Boosted Human Re-identification using Riemannian Manifolds

Sławomir Bąk, Etienne Corvée, Francois Brémond, Monique Thonnat

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

Abstract

This paper presents an appearance-based model to address the human re-identification problem. Human re-identification is an important and still unsolved task in computer vision. In many systems there is a requirement to identify individuals or determine whether a given individual has already appeared somewhere in a network of cameras. The human appearance obtained in one camera is usually different from the ones obtained in another camera. In order to re-identify people a human signature should handle difference in illumination, pose and camera parameters. The paper focuses on a new appearance model based on Mean Riemannian Covariance (MRC) patches extracted from tracks of a particular individual. A new similarity measure using Riemannian manifold theory is also proposed to distinguish sets of patches belonging to a specific individual. We investigate the significance of the MRC patches based on their reliability extracted during tracking and their discriminative power obtained by a boosting scheme. The methods are evaluated and compared with the state of the art using benchmark video sequences from the ETHZ and the i-LIDS datasets. The re-identification performance is presented using the cumulative matching characteristic (CMC) curve. We demonstrate that the proposed approach outperforms state of the art methods. Finally, the results of our approach are shown on two other more pertinent datasets.

Keywords: Re-identification, Covariance Matrix, Riemannian Manifold, Human Detection, Appearance Matching, Boosting

1. Introduction

Recently, cameras spread out across various domains that range from personal computers, video games, home surveillance applications, to large camera networks which facility access to sports venue, monitored environments (*e.g* airports, metro stations, car parks *etc.*). A natural consequence of such situation is a need for an automated extraction of high-level semantic information from extremely large volumes of recorded video data. In many surveillance systems, detection and tracking of moving objects constitute the main problem. The number of targets and occlusions produce ambiguity which introduces a requirement for reacquiring objects which have been lost during tracking. However, the ultimate goal of any surveillance system is not to track and reacquire targets, but to understand a scene and to determine whether a given person of interest has already been observed over a network of cameras. It means that an essential idea of tracking an object is a knowledge about its *identity*.

Human re-identification enables recognizing an individual in different disjoint camera views. Beside detection and tracking of an individual in one camera, human re-identification is necessary to associate different appearances between different cameras. In small video surveillance systems, this problem can be approached using temporal and visual information. One of the earliest works on associating appearances from disjoint cameras is [1]. A Bayesian framework has been proposed to fuse cues including inter-camera time intervals, location of exit/entrances, velocities of objects and a simple color histogram representation to perform re-identification. Nevertheless, in order to recognize people in large systems (*e.g.* metro stations, airports) where temporal transition time between cameras may differ greatly, matching of individuals has to rely on visual features alone.

The human re-identification problem can be defined as a determination whether a given person of interest has already been observed over a network of cameras. This issue is also known as *the person re-identification* problem. Person re-identification can be considered on different levels depending on information cues which are currently available in the system. Biometrics such as iris [2], face [3], gait [4] or their combinations [5] can be used to recognize identities. Nevertheless, in most video surveillance scenarios such detailed information is not available due to video low-resolution or difficult segmentation (crowded environments). Therefore a robust modelling of a global appearance of an individual is necessary to re-identify a given person of interest. In these identification techniques clothing is the most reliable information about an identity of an individual (there is an assumption that individuals wear the same clothes between different sightings). This approaches are referred to as *appearance-based person re-identification* which is the main topic of this paper.

An appearance-based approaches relies on modelling a human signature using tracking and detection results. The re-identification techniques which exploit an appearance information using only one image are referred to as *single-shot* approaches and until now they were the most popular methods. Currently researches try to improve identification accuracy by integrating information over many images. The group of methods which employs multiple images of the same person as a training data is called *multiple-shot* approaches.

As the re-identification concerns a large sets of individuals acquired from different cameras, it is necessary to provide a *distinctive* and *invariant* signature. It has to be based on *discriminative* features to allow browsing the most similar signatures over a network of cameras. It can be achieved by signature matching which has to handle differences in illumination, pose and camera parameters. Note that, inter-camera variations in lighting conditions, differences in illuminations, different camera parameters, changes in object orientation and object pose make this task extremely difficult. Besides, occlusions (caused by other people or objects of the scene) and self-occlusions (caused by body parts) make the re-identification problem one of the hardest tasks in the video surveillance domain.

The main topic of this paper is a novel *multiple-shot* approach which builds a specific human signature model to re-identify a given individual. In our approach a human detection algorithm is used to find out people in video sequences (Section 4.1). Then, the detected individual is tracked to gather as many frames as possible. The appearances obtained from the tracking results are used to extract discriminative signature. This paper makes the following contributions:

- We offer a new set of Mean Riemannian Covariance (MRC) patches extracted using Riemannian manifold theory. This mean covariance matrix keeps not only an information about feature distribution but also carries out an essential information about temporal changes of an appearance (Section 5.3).

- We propose two methods to select and to enhance discriminative MRC patches which can represent the human signature. We investigate the significance of these MRC patches in the two following ways (Section 5.4): (1) we take advantage of patch reliability obtained during tracking; (2) a boosting scheme is developed to select the most discriminative patches reducing the amount of ambiguity among the human class.

- We introduce a new similarity measure between signatures which is able to hold discriminative power coming from the relative position of the MRC patches (Section 5.5).
- We extract new sets of individuals from i-LIDS data to investigate more carefully advantages of using tracking results in building the *multiple-shot* signature. These publicly available datasets finally satisfy all requirements of the *multiple-shot* person re-identification (Section 6).

2. Related Work

Recently, the person re-identification problem became one of the most important tasks in video surveillance. There is a natural consequence of an invention of robust human detection algorithms to extend approaches for recognition purposes. Considering the *appearance based* methods, we divide them into two main groups, *single-shot* approaches and *multiple-shot* approaches.

As to *single-shot* approaches, in [6] the clothing color histograms taken over the head, shirt and pants regions together with the approximated height of the person were used as the discriminative feature. Similarly, clothing segmentation together with facial features [7] were employed to recognize individuals. In [8] an image of the pedestrian is segmented into ten equally spaced horizontal strips. The median HSL color of the foreground pixels of each of these ten strips is then used as the set of features defining signature. This method is used for detecting people loitering by re-identification the same person being present for a long time. The human signature represented by color patches along edges extracted from an appearance was proposed in [9]. The similarity of patches together with their geometric constraints are encoded in matching function. In [10] a generative and discriminative models are combined together with online learning strategy to perform re-identification. The human appearance is represented by histograms computed over upper body, lower body and full body segment. The histograms were composed of color features, autocorrelograms and a bag of features based on SIFT [11] descriptor. In [12] a human body parts detector was applied to establish the correspondence between appearances. First, specialized HOG-based detectors [13] identify body parts. Then, covariance descriptor [14] is used to represent appearance of detected body parts. Finally, a spatial pyramid scheme [15] is adopted to improve discriminative matching of two collections of body features. Shape and appearance context model is proposed in [16]. 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, the ensemble of localized features (*ELF*) [17] has been proposed. 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. Enhancement of discriminative power of each individual signature with respect to the others were also the main issue in [18]. Pairwise dissimilarity profiles between individuals have been learned and adapted into nearest neighbour classification. Similarly, in [19], a rich set of feature descriptors based on color, textures and edges have been used to reduce the amount of ambiguity among human class. The high-dimensional signature was transformed into a low-dimensional discriminant latent space using a statistical tool called Partial Least Squares (PLS) in one-against-all scheme. Nevertheless in both methods, an extensive learning phase based on the pedestrians to re-identify is necessary to extract discriminative profiles what makes the approaches non-scalable. The person re-identification problem has been reformulated as a ranking problem in [20]. The authors presented extensive evaluation of learning approaches and show that a ranking relevance based model can improve the reliability and accuracy. Furthermore, a context information can be used to improve recognition accuracy. Visual information coming from surrounding people have been used in [21] to reduce ambiguity in person identification. This method shows that group association between two or more people can give valuable information about identity of an individual. Similarly, in [22] group context information handled by covariance descriptor [14] improves the performance of person re-identification.

Concerning *multiple-shot* approaches, in [23] the spatiotemporal graph was generated for a ten consecutive frames for grouping spatiotemporally similar regions. Then, a clustering method is applied to capture the local descriptions over time and improve matching accuracy. Bag of features based on SIFT [11] descriptor together with online learning were proposed in [24]

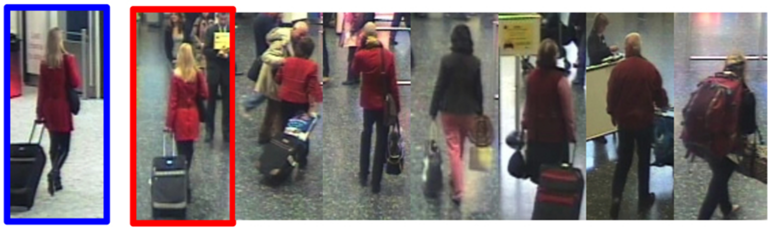


Figure 1: The results of the query. The first image on the left is the query image. The true match is on the first position in the list.

to improve matching accuracy. In [25] re-identification is performed using SURF [26] interest points collected during short video sequences. The interest points are stored in KD-tree to speed-up the query processing time. In [27], the AdaBoost was applied to extract the most discriminative and invariant haar-like features. Here, again one-against-all learning scheme was used to catch human dissimilarities. In [28], the authors proposed to combine three features: 1) chromatic content (HSV histogram); 2) maximal stable colour regions (MSCR [29]) and 3) recurrent highly structured patches (RHSP). The extracted features were weighted by the distance with respect to the vertical axis to minimize effects of pose variations. Recurrent patches were also proposed in [30]. Epitome analysis was used to extract highly informative patches form the set of images. Finally, in [31] re-identification is performed in aerial images. Here, the PageRank algorithm is applied to extract the most informative regions to distinguish individuals. First, the multiple images are integrated using undirected graph and then PageRank assign higher weights to prominent regions and degrade noisy regions by assigning them lower weights.

3. Problem Definition

We lay the problem as the following. We generate human signature for each person detected and tracked in our video surveillance system. Let us denote a signature as $\mathfrak{s}_i^c$, where i encodes the person identity and c denotes the camera. The task is to find for each signature its corresponding signature in another camera. It is realized by querying the database of signatures $\mathfrak{s}_j^{c'}$, where $c \neq c'$ with signature of interest $\mathfrak{s}_i^c$. The results of the query is the list of the most similar signatures ordered by increasing dissimilarity (see Fig. 1). The position in the list of the true match is called the rank score.

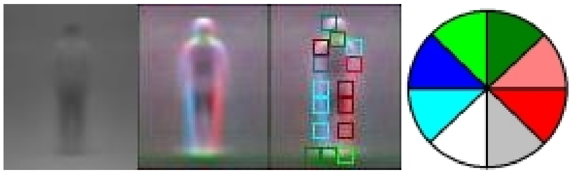


Figure 2: Mean human image with corresponding edge magnitudes and the 15 most dominant cells. From the left: the first image shows the mean human image calculated over all positive samples in the database; the second image shows the corresponding mean edge magnitude response; the third image shows this later image superposed with the 15 most dominant cells of size 8×8 pixels. The cell bounding boxes are drawn with a color set by their most dominant edge orientation with the scheme defined in the last image.

The underlying challenge of the problem arises from significant differences in illumination, pose and camera parameters. The same object acquired under different cameras shows color dissimilarities. Even identical cameras which have the same optical properties and are working under the same lighting conditions may not match in their color responses. Moreover, the pose variations due to camera and view point change as well as the articulation of a human body lead to significant differences in appearances of the same individual observed from different cameras.

4. Overview of the approach

A typical way of extracting appearance models in automatic surveillance systems is by first detection of objects in an image, and then by tracking them using different strategies. In this work, we focus only on human appearances. For human detection, we use HOG-based detector which philosophy is similar to [13]. After detection and tracking, an invariance to differences in ambient illumination is achieved by color normalization.

4.1. Human detection and tracking

We use a *Histogram of Oriented Gradient* (HOG) based technique to automatically detect humans in different scenes before their visual signatures are extracted for re-identification purposes. Our HOG technique is adapted from the face detection method [32] to detect human silhouette. The detection algorithm extracts histograms of gradient orientation, using a Sobel convolution kernel, in a multi-resolution framework. The technique was originally designed to detect faces using the assumption that facial features remain

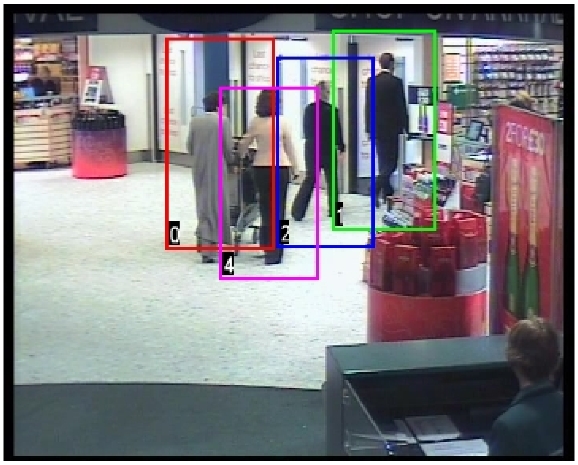


Figure 3: Examples of detected people.

approximatively at the same location. However, location of human silhouette features do not remain constant in template with the varying poses (*e.g.* knees are constantly changing position when walking; a shoulder changes position from walking to slightly bending when pushing a trolley). Hence, we modified algorithm to detect humans using cells located at specific locations around the human silhouette as shown in Figure 2. The most dominant cells used to characterize human shapes are the 15 most dominant cells selected among 252 cells covering the human sample area. These most dominant cells are the cells having the closest HOG vector to the mean HOG vector calculated over the vectors (of the corresponding cell) from a human database. The system was trained using $10,000$ positive and $20,000$ negative image samples from the NICTA database [33]. Figure 3 shows an example of several detected persons in dynamically occluded scenario. Once a human has been detected and tracked, the next step is to handle color dissimilarities caused by camera illumination differences. Thus, we use color normalization technique called histogram equalization (see Section 4.2).

4.2. Color normalization

One of the most challenging problem using the color as a feature is that images of the same object acquired under different cameras show color dissimilarities. Even identical cameras which have the same optical properties and are working under the same lighting conditions may not match in their color responses. Hence, color normalization procedure has been carried out

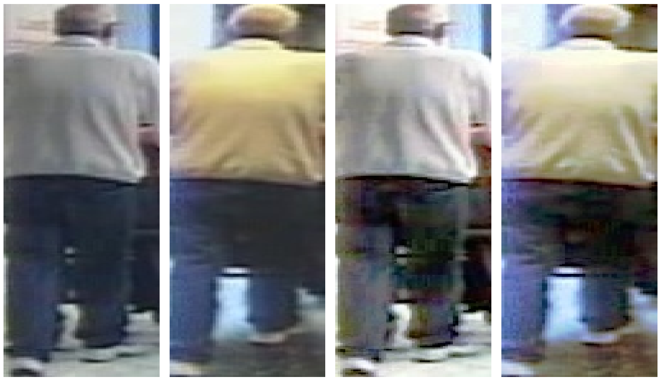


Figure 4: The first two columns show original images of the same person captured from different cameras in different environments. The last two columns show these images after histogram equalization.

in order to obtain invariant signature. We use a technique called histogram equalization [34]. The aim was to increase the overall contrast in the image by brighting dark areas of an image, increasing the detail in those regions. Histogram equalisation achieves this aim by stretching range of histogram to be as close as possible to an uniform histogram. The approach is based on the idea that amongst all possible histograms, an uniformly distributed histogram has maximum entropy [35]. Maximizing the entropy of a distribution we maximize its information and thus histogram equalization maximizes the information content of the output image. We apply the histogram equalization to each of the color channels (RGB) to maximize the entropy in each of those channels and obtain the invariant image. Figure 4 illustrates the effect of applying the histogram equalization technique to images of the same individual captured by different cameras. These images highlight the fact that a change in illumination leads to a significant change in the colors captured by the camera. The last two columns show result images after applying histogram equalization procedure. It is clear that the resulting images are much more similar than the two original images.

5. Human appearance

In this section we define our appearance model based on the novel MRC patches extracted from tracks of a specific individual. The input of our approach is a set of cropped images corresponding to human detection and tracking results, normalized by the histogram equalization applied to each of the color channels (RGB). From such normalized images we extract the MRC patches (Section 5.3). Nevertheless, before explaining details concerning the proposed MRC patches, we present a brief overview of the covariance descriptor and a short introduction to Riemannian geometry.

5.1. Covariance Descriptor

Here, we present an overview of the covariance descriptor [14] and its specialization in our approach.

Let I be an image. The method can be generalized to any type of image such as one dimensional intensity image, three channel color image or even other type of images, *e.g* infrared. Let F be the d-dimensional feature image extracted from I

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

where the function ϕ can be any mapping such as color, intensity, gradients, filter responses, *etc.* For a given rectangular region $Reg \subset F$, let $\{f_k\}_{k=1...n}$ be the d-dimensional feature points inside Reg. The region Reg is represented with the $d \times d$ covariance matrix of the feature points

$$C_{Reg} = \frac{1}{n-1}\sum_{k=1}^{n}(f_k - \bar{f})(f_k - \bar{f})^T \tag{2}$$

where $\bar{f}$ is the mean of the points. Such defined covariance matrix can be computed in an efficient way for each subregion of an image using integral images [14].

In our approach we define the mapping $\phi(I, x, y)$ as

$$\left[x, y, R_{xy}, G_{xy}, B_{xy}, \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right] \tag{3}$$

where x and y are pixel location, R_{xy}, G_{xy}, B_{xy} are RGB channel values and ∇ and θ corresponds to gradient magnitude and orientation in each channel, respectively.

The input image region is mapped to $d = 11$ dimensional feature image. Thus, the covariance region descriptor is represented by an 11×11 matrix. The descriptor encodes information of the variances of the defined features inside the region, their correlations with each other and spatial layout. It is shown that the performance of the covariance features is superior to other methods as rotation and illumination changes are absorbed by covariance matrix. Moreover, we enhance robustness to local illumination variations in an image by using *the Pearson Product-Moment Correlation Coefficients* (PMCC) as elements of covariance matrices (such matrix is also often called *correlation matrix*). In a correlation matrix every element (i, j) is normalized by the product of their standard deviations

$$\hat{C}_{Reg}(i,j) = \frac{C_{Reg}(i,j)}{\sigma_i \sigma_j}. \tag{4}$$

Since now, every matrix in this paper is assumed to be such normalized covariance matrix.

5.2. Riemannian Geometry

Covariance matrix as a positive definite and symmetric matrix can be seen as a tensor. The main problem is that such defined tensor space is a manifold that is not a vector space with the usual additive structure. A manifold is a topological space which is locally similar to an Euclidean space. A neighbourhood homomorphic to an open subset $\Re^m$ can be defined for every point on the manifold. As symmetric positive definite matrices constitute a convex half-cone in the vector space of matrices, many usual operations (like the mean) are stable in this space. Nevertheless problems arise when tensors are extracted from real data (estimated symmetric matrix could have negative eigenvalues). Therefore manifold is specified as Riemannian to determine a powerful framework using tools from differential geometry [36]. A Riemannian manifold $\mathcal{M}$ is a differentiable manifold in which each tangent space has an inner product which varies smoothly from point to point. Since covariance matrices can be represented as a connected Riemannian manifold we apply operations such as the distance and the mean computation using this differential geometry.

5.2.1. Covariance Matrix Distance

The space of positive definite and symmetric covariance matrices can be formulated as a connected Riemannian manifold. Hence, we use the distance definition proposed by [37] to compute the dissimilarity between two covariance matrices

$$\rho(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)} \tag{5}$$

where $\lambda_k(C_i, C_j)_{k=1...d}$ are the generalized eigenvalues of C_i and C_j, determined by

$$\lambda_k C_i x_k - C_j x_k = 0, \qquad k = 1 \dots d \tag{6}$$

and $x_k \neq 0$ are the generalized eigenvectors.

5.2.2. Mean Covariance

Let $C_1, \dots, C_N$ be a set of covariance matrices. The Karcher or Fréchet mean is the set of tensors minimizing the sum of squared distances. In the case of tensors, the manifold has a non-positive curvature, so there is a unique mean value μ

$$\mu = arg \min_{C \in \mathcal{M}} \sum_{i=1}^{N} \rho^2(C, C_i). \tag{7}$$

where ρ is the covariance matrix distance (Eq. 5).

Since covariance matrices lay in a Riemannian manifold we use the intrinsic Newton gradient descent algorithm to compute the approximation mean covariance at step $t+1$

$$\mu_{t+1} = exp_{\mu_t}\left[\frac{1}{N}\sum_{i=1}^{N} log_{\mu_t}(C_i)\right] \tag{8}$$

where exp_{μ_t} and log_{μ_t} are specific operators uniquely defined on the Riemannian manifold. This iterative gradient descent algorithm usually converges very fast (in experiments 5 iterations were enough, which is similar to [36]).

5.3. MRC Patches

In this section we define *the Mean Riemannian Covariance* (MRC) patches and explain their merits. Once a human has been detected and color has been normalized, we scale every cropped image into a fixed size $W \times H$. The MRC patch corresponds to a square region (of size $\frac{W}{4} \times \frac{W}{4}$ and $\frac{W}{2} \times \frac{W}{2}$). We assume that such MRC patches can form the patch combinations (see Fig. 5). In our approach we only consider the patch combinations built using maximally two MRC patches. The patch combination may consist of two patches with the same size as well as with different size. Different patterns of a human appearance are captured from a window $W \times H$. We extract patches of size $\frac{W}{4} \times \frac{W}{4}$ shifted horizontally and vertically by $\frac{W}{8}$ and patches of size $\frac{W}{2} \times \frac{W}{2}$ shifted horizontally and vertically by $\frac{W}{4}$. The position of the MRC patch on the fixed size window and the spatial correlation between patches is essential to carry out discriminative power.

There have already been proposed some statistics on covariance functions which take into account spatial and temporal changes [38, 39]. Nevertheless, these approaches assume that covariance function is a stationary (or weakly stationary) process in space and time. Unfortunately in our case we can not make such strong assumptions as visual features mostly do not meet this requirement in a video sequence. Moreover, noisy human detections and gaps in tracking prevent the covariance functions to be separable in space and time. As there is hard to apply these statistics methods to our covariance matrices we have decided to use the mean covariance computed on a Riemannian manifold as a descriptor of a region in a video sequence.

Let $C_1^p, \ldots, C_N^p$ be a set of covariance matrices extracted during tracking of N frames corresponding to image square regions at position of the patch p. We define the MRC patch as the mean covariance of these covariance matrices (see Section 5.2.2) computed using a Riemannian space (see Fig. 6). The mean covariance matrix as an intrinsic average blends all extracted matrices. This mean covariance matrix keeps not only an information about features distribution but also carries out an essential information about temporal changes of the appearance related to the position of the patch p. The MRC patch (the mean covariance matrix together with its position) is fundamental in our approach. We claim that this descriptor is extremely efficient for re-identification purposes which is confirmed in experimental results. Now we introduce a new reliability measure which describes invariance of the patch.

5.3.1. Patch Reliability

Let $R_1^p, \ldots, R_N^p$ be a set of image square regions extracted during tracking of N frames at position of the patch p. As we already know, for each such region we extract covariance C_i^p which encodes information of the variances of the defined features (f) inside the region and their correlations.

Together with the mean covariance computation we define the reliability measure $\mathfrak{R}$ which describes invariance of the tracked image region

$$\mathfrak{R} = 1 - \frac{\sigma}{\max \sigma}, \quad \sigma = \frac{1}{N-1} \sum_{k=1}^{N} (\bar{f}_k - \bar{\Gamma})^2 \qquad (9)$$

where $\bar{\Gamma}$ is the mean feature vector along the N tracking regions and $\bar{f}_k$ is the mean feature vector corresponding to the region k, $\max \sigma$ corresponds to

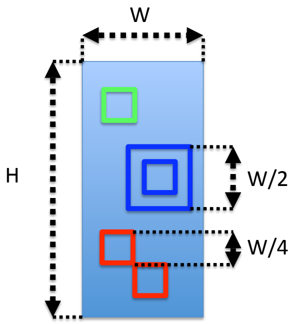


Figure 5: MRC Patches. Green patch corresponds to the single square region (top). Blue patch is an example of combination (small - big) of two patches with the different size (middle). Red patch illustrates combination (small - small) of patches with the same size (bottom).

the maximal value of σ for the specific individual. As we assume that only partial occlusions may occur, the reliability works similarly to background subtraction. Here, our idea is to remove the most variable features because we assume that these features are the noisiest (containing background). See Fig. 8 (b), the most unreliable features correspond to legs and background.

5.4. MRC Patch Selection

We represent the human signature by a set of the MRC patches. The human signature should be built on patches which are discriminative and reliable. As already mentioned patches can form combinations. In this case we call it patch combination. These combinations together with a single MRC patch represent the MRC patch space. From this MRC patch space we select the most significant MRC patches. We propose two ways for the MRC patch selection to show how a learning phase can improve performance.

The first method is based on the patch reliability measure. We select the most reliable patches with the highest reliability among others. If the patch combination is considered, the reliability is an average of reliabilities of both patches. This selection method allows us to create the human signature without any learning phase.

In the second method, we extract discriminative features using boosting where we follow one-against all scheme. We lay the re-identification problem as classification where we separate the specific individual class from the rest

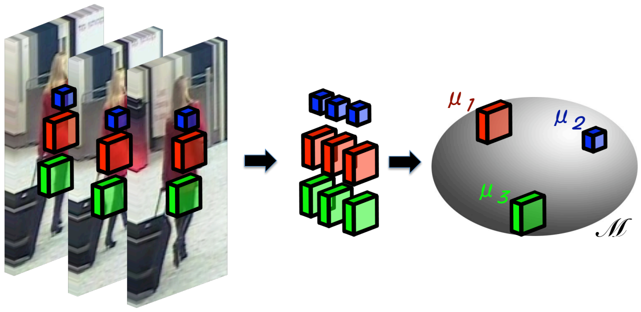


Figure 6: Computation of three the MRC patches. Covariances gathered from tracking results are used to compute the mean covariance using Riemannian manifold space (depicted with the surface of the sphere). The mean covariance forms the MRC patch.

of humans. Here, we assume that discriminative features of an individual extracted from one camera correspond to discriminative features of the same individual extracted from another camera. Let us assume that we want to *learn* the signature $\mathfrak{s}^c_j$. Each signature has an associated set of relevant observations (tracking results of the human j in N frames from camera c) represented by set of images $\mathfrak{I}^{+c}_j = \{I^c_{j,1}, \dots I^c_{j,N}\}$. This set is used as positive samples in a learning phase. Negative samples are obtained from images extracted from the same camera during tracking the rest of humans: $\mathfrak{I}^{-c}_j = \{I^c_{i,k}\}_{i \neq j}$.

Here we take advantage of improved boosting algorithm using confidence-rated predictions [40]. An essential advantage of this approach is that a weak hypothesis generates not only a predicted classification but also a self-rated confidence score which estimates the confidence of its prediction. Later, this confidence value is used to define the similarity measure between two signatures. In our approach weak hypotheses are searched on Riemannian manifold space where a weak classifier is represented by the MRC patch extracted from positive samples. The manifold is a rather complicated space compared with euclidean. As an example, let us consider a linear classifier on $2D$ euclidean space. Having two points in this space, we can easily define a line which separates these points dividing the space into two. When we take a manifold as a space (let us imagine a torus) there is no way to separate a space into two. The first thought would be to map the manifold space to a higher dimensional euclidean space, which can be considered as flattening the

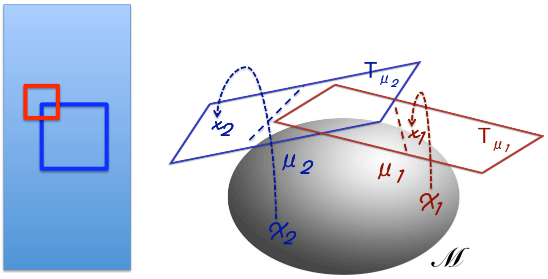


Figure 7: MRC Classifier. The manifold is depicted with the surface of the sphere, and the planes are the tangent spaces of the mean covariances. The samples $\mathfrak{X}_i$are projected to tangent space T_{μ_i}at means via log_μ operation.

manifold. Nevertheless, in general case, there does not exist such mapping into flattened space to preserve the distances between the points on the manifold. The global structure of the points will be disturbed. Therefore the evaluated samples are projected to the tangent spaces at the computed means, where the weak classifiers are learned. At each iteration, the weights of samples are adjusted through boosting. Then, we map the points to the tangent space at the mean and learn a weak classifier on this vector space. We do not go into details concerning learning on the manifold space as it is not the main topic of this paper. The interested reader is pointed to [41].

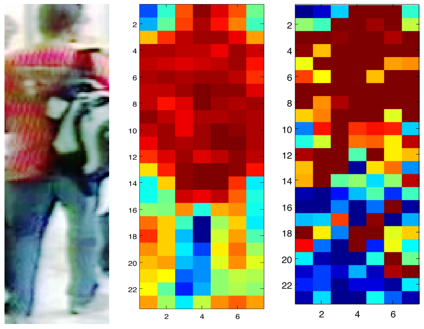
5.4.1. MRC Patch as a Weak Classifier

We define a weak classifier as a function h based on distance computation on Riemannian manifold space (see Fig. 7). Image I is classified by a weak classifier built on patch p using the threshold function defined as

$$h(p, I) = \begin{cases} \alpha = \frac{1}{2}\ln(\frac{W_+^+}{W_-^+}) & if \ \sum_j w_j \rho(\mu_j, C_j) \leq T_p \\ \beta = \frac{1}{2}\ln(\frac{W_+^-}{W_-^-}), & otherwise \end{cases} \tag{10}$$

where W_+^+, W_-^+, W_+^-, W_-^- correspond to weighted sum of true positive, false positive, false negative and true negative samples, respectively. Threshold T_p and weights w_j are obtained by minimizing error of weak classifier during learning; μ_j is the mean covariance of the MRC patch classifier and covariance C_j corresponds to the image region of I.

Each the MRC patch and the MRC patch combination extracted during tracking forms a weak classifier. From the set of weak classifiers, boosting



(a) I (b) $\mathfrak{R}$ (c) α

Figure 8: Illustration of patch significance: (a) one of many frames obtained during tracking; (b) reliability map obtained by the first method; (c) confidence map obtained by boosting. Colours correspond to significance of patches (for clarity only $\frac{W}{4} \times \frac{W}{4}$ patches, shifted by $\frac{W}{8}$ pixels are illustrated, red indicates the highest significance, blue the lowest).

algorithm selects the ones which together form a strong classifier. Details of boosting algorithm using confidence-rated predictions can be found in [40].

We use the chosen weak classifiers in the MRC patch matching. In Fig. 8 we present reliability and confidence maps, based on reliability $\mathfrak{R}$ and confidence α, respectively.

5.5. MRC Patch Matching

Given extracted human signatures, we introduce a way to effectively distinguish individuals. The human signature is represented by a set of the relevant MRC patches (extracted using one of the aforementioned selection methods). The MRC patch consists of the mean covariances with its positions. The position of the patch is essential to keep discriminative power of the human signature. In general, the matching of two signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is carried out by maximizing the similarity measure. We shift one signature over another to reduce body alignment issues. When we shift signature (see Fig. 9) we preserve relative position between patches to avoid wasting of discriminative property of the patch position. In our experiments the signature is shifted over another not more than $\frac{W}{4}$ pixels to maximize similarity. The similarity between two human signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is defined as

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \frac{1}{\mid K \mid} \sum_{i \in K} \frac{\beta_{A_i} + \beta_{B_i}}{\rho(\mu_{A_i}, \mu_{B_i})} \tag{11}$$

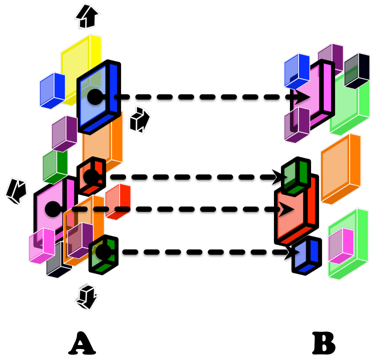


Figure 9: The similarity between two human signatures. Every signature is a set of the MRC patches. Signature A is shifted left/right/up and down to find out the best corresponding patches in signature B (position of a patch determines matching). Connections in the figure represent corresponding patches. Some connections are suppressed for clarity.

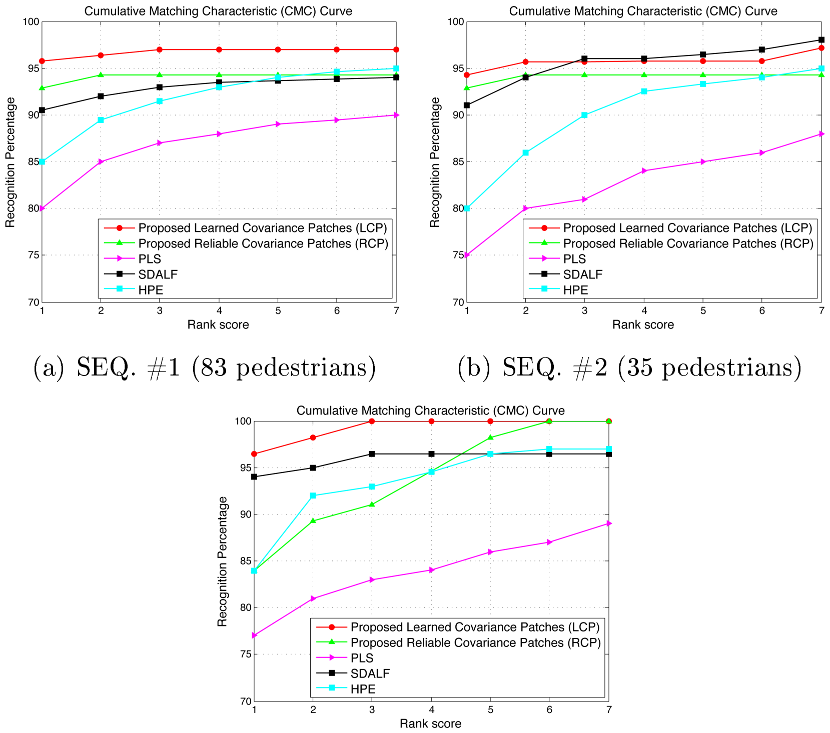
where K stands for the set of corresponding patches in signature $\mathfrak{s}_A$ and signature $\mathfrak{s}_B$; ρ is the covariance matrix distance; β is a reliability ($\mathfrak{R}$) or a confidence (α) value of the corresponding patches depending on selection method.

6. Experimental results

We evaluate our methods on the ETHZ and the i-LIDS dataset. Moreover, we extracted new sets of individuals from i-LIDS data to create more pertinent sets of individuals. The performance is shown using the Cumulative Matching Characteristic (CMC) curve suggested in [42] as the validation method for the re-identification problem. The CMC curve represents the expectation of finding the correct match in the top n matches.

Experimental setup

Every human image is scaled into a fixed size of 64 × 192 pixels. We generate the MRC patches of 16 × 16 pixels with 8 pixels step (it gives 161 patches). We generate the MRC patches of 32 × 32 pixels with 16 pixels step (it gives 33 patches). Then, we generate combination of the MRC patches (small - small, small - big, big - big, we limit space of patches to those which are situated with the 16 pixels step). Finally, in total we have 3.401 MRC weak classifiers. In the evaluation the proposed selection method based only on reliability of patches is referred to as *Reliable Covariance Patches (RCP)* method and the method based on a boosting scheme is denoted as *Learned Covariance Patches (LCP)* method.



(a) SEQ. #1 (83 pedestrians)

(b) SEQ. #2 (35 pedestrians)

(c) SEQ. #3 (28 pedestrians)

Figure 10: CMC curves obtained on ETHZ dataset. Our descriptors are noted as LCP and RCP respectively. We compare our methods with the results of HPE [30], SDALF [28] and PLS [19].

6.1. ETHZ dataset

ETHZ dataset was originally used for human detection [43]. In [19] this data has been adjusted for re-identification purposes[1]. The modified dataset consists of three sequences: SEQ. #1 contains 83 pedestrians, SEQ. #2 contains 35 pedestrians and SEQ. #3 contains 28 pedestrians. The main drawback of this dataset is that the re-identification is performed with the same camera. Since the human images are very similar we randomly pick up a set of 10 consecutive frames from the beginning and from the end of each sequence to maximize challenging aspects and to be comparable with

[1] ETHZ Dataset for Appearance-Based Modelling: *http://www.umiacs.umd.edu/ schwartz/datasets.html*

the multi-image signatures from [28]. We compare our methods with HPE [30], PLS [19] and SDALF [28] (see Fig. 10). In SEQ. #1 our approaches outperform state of the art methods. SEQ. # 2 seems to be more challenging despite the fact that it has less number of individuals. Unfortunately, it is difficult to say which method performs the best in this sequence. Our proposed approaches perform better for the top rank, nevertheless after a few ranks curves are crossing. In SEQ. #3 LCP significantly outperforms all descriptors. LCP matching rate for top rank is 96.42%. LCP method always outperforms RCP. The results show that discriminative learning improves performance.

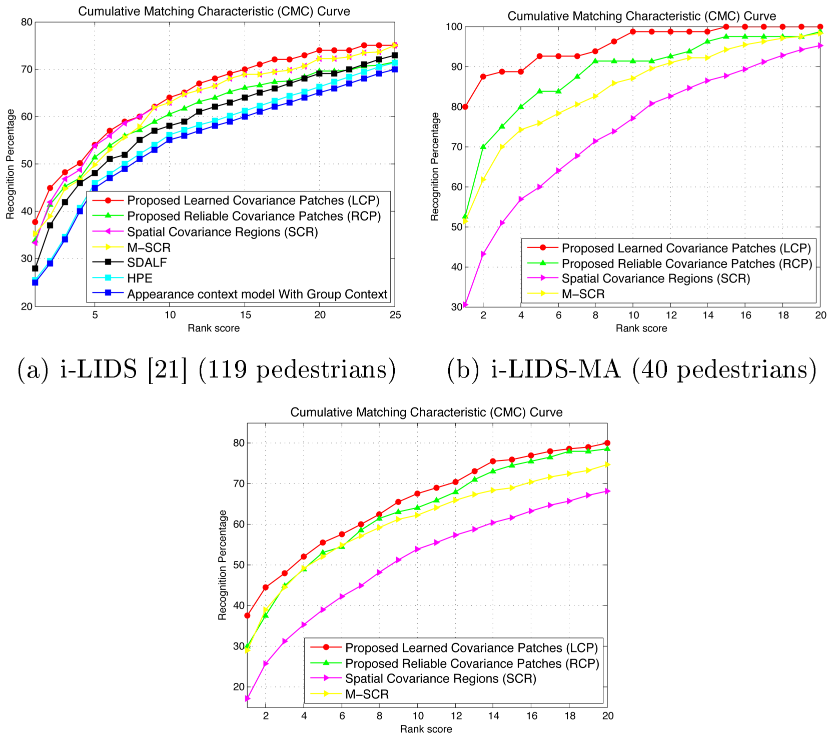
In our belief, despite such challenging aspects as illuminations changes and occlusions the ETHZ dataset is not challenging enough to evaluate re-identification approaches. One of the most challenging issues in the re-identification problem is due to different camera settings, different color responses, different camera view points and different environments, which is not in this case. Hence, we have also evaluated our approach on images from more challenging i-LIDS (MCTS) dataset.

6.2. i-LIDS dataset

The experiments are performed on images from the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset with multiple camera views. The evaluation dataset contains 476 images with 119 individuals automatically extracted by [21]. This dataset is very challenging since there are many occlusions and often only the top part of the person is visible. Unfortunately, this dataset does not fit very well for *multiple-shot* signature because the number of images per individual is very low. Hence, we applied simple affine transformation on the image to obtain multiple images (coordination of transformation matrix were changed by 5% and rotation angle was in range of $[-6°; 6°]$). As every transformation of original image is allowed we claim that this solution is fair with the state of the art. On such generated data we follow the scheme of evaluation presented in [21].

We compared our approach with methods which obtained the best performance on this dataset: SCR [12], Appearance Context [21], HPE [30] and SDALF [28][2]. As SCR belongs to *single-shot* approaches, we extended SCR

[2] For *multiple-image* signature, the authors in [28] assume that for each individual 4 images in average are given and test their approach using 3 images to build signature. Nevertheless, for 55 individuals (from 119) there is only maximum 3 images given. If the



(a) i-LIDS [21] (119 pedestrians)

(b) i-LIDS-MA (40 pedestrians)

(c) i-LIDS-AA (100 pedestrians)

Figure 11: CMC curves obtained on i-LIDS datasets; (a) We compare our methods with the results of SCR [12] and M-SCR, HPE [30], SDALF [28] and Appearance context model with Group Context [21]; (b) and (c) We compare our methods with the results of SCR [12] and M-SCR.

to *multiple-shot* approach by applying *"set matching"* (the minimal distance between pair of images). This makes our evaluation fairer to SCR method. The extended SCR method is noted as M-SCR.

Our evaluation results together with the state of the art approaches are

authors of [28] really used 3 images per signature, for 46% of the database they do not have *probe* images (the evaluation is done only on the part of the database) or they used exactly the same images to create a gallery and probe signatures that automatically influence true matching. Both solutions are definitely wrong and induce incomparable results (for at least 46% of individuals the comparison in [28] is definitely not fair with the state of the art). This is the reason why we do not follow this evaluation scheme for *multiple-image* signature.

presented in Fig. 11 (a). It is worth noting that the performance is not very high because the person images from the i-LIDS data are very challenging since they were captured from non-overlapping multiple camera views subject to significant occlusions and large variations in both view angle and illumination. Our LCP outperforms all descriptors. Nevertheless the difference between SCR [12] and M-SCR is not significant. We think that this is a consequence of scarce amount of images per pedestrian in this database. Therefore we have extracted two new sets of individuals from i-LIDS data to investigate more carefully advantage of using tracking results in building the human signatures. These datasets finally satisfy all requirements of *multiple-shot* person re- identification.

i-LIDS-MA

First dataset contains 40 individuals extracted from two cameras. For each individual 46 frames are annotated manually from both cameras. Therefore we have $40 \times 2 \times 46 = 3680$ annotated images. We denote this dataset as i-LIDS-MA. For each pedestrian we create human signature using $N = 1$ (for SCR [12]) or $N = 10$ (for M-SCR, RCP, LCP) randomly selected images. Then, every signature is used as a query to the gallery set of signatures from different camera. The procedure has been repeated 10 times and average CMC curves together with our results are displayed in Fig. 11 (b). As we can notice our LCP outperforms significantly SCR. LCP matching rate for top rank is 80% in comparison with SCR which obtained 30%. Curves of RCP and M-SCR are between others. Also our RCP is a little bit better than M-SCR. We outperform SCR because our human signature is based on the MRC patches which take advantage of tracking results and keep an information about temporal changes of the appearance. Moreover, the results show that the discriminative learning phase improves performance. Nevertheless, this manually annotated dataset does not reflect real video surveillance scenario where humans are detected and tracked automatically. Consequently, we applied HOG-based human detector and tracker to obtain 100 individuals seen from both cameras. In this case, detection and tracking results are noisy which makes the dataset more challenging. We name this dataset as i-LIDS-AA.

i-LIDS-AA

This dataset contains 100 individuals. For each individual we extracted a different number of frames depending on the tracking difficulties. In total, the dataset contains 10754 images. The performance on this dataset is shown in Fig. 11 (c). The evaluation scheme was the same as for i-LIDS-MA dataset. The results show again that our descriptors outperform significantly SCR and there are also better than M-SCR. Nevertheless the performance is not very high in comparison with the results obtained on i-LIDS-MA. It shows one of the main limitations that our approach performance directly depends on human detection results (*e.g.* detected bounding boxes not accurately centred around the people, only part of the people are detected due to occlusion). However, the results show that despite this limitation our descriptors perform better than the state of the art approaches.

7. Discussion

The results definitively show that the boosting scheme improves performance. Nevertheless, discriminative approaches are often accused of non-scalability (like [18, 19]). It is true that in these approaches (in our as well) an extensive learning phase is necessary to extract discriminative signatures. These approaches are difficult to apply to real scenarios where new people appear continuously. First, the learning phase prevents the signature generation to be real-time. Second, every time when a new signature is created we have to update all signatures in the database (one-against-all scheme). For example, in PLS [19] there is a requirement to have all the gallery signatures beforehand, in order to estimate the weights on the appearance model. If one pedestrian is added the weights must be recomputed which makes the approach not-suitable for video surveillance systems.

As a solution to scalability issues we can propose to extract a *reference dataset* which can be used as negative samples for learning a discriminative signature. This *reference dataset* can be chosen offline to be the most representative. Nevertheless, in this case, a time consumption issue still remains. As in the system, there is no constraints for a *real-time* signature generation, there already exist learning approaches which can operate in reasonable time. Moreover, recently, driven by the insatiable market demand for real-time, the programmable *Graphic Processor Unit* (GPU) has evolved into a highly parallel, multithreaded, manycore processor with tremendous computational power and very high memory bandwidth. These new hardware architectures such as NVIDIA [44] are investigated to speed up the computation of pattern recognition problems. Currently, step by step, well known learning algorithms [45] or time consuming descriptors [46] are ported to such specialized architectures. As image and media processing demand a lot of computation power, this direction seems to fit perfectly for more sophisticated approaches. We claim that the usage of high-performance computing can be a solution to make a discriminative learning more suitable for video-surveillance systems.

8. Conclusions

We have proposed a new approach for the human re-identification problem. An extensive evaluation has been performed on the ETHZ and the i-LIDS datasets. It has been shown that the proposed MRC patches computed using a Riemannian manifold theory can extract an essential information about appearance of the human and its variability. The experiments show that a joint combination of these distinctive patches constructs a robust invariant human signature which can handle differences in camera parameters. Finally, we have proposed to evaluate the re-identification approaches on two more pertinent sets of individuals.

In the future work we will consider how to minimize the influence of noisy human detection and tracking on matching human signatures. Also we are planing to consider 2D/3D body parts modelling to improve matching of different poses of individuals. Finally, we are going to investigate performance of GPU-based discriminative learning to make the approach suitable for real-time video surveillance systems.

References

[1] O. Javed, Z. Rasheed, K. Shafique, M. Shah, Tracking across multiple cameras with disjoint views, in: Proceedings of the 9th IEEE International Conference on Computer Vision, ICCV, IEEE Computer Society, 2003, pp. 952–957.

[2] H. Proenca, Iris recognition: On the segmentation of degraded images acquired in the visible wavelength, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1502–1516.

[3] M. Bauml, K. Bernardin, M. Fischer, H. K. Ekenel, R. Stiefelhagen, Multi-pose face recognition for person retrieval in camera networks, in: Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS, 2010.

[4] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 1505–1518.

[5] R. Chellappa, A. K. Roy-Chowdhury, A. Kale, Human identification using gait and face, in: Proceedings of the 20th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2007, pp. 1–2.

[6] U. Park, A. Jain, I. Kitahara, K. Kogure, N. Hagita, Vise: Visual search engine using multiple networked cameras, in: Proceedings of the 18th International Conference on Pattern Recognition, ICPR, IEEE Computer Society, 2006, pp. 1204–1207.

[7] A. C. Gallagher, T. Chen, Clothing cosegmentation for recognizing people, in: Proceedings of the 21st Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2008, pp. 1–8.

[8] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, A. Isaacs, Detection of loitering individuals in public transportation areas, IEEE Transactions on Intelligent Transportation Systems (2005) 167–177.

[9] Y. Cai, K. Huang, T. Tan, Human appearance matching across multiple non-overlapping cameras., in: Proceedings of the 19th International Conference on Pattern Recognition, ICPR, IEEE Computer Society, 2008.

[10] L. Hu, Y. Wang, S. Jiang, Q. Huang, W. Gao, Human reappearance detection based on on-line learning, in: Proceedings of the 19th International Conference on Pattern Recognition, ICPR, IEEE Computer Society, 2008, pp. 1 –4.

[11] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[12] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Person re-identification using spatial covariance regions of human body parts, in: Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS, IEEE Computer Society, 2010.

[13] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 18th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2005, pp. 886–893.

[14] O. Tuzel, F. Porikli, P. Meer, Region covariance: A fast descriptor for detection and classification, in: Proceedings of the 9th European Conference on Computer Vision, ECCV, Springer-Verlag, 2006, pp. 589–600.

[15] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the 19th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2006, pp. 2169–2178.

[16] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, P. Tu, Shape and appearance context modeling, in: Proceedings of the 11th International Conference on Computer Vision, ICCV, IEEE Computer Society, 2007, pp. 1–8.

[17] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the 10th European Conference on Computer Vision, ECCV, Springer-Verlag, 2008, pp. 262–275.

[18] Z. Lin, L. S. Davis, Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance, in: Proceedings of the 4th International Symposium on Advances in Visual Computing, ISVS, Springer-Verlag, 2008, pp. 23–34.

[19] W. R. Schwartz, L. S. Davis, Learning discriminative appearance-based models using partial least squares, in: Proceedings of the 22nd Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI, IEEE Computer Society, 2009, pp. 322–329.

[20] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: Proceedings of the 21st British Machine Vision Conference, BMVC, BMVC Press, 2010, pp. 21.1–21.11.

[21] W.-S. Zheng, S. Gong, T. Xiang, Associating groups of people, in: Proceedings of the 20th British Machine Vision Conference, BMVC, BMVC Press, 2009.

[22] Y. Cai, V. Takala, M. Pietikainen, Matching groups of people by covariance descriptor, in: Proceedings of the 20th International Conference on Pattern Recognition, ICPR, IEEE Computer Society, 2010, pp. 2744–2747.

[23] N. Gheissari, T. B. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: Proceedings of the 19th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2006, pp. 1528–1535.

[24] L. F. Teixeira, L. Corte-Real, Video object matching across multiple independent views using local descriptors and adaptive learning, Pattern Recognition Letters 30 (2009) 157–167.

[25] O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux, Person reidentification in multi-camera system by signature based on interest point descriptors collected on short video sequences, in: Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC, IEEE Computer Society, 2008, pp. 1–6.

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

[27] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Person re-identification using haar-based and dcd-based signature, in: Proceedings of the 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems in conjunction with 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AMMCSS, IEEE Computer Society, 2010.

[28] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the 23th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2010, pp. 2360–2367.

[29] P.-E. Forssén, Maximally stable colour regions for recognition and matching, in: Proceedings of the 20th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2007.

[30] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, V. Murino, Multiple-shot person re-identification by hpe signature, in: Proceedings of the 20th International Conference on Pattern Recognition, ICPR, IEEE Computer Society, 2010, pp. 1413–1416.

[31] O. Oreifej, R. Mehran, M. Shah, Human identity recognition in aerial images, in: Proceedings of the 23th Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2010, pp. 709–716.

[32] E. Corvee, F. Bremond, Combining face detection and people tracking in video sequences, in: Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention, ICDP, IEEE Computer Society, 2009.

[33] G. Overett, L. Petersson, N. Brewer, L. Andersson, N. Pettersson, A new pedestrian dataset for supervised learning, IEEE Intelligent Vehicles Symposium (2008).

[34] S. D. Hordley, G. D. Finlayson, G. Schaefer, G. Y. Tian, Illuminant and device invariant colour using histogram equalisation, Pattern Recognition (2005).

[35] R. C. Gonzalez, R. E. Woods, Digital Image Processing, Addison-Wesley Longman Publishing Co., Boston, MA, USA, 2001.

[36] X. Pennec, P. Fillard, N. Ayache, A riemannian framework for tensor computing, International Journal on Computer Vision 66 (2006) 41–66.

[37] W. Förstner, B. Moonen, 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.

[38] N. Cressie, H. cheng Huang, Classes of nonseparable, spatio-temporal stationary covariance functions, Journal of the American Statistical Association (1999) 1330–1340.

[39] M. Fuentes, Testing for separability of spatial-temporal covariance functions, Journal of Statistical Planning and Inference (2006) 447 – 466.

[40] R. E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine Learning (1999) 297–336.

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

[42] D. Gray, S. Brennan, H. Tao, Evaluating Appearance Models for Recognition, Reacquisition, and Tracking, in: Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, PETS, IEEE Computer Society, 2007.

[43] A. Ess, B. Leibe, L. Van Gool, Depth and appearance for mobile scene analysis, in: Proceedings of the 11th International Conference on Computer Vision, ICCV, IEEE Computer Society, 2007, pp. 1–8.

[44] E. Lindholm, J. Nickolls, S. Oberman, J. Montrym, Nvidia tesla: A unified graphics and computing architecture, IEEE Micro (2008) 39–55.

[45] B. Catanzaro, N. Sundaram, K. Keutzer, Fast support vector machine training and classification on graphics processors, in: Proceedings of the 25th International Conference on Machine learning, ICML, ACM, 2008, pp. 104–111.

[46] S. Bak, K. Kurowski, K. Napierala, Human re-identification system on highly parallel gpu and cpu architectures, in: Proceedings of the 4th International Conference on Multimedia Communications, Services and Security, MCSS, Communications in Computer and Information Science, Springer, 2011.