the-art methods clearly showing superior performance.

Keywords: Person re-identification, Fisher Vector, Adaptive boosting, Likelihood ratio, Similarity ranking

1. Introduction

The problem to automatically retrieve a selected person from video streams is of fundamental importance to video analysis. Applications vary from searching for suspicious individuals in a network of surveillance cameras, to maintaining person identity from one camera to the other for behavior analysis. Several factors contribute making the problem very hard, in fact a person's appearance can vary greatly through scenes due to changes in viewpoints, illumination conditions, pose and orientation, or to the possible usage of different acquisition devices. Other disturbing factors are the presence of shadows, occlusions, or individuals in the scene with similar appearance.

Person re-identification consists of matching observations of individuals across disjoint camera views. In very recent years, this problem has received a considerable attention, and various surveys and reviews are available, pointing

[1] Corresponding author. e-mail: modena@fbk.eu. Phone: +390461314508 Fax: +39 0461314501

out different aspects of this challenging topic [1, 2, 3, 4, 5, 6]. For this reason, we direct the reader to these papers for a detailed discussion on the challenges posed by the problem, and for an overview of state-of-the-art methods along with their performance on publicly available data-sets.

Broadly speaking, in order to address the problem, people have to be detected in videos and be represented by descriptors which aim to capture their visual appearance. The descriptors are then used to compare different individuals and to determine the correspondence among them. Re-identification methods proposed in literature usually avoid to consider the detection phase and assume to work with images whose content is restricted to a bounding box around the person. They differ on the descriptor construction that can refer to a single view of the person (single-shot methods) or to multiple views obtained by briefly tracking (tracklet) the person's movements (multi-shot methods), and on the comparison of descriptors, which can be direct (unsupervised) or based on similarity measures learned using a set of labelled samples (supervised).

Although re-identification can be regarded as a binary classification problem over pairs of people descriptors, it is clear that a binary answer (same person or not) becomes harder as the gallery size increases. Thus the evaluation of a re-identification system is accomplished by regarding re-identification as a ranking problem rather than a classification one: the algorithms return a sorted list of candidates and the best performance is obtained if the correct correspondence is in most cases at, or close to, the first position of the returned list.

Using a standard taxonomy, the method proposed in this paper is a supervised single-shot recognition method. The major novelty of the proposed method, named *BFiVe*, consists of combining the power of Fisher Vector descriptors with the ability of boosting procedures to select the most appropriate local descriptors to build a strong scoring function.

Starting from low-level features computed at pixel level in regions obtained from a coarse to fine image subdivision, an image is initially represented by a family of local descriptors based on Fisher Vectors that are then dimensionally reduced to an optimal size using Principal Component Analysis. In the training phase, a pool of weak scoring functions is generated using the local descriptors. Finally, the construction of a strong scoring function by means of an adaptive boosting procedure is performed using a minimum error procedure on the weak learners. The error is computed by analysing the position of the right match in the ranked output of the weak scoring function. In this way, the regions that better contribute to collocate the right match in the very first positions weigh more in the global scoring function.

Previously published methods, to the best of our knowledge, aggregate local descriptors in order to build a single image descriptor, and learn a single metric to provide the final ranking. The novelty of *BFiVe* is that it learns a proper metric for each subimage, *i.e.* there are as many learnt rankers as the regions the image is divided into. A second learning step is performed using a ranking-based boosting approach, which combines local rankers to establish the final ranking function.

The proposed method has been experimentally validated on four challenging data-sets: VIPeR, 3DPeS, PRID 2011 and i-LIDS-119. The obtained figures clearly outperform the best previously published results on all of them.

The rest of the paper is organized as follows. Section 2 briefly describes the state-of-the-art methods included in the supervised single-shot category. Section 3 presents synthetically the *BFiVe* method. Sections 4 and 5 explain the techniques we propose for the description of images and for the learning of the scoring functions, respectively, while Section 6 illustrates the on-line usage of the method. Section 7 presents the experimental validation of *BFiVe* including a comparison with the state-of-the-art, the methodology followed for parameters selection, and an analysis of the computational complexity. Section 8 analyzes several aspects of the proposed method, discussing its main features. Section 9 concludes the paper.

2. Related works

In this section, we review several works in recent literature that fall into the supervised, single-shot re-identification category. Methods in this class are characterized by specific features used to describe the images and by specific procedures that make use of a labelled data-set to learn a metric by enforcing small distances among data of the same class (images depicting the same person). The usage of common data-sets and evaluation protocols is mandatory for a direct and meaningful comparison of the method's performance.

In Ma *et al.* [7], the color image is firstly divided into large, fixed, non-overlapping rectangular regions and each pixel is described by simple feature vectors. The feature vectors of the pixels that fall in each region are encoded and aggregated into Fisher Vectors, which are then concatenated and dimensionally reduced with Principle Component Analysis (PCA) to obtain the final signature of the image. Using Pairwise Constrained Component Analysis (PCCA) [8] a similarity metric, sLDFV, is learnt, *i.e.* a projection into a low-dimensional space where distances between pairs of signatures respect the desired matching constraints.

In Pedagadi *et al.* [9] images are described by very high dimensional features based on local color histograms and their statistics in HUV and HSV color spaces, separately. The feature vectors can be exploited in an efficient way using a dimensionality reduction approach that combines unsupervised and supervised techniques, namely PCA and Local Fisher discriminative analysis (LF). The Euclidean metric is then used for the comparison. In the same paper, a novel statistic is introduced to characterize re-identification performance, called Proportion of Uncertainty Removed (PUR) index. It is invariant to test set size, and we use it to evaluate our method's performance.

In [10, 11, 12], the main focus is on metric learning rather than on feature selection specific to the re-identification task. In [10], a Support Vector Machine framework is proposed to obtain an optimized metric for nearest neighbor classification called Large Margin Nearest Neighbor with Rejection (LMNN-R), *i.e.* the classifier returns no matches if all neighbors are beyond a certain distance.

The signature is built by applying a PCA reduction to the concatenation of histograms of color channels (RGB and HSV) extracted from a grid of rectangular overlapping windows.

In [11], relaxed pairwise distance metric learning, RP-MeL, is used to address the problem of maximizing the probability that a pair of images depicting the same person has a smaller distance than a pair of different individuals. Once the metric has been learnt, only linear projections are necessary at search time, where a nearest neighbor classification is performed. The image descriptor is obtained by merging local color and texture features computed on overlapping rectangular regions, then reduced with PCA. In [12], a "keep it simple and straightforward metric" (KISSME) was introduced to learn a distance metric from equivalence constraints. The method is applied on a variety of challenging benchmarks including person re-identification across spatially disjoint cameras, using the same descriptors as [11].

The KISSME metric learning algorithm is also used by Ma *et al.* in [13] to improve the discriminative ability of their proposed descriptors: To gain robustness to illumination variations, scale and shifts, the image representation relies on the combination of biologically inspired features [14] based on covariance descriptors. This approach, named kBiCov, that focuses on feature selection and on metric learning, produces one of the best results currently present in literature.

In the re-identification task, one of the main problems is the different responses of the camera due to sensor variability, illumination changes, and aiming angle. Hirzer *et al.* [15], address the 'different camera properties problem' by learning a transition function from one camera to an other. This is realized by learning a Mahalanobis metric using pairs of images coming from different cameras. The mean color values from small image regions are combined with a histogram of Local Binary Patterns to represent an image, and then pairwise sample differences are learnt for re-identification, considering correspondent people and also impostors that invade the perimeter of a given pair (Efficient Impostor-based Metric Learning, EIMeL).

In [28] the authors formulate a relative distance comparison (RDC) model, to maximise the likelihood of a pair of true matches that have a relatively smaller distance compared to an incorrect matching pair in a soft discriminant manner. The descriptors are obtained by dividing the images into six horizontal stripes. For each stripe, color features and texture features are extracted, giving rise to an image descriptor vector in a 2784 dimensional feature space. The model is based on logistic functions which are learnt with an iterative optimization algorithm on subsets of the data and then combined in an ensemble way to obtain the final RDC.

Li and Wang [16] propose locally aligned feature transforms, LAFT, for matching people across camera views that can have complex cross-view variations. Images to be matched are softly assigned to different local experts of a gating network according to the similarity of cross-view transforms, then they are projected to a common feature space and matched with a locally learnt discriminative metric.

An original framework is proposed in [17], where a reference set of images is used to generate reference-based descriptors for probe and gallery people. The starting signatures are built from color and texture features following the approach in [11]. In the training phase, a reference set of image pairs is used to learn a subspace in which the data of the same subjects from different cameras are maximally correlated using Regularized Canonical Correlation Analysis (RCCA). The so-called reference descriptors (RDs) of probe and, respectively, gallery images are then obtained by projecting the original feature vectors into the RCCA subspace using the two learnt matrices. Re-identification is performed by comparing the RDs of the probes and the RDs of the gallery images. In this way, a direct comparison of probes and gallery images is avoided.

3. Outline of the *BFiVe* method

In this section, we provide an overview of the proposed re-identification method, which is outlined in Figure 1. As labelled data-sets are crucial for developing supervised methods, we briefly explain their typical structure and usage in the context of single-shot re-identification algorithms. Such data-sets consist of a set of N_P individuals each depicted in two images, typically taken from different cameras: $D = \{(I_1^a, I_1^b), (I_2^a, I_2^b), \ldots, (I_{N_P}^a, I_{N_P}^b)\}$, where a and b indicate the first and the second view, respectively.

In order to train the algorithm and to evaluate system performance, the data-set is (randomly) split into two disjoint parts D_L and D_T, called *learning set* and *test set*, with cardinality N_L and N_T, respectively. The learning set is used to train the re-identification system and the test set is used to evaluate the performance on people and images never seen during the training phase. To perform this last step, the view pairs of the test set of individuals are split and assigned to two different sets, called *probes* and *gallery*. For each probe image, the system provides a ranking of the gallery images based on their similarity to the probe. By knowing the gallery individuals corresponding to the probes it is possible to evaluate the quality of the rankings and compare performance of the different methods.

The supervised phase consists of three steps: image description, training of weak learners, and adaptive boosting (Figure 1 top). The input is a set of labelled samples along with some system parameters, while the output is a *scoring function*, that associates to each couple of images a score expressing the likelihood that the two images depict the same individual. In the first step, the image is regarded as a set of local regions, called *receptive fields*. For each receptive field, color and gradient low-level features are extracted at pixel level, decorrelated, and then encoded by means of Fisher Vectors. This is a technique based on the Fisher Kernel [18], popular in image classification [19] and used in [7] for person re-identification. The receptive field descriptors are then dimensionally reduced after applying PCA.

In the second step, using the labelled learning set, the local descriptors are employed to estimate a weak scoring function, or weak learner, for each receptive field. The definition of a weak scoring function is based on: (i) the differences

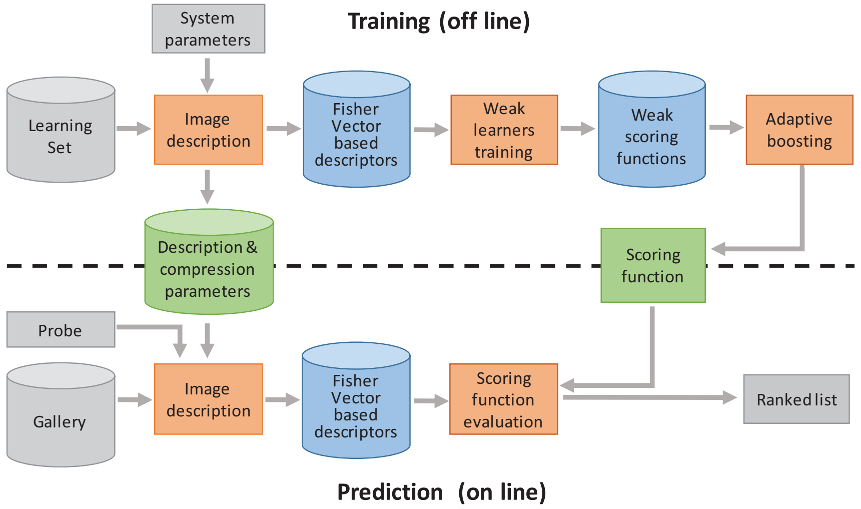


Figure 1: A schematic view of the proposed method.The learning phase takes place off-line (top) using a data-set of labelled samples (learning set) and system parameters. Through three main steps (orange boxes), a scoring function is generated and a set of internal parameters, which are used on-line (bottom) where a probe image, is compared with the gallery images and a ranked list is produced. Best viewed in color.

between correspondent local descriptors of each image pair in the learning set, and (ii) on the comparison of the difference distributions coming from image pairs depicting the same individual and different individuals, respectively.

In the third step of the learning phase we compute a scoring function, Ω, defined as a linear combination of a subset of the weak learners that are selected, along with their coefficients, through an adaptive boosting procedure.

The on-line usage of the re-identification system, Figure 1 (bottom), involves a probe image and a set of gallery images. The images are described by means of a set of Fisher Vectors using the internal parameters learnt in the off-line phase. The learnt function Ω, which associates a similarity score to a couple of images, permits the system to sort the gallery images with respect to their similarity to the probe.

4. Image description

In this section, we explain our technique for the description of an image depicting a single individual.

4.1. Receptive fields

As a preliminary step, images are re-scaled to a prefixed size $M \times N$. Next, they are regarded as a cover of *receptive fields*, connected regions characterized by various sizes and shapes. As shown in Figure 2, in our case, receptive fields are overlapping rectangular regions covering the image at different levels (i, j). The pair (i, j) indicates the number of parts the image is split into, respectively, along the vertical and horizontal direction.

At level (i, j), the rectangle's size is $M/i \times N/j$ and the row and column coordinates of the top-left corner span the range $[0, M(i-1)/i]$ with step $\frac{M}{2i}$ and the range $[0, N(j-1)/j]$ with step $\frac{N}{2j}$, respectively (dots in Figure 2). The number of rectangles at level (i, j) is $(2i-1)(2j-1)$. The number of receptive fields $N_{\texttt{RF}}$ in the example in Figure 2 is 104.

We emphasize the fact that the region set includes tiles with different sizes, such as the whole image and small overlapping cells, obtained by dividing the image into regions in a pyramidal way.

4.2. Low-level feature extraction

Input images are firstly converted from RGB into HSL and into YC_bC_r color spaces and the following ordered set of $N_{\mathcal{C}} = 5$ channels is considered: $\mathcal{C} = \{H, S, L, C_b, C_r\}$. We ignore the Y channel because it is strongly correlated to L. For each color channel $c \in \mathcal{C}$, we compute the gradient maps c_x and c_y along the horizontal and vertical axis, respectively, and the second order derivatives c_{xx}, and c_{yy}. For each image component c, each pixel $p = (x_p, y_p)$ is described, as in [7], by the following 7-dimensional low-level feature vector:

$$D_c(p) = (x_p, y_p, c(p), c_x(p), c_y(p), c_{xx}(p), c_{yy}(p)).$$

4.3. Fisher encoding

The Fisher Vector encoding method aims to fit a generic probabilistic model $P(X; \Theta)$ to the data – where Θ represents the model parameters and X the data – and then to characterize the data by its deviation from the generative model. The deviation is measured using the derivative of the data log-likelihood with respect to the model parameters, called 'score':

$$G(X, \Theta) = \frac{\partial}{\partial \Theta} \ln P(X; \Theta). \quad (1)$$

The covariance matrix of the score vector is known as the Fisher Information Matrix (FIM). The Fisher Vector of X is defined [19] as the normalization of the score in Eq. 1, obtained by applying the triangular matrix of the FIM's inverse Cholesky decomposition. The Fisher Vector is used as a signature for the data, which can be classified using a discriminative classifier. Generally, mixture models are chosen as generative models because of their attractive flexibility for the estimation of underlying density. The components of the convex combination are themselves densities with vector valued parameters. We assume that the data points are generated from a mixture of a finite number of multivariate

Gaussian distributions, *i.e.* we model the density distribution of data with the classical Gaussian Mixture Model (GMM), and, following [19], we assume that the covariance matrices of the components are diagonal. This generative model also offers the advantage of tractability in computing the needed gradients.

For each receptive field $R \in \mathcal{R} = \{R_1, \ldots, R_{N_{\mathtt{RF}}}\}$ and for each channel $c \in \mathcal{C}$, we model the distribution of the data $\{D_c(p)\}_{p \in R}$ by means of a mixture of K Gaussians:

$$g_{\Theta(R,c)}(x) = \sum_{i=1}^{K} w_i g_i(x; \mu_i, \sigma_i) \tag{2}$$

where w_i are the weights of the different components and the parameter vector is $\Theta(R, c) = (\mu_1, \ldots, \mu_K, \sigma_1 \ldots, \sigma_K)$, with μ_i indicating the means of the K multivariate Gaussians, and σ_i their diagonal covariance matrices. The mixture parameters $\Theta(R, c)$ are estimated by using a maximum likelihood approach over a subset of images randomly selected from the learning set D_L.

As explained in [19], it is good practice to reduce the dimension of low-level descriptors using PCA before fitting the Gaussian Mixture Model. In our case, we apply principal component mapping to decorrelate the features without reducing the space dimension, since it is already low. In this way, the diagonal covariance matrices assumption made by the considered model, is better fulfilled.

For each receptive field R and for each channel c, we then compute the Fisher Vector $\mathbf{f}_{R,c}$ using the estimated generative model. This descriptor has the advantage to have a fixed number of components, independently of the number of pixels in the receptive field. The dimension of $\mathbf{f}_{R,c}$ is given by the number of the Gaussian mixture parameters involved in the Fisher Vector computation, $2K$, times the dimension d of the low-level feature vectors ($d = 7$, in our case). The final descriptor $\mathbf{F}_R$, for a receptive field R, is the vector built by concatenating the $N_{\mathcal{C}}$ vectors $\mathbf{f}_{R,c}$ with $c \in \mathcal{C}$. Its dimension is $N_{\mathtt{fv}} = 2\,K\,d\,N_{\mathcal{C}}$. In our experiments, we use $K = 16$ which yields a local descriptor with $N_{\mathtt{fv}} = 1120$ components. The parameters of the GMMs are stored in a repository to be used in the prediction phase.

4.4. Dimensionality reduction

Dimensionality reduction is commonly used as a preprocessing step before training a supervised learner. One might expect that the dimensionality reduction influences the generalization performance because some information is discarded. In line with other works, *e.g.* [13], we found through experimentation that by applying PCA dimensionality reduction to descriptors gives rise to the double benefit of reducing the size of the vector and thus the complexity of the method, and increasing system performance.

PCA computes the linear transformation that projects the training descriptors into a variance-maximizing subspace. Although PCA operates in an unsupervised setting, without using the labels from the training set, it still exhibits useful properties in the loop of the recognition process because of (i) de-noising the information carried by the descriptor and (ii) decorrelating the data. An excessive reduction of the descriptor dimension causes a loss of information carried by $\mathbf{F}_R$, negatively affecting re-identification performance. Selecting the correct number of principal components is crucial to the success of PCA in representing the data-set. The most suitable number ℓ of principal components (PCs) to be considered for the dimensionality reduction can be obtained by applying multiple rounds of cross-validation using different partitions of the learning set. ℓ is chosen among a reasonable set of first principal directions $\mathcal{L} = \{\ell_1, \dots, \ell_{N_{\mathcal{L}}}\}$. We perform PCA computation by means of an iterative method based on Expectation Maximization [20], that permits us to compute only the desired number of principal components without computing all of them.

For each receptive field R, we compute the reduced Fisher Vector $\hat{\mathbf{F}}_R \in \mathbb{R}^{\ell}$. The parameters involved in the dimensionality reduction step, *i.e.* mean vector, scale and projection matrix, for each receptive field, are stored in the mentioned repository to be used during prediction.

Summing up, at this stage the description of an image I is obtained as a set of local functions:

$$\text{Ł}_R : I \mapsto \hat{\mathbf{F}}_R \in \mathbb{R}^{\ell} \tag{3}$$

with $R \in \mathcal{R}$ and ℓ representing the chosen dimension for descriptor reduction. The computation of the image descriptor set makes use of the following data:

- the projection matrices to decorrelate the low-level feature vectors; they are $N_{\mathcal{C}} \times N_{\mathtt{RF}}$;
- the GMM parameter vectors $\Theta(R, c)$ for each $c \in \mathcal{C}$;
- the parameters for the Fisher Vector PCA reduction to ℓ-dimensional subspace: mean vector, scale and projection matrix.

5. Scoring function learning

Boosting is a general iterative method that combines a set of weak classifiers (or learners) to form a strong classifier. The final classifier is a linear combination of the selected weak learners, each weighted by a coefficient estimated by the boosting procedure. The core idea is to assign a weight to each training sample in order to change their importance during the procedure. Hard to classify samples tend to have higher weights than the others. In fact, misclassified samples increase the error according to their weight. At each iteration, samples are re-weighted according to the result of their classification.

Re-identification, however, is regarded as a ranking problem rather than a classification one. Adaptive boosting methods have been proposed to build strong ranking functions starting from a set of weak rankers [21]. In this case, the goal is to form a ranking function that respects at best a set of pairwise constraints among samples.

Our goal is to learn a scoring function that associates to each pair of images the likelihood that they depict the same person. Therefore, samples are represented by couples of images and the pairwise constraints try to force image pairs depicting the same person to have a higher score with respect to pairs depicting different people. In the next two subsections, we (i) define the weak learners used in the first learning step of *BFiVe* and (ii) describe the second learning step, *i.e.* the boosting procedure to obtain the strong ranker.

5.1. Training of weak learners

A weak learner (or weak ranker) is a scoring function that associates to a pair of images the likelihood they depict locally the same person. We train $N_{\mathtt{RF}}$ weak learners, one for each receptive field R that is described by means of a ℓ-dimensional vector. Let I_i^v be an image, where v is the view, a or b, and $i \in \mathcal{I}_L = \{1, \ldots, N_{\mathrm{L}}\}$, the set of indices of the learning set elements. The images I_i^v have been described by means of a family of ℓ-dimensional vectors $\{\text{Ł}_R(I_i^v)\}_{R \in \mathcal{R}}$. As we are focusing on a fixed R, for the sake of notation simplicity, let us denote in this subsection, the vector $\text{Ł}_R(I_i^v)$ simply with $\boldsymbol{x}_i^v$.

Let $\mathcal{S}_R$ and $\mathcal{D}_R$ be the sets containing the differences of local descriptors of image pairs depicting, respectively, the same individual (similar) and different individuals (dissimilar):

$$\mathcal{S}_R = \{\boldsymbol{x}_i^a - \boldsymbol{x}_i^b,\, \boldsymbol{x}_i^b - \boldsymbol{x}_i^a \mid i \in \mathcal{I}_L\}$$

$$\mathcal{D}_R = \{\boldsymbol{x}_i^v - \boldsymbol{x}_j^w | i, j \in \mathcal{I}_L, i \neq j; v, w \in \{a, b\}\}.$$

Both the vector sets are modelled by means of multivariate Gaussian distributions:

$$\begin{aligned} P(\boldsymbol{x}|\mathcal{S}_R) &= (2\pi)^{-\frac{\ell}{2}} |\boldsymbol{\Sigma}_\mathcal{S}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\boldsymbol{x}^t \boldsymbol{\Sigma}_\mathcal{S}^{-1} \boldsymbol{x}} \\ P(\boldsymbol{x}|\mathcal{D}_R) &= (2\pi)^{-\frac{\ell}{2}} |\boldsymbol{\Sigma}_\mathcal{D}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\boldsymbol{x}^t \boldsymbol{\Sigma}_\mathcal{D}^{-1} \boldsymbol{x}} \end{aligned}$$

with variance $\boldsymbol{\Sigma}_\mathcal{S}$ and $\boldsymbol{\Sigma}_\mathcal{D}$, respectively, which are $\ell \times \ell$ symmetric, positive semi-definite matrices. The means of the distributions are the null vector, as for each vector difference $\boldsymbol{x}$ in $\mathcal{S}_R$, or in $\mathcal{D}_R$, also the opposite vector $-\boldsymbol{x}$ belongs to the same set.

The scoring function characterizing the weak learner is directly derived from the log-likelihood ratio $\log(P(\boldsymbol{x}|\mathcal{S}_R)/P(\boldsymbol{x}|\mathcal{D}_R))$. Having modeled the probabilities as multivariate Gaussian [22], we have:

$$\begin{aligned} &\log \frac{P(\boldsymbol{x}|\mathcal{S}_R)}{P(\boldsymbol{x}|\mathcal{D}_R)} = \\ &\qquad \log \frac{(2\pi)^{-\frac{\ell}{2}} |\boldsymbol{\Sigma}_\mathcal{S}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\boldsymbol{x}^t \boldsymbol{\Sigma}_\mathcal{S}^{-1} \boldsymbol{x}}}{(2\pi)^{-\frac{\ell}{2}} |\boldsymbol{\Sigma}_\mathcal{D}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\boldsymbol{x}^t \boldsymbol{\Sigma}_\mathcal{D}^{-1} \boldsymbol{x}}} = \\ &\qquad \frac{1}{2}(-\log|\boldsymbol{\Sigma}_\mathcal{S}| - \boldsymbol{x}^t \boldsymbol{\Sigma}_\mathcal{S}^{-1} \boldsymbol{x} + \log|\boldsymbol{\Sigma}_\mathcal{D}| + \boldsymbol{x}^t \boldsymbol{\Sigma}_\mathcal{D}^{-1} \boldsymbol{x}). \end{aligned} \tag{4}$$

By eliminating offset and scale, which do not affect the ranking, we define the following function:

$$\begin{aligned} \hat{\Omega}_R \quad &: \quad \mathbb{R}^\ell \to \mathbb{R} \\ \hat{\Omega}_R(\boldsymbol{x}) \quad &= \quad \boldsymbol{x}^t(\boldsymbol{\Sigma}_\mathcal{D}^{-1} - \boldsymbol{\Sigma}_\mathcal{S}^{-1})\boldsymbol{x}. \end{aligned}$$

Finally, the weak scoring function Ω_R of two images I, J is defined as follows:

$$\Omega_R(I, J) = \hat{\Omega}_R(\text{Ł}_R(I) - \text{Ł}_R(J)). \tag{5}$$

It is completely determined by the matrix $\mathbf{M}_R = \mathbf{\Sigma}_{\mathcal{D}}^{-1} - \mathbf{\Sigma}_{\mathcal{S}}^{-1}$.

Summing up, for each receptive field R the weak learner training procedure generates the $\ell \times \ell$ matrix $\mathbf{M}_R$ that defines the scoring function Ω_R.

5.2. Adaptive boosting

The adaptive boosting algorithm is detailed in Algorithm 1. The input consists of the weak learners Ω_R, the learning set images D_L and the maximum number of iterations $N_{\mathtt{loop}}$.

The sample set consists of pairs of images $s_{i,j} = \{(I_i^a, I_j^b)\}_{i,j \in \mathcal{I}_L}$ taken from D_L (line 4). The initial weights $w_{i,j}$ assigned to samples whose images depict the same person are set to $0.6/N_{\mathtt{L}}$, while for the other samples the weight is set to $0.4/(N_{\mathtt{L}}^2 - N_{\mathtt{L}})$ (lines 6-8). In this way, we initially give more importance to ranking errors of similar pairs with respect to the others.

Starting from the set of weak learners $\{\Omega_R \mid R \in \mathcal{R}\}$, the iterative boosting procedure selects, at each k-th iteration, the learner Ω_{R_k} that produces the minimum error E on the training samples (lines 15-19).
We define the *error function* E of the weak learner as the sum of the weights associated to the incorrectly ranked samples. Formally:

$$E = \sum_i \left(\Xi(i) w_{i,i} + \sum_j \chi(i,j) w_{i,j} \right) \tag{6}$$

where

$$\chi(i,j) = \begin{cases} 1 & \text{if } \Omega_R(I_i^a, I_j^b) \geq \Omega_R(I_i^a, I_i^b) \\ 0 & \text{otherwise} \end{cases}$$

$$\Xi(i) = \begin{cases} 1 & \text{if } \sum_j \chi(i,j) > 1 \\ 0 & \text{otherwise} \end{cases}$$

Let Ω_{R_k} be the ranker that gives rise to the minimum error E_{min} at iteration k. Ω_{R_k} is selected to be part of the final strong learner and its coefficient α_k is computed as a function of the minimum error (lines 20-21). At each iteration, the weights of the samples are updated depending on the outcome of their ranking and then normalized to sum to one (lines 23-26). Note that the same weak learner can be selected in more than one iteration.

The final scoring function is a linear combination of the scalar functions selected by the boosting procedure, each one weighted by the associated coefficient:

$$\Omega(I, J) = \sum_k \alpha_k \Omega_{R_k}(I, J). \tag{7}$$

Summing up, the training procedure gives rise to a subset of selected receptive fields $\{R_k\} \subseteq \mathcal{R}$, each one specifing a weak scoring function Ω_{R_k}, along with the associated coefficients α_k. They allow the computation of the final scoring function in Eq. 7.

Algorithm 1 Adaptive boosting procedure.

```
**Input:** $D_L$, $N_{\mathtt{loop}}$, $\Omega_R$
**Output:** $\Omega\_list$
$\diamond$ *build the training samples*
$S \leftarrow \{s_{i,j} = (I_i^a, I_j^b)\}_{i,j \in \mathcal{I}_L}$
$\diamond$ *initialize weights of samples*
**for all** $i, j \in \{1, \ldots, N_{\mathtt{L}}\}$ **do**
    $w_{i,j} = \begin{cases} 0.6/N_{\mathtt{L}} & \text{if i = j} \\ 0.4/(N_{\mathtt{L}}^2 - N_{\mathtt{L}}) & \text{otherwise} \end{cases}$
**end for**
$\diamond$ *initialize the scoring functions list*
$\Omega\_list \leftarrow \emptyset$
$k \leftarrow 0$
**while** $k < N_{\mathtt{loop}}$ **do**
    $\diamond$ *initialize the minimum error*
    $E_{min} \leftarrow \infty$
    **for all** $(R)$ **do**
        select scoring function $\Omega_R$
        compute its error $E$ on $S$ as in Eq. 6
        update $E_{min}$ and best $(R_k)$
    **end for**
    $\alpha_k \leftarrow \frac{1}{2} log((1 - E_{min})/E_{min})$
    $\Omega\_list \leftarrow \Omega\_list \cup \{(\Omega_{R_k}, \alpha_k)\}$
    $\diamond$ *update weights of training samples*
    **for all** $i, j \in \{1, \ldots, N_{\mathtt{L}}\}$ **do**
        $w_{i,j} = \begin{cases} w_{i,j} e^{-\alpha_k} & s_{i,j} \text{ correctly ranked} \\ w_{i,j} e^{\alpha_k} & \text{otherwise} \end{cases}$
    **end for**
    normalize weights to sum 1.0
    $k \leftarrow k + 1$
**end while**
```

6. Prediction

The on-line phase involves a cropped image depicting a person (probe I_p) that has to be compared to a set of images (gallery: $\{I_{g_1}, \ldots, I_{g_m}\}$), typically stored in a repository along with their descriptors. The local descriptors of each I_{g_i} are computed relatively only to the subset of random fields R_k involved in the learnt scoring function (Eq. 7). The computation of the local descriptors uses the internal parameters stored during the training phase: projection matrices to decorrelate the low-level features, GMM parameters for the Fisher Vector computation, and PCA dimensionality reduction matrices.

The descriptor of the probe image I_p is computed in the same way. Next, the scoring function Ω is applied to each pair (I_p, I_{g_i}), with $i = 1, \ldots, m$, yielding a list of scores that enables the system to rank the gallery images with respect to their similarity to the probe.

7. Experimental results

The proposed method has been tested on four data-sets, namely VIPeR [23], 3DPeS [24], PRID 2011 [25], and i-LIDS-119 [26], following the experimental protocol used by the large majority of the works that adopt them. We compare our method with the results of the best state-of-art techniques which fall in the supervised single-shot category, using the figures declared by their authors.

Re-identification methods are classically evaluated by comparing their Cumulative Matching Characteristic (CMC) curve. The curve synthesizes the quality of the gallery image rankings produced by the algorithm for each probe in the test set. The CMC curve provides, for a given rank r (on the horizontal axis), the probability that the rank of the correct person falls in the first r positions of the ranking output by the re-identification system (on the vertical axis). By knowing the correct correspondence between probe and gallery images it is possible to create a histogram h where the r-th bin counts how many times the correct image has rank r. The histogram is then normalized by dividing every bin by $N_{\mathtt{T}}$. The CMC curve is the cumulative histogram of h. Often, results are presented in tables where only the probabilities corresponding to some selected ranks are reported. Other indices used to compare re-identification methods are nAUC (normalized Area Under Curve) and PUR (Proportion of Uncertainty Removed) [9]. The first one represents the normalized area under the CMC curve, while the second computes the uncertainty reduction in re-identification after the ranking computation.

The VIPeR data-set can be considered the standard data-set for person re-identification. Almost all recent works in this field compare their results using VIPeR. It contains 1264 images depicting 632 individuals, each one observed from two different point of views. The images are size normalized to 128×48. The data-set is characterized by relevant variations in viewpoint and illumination, causing strong differences in people appearance.

The data-set 3DPeS contains various video sequences taken from a real surveillance network, composed of 8 cameras, monitoring a section of a University campus. Data was collected during several days and is characterized by strong illumination variations. A selection of snapshots from the database has been extracted specifically to validate re-identification algorithms. There are 1012 snapshots of 200 individuals. Only 192 of them appear in at least two images. The images are not size normalized, the rows vary from 88 to 362 and the columns from 31 to 272, while the average size is about 158×74.

The PRID 2011 data-set contains images of several individuals taken by two surveillance cameras, named A and B. Images taken from camera A and camera B depict, respectively, 385 and 749 individuals, with 200 of them appearing in both views. The main difficulty related to this data-set comes from the fact that there are significant differences in people pose, illumination conditions and background characteristics. The images are size normalized to 128×64.

The i-LIDS for re-identification data-set, also known as i-LIDS-119, was built from the i-LIDS Multiple-Camera Tracking Scenario. It contains 476 images captured by non-overlapping cameras, representing 119 people. The number of images for each individual varies from 2 to 8 and the image dimensions from 32×76 to 115×294. In addition to pose changes and illumination variations, people in this data-set are also subject to occlusion and often only the top part of the person is visible.

Examples of image pairs depicting the same person, taken from different data-sets, are visualized in Figure 3.

7.1. Parameters selection

The presented method depends on some parameters that affect, to a different extent, the re-identification performance. The involved parameters are: N_P, $N_{\mathtt{L}}$, $N_{\mathtt{T}}$, $N_{\mathtt{RF}}$, $M \times N$, K, ℓ, $N_{\mathtt{loop}}$. Some parameter values depend on the data-set (N_P) or are mandatory to make a fair comparison with state-of-the-art methods ($N_{\mathtt{L}}$, $N_{\mathtt{T}}$), other have been fixed at design time after some preliminar tests ($N_{\mathtt{RF}}$, $M \times N$, K), proving a good trade-off between performance and computational complexity.

The value of ℓ, $N_{\mathtt{loop}}$ parameters have been estimated during a cross validation test based on image pairs in the learning set. To this purpose, we randomly split the learning set into two parts: a reduced training set and a cross validation set, respectively containing approximately 3/4 and 1/4 of the number of learning set couples. Table 1 reports the number of elements in the two parts for the considered data-sets.

Parameter	VIPeR	3DPeS	PRID 2011	i-LIDS-119
$N_{\mathtt{L}}$	316	96	100	89
N_{LR}	250	72	75	66
N_{LC}	66	24	25	23

Table 1: Number of people in the learning set, $N_{\mathtt{L}}$, in the reduced learning set, N_{LR} and in the cross validation set, N_{LC}, for each data-set.

Selecting the most suitable number ℓ of principal components is relevant to the system performance, and it will be discussed in Section 8. ℓ is chosen among a reasonable set of first principal directions $\mathcal{L} = \{\ell_1, \ldots, \ell_{N_{\mathcal{L}}}\}$ by repeating the following procedure:

1. compute Fisher descriptors on the reduced learning set;
2. reduce dimensionality by selecting the first ℓ PCs;
3. build the corresponding weak-learners using the reduced learning set;
4. apply the boosting procedure;
5. evaluate performance on the cross validation set, and keep track of the best result.

Parameter	VIPeR	3DPeS	PRID 2011	i-LIDS-119
N_P	632	192	$200^{(*)}$	119
$N_{\mathtt{L}}$	316	96	100	89
$N_{\mathtt{T}}$	316	96	$100^{(*)}$	$30^{(**)}$
$M \times N$	128×64	128×64	128×64	128×64
$N_{\mathtt{RF}}$	104	104	104	104
K	16	16	16	16
ℓ	90	40	40	40
$N_{\mathtt{loop}}$	59	58	60	38

Table 2: The table shows the values of the parameters involved in the proposed algorithm for the considered data-sets. In order to make a fair comparison with state-of-the-art methods, 549 extra images are added to the gallery in (*) and about 90 images populate the probes part of the test set in (**).

The dimension ℓ giving the best performance in terms of average PUR index (in the first 300 boosting iterations) has been selected. Figure 4 plots the average PUR index on the cross validation set over 30 random splits across different PCA reductions, for all the considered data-sets.

The value of the $N_{\mathtt{loop}}$ parameter is estimated analogously, although it is not critical from a certain point onwards. Figure 5 shows the behaviour of the PUR index with respect to the number of boosting iterations ($N_{\mathtt{loop}}$). For all the data-sets, the plots increase quickly to reach a stable value. Therefore, we selected the number of boosting iterations corresponding to the first local maximum in the steady state. Table 2 reports the estimated values of ℓ and $N_{\mathtt{loop}}$ along with the those of the other system parameters.

Concerning the computation of the Fisher Vectors $\mathbf{f}_{R,c}$ we used the GMM-Fisher Library by J. Sanchez which is a sub-library of *Encoding Methods Evaluation Toolkit* [27]. In the vector normalization step of $\mathbf{f}_{R,c}$ we used the library default parameters ($\alpha = 0.5$ and norm $L_p = L_2$) *i.e.* the standard power normalization, which consists of transforming each element of the vector by the square root of its absolute value, then followed by the l_2 normalization, which consists of rescaling the vector to have unit l_2-norm.

7.2. Results

In this section, we report the figures obtained by *BFiVe* on four data-sets. In order to increase the reliability of the results, training and tests are repeated 30 times using different random partitions[2] and the average scores are reported in tables and also presented by means of CMC curves. They express the probabilities that the correct person falls within the first positions in the ranking

[2] To permit fair comparisons we provide the list of images for the different partitions of the considered data-sets at tev.fbk.eu/bfivesplits

provided by our method and related state-of-the-art methods, that were briefly described in Section 2.

VIPeR. Following the standard protocol for this data-set, it is randomly split into two sets of 316 image pairs, the former used to train the re-identification module and the latter to evaluate its performance. Table 3 and Figure 6 compare the performance of *BFiVe* with respect to the most relevant recent state-of-the-art methods. Our method clearly outperforms the others in all the rankings.

3DPeS. The set of 192 people appearing in at least two images is randomly split into two halves which populate the learning and the test set, respectively. Among the images available for each person, two are randomly selected to be part of the learning or test set. The selected images have been normalized to size 128 × 64. Table 4 and Figure 7 show how our method outperforms the state-of-the art. As the 3DPeS data-set has been made available only recently, the comparison is limited to three methods (figures taken from [9]).

PRID 2011. The data-set includes 200 individuals observed by both cameras. They are randomly split into two subsets of equal size that compose the learning and the test set. Only a few works have evaluated their methods on this data-set using a protocol which includes in the gallery set all the 549 images depicting people taken from camera B but not from camera A. As a consequence the probe set consists of 100 images while the gallery contains 649 images. Table 5 and Figure 8 compare the performance of *BFiVe* with respect to that methods. Our algorithm outperforms the others in all the rankings on this data-set, too.

i-LIDS-119. Following the protocol presented in [28], the set is divided in two parts: p people for the test set and $119 - p$ for the learning set. As each individual is depicted in a variable number of images (from 2 to 8) taken from different cameras, one image is randomly selected as a gallery image, while the remaining views form the probe set. As a consequence the gallery set consists of p images while the probe set contains a variable number of images, around $3p$. We used the training data in a single-shot fashion, *i.e.* two views have been randomly selected, one for the gallery and one for the probes of the learning set.

Table 6 and Figure 9 compares the performance of *BFiVe* with respect to recent state-of-the-art methods using the same protocol, with $p = 30$. *BFiVe* clearly outperforms the others in all the rankings.

7.3. Computational complexity

In this section, we present an analysis of the computational complexity of the proposed method. The complexity, reported in Table 7, is expressed as a function of the system parameters along with the following quantities:

- the number of iterations of the EM algorithm to estimate the Gaussian Mixture Models ($N_{\texttt{EM}}$);
- the number of different weak learners involved in the final scoring function ($N_{\texttt{W}}$) whose maximum value is given by $\min(N_{\texttt{loop}}, N_{\texttt{RF}})$;

Method	*Rank*				
(nr of random splits)	1	10	20	50	100
BFiVe (30)	**38.9**	**81.9**	**91.1**	**98.3**	**99.9**
kBiCov[13] (10)	31.1	70.7	82.4	-	-
RCCA[17] (10)	30	75	87	96	99
LAFT[16] (100)	29.6	69.3	81.3	96.8	-
RP-MeL[11] (10)	27	69	83	95	99
sLDFV[7] (100)	26.5	70.9	84.6	-	-
LF[9] (100)	24.2	67.1	81.4	94.1	-
EIMeL[15] (10)	22	63	78	93	98
KISSME[12] (100)	20	62	75	92	-
RDC[28] (10)	15.7	53.9	70.1	-	-

Table 3: **VIPeR** - The table shows the probability (in percentage) that the correct person appears in the first *Rank* $(1, 10, 20, 50, 100)$ positions in the similarity ranked list. The figures are computed on test sets as averages over a number (shown in parenthesis) of random partitions of the data-set into training and test. The nAUC for *BFiVe* is 0.981, while for kBiCov it is 0.965. The PUR index of the proposed method is 0.574.

Method	*Rank*				PUR
(nr of random splits)	1	10	25	50	index
BFiVe (30)	**41.7**	**73.4**	**86.7**	**95.9**	**0.464**
LF[9] (100)	33.4	70.0	84.8	95.1	0.349
KISSME[12] (100)	22.9	62.2	80.7	93.2	0.255
LMNN-R[10] (100)	23.0	55.2	73.4	88.9	0.211

Table 4: **3DPeS** - The table shows the probability (in percentage) that the correct person appears in the first *Rank* $(1, 10, 25, 50)$ positions in the similarity ranked list. The nAUC of the proposed method is 0.905. The last column reports the PUR index of the referenced methods as published in [9].

- the total number of pixels in all the receptive fields ($N_{\mathtt{PIX}}$);
- the total number of pixels in the receptive fields involved in the final scoring function ($N_{\mathtt{pix}}$).

Their mean and standard deviation values over 30 tests are presented in Table 8 for the considered data-sets.

8. Discussion

In this section, we present a discussion about the main features of the proposed method and emphasize those that mostly contribute to its good performance. Furthermore, we perform an analysis of the contribution of the receptive fields and their associated scoring functions to the final strong ranker. Finally, we show some examples where the system performs poorly.

Method	*Rank*				
(nr of random splits)	1	10	20	50	100
BFiVe (30)	**19.6**	**52.7**	**65.2**	**79.4**	**96.1**
RP-MeL[11] (10)	15	42	54	70	80
EIMeL[15] (10)	15	38	50	67	80

Table 5: **PRID 2011** - The table shows the probability (in percentage) that the correct person appears in the first *Rank* $(1, 10, 20, 50, 100)$ positions in the similarity ranked list. The nAUC of the proposed method is 0.947 and the PUR index is 0.485.

Method	*Rank*			
(nr of random splits)	1	5	10	20
BFiVe (30)	**48.09**	**74.85**	**87.37**	**97.65**
RDC[28] (10)	44.05	72.74	84.69	96.29
kBiCov[13] (10)	39.17	68.19	82.10	95.26

Table 6: **i-LIDS-119** (test with $p = 30$) - The table shows the probability (in percentage) that the correct person appears in the first *Rank* $(1, 5, 10, 20)$ positions in the similarity ranked list. The nAUC of the proposed method is 0.890 and the PUR index is 0.426.

In common with other works in literature, *BFiVe* (i) is based on local descriptors extracted from several regions that cover the image (receptive fields), (ii) exploits the descriptive power of Fisher Vectors, and (iii) adopts a scoring function based on the well-known likelihood ratio discriminant function. The most relevant difference with respect to other works is that the scoring function is not based on the concatenation of local descriptors, but instead many local scoring functions are learnt, one for each receptive field, and these scoring functions are then combined using a boosting procedure. Moreover, we show that the introduction of a PCA-based reduction of the local descriptors provides a remarkable benefit to system performance.

8.1. Combination of scoring functions

We start by comparing the performance of our boosting method to combine local scoring functions, with respect to learning a single function built on the concatenation of local descriptors. To this end, the design of a simplified experiment has been necessary by considering that the concatenation of 104 receptive field descriptors, reduced to a dimension of, for example, 60 PCA, gives rise to a vector in a 6240-dimensional space, and that training our scoring function requires the system to compute and invert two covariance matrices in such high-dimensional space. We therefore considered a subset of 11 receptive fields – those belonging to levels $(1 \times 1), (2 \times 1), (4 \times 1)$ – each described by a vector projected on ℓ-dimensional PCA subspaces, with ℓ varying from 5 to 30 with a step of 5.

Figure 10 (left), compares the PUR index of the ranking obtained on the VIPeR data-set while using two different ways of combining the local descriptors

Prediction

Low-level features extraction and decorrelation	$O(N_{\mathtt{pix}}\, N_{\mathcal{C}}\, d^2)$
Fisher vector computation	$O(N_{\mathtt{fv}}\, N_{\mathtt{pix}})$
Fisher vector PCA reduction	$O(N_{\mathtt{fv}}\, N_{\mathtt{W}} \ell)$
Score computation	$O(N_{\mathtt{W}} \ell^2)$

Training

Low-level feature extraction and decorrelation	$O(N_{\mathtt{L}}\, N_{\mathtt{pix}}\, N_{\mathcal{C}}\, d^2)$
GMM data modelling	$O(N_{\mathtt{L}}\, N_{\mathtt{fv}}\, N_{\mathtt{PIX}}\, N_{\mathtt{EM}})$
Fisher vector computation	$O(N_{\mathtt{L}}\, N_{\mathtt{fv}}\, N_{\mathtt{PIX}})$
Fisher vector PCA reduction	$O(N_{\mathtt{L}}\, N_{\mathtt{RF}}\, N_{\mathtt{fv}}\, \ell)$
Building weak learners	$O(N_{\mathtt{RF}}\, (N_{\mathtt{L}}^2 + \ell^{2.373}))$
Boosting	$O(N_{\mathtt{loop}}\, N_{\mathtt{RF}}\, N_{\mathtt{L}}^2)$

Table 7: The computational complexity of the main steps of the proposed method. Figures are expressed in terms of the values of the system parameters. The off-line and on-line steps are reported separately.

Param	VIPeR		3DPeS		PRID 2011		i-LIDS-119	
	mean	std	mean	std	mean	std	mean	std
$N_{\mathtt{EM}}$	15.2	3.4	14.6	2.4	14.1	2.9	15.4	2.4
$N_{\mathtt{W}}$	29.2	3.4	20.4	4.1	23.8	2.7	20.6	2.3
$N_{\mathtt{PIX}}$	125440	-	125440	-	125440	-	125440	-
$N_{\mathtt{pix}}$	39987	6481	24917	7744	32700	7442	33775	6439

Table 8: Parameters values (mean and standard deviation) that affect the computational complexity of the proposed method, measured in 30 tests performed in the experimental session on different data-sets.

of the receptive fields. The first one corresponds to the *BFiVe* method (Boosting), while the second one corresponds to the concatenation of the descriptors into a single vector (Concatenation) which encloses the same information.

An analogous behaviour is obtained if the dimensionality reduction is performed by random feature selection instead of PCA. The graph in Figure 10 (right) compares the PUR index of boosting ranking, obtained on VIPeR data-set, while using two different ways of reducing the dimensionality of descriptors – namely principal components (PCA) and random selection (Random). Again, only 11 receptive fields are considered in this experiment.

Results clearly show that the integration of local scoring functions is better than concatenation, even with a different dimensionality reduction method. Furthermore, by comparing PUR figures on the left and on the right in Figure 10, we can observe that in this task PCA dimensionality reduction is significantly more suitable than a random feature selection.

8.2. PCA dimensionality reduction

We analyze the impact on system performance of the PCA-based reduction of descriptors dimension and in particular of the value of the ℓ parameter. In order to be independent of the boosting algorithm, only the receptive field at level (1×1), *i.e.* covering the whole image, has been considered in the tests. The plots in Figure 11 show the ranking performance on VIPeR data-set while varying the number of principal components used to describe the receptive field. Performance on both the training and the test set are reported in terms of recognition rate at Rank 1, Rank 10 and Rank 50, and in terms of PUR index.

As expected, the PUR index on the training set increases while the number of retained components increases, while on the test set it reaches a maximum around the value $\ell = 100$ and then drops for larger ℓ values. This behaviour is caused by the fact that as ℓ increases the same happens to the number of parameters of the Gaussians that model the similar and dissimilar data. As a consequence, from a certain point on, overfitting takes place and the generalization ability of the learnt scoring function drops. This is even more evident if we consider the plots in Figure 12. Here the scoring function associated to the single receptive field at level (1×1) is used as a similar/dissimilar classifier: a pair of images, whose difference descriptor vector is $\boldsymbol{x}$, is classified as representing the same individual if $P(\boldsymbol{x}|\mathcal{S}) > P(\boldsymbol{x}|\mathcal{D})$, *i.e.* they are 'similar' if the value in Eq. 4 is positive, 'dissimilar' otherwise.

The plots show *true positive* and *true negative* accuracies, for training and test set, at different numbers of principal components. At around $\ell = 100$, we observe a sudden drop of true positive accuracy in the test set, again due to overfitting the training set, causing a deterioration of the ranking performance.

In order to evaluate the sensitivity of the method to different PCA dimensionality reductions and to highlight the effectiveness of parameter estimation, we computed the CMC curves at different values of parameter ℓ on the test sets for each data-set. Figure 13 reports, as an example, the result of the study on the ViPER data-set. We observe that the dimension selected by the cross-validation procedure provides good performance, close to the best ($\ell = 110$ provides slightly better PUR index with respect to $\ell = 90$).

8.3. Contribution of receptive fields

Receptive fields contribute to a different extent (including not contribute at all) to the final ranker, depending on the selection process during the boosting stage. For this reason, we investigate how weak learners contribute to the building of the strong learner.

The average number of different weak learners involved in the final scoring function $N_{\mathtt{W}}$ is available in Table 8. As an example, in the case of VIPeR they are only about 29.2 (average over 30 random splits) out of the 104 in the pool. Figure 14 shows the receptive fields that produced the first 10 more relevant scoring functions (in terms of α) for a random partition of ViPER data-set. In this partition, the different weak learners involved in the final scoring function are 31.

Figure 15 shows the percentage of the receptive fields, level per level, involved in the final scoring function (average over 30 random splits). It is evident in all data-sets that small sized regions are preferred over the largest one.

Furthermore, we analyzed the contribution to the final ranker at pixel level. Figure 16 graphically shows how many times the image pixels have been used in the final scoring function. The picture is a mean over 30 random splits of the considered data-sets. It can be observed that relevant information is localized in two main areas for VIPeR and PRID 2011, one area for 3DPES and one for i-LIDS-119. In 3DPES, relevant information is more localized with respect to the other data-sets.

8.4. Failure examples

Finally, in Figure 17 we show, for each data-set, an example of where the proposed algorithm fails to rank the correct person in the first 10 positions. As we observe from these examples, major difficulties for the algorithm arise from people having similar clothing. Concerning the i-LIDS-119 data-set, occluded subjects pose a major problem.

9. Conclusions

In this paper, a supervised learning approach for single-shot person re-identification is proposed. The descriptor of a person image consists of a set of local region descriptors based on Fisher Vectors extracted from a coarse to fine image partition, starting from color and gradient features of the pixels in the region.

The training phase acts in two steps. In the first step, each region descriptor is used to define a weak scoring function, or weak ranker, that, when applied to a pair of images, produce a similarity score between them. In this way, it is possible to sort a database of known people with respect to the likelihood they depict the same person by only regarding corresponding portions of the images.

The second step consists of a boosting procedure, that is performed to select a subset of the weak scoring functions to minimize a ranking error computed on the learning set. The error takes into account the relative position in the ranking of image pairs depicting the same person with respect to those depicting different individuals. The final scoring function is a weighted combination of the selected weak ranking functions. The function is applied on a test set to evaluate the proposed method.

We experimentally demonstrated that using weak learners with a boosting strategy outperforms the use of a single learner based on a descriptor containing the same information.

The experiments, conducted on publicly available data-sets that are typically used in single-shot re-identification papers, show that the proposed method outperforms the state-of-the-art: the recognition rate at rank 1 is 38.9% on VIPeR, 41.7% on 3DPeS, 19.6% on PRID 2011, and 48.1% on i-LIDS-119.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that contributed to improving the quality of the work. Furthermore, they express their gratitude to Paul Chippendale for his careful reading of the paper.
This work has been partially developed within a research collaboration between Fondazione Bruno Kessler and BV TECH Ricerca S.r.l.

References

[1] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011.

[2] T. D'Orazio and G. Cicirelli. People re-identification and tracking from multiple cameras: A review. In *IEEE International Conference on Image Processing*, pages 1601–1604, Orlando, FL, USA, 2012.

[3] R. Satta. Appearance Descriptors for Person Re-identification: a Comprehensive Review. July arXiv:1307.5748v1, Department of Electrical and Electronic Engineering, University of Cagliari, Italy, 2013.

[4] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 42(2):1–37, 2013.

[5] A. Bedagkar-Gala and S.K. Shah. A Survey of Approaches and Trends in Person Re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[6] M.A. Saghafi, A. Hussain, H.B. Zaman, and M.H.M. Saad. Review of person re-identification techniques. *IET Computer Vision*, 8(1):20, 2014.

[7] B. Ma, Y. Su, and F. Jurie. Local Descriptors encoded by Fisher Vectors for Person Re-identification. In *International Workshop on Re-Identification, in conjunction with ECCV*, pages 413–422, Florence, Italy, 2012.

[8] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2672, Providence, RI, USA, 2012.

[9] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325, Portland, OR, USA, 2013.

[10] M. Dikmen, E. Akbas, T.S. Huang, and N. Ahuja. Pedestrian Recognition with a Learned Metric. In *Asian Conference on Computer Vision*, pages 501–512, Queenstown, New Zealand, 2010.

[11] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof. Relaxed Pairwise Learned Metric for Person Re-identification. In *European Conference an Computer Vision*, pages 780–793, Florence, Italy, 2012.

[12] M. Köstinger, M. Hirzer, Wohlhart, P.M. Roth, and H. Bischof. Large Scale Metric Learning from Equivalence Constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, Providence, RI, USA, 2012.

[13] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6-7):379–390, 2014.

[14] M. Riesenhuber and T. Poggio. Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[15] M. Hirzer, P. Roth, and H. Bischof. Person Re-identification by Efficient Impostor-based Metric Learning. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 203–208, Beijing, China, 2012.

[16] W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, Portland, OR, USA, 2013.

[17] L. An, M. Kafai, S. Yang, and B. Bhanu. Reference-based person re-identification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 244–249, Krakow, Poland, 2013.

[18] T.S. Jaakkola and D. Haussler. Exploiting Generative Models in Discriminative Classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493, Denver, CO, USA, 1998.

[19] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[20] S. Roweis. EM Algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems 10*, pages 626–632, Denver, CO, USA, 1997.

[21] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[22] B. Moghaddam, T. Nastar, and A. Pentland. A Bayesian Similarity Measure for Direct Image Matching. In *International Conference on Pattern Recognition*, pages 350–358, Vienna, Austria, 1996.

[23] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition,and Tracking. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 262–275, Rio de Janeiro, Brazil, 2007.

[24] D. Baltieri, R. Vezzani, and R. Cucchiara. 3DPes: 3D People Dataset for Surveillance and Forensics. In *International ACM Workshop on Multimedia Access to 3D Human Objects*, pages 59–64, Scottsdale, AZ, USA, 2011.

[25] M. Hirzer, C. Beleznai, P.M. Roth, and H. Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In *Scandinavian Conference on Image Analysis*, pages 91–102, Lund, Sweden, 2011.

[26] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *British Machine Vision Conference*, pages 23.1–23.11, London, UK, 2009.

[27] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, pages 76.1–76.12, Dundee, UK, 2011.

[28] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by Relative Distance Comparison, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

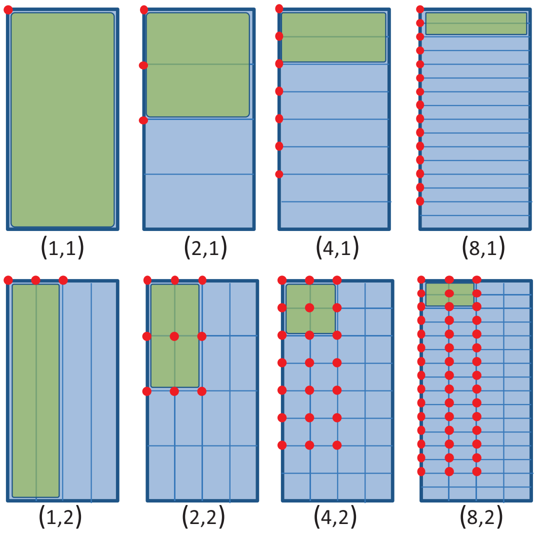


Figure 2: Image coverage by means of *receptive fields*. The number of parts in which the image is split along the vertical and horizontal direction is indicated under each subdivision. At each level, the highlighted rectangle represents the region shape and the dots represent the possible locations of its top-left corner.

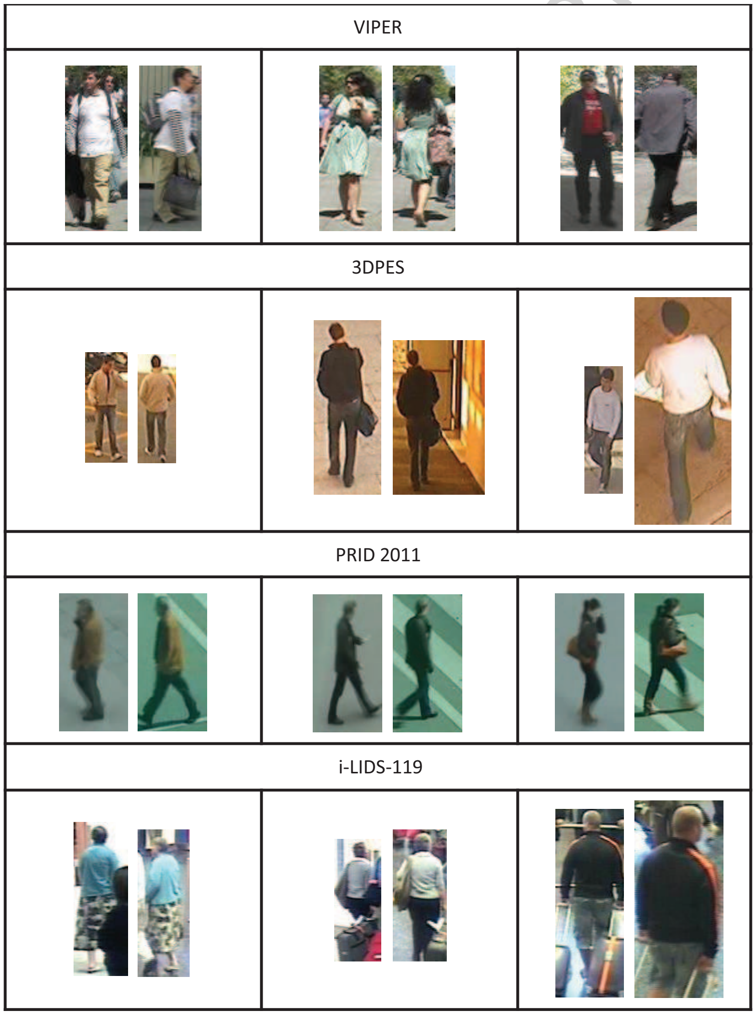


Figure 3: Some shots taken from the data-sets used in the experimental session. Images in the same box depict the same person.

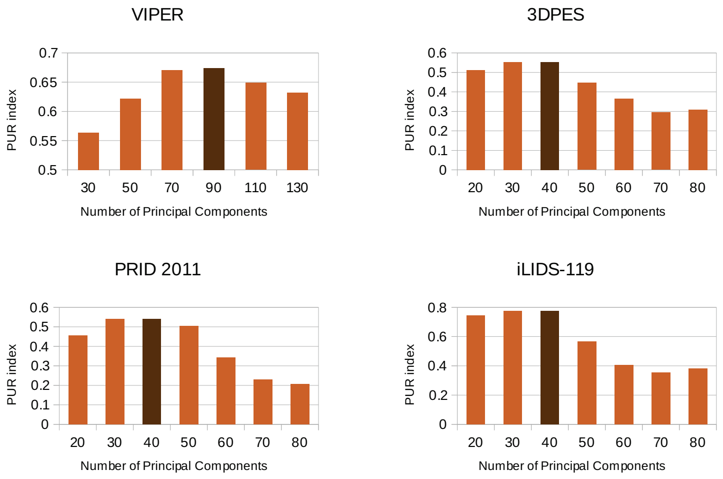


Figure 4: For each data-set: average PUR index, over 30 random splits, computed on the cross-validation set across a different number of considered PCs. The dimension ℓ giving the best PUR indices is highlighted.

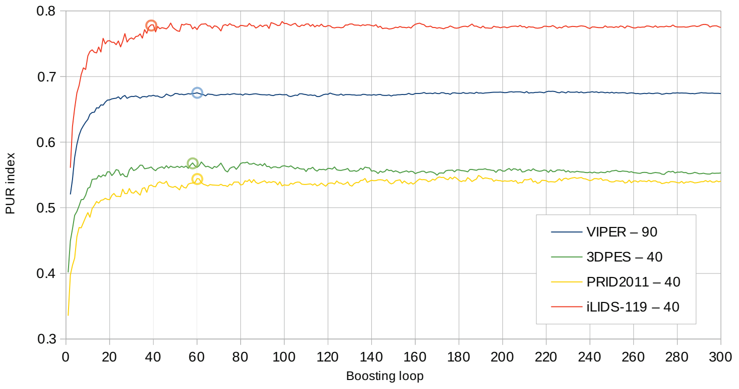


Figure 5: Average PUR index, computed over 30 cross-validation tests, across boosting loops for descriptors reduced to the best dimension (reported after the data-set name). The selected $N_{\mathtt{loop}}$ are highlighted with a small circle.

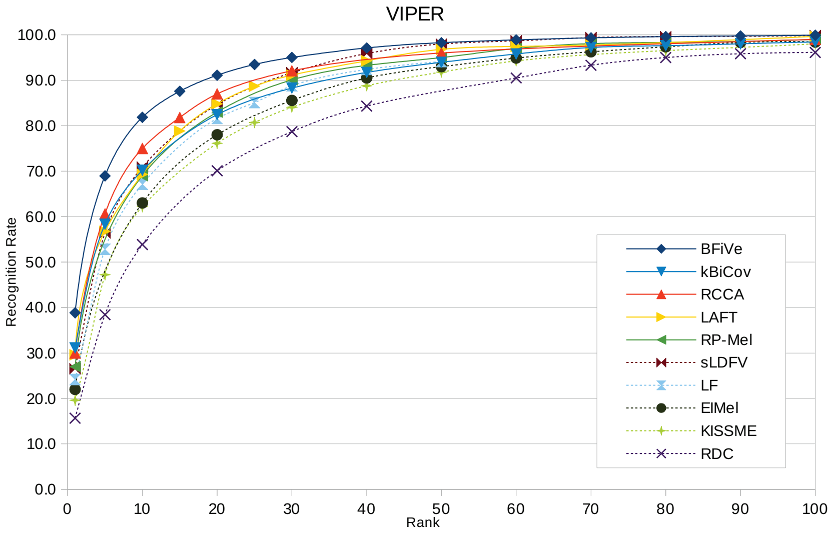


Figure 6: ViPER - CMC curves to compare *BFiVe* with several other methods.

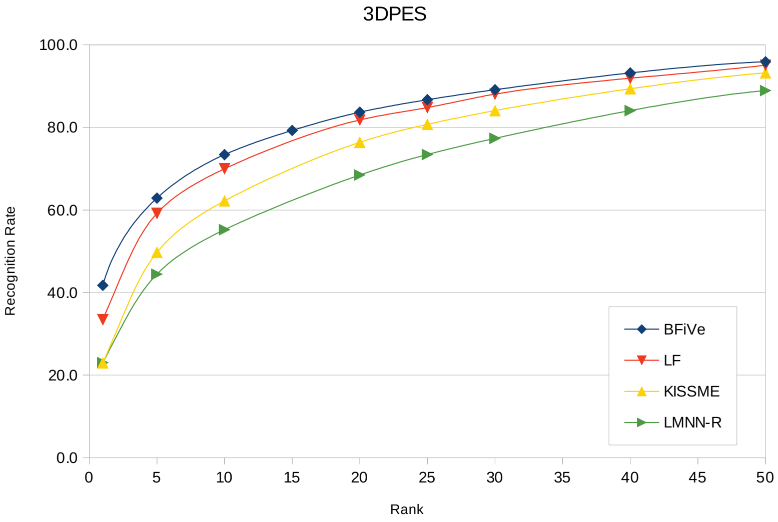


Figure 7: 3DPeS - CMC curves to compare *BFiVe* with other state-of-the-art methods.

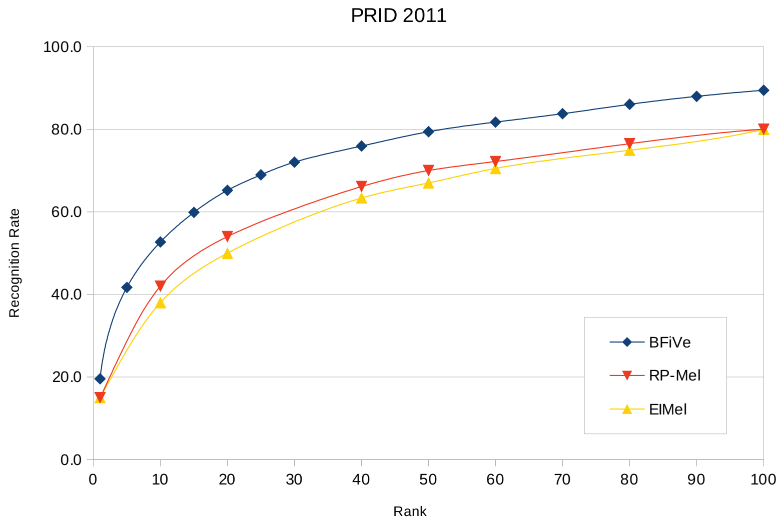


Figure 8: CMC curves on PRID-2011 to compare *BFiVe* performance with two methods on the same data-set.

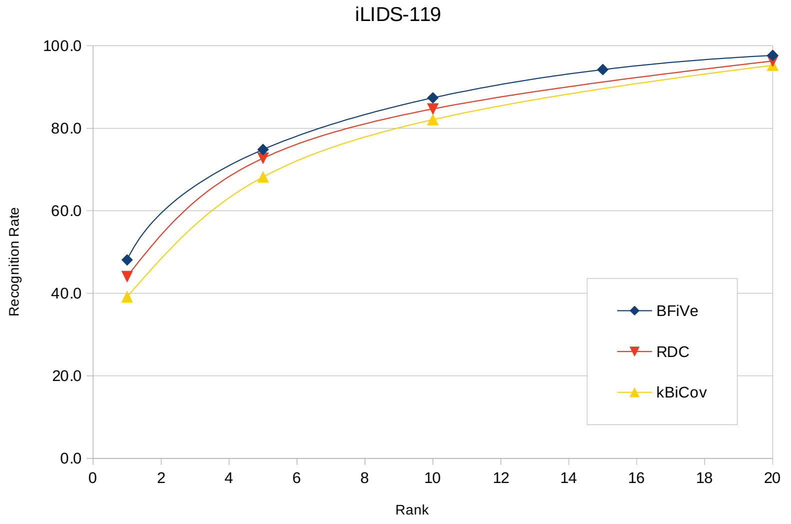


Figure 9: CMC curves on i-LIDS-119 data-set to compare *BFiVe* performance with two other methods.

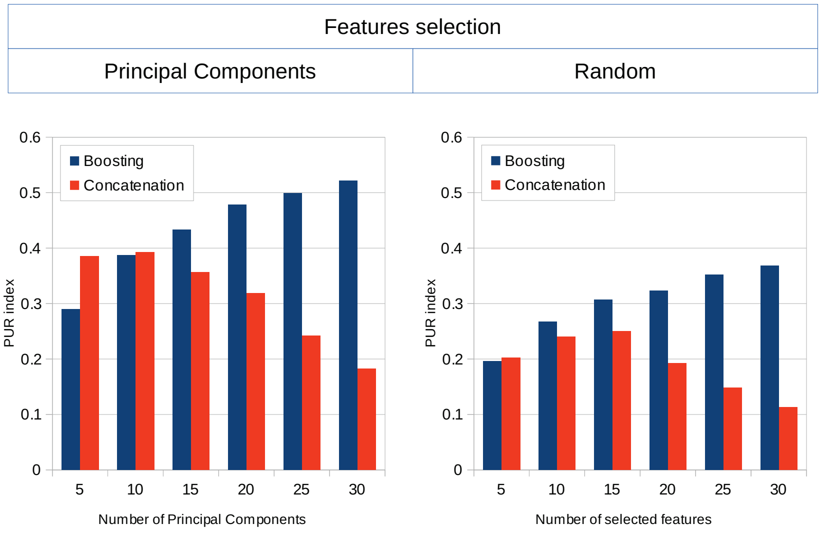


Figure 10: Ranking result comparisons with different combination of local descriptors (Boosting vs Concatenation) and different modes of dimensionality reduction of local descriptors (PCA vs Random).

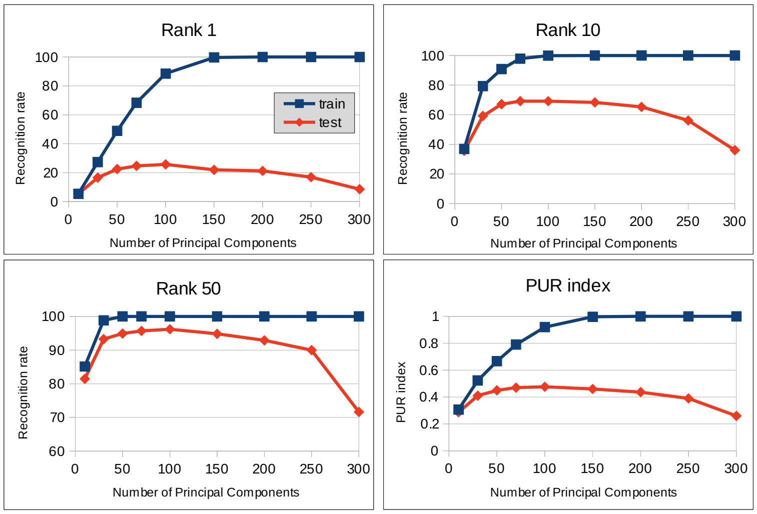


Figure 11: The plots show the recognition performance of one local descriptor, on training and on test sets, versus the number of retained principal components in the descriptor dimensionality reduction. Experiments were conducted on the VIPeR data-set using the receptive field at level (1 × 1).

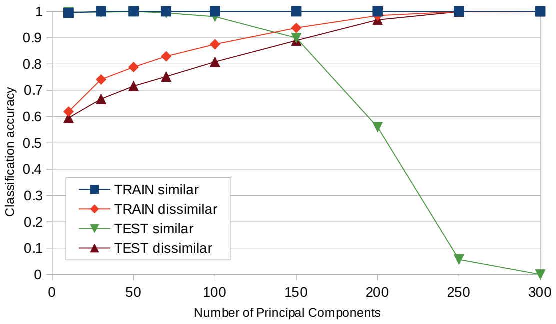


Figure 12: The figure shows the classification performance of the scoring function associated to a single receptive field used as a similar/dissimilar classifier. It shows the accuracy on training and test sets versus the number of retained principal components. Experiments were conducted on the VIPeR data-set using the receptive field at level (1×1).

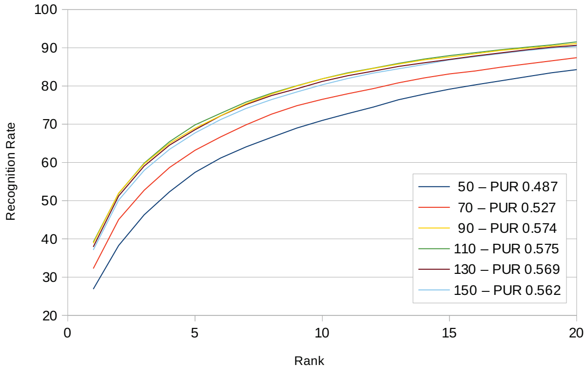


Figure 13: Average CMC curves on the VIPeR test sets. The curves and the associated PUR indices show the *BFiVe* performance at different PCA dimensionality reductions, varying in the range 50–150.

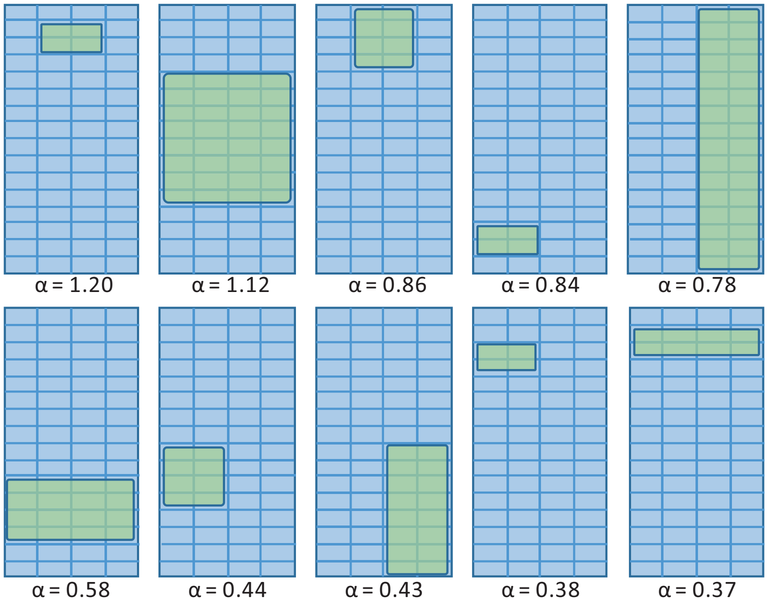


Figure 14: Receptive fields associated to the ten most relevant scoring functions, from a total of 31, generated by the training procedure on a random split of VIPeR. The corresponding α coefficients are also reported.

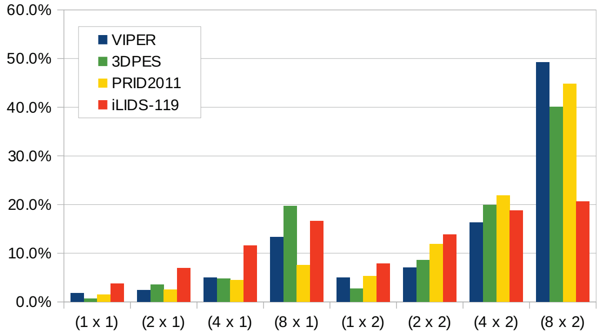


Figure 15: Percentage usage of receptive fields, grouped per level ($i \times j$), for four data-sets (averaged over 30 random splits). Smaller sized regions are preferred over the largest one (1×1).

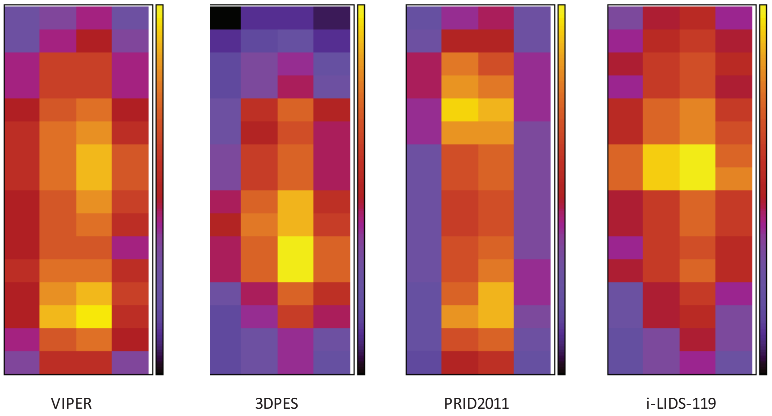


Figure 16: Visualization of the relative contribution of the different image regions to the final scoring function (averaged over 30 random splits) for four data-sets (from left to right: VIPeR, 3DPeS, PRID 2011 and i-LIDS-119). Heatmap representation, best viewed in color.

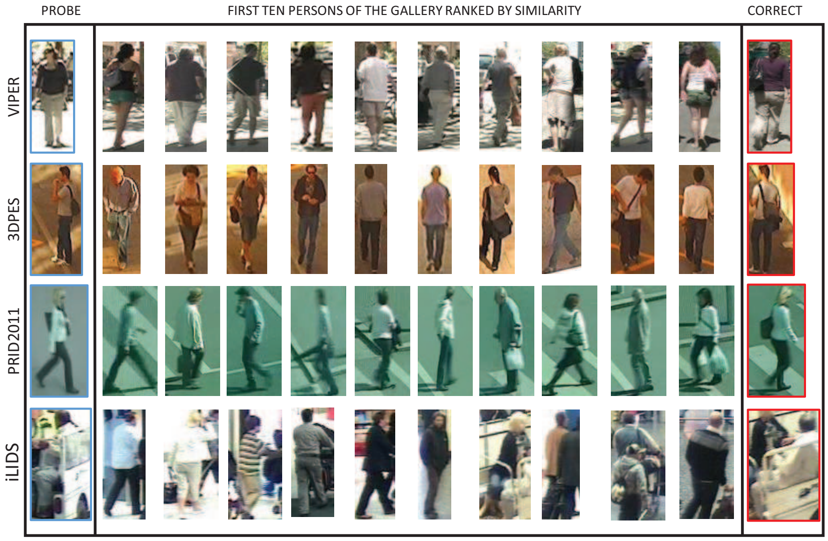


Figure 17: Examples of images in which the algorithm ranks the correct person from the gallery outside the first ten positions. For each considered data-set, the figure reports a probe image, on the left, and the first ten ranked people of the gallery. The last column shows the correct correspondence, which is positioned at a higher rank (11th position for VIPeR, 15th for 3DPeS, 16th for PRID 2011, 19th for i-LIDS-119).

Highlights

We propose BFiVe, a new supervised algorithm for single-shot person re-identification.

The descriptors are a set of compressed local Fisher vectors extracted from a coarse to fine image subdivision.

In the training step each region gives rise to a learnt weak ranking function.

The ranking function of the image gallery is obtained by a boosted selection of a weak learner subset.

The matching rate at rank 1 on VIPeR is 38.9%, on 3DPes 41.7%, on PRID-2011 19.6%, on i-LIDS-119 48.1%.