Learning to Match Appearances by Correlations in a Covariance Metric Space

Slawomir Bak, Guillaume Charpiat, Etienne Corvee, Francois Bremond, Monique Thonnat

Learning to Match Appearances by Correlations in a Covariance Metric Space

Sławomir Bąk, Guillaume Charpiat, Etienne Corvée, François Brémond, Monique Thonnat

INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex - France
firstname.surname@inria.fr

Abstract. This paper addresses the problem of appearance matching across disjoint camera views. Significant appearance changes, caused by variations in view angle, illumination and object pose, make the problem challenging. We propose to formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects. Learning is performed in a covariance metric space using an entropy-driven criterion. Our main idea is that different regions of the object appearance ought to be matched using various strategies to obtain a distinctive representation. The proposed technique has been successfully applied to the person re-identification problem, in which a human appearance has to be matched across non-overlapping cameras. We demonstrate that our approach improves state of the art performance in the context of pedestrian recognition.

Keywords: covariance matrix, re-identification, appearance matching

1 Introduction

Appearance matching of the same object registered in disjoint camera views is one of the most challenging issues in every video surveillance system. The present work addresses the problem of appearance matching, in which non-rigid objects change their pose and orientation. The changes in object appearance together with inter-camera variations in lighting conditions, different color responses, different camera viewpoints and different camera parameters make the appearance matching task extremely difficult.

Recent studies tackle this topic in the context of pedestrian recognition. Human recognition is of primary importance not only for people behavior analysis in large area networks but also in security applications for people tracking. Determining whether a given person of interest has already been observed over a network of cameras is referred to as *person re-identification*. Our approach is focused on, but not limited to, this application.

Appearance-based recognition is of particular interest in video surveillance, where biometrics such as iris, face or gait might not be available, *e.g.* due to sensors' scarce resolution or low frame-rate. Thus, appearance-based techniques rely

on clothing assuming that individuals wear the same clothes between different sightings.

In this paper, we propose a novel method, which *learns* how to match appearance of a specific object class (*e.g.* class of humans). Our main idea is that different regions of the object appearance ought to be matched using different strategies to obtain a distinctive representation. Extracting region-dependent features allows us to characterize the appearance of a given object class in a more efficient and informative way. Different kinds of features characterizing various regions of an object is fundamental to our appearance matching method.

We propose to model the object appearance using covariance descriptor [1] yielding rotation and illumination invariance. Covariance descriptor has already been successfully used in the literature for appearance matching [2, 3]. In contrast to these approaches, we do not define *a priori* feature vector for extracting covariance, but we learn which features are the most descriptive and distinctive depending on their localization in the object appearance. Characterizing a specific class of objects (*e.g.* humans), we select only essential features for this class, removing irrelevant redundancy from covariance feature vectors and ensuring low computational cost. The output model is used for extracting the appearance of an object from a set of images, while producing a descriptive and distinctive representation.

In this paper, we make the following contributions:

- We formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects (Section 3).
- By using a combination of *small covariance matrices* (4×4) between few relevant features, we offer an efficient and distinctive representation of the object appearance (Section 3.1).
- We propose to learn a general model for appearance matching in a *covariance metric space* using *correlation-based feature selection* (CFS) technique (Section 3.2).

We evaluate our approach in Section 4 before discussing future work and concluding.

2 Related Work

Recent studies have focused on the appearance matching problem in the context of pedestrian recognition. Person re-identification approaches concentrate either on *distance learning* regardless of the representation choice [4, 5], or on feature modeling, while producing a distinctive and invariant representation for appearance matching [6, 7]. Learning approaches use training data to search for strategies that combine given features maximizing inter-class variation whilst minimizing intra-class variation. Instead, feature-oriented approaches concentrate on an invariant representation, which should handle view point and camera changes. Further classification of appearance-based techniques distinguishes the

single-shot and the *multiple-shot* approaches. The former class extracts appearance using a single image [8, 9], while the latter employs multiple images of the same object to obtain a robust representation [6, 7, 10].

Single-shot **approaches:** In [9], a model for shape and appearance of the object is presented. A pedestrian image is segmented into regions and their color spatial information is registered into a co-occurrence matrix. This method works well if the system considers only a frontal viewpoint. For more challenging cases, where viewpoint invariance is necessary, an ensemble of localized features (ELF) [11] is selected by a boosting scheme. Instead of designing a specific feature for characterizing people appearance, a machine learning algorithm constructs a model that provides maximum discriminability by filtering a set of simple features. In [12], pairwise dissimilarity profiles between individuals are learned and adapted for nearest neighbor classification. Similarly, in [13], a rich set of feature descriptors based on color, textures and edges is used to reduce the amount of ambiguity among human class. The high-dimensional signature is transformed into a low-dimensional discriminant latent space using a statistical tool called Partial Least Squares (PLS) in a *one-against-all* scheme. However, both methods demand an extensive learning phase based on the pedestrians to re-identify, extracting discriminative profiles, which makes the approaches non-scalable. The person re-identification problem is reformulated as a ranking problem in [14]. The authors present extensive evaluation of learning approaches and show that a ranking-based model can improve the reliability and accuracy. Distance learning is also the main topic of [5]. A probabilistic model maximizes the probability of true match pair having a smaller distance than that of a wrong match pair. This approach focuses on maximizing matching accuracy regardless of the representation choice.

Multiple-shot **approaches:** In [15], every individual is represented by two models: descriptive and discriminative. The discriminative model is learned using the descriptive model as an assistance. In [10], a spatiotemporal graph is generated for ten consecutive frames to group spatiotemporally similar regions. Then, a clustering method is applied to capture the local descriptions over time and to improve matching accuracy. In [7], the authors propose the feature-oriented approach which combines three features: (1) chromatic content (HSV histogram); (2) maximally stable color regions (MSCR) and (3) recurrent highly structured patches (RHSP). The extracted features are weighted using the idea that features closer to the bodies' axes of symmetry are more robust against scene clutter. Recurrent patches are presented in [6]. Using epitome analysis, highly informative patches are extracted from a set of images. In [16], the authors show that features are not as important as precise body parts detection, looking for part-to-part correspondences.

Learning approaches concentrate on distance metrics regardless of the representation choice. In the end, those approaches use very simple features such as color histograms to perform recognition. Instead of learning, feature-oriented approaches concentrate on the representation without taking into account discriminative analysis. In fact, learning using a sophisticated feature representa-

tion is very hard or even unattainable. In [15], the authors deteriorate covariance feature to apply learning. As covariances do not live on Euclidean space, it is difficult to perform learning on an unknown manifold without a suitable metric.

This work overcomes learning issues by considering a *covariance metric space* using an entropy-driven technique. We combine advantages of a strong descriptor with the efficient selection method, thus producing a robust representation for appearance matching.

3 The approach

Our appearance matching requires two operations. The fist stage overcomes color dissimilarities caused by variations in lighting conditions. We apply the *histogram equalization* [17] technique to the color channels (RGB) to maximize the entropy in each of these channels and to obtain camera-independent color representation. The second step is responsible for appearance extraction (Section 3.3) using a general model learned for a specific class of objects (*e.g.* humans). The following sections describe our feature space and the learning, by which the appearance model for matching is generated.

3.1 Feature Space

Our object appearance is characterized using the *covariance descriptor* [1]. This descriptor encodes information on feature variances inside an image region, their correlations with each other and their spatial layout. The performance of the covariance features is found to be superior to other methods, as rotation and illumination changes are absorbed by the covariance matrix.

In contrast to [1, 2, 15], we do not limit our covariance descriptor to a single feature vector. Instead of defining *a priori* feature vector, we use a machine learning technique to select features that provide the most descriptive representation of the appearance of an object.

Let $L = \{R, G, B, I, \nabla_I, \theta_I, \dots\}$ be a set of feature layers, in which each layer is a mapping such as color, intensity, gradients and filter responses (texture filters, *i.e.* Gabor, Laplacian or Gaussian). Instead of using covariance between all of these layers, which would be computationally expensive, we compute covariance matrices of a few relevant feature layers. These relevant layers are selected depending on the region of an object (see Section 3.2). In addition, let layer $\mathfrak{D}$ be a distance between the center of an object and the current location. Covariance of distance layer $\mathfrak{D}$ and three other layers l ($l \in L$) form our descriptor, which is represented by a 4×4 covariance matrix. By using distance $\mathfrak{D}$ in every covariance, we keep a spatial layout of feature variances, which is rotation invariant. State of the art techniques very often use pixel location (x, y) instead of distance $\mathfrak{D}$, yielding better description of an image region. Conversely, among our detail experimentation, using $\mathfrak{D}$ rather than (x, y), we did not decrease the recognition accuracy in the general case, while decreasing the number of features in the covariance matrix. This discrepancy may be due to the fact that we hold spatial

Fig. 1. A meta covariance feature space. Example of three different covariance features. Every covariance is extracted from a region (P), *distance* layer ($\mathfrak{D}$) and three channel functions (*e.g.* bottom covariance feature is extracted from region P_3 using layers: $\mathfrak{D}$, I-intensity, ∇_I-gradient magnitude and θ_I-gradient orientation).

information twofold, (1) by location of a rectangular sub-region from which the covariance is extracted and (2) by $\mathfrak{D}$ in covariance matrix. We constraint our covariances to combination of 4 features, ensuring computational efficiency. Also, bigger covariance matrices tend to include superfluous features which can clutter the appearance matching. 4×4 matrices provide sufficiently descriptive correlations while keeping low computational time needed for calculating generalized eigenvalues during distance computation.

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

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

3.2 Learning in a Covariance Metric Space

Let $\mathbf{a}_i^c = \{\mathbf{a}_{i,1}^c, \mathbf{a}_{i,2}^c, \ldots \mathbf{a}_{i,m}^c\}$ be a set of relevant observations of an object i in camera c, where $\mathbf{a}_{ij}^c$ is a n-dimensional vector composed of all possible covariance

features extracted from image j of object i in the n-dimensional meta covariance feature space $\mathfrak{C}$. We define the distance vector between two samples $\mathbf{a}_{i,j}^{c}$ and $\mathbf{a}_{k,l}^{c'}$ as follows

$$\delta(\mathbf{a}_{i,j}^{c}, \mathbf{a}_{k,l}^{c'}) = \left[\rho(\mathbf{a}_{i,j}^{c}[z], \mathbf{a}_{k,l}^{c'}[z])\right]_{z\in\mathbb{Z}}^{T}, \tag{1}$$

where ρ is the geodesic distance between covariance matrices [18], and $\mathbf{a}_{i,j}^{c}[z]$, $\mathbf{a}_{k,l}^{c'}[z]$ are the corresponding covariance matrices (the same region P and the same combination of layers). The index z is an iterator of $\mathfrak{C}$.

We cast the appearance matching problem into the following *distance learning* problem. Let δ^+ be distance vectors computed using pairs of relevant samples (of the same people captured in different cameras, $i = k$, $c \neq c'$) and let δ^- be distance vectors computed between pairs of related irrelevant samples ($i \neq k$, $c \neq c'$). Pairwise elements δ^+ and δ^- are distance vectors, which stand for positive and negative samples, respectively. These distance vectors define a *covariance metric space*. Given δ^+ and δ^- as training data, our task is to find a general model of appearance to maximize matching accuracy by selecting relevant covariances and thus defining a distance.

Riemannian geometry: Covariance descriptors as positive definite symmetric matrices lie on a manifold that is not a vector space (they do not lie on Euclidean space). Specifying the covariance manifold as Riemannian we can apply differential geometry [19] to perform usual operations such as mean or distance. However, learning on a manifold space is a difficult and unsolved challenge. Methods [2, 20] perform classification by regression over the mappings from the training data to a suitable tangent plane. Defining tangent plane over the Karcher mean of the positive training data points, we can preserve a local structure of the points. Unfortunately, the models extracted using means of the positive training data points tend to overfit. These models concentrate on tangent planes obtained from training data and do not have generalization properties. In [15], authors avoid this problem by casting covariance matrices into *Sigma Points* that lie on an approximate covariance space.

Neither using tangent planes over the Karcher means extracted from training data, nor casting covariances into *Sigma Points* satisfies our matching model. As we want to take full advantage of the covariance manifold space, we propose to extract a general model for appearance matching by identifying the most salient features. Based on the hypothesis: *"A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other"* [21], we build our appearance model using covariance features f_z ($f_z \in \mathfrak{C}$) chosen by a *correlation-based feature selection* technique.

Correlation-based Feature Selection (CFS): *Correlation-based feature selection* (CFS) [21] is a filter algorithm that ranks feature subsets according to a correlation-based evaluation function. This evaluation function favors feature subsets which contain features highly correlated with the class and uncorrelated with each other. In our *distance learning* problem, we define positive and negative class by δ^+ and δ^-, as relevant and irrelevant pairs of samples (see Section 3.2). Further, feature $f_z \in \mathfrak{C}$ is characterized by a distribution of the

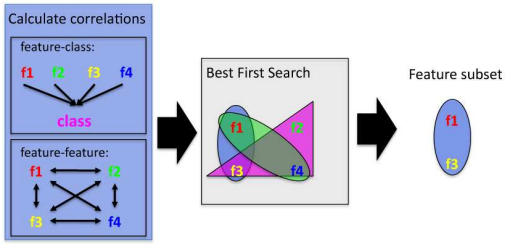


Fig. 2. Correlation-based feature selection.

zth elements in distance vectors δ^+ and δ^-. The feature-class correlation and the feature-feature inter-correlation is measured using a symmetrical uncertainty model [21]. As this model requires nominal valued features, we discretize f_z using the method of Fayyad and Irani [22]. Let X be a nominal valued feature obtained by discretization of f_z.

We assume that a probabilistic model of X can be formed by estimating the probabilities of the values $x \in X$ from the training data. The uncertainty can be measured by entropy $H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$. A relationship between features X and Y can be given by $H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y)$. The amount by which the entropy of X decreases reflects additional information on X provided by Y and is called the *information gain* (*mutual information*) defined as $Gain = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.

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

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

$$\mathfrak{M}(\mathfrak{S}) = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k\,(k+1)\,\overline{r_{ff}}}}, \tag{2}$$

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

$$\overline{r_{cf}} = \frac{1}{k} \sum_{f_z \in \mathfrak{S}} r_{cf_z}, \qquad \overline{r_{ff}} = \frac{2}{k\,(k-1)} \sum_{\substack{f_i, f_j \in \mathfrak{S} \\ i<j}} r_{f_i f_j}, \tag{3}$$

where c is the class, or relevance feature, which is $+1$ on δ^+ and -1 on δ^-. The numerator in Eq. 2 indicates predictive ability of subset $\mathfrak{S}$ and the denominator stands for redundancy among the features.

Equation 2 is the core of CFS, which ranks feature subsets in the search space of all possible feature subsets. Since exhaustive enumeration of all possible feature subsets is prohibitive in most cases, a heuristic search strategy has to be applied. We have investigated different search strategies, among which *best first search* [23] performs the best.

Best first search is an AI search strategy that allows backtracking along the search path. Our *best first* starts with no feature and progresses forward through

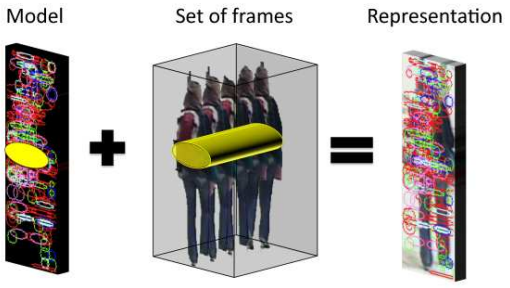


Fig. 3. Extraction of the appearance using the model (the set of features selected by CFS). Different colors and shapes in the model refer to different kinds of covariance features.

the search space adding single features. The search terminates if T consecutive subsets show no improvement over the current best subset (we set $T = 5$ in experiments). By using this stopping criterion we prevent the best first search from exploring the entire feature subset search space. Fig. 2 illustrates CFS method. Let Π be the output of CFS that is the feature subset of $\mathfrak{C}$. This feature subset Π forms a model that is used for appearance extraction and matching.

3.3 Appearance Extraction

Having the general model Π for a specific object class (*e.g.* humans), we compute the appearance representation using a set of frames (see Fig. 3). In the context of person re-identification, our method belongs to the group of *multiple-shot* approaches, where multiple images of the same person are used to extract a compact representation. This representation can be seen as a *signature* of the multiple instances.

Using our model Π, a more straightforward way to extract appearance would be to compute covariance matrices of a video volume directly for every $f_z \in \Pi$. However, using volume covariance we loose information on real feature distribution (time feature characteristics are merged). Thus, similarly to [2] we compute Karcher means using a Riemannian manifold. The mean covariance matrix as an intrinsic average blends appearance information from multiple images. For every feature $f_z \in \Pi$ we compute the mean covariance matrix. The set of mean covariance matrices stands for an appearance representation of an object (*signature*).

3.4 Appearance Matching

Let $\mathfrak{A}$ and $\mathfrak{B}$ be the object signatures. The signature consists of mean covariance matrices extracted using set Π. The similarity between two signatures $\mathfrak{A}$ and $\mathfrak{B}$ is defined as

$$S(\mathfrak{A}, \mathfrak{B}) = \frac{1}{|\Pi|} \sum_{i \in \Pi} \frac{1}{\max(\rho(\mu_{\mathfrak{A},i}, \mu_{\mathfrak{B},i}), \epsilon)}, \tag{4}$$

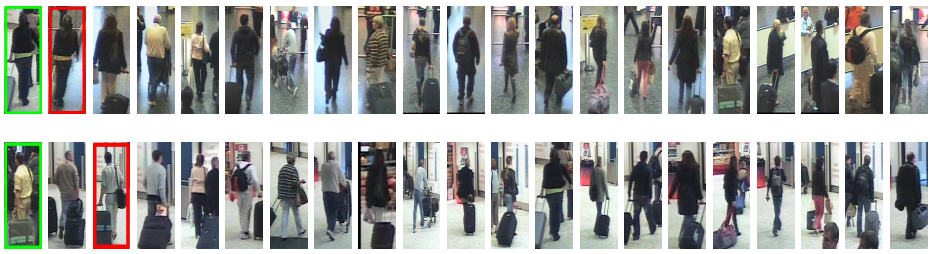


Fig. 4. Example of the person re-identification on i-LIDS-MA. The left-most image is the probe image. The remaining images are the top 20 matched gallery images. The red boxes highlight the correct matches.

where ρ is a geodesic distances [18], $\mu_{\mathfrak{A},i}$ and $\mu_{\mathfrak{B},i}$ are mean covariance matrices extracted using covariance feature $i \in \mathit{\Pi}$ and $\epsilon = 0.1$ is introduced to avoid the denominator approaching to zero. Using the average of similarities computed on feature set $\mathit{\Pi}$ the appearance matching becomes robust to noise.

4 Experimental Results

We carry out experiments on 3 i-LIDS datasets[1]: i-LIDS-MA [2], i-LIDS-AA [2] and i-LIDS-119 [24]. These datasets have been extracted from the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset with multiple non-overlapping camera views. These datasets tackle the person re-identification problem, where appearances of the same person acquired by different cameras, has to be matched. The results are analyzed in terms of recognition rate, using the *cumulative matching characteristic* (CMC) [25] curve. The CMC curve represents the expectation of finding the correct match in the top n matches (see Fig. 4).

4.1 Experimental setup

Feature space: We scale every human image into a fixed size window of 64×192 pixels. The set of rectangular sub-regions $\mathbf{P}$ is produced by shifting 32×8 and 16×16 pixel regions with 8 pixels step (up and down). It gives $|\mathbf{P}| = 281$ overlapping rectangular sub-regions. We set $L = \{(l, \nabla_l, \theta_l)_{l=I,R,G,B}, G_{i=1\ldots4}, \mathcal{N}, \mathfrak{L}\}$, where I, R, G, B refer to intensity, red, green and blue channel, respectively; ∇ is the gradient magnitude; θ corresponds to the gradient orientation; G_i are Gabor's filters with parameters $\gamma, \theta, \lambda, \sigma^2$ set to $(0.4, 0, 8, 2)$, $(0.4, \frac{\pi}{2}, 8, 2)$, $(0.8, \frac{\pi}{4}, 8, 2)$ and $(0.8, \frac{3\pi}{2}, 8, 2)$, respectively; $\mathcal{N}$ is a gaussian and $\mathfrak{L}$ is a laplacian filter. A learning process involving all possible combinations of three layers would not be computationally tractable (229296 covariances to consider in section 3.2). Thus instead, we experimented with different subsets of combinations and selected a

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

reasonably efficient one. Among all possible combinations of the three layers, we choose 10 combinations ($C_{i=1...10}$) to ensure inexpensive computation. We set C_i to (R,G,B), $(\nabla_R,\nabla_G,\nabla_B)$, $(\theta_R,\theta_G,\theta_B)$, (I,∇_I,θ_I), (I,G_3,G_4), $(I,G_2,\mathfrak{L})$, $(I,G_2,\mathcal{N})$, $(I,G_1,\mathcal{N})$, $(I,G_1,\mathfrak{L})$, (I,G_1,G_2), respectively, separating color and texture features. Similar idea was already proposed in [11]. Note that we add to every combination C_i layer $\mathfrak{D}$, thus generating our final 4×4 covariance descriptors. The dimension of our meta covariance feature space is $n = |\mathfrak{C}| = 10 \times |\mathbf{P}| = 2810$.

Learning and testing: Let us assume that we have $(p+q)$ individuals seen from two different cameras. For every individual, m images from each camera are given. We take q individuals to learn our model, while p individuals are used to set up the gallery set. We generate positive training examples by comparing m images of the same individual from one camera with m images from the second camera. Thus, we produce $|\delta^+| = q \times m^2$ positive samples. Pairs of images coming from different individuals stand for negative training examples, thus producing $|\delta^-| = q \times (q-1) \times m^2$.

Acronym: We name our approach *COrrelation-based Selection of covariance MATrIces* (COSMATI).

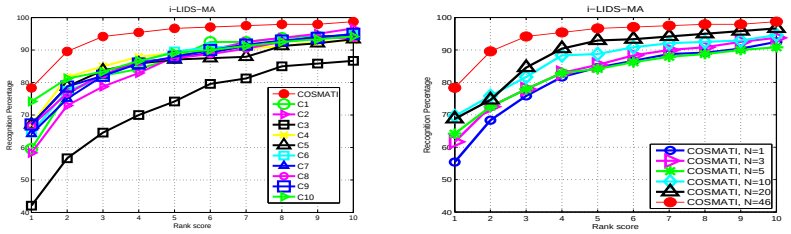
4.2 Results

i-LIDS-MA [2]: This dataset consists of 40 individuals extracted from two non-overlapping camera views. For each individual a set of 46 images is given. The dataset contains in total $40 \times 2 \times 46 = 3680$ images. For each individual we randomly select $m = 10$ images. Then, we randomly select $q = 10$ individuals to learn a model. The evaluation is performed on the remaining $p = 30$ individuals. Every signature is used as a query to the gallery set of signatures from the other camera. This procedure has been repeated 10 times to obtain averaged CMC curves.

COSMATI *vs.* C_i: We first evaluate the improvement in using different combinations of features for the appearance matching. We compare models based on a single combination of features with the model, which employs several combinations of features. From Fig. 5(a) it is apparent that using different kinds of covariance features we improve matching accuracy.

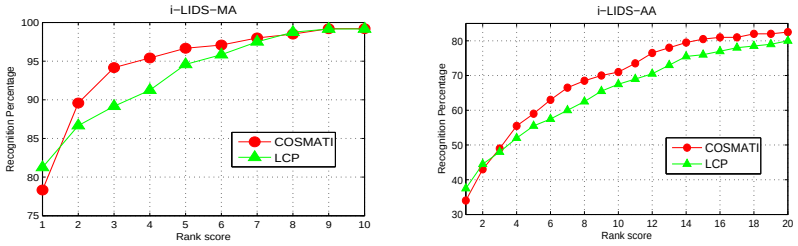
COSMATI w.r.t. the number of shots: We carry out experiments to show the evolution of the performance with the number of given frames per individual (Fig. 5(b)). The results indicate that the larger number of frames, the better performance is achieved. It clearly shows that averaging covariance matrices on a Riemannian manifold using multiple shots leads to a much better recognition accuracy. It is worth noting that $N \sim 50$ is usually affordable in practice as it corresponds to only 2 seconds of a standard 25 framerate camera.

COSMATI *vs.* LCP: We compare our results with LCP [2] method. This method employs *a priori* 11×11 covariance descriptor. Discriminative characteristics of an appearance are enhanced using *one-against-all* boosting scheme. Both methods are evaluated on the same subsets ($p = 30$) of the original data. Our method achieves slightly better results than LCP as shown in Fig. 6(a). It



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

Fig. 5. Performance comparison: (a) with models based on a single combination of features; (b) w.r.t. the number of given frames



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

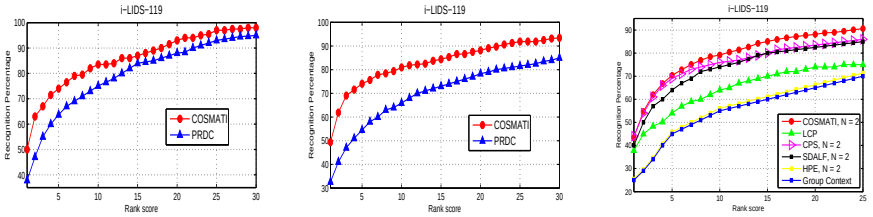
Fig. 6. Performance comparison with LCP [2] using CMC curves.

shows that using specific descriptors for different regions of the object, we are able to obtain equally distinctive representation as LCP, which uses an exhaustive *one-against-all* learning scheme.

i-LIDS-AA [2]: This dataset contains 100 individuals automatically detected and tracked in two cameras. Cropped images are noisy, which makes the dataset more challenging (*e.g.* detected bounding boxes are not accurately centered around the people, only part of the people is detected due to occlusion).

COSMATI *vs.* LCP: Using the models learned on i-LIDS-MA, we evaluate our approach on 100 individuals. Results are illustrated in Fig. 6(b). Our method again performs better than LCP. It is relevant to mention that once our model is learned offline, it does not need any additional discriminative analysis. Specifying informative regions and their characteristics (features) we obtain a distinctive representation of the object appearance. In contrast to [2] our offline learning is scalable and do not require any *reference* dataset.

i-LIDS-119 [24]: i-LIDS-119 does not fit very well for *multiple-shot* case, because the number of images per individual is very low (in average 4). However, this dataset was extensively used in the literature for evaluating the person re-identification approaches. Thus, we also use this dataset to compare with state of the art results. The dataset consist of 119 individuals with 476 images. This dataset is very challenging since there are many occlusions and often only the top part of the person is visible. As only few images are given, we extract our signatures using maximally $N = 2$ images.



(a) p = 50 (b) p = 80 (c) p = 119

Fig. 7. Performance comparison using CMC curves, p is the size of the gallery set (larger p means smaller training set): (a),(b) our *vs.* PRDC [5]; (c) our *vs.* LCP [2], CPS [16], SDALF [7], HPE [6], Group Context [13]

COSMATI *vs.* PRDC [5]: We compare our approach with PRDC method. This method also requires offline learning. PRDC focuses on distance learning that can maximize matching accuracy regardless of the representation choice. We reproduce the same experimental settings as [5]. All images of p randomly selected individuals are used to set up the test set. The remaining images form the training data. Each test set is composed of a gallery set and a probe set. In contrast to [5], we use multiple images (maximally $N = 2$) to create the gallery and the probe set. The procedure is repeated 10 times to obtain reliable statistics. Our results (Fig. 7(a,b)) show clearly that COSMATI outperforms PRDC. The results indicate that using strong descriptors, we can significantly increase matching accuracy.

COSMATI *vs.* LCP [2], CPS [16], SDALF [7], HPE [6] and Group Context [13]: We have used models learned on i-LIDS-MA to evaluate our approach on the full dataset of 119 individuals. Our CMC curve is generated by averaged CMC over 10 trials. Results are presented in Fig. 7 (c). Our approach performs the best among all considered methods. We believe that it is due to the informative appearance representation obtained by CFS technique. It clearly shows that a combination of the strong covariance descriptor with the efficient selection method produces distinctive models for the appearance matching problem.

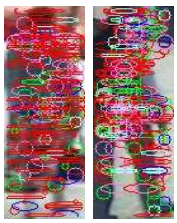
Model Analysis: The strength of our approach is the combination of many different kinds of features into one similarity function. Table 1 shows the percentage of different kinds of covariance features embedded in all our models, which were used during the evaluation. Unlike [11], our method concentrates more on texture filters than on color features. It appears that the higher the resolution of images, the more frequent the usage of texture filters is. Figure 8 illustrates that models extracted on higher resolution employ more texture features.

Using multiple images, a straightforward way to extract an appearance would be to use a *background subtraction* algorithm to obtain foreground regions. Unfortunately in many real scenarios, consecutive frames are not available due to gaps in tracking results. Having a still image, we could apply the extended graph-cut approach: GrabCut [26] to obtain a human silhouette. This approach can be driven by cues coming from detection result (a rectangle around the desired object). Surprisingly, employing GrabCut in our framework did not increase

matching accuracy. The main reason for this is that GrabCut often mis-segments significant part of foreground region. Further, by learning a model, our approach already focuses on features corresponding to foreground.

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
9.61 %	5.30%	4.62%	4.37%	5.47%	12.70 %	14.29%	17.35%	12.12%	14.16 %

Table 1. A table showing the percent of different covariance features embedded in models



(a) (b)

Fig. 8. Extracted models for (a) lower and (b) higher resolutions: red indicates color features which turn out to be more prominent in (a).

Computation complexity: In our experiments, for $q = 10$ and $m = 10$ we generate $|\delta^+| = 1000$ and $|\delta^-| = 9000$ training samples. Learning on 10.000 samples takes around 20 minutes on Intel quad-core 2.4GHz. The model is composed of 150 covariance features in average. The calculation of generalized eigenvalues of 4×4 covariance matrices (distance computation) takes $\sim 2\mu s$ without applying any hardware-dependent optimization routines (*e.g.* LAPACK library can perform faster using *block operations* optimized for architecture).

5 Conclusion

We proposed to formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects. Our strategy is based on the idea that different regions of the object appearance ought to be matched using various strategies. Our experiments demonstrate that: (1) by using different kinds of covariance features w.r.t. the region of an object, we obtain clear improvement in appearance matching performance; (2) our method outperforms state of the art methods in the context of pedestrian recognition. In the future, we plan to integrate the notion of motion in our recognition framework. This would allow to distinguish individuals using their behavioral characteristics and to extract only the features which surely belong to foreground region.

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

References

1. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: ECCV. (2006) 589–600
2. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Boosted human re-identification using riemannian manifolds. Image and Vision Computing (2011)

3. Oncel, F.P., Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based on lie algebra. In: CVPR. (2006)
4. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: ACCV. (2010) 501–512
5. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR. (2011)
6. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: ICPR. (2010) 1413–1416
7. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. (2010)
8. Park, U., Jain, A., Kitahara, I., Kogure, K., Hagita, N.: Vise: Visual search engine using multiple networked cameras. In: ICPR. (2006) 1204–1207
9. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: ICCV. (2007) 1–8
10. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR. (2006) 1528–1535
11. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV. (2008) 262–275
12. Lin, Z., Davis, L.S.: Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: ISVC. (2008) 23–34
13. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: SIBGRAPI. (2009) 322–329
14. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC. (2010) 21.1–21.11
15. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: SCIA. (2011) 91–102
16. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC. (2011) 68.1–68.11
17. Hordley, S.D., Finlayson, G.D., Schaefer, G., Tian, G.Y.: Illuminant and device invariant colour using histogram equalisation. Pattern Recognition **38** (2005)
18. Förstner, W., Moonen, B.: A metric for covariance matrices. In: Quo vadis geodesia ...?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday, TR Dept. of Geodesy and Geoinformatics, Stuttgart University. (1999)
19. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. Int. J. Comput. Vision **66** (2006) 41–66
20. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. **30** (2008) 1713–1727
21. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, Department of Computer Science, University of Waikato (1999)
22. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: IJCAI. (1993) 1022–1027
23. Rich, E., Knight, K.: Artificial Intelligence. McGraw-Hill Higher Education (1991)
24. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC. (2009)
25. Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. PETS (2007)
26. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. In: SIGGRAPH. (2004) 309–314