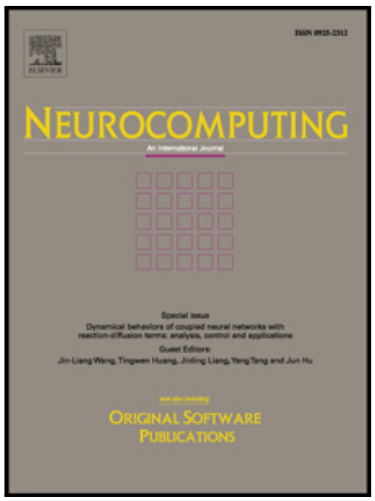
Fusion of appearance and motion-based sparse representations for multi-shot person re-identification

Mohamed Ibn Khedher, Mounim El Yacoubi, Bernadette Dorizzi



Fusion of Appearance and Motion-based Sparse Representations for Multi-shot Person Re-identification

Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi

PII: S0925-2312(17)30430-7
DOI: 10.1016/j.neucom.2016.11.073
Reference: NEUCOM 18180

To appear in: *Neurocomputing*

Received date: 13 June 2016
Revised date: 14 October 2016
Accepted date: 29 November 2016

Please cite this article as: Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi, Fusion of Appearance and Motion-based Sparse Representations for Multi-shot Person Re-identification, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.11.073

Fusion of Appearance and Motion-based Sparse Representations for Multi-shot Person Re-identification

Mohamed Ibn Khedher, Mounim A. El-Yacoubi and Bernadette Dorizzi

SAMOVAR, Telecom SudParis, CNRS, University of Paris-Saclay, France
mohamed.ibn_khedher@telecom-sudparis.eu, mounim.el_yacoubi@telecom-sudparis.eu, bernadette.dorizzi@telecom-sudparis.eu

Abstract

We present in this paper a multi-shot human re-identification system from video sequences based on interest points (IPs) matching. Our contribution is to take advantage of the complementary of person's appearance and style of its movement that leads to a more robust description with respect to various complexity factors. The proposed contributions include person's description and features matching. For person's description, we propose to exploit a fusion strategy of two complementary features provided by appearance and motion description. We describe motion using spatiotemporal IPs, and use spatial IPs for describing the appearance. For feature matching, we use Sparse Representation (SR) as a local matching method between IPs. The fusion strategy is based on the weighted sum of matched IPs votes and then applying the rule of majority vote. This approach is evaluated on a large public dataset, PRID-2011. The experimental results show that our approach clearly outperforms current state-of-the-art.

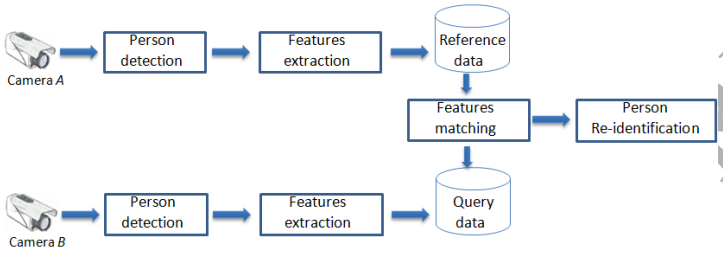
Keywords: Person re-identification, fusion, motion, appearance, sparse representation, dynamic dictionary.

1. Introduction

Person re-identification is an important video monitoring task for tracking people over multiple cameras . If a person crosses the view field of a camera-A, the goal is to check whether he/she passed through the view field of a second

camera-B (Figure 1). Typical applications include video surveillance in sensitive access control facilities, shopping centers or hospitals, where tracking a person could be cast as a re-identification problem.

Figure 1: Steps of a generic re-identification system



Most state of the art methods have addressed the re-identification problem at the frame level, and temporal information, despite its successful use in Human Action Recognition (HAR), is not considered. In HAR context, actions are typically acquired by a single camera in similar view angle and lighting settings. In such conditions, motion descriptors can highly discriminate the actions irrespective of person identity [1, 2]. Person re-identification is the opposite task of re-identifying the person irrespective of the action performed. This explains why temporal information has been overlooked in current systems. In spite of being less discriminative of person identity, we believe, nonetheless, that temporal information may be important to detect salient body parts to be exploited for re-identification.

Our contribution is threefold: first, we exploit the temporal information available in video sequences, through 3D Interest Points (IPs), for the first time in the person re-identification task. Second, we propose a new fusion mechanism leveraging the two feature types: spatial (appearance) and temporal (motion features), for improved accuracy. Finally, our approach relies on sparse representation for more robust IP matching as this is shown by our experimental validation on public datasets acquired under uncontrolled conditions.

Regarding the temporal information, as it discriminates actions and not people, we propose to harness it not to characterize the person motion, but

rather to discard static information (object, background, etc.) that usually hinders silhouette matching, and, therefore, to focus only on moving parts in the video sequence, for subsequent person identification, the hypothesis being that the moving parts are mainly associated with body parts.

Second, beside the temporal-based person description, we propose to use a purely static frame-based description, based on our work in [3]. From the person representation standpoint, we use spatial (2D) and spatial-temporal (3D) IPs as local descriptors. Relying on these IPs offers a tradeoff between extracting only salient points from images or videos on one hand and keeping sufficient and relevant information to describe a person on the other. As the two descriptions are produced from two different channels (static and temporal), we combine them for a robust person representation. The fusion strategy is based on the weighted sum of matched IPs votes and then applying the majority vote rule.

Third, we propose a novel framework for IPs matching via sparse representation (SR), where each query IP is expressed as a linear combination vector of reference IPs. This representation yields a sparse vector whose nonzero entries correspond to the weights of reference IPs that match the query IP. As the SR dictionary is usually built from all reference vectors, its size can be huge for large datasets. For efficiency, we select a dynamic and reduced dictionary from the reference dataset . The dictionary is dynamic as it changes for each query IP, and is reduced since it is composed of only its N closest reference IPs. To accelerate the search of nearest neighbors, we use KD-tree [4, 5] for an efficient representation of the dictionary. Once the sparse representation of the query IP is obtained, the query IP is reconstructed each time in turn, using one of the reference IPs retrieved by SR and discarding the others. The query IP is then assigned to the reference identity minimizing its reconstruction. After all query IPs are assigned in this way to their reference identities, the majority vote rule is applied to classify the query sequence, i.e. to perform person re-identification.

The justification of a *local* IP matching method, where each IP is matched independently of the others, is that it is more robust for matching two silhouettes from the same person, corrupted by background or carried objects. Take a

motivating example where the reference image corresponds to a person pushing a stroller while the query image corresponds to the same person without the stroller. The local matching of salient points (features) is more likely to lead to a good match as the silhouette features are not corrupted by the stroller descriptors. Second, as we are representing a ROI by a set of local features, matching features locally enables the design of smart filtering post processing with the aim to keep the best matched features pairs only and disregard other unreliable pairs (e.g. stroller feature point matched to a silhouette feature point).

The motivation behind SR-based matching is that, considering video sequences, any relevant query IP is likely to have several similar IPs in the reference video. For instance, if an IP is detected on the face of a query frame, the number of similar reference IPs is proportional to the number of reference frames and it is important to use all of them to help re-identification. This can be ensured through SR which dynamically adapts the number of relevant reference IPs depending on the query IP by finding their sparsest linear combination. Second, because of the large differences between the two cameras settings, IP descriptors might get partially transformed due to changes in camera view angle or lighting. The kNN approach is non convenient in this case as these IPs are significantly different from the corresponding reference points. SR is more suitable as it does not search for close reference points each taken separately, but for the sparsest combination of reference IPs that is close to the reference point even though no reference IP is close by itself to the query IP.

Table 1: List of acronyms

Acronym	**Description**
BoW	Bag of visual Words
IP	Interest Point
KDTree	K-Dimensional Tree
kNN	k Nearest Neighbors
ROI	Region Of Interest
SR	Sparse Representation
SURF	Speeded Up Robust Features

The rest of the paper is organized as follows. In section 2, the state of the

art for people re-identification is presented. Our approach is detailed in section 3. Section 4 gives an overview of extracted features: static, namely SURF and spatio-temporal, namely Cuboids. Section 5 describes the principle of sparse representation in the context of interest points matching. The fusion scheme is presented in section 6. Afterward, results are presented in section 7. Section 8 presents the phase of filtering unreliable IPs or matched IP pairs and finally a conclusion is drawn and future directions are discussed in section 9. Table 1 shows a list of the most used acronyms in the rest of this paper.

2. Related work

We categorize the re-identification approaches into supervised and unsupervised ones. The first category consists of dividing the database into training and testing parts. The training set consists of pairs of a *priori* matched images. It allows to optimize the parameters of the re-identification model. Unlike the first category, an unsupervised approach consists of re-identifying people without any prior information. We present next a state of the art of re-identification approaches and on the use of sparse representation in re-identification task.

2.1. Unsupervised approaches

The choice of descriptors and their combination to obtain a robust person representation is a challenge for unsupervised methods. This representation can be based on two categories: interest points and on division into regions.

For the former category, in [5], a person model is represented by IPs, namely SURF, collected over a short video sequence. For matching each query SURF, a vote is added to the person associated with the nearest reference SURF. The person with the majority of votes is claimed as the re-identified one. In [6], the authors propose a three stage system, i.e. person detection, tracking and re-identification. The person detection and tracking steps are based on the Implicit Shape Model (ISM), while the re-identification is based on matching of BoW of descriptors formed by addition of SIFT interest point (for "Scale Invariant Features Transform") and a spatial description (the position of IP in the image).

For the latter category, in [7] the ROI is segmented into K horizontally bands, and then represented by the concatenation of a color description extracted from its K-bands. For matching, a query ROI is represented as a sparse combination of all references ROIs. Then, a reconstruction error is calculated for each reference identity using only the coefficients corresponding to this identity. The query ROI is assigned to the reference identity minimizing the reconstruction error. The authors used the same matching algorithm presented in [8] for face recognition, as the input to SR was the whole ROI. In [9], the human body is decomposed via two horizontals axes, into three parts that generally correspond to head, torso and legs. Each part is represented by a combination of three features related to texture and color. The similarity between two images is defined as the weighted sum of the three associated distances. In [10], the authors represent the person by a combination of descriptive and discriminative descriptors. The matching is carried out in two steps. In the first, the descriptive model is used to generate the most similar 50 reference images to the query. In the second step, the discriminative model is applied to refine the first classification based on the Adaboost algorithm. For feature extraction, the authors use the covariance image descriptor [11] and the Haar wavelet. In [12], the re-identification problem is considered as a ranking problem. For person description, the image is divided into overlapping horizontal strips. From each strip, Hue-Saturation and RGB histograms are extracted. Moreover, a HOG (for "Histogram of Oriented Gradient") descriptor is extracted and concatenated with the two previous features to form the final descriptor. For matching, the authors use an iterative sparse representation to rank the reference people.

2.2. Supervised approaches

Supervised approaches require the existence of a training set composed of a *priori* matched pairs of images. Most of these approaches are based on metric learning algorithms. The latter consist of learning the parameters of the metric defining the similarity between two images. The choice of the metric at i) minimizing the similarity between images pairs associated with different per-

sons (negative image pairs) and *ii*) maximizing similarity between images pairs associated with the same person (positive image pairs).

Regarding learning metrics, an "Ensemble of Localized Features" (ELF) is introduced in [13]. The main idea consists of combining a set of features to represent a person and then estimating a weight for each of them. These weights are estimated over a training dataset using the Adaboost algorithm. The similarity function between two images is based on the learned features weights. Prosser et al. [14] formulated person re-identification as a ranking problem. The Authors used different ranking algorithms such as RankBoost or RankSVM to learn the pairwise similarity metric. In [15], the authors learn the Mahalanobis distance and then a query person is matched via the k-nearest-neighbor algorithm. The proposed solution is called Large Margin Nearest Neighbor with Rejection (LMNN-R) ("rejection" means the classifier returns no matches if all neighbors are beyond a certain distance). In [16], the re-identification task is formulated as a problem of "Probabilistic Relative Distance Comparison" (PRDC). The distance probabilities are learned by maximizing distances between negative images pairs and minimizing distance between positive image pairs. In [17], the similarity between an image pair is measured using the cosine distance between their projections on the RCCA space (RCCA for "regularized Canonical Correlation Analysis Method" [18]). The latter method constructs a projection space where correlations between positive images are maximal.

In this paper, we adopt both the unsupervised and supervised schemes. The first allows us to compare the performance of our system with the state of the art on video sequences that is mainly based on unsupervised setting. The second one, i.e. supervised scheme, allows us to design a trainable IP filtering scheme that optimizes our SR-based matching algorithm.

2.3. Sparse representation in re-identification

Sparse representation (SR) of signals has been studied since two decades [19, 20], but it became popular in computer vision, only recently after its successful application to face recognition [8]. Since then, it also has been used in other

classification tasks such as gait recognition, speech recognition and person re-identification [7, 12]. In the latter two works, SR is used to match silhouettes. The use of sparse representation in this work is different from [7] from two standpoints. First, SR in [7] is global (Silhouette representation as a whole), while ours is local (SR for each IP in the silhouette). Second, in [7], for each SR, the dictionary consists of all reference samples, while in our approach we use a reduced and dynamic dictionary of few selected reference IPs. The SR considered in [12] is also global like the one presented in [7]. Moreover, in [12], an iterative and weighted sparse representation is applied: at each round, the coefficients contributing little to the reconstruction are eliminated; SR weights are then updated and SR is again computed to rank the other identities.

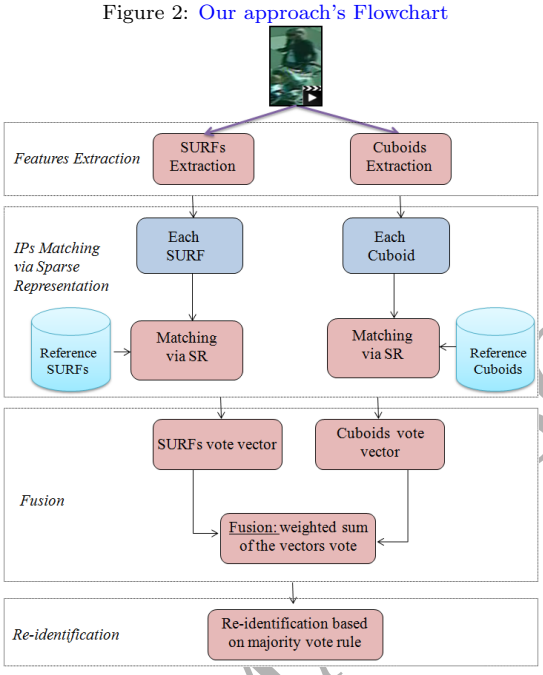
3. The Proposed Approach

Our approach basically consists of four stages: 1) Feature extraction, 2) IPs Matching via sparse representation, 3) Fusion and 4) Person re-identification based on majority vote rule. Figure 2 shows the flowchart of our approach.

First, for each input video, two sets of features, spatial and spatiotemporal, are extracted to model the signature (description) of a person identity: SURF and Cuboids. SURF analyzes each image from the input video and generates 2D spatial IPs. Then, an appearance descriptor around each IP is computed. Cuboids, on the other hand, is applied on the whole video to output 3D spatiotemporal IPs. Then, an appearance descriptor is extracted around each IP. Thus, the spatiotemporal features are pseudo-motion related as their detection relies on motion while their description is based on appearance.

The IP matching task is performed via sparse representation. Each test IP is matched by expressing it as a linear combination vector of a set of references IPs of the same type (SURF or Cuboids). This set is called a dictionary. The obtained representation corresponds to a sparse vector whose nonzero entries are related to the weights of reference IPs. This representation is harnessed for query matching, by assigning the query to the reference IP minimizing the reconstruction error of this query. In our IP-based SR, the dictionary would

Figure 2: Our approach's Flowchart



consist of all reference IPs. For large video datasets, such a dictionary would include millions of IPs, thus hindering the feasibility of sparse representation. To overcome this issue, a dynamic and reduced dictionary is selected for each IP from the reference set.

Third, for each query SURF, a vote is added to the person associated with the reference identity the query is matched to. In this way, a vote vector of dimension equal to the number of reference identities is generated. Similarly, all query Cuboids are matched and a second vote vector is generated. Then, a fusion of the two vote vectors is performed by a weighed sum the parameters of which are derived based on the density of each type of IP (spatial and spatiotemporal) in the input video sequence. Finally, the reference person obtaining the majority of votes is claimed as the re-identified person.

A pseudo-code of re-identifying one query sequence from a set of reference sequences is presented in Algorithm 1.

Algorithm 1 Re-identify a test sequence from a set of reference sequences

Require: - A reference dataset of M identities (persons)
- A query video sequence

Ensure: - Identity of the query video sequence

****

// Feature extraction

Extract SURFs and Cuboides from the reference sequences

Extract SURFs and Cuboides from the query sequence

****

// SURF Matching via Sparse Representation

for each query SURF **do**

Find the matched reference identity ref_id via Sparse Representation

Increment the number of votes associated with ref_id

end for

$V \leftarrow$ each component $V(i)$ corresponds to the sum of votes associated with the reference identity i

****

// Cuboides Matching via Sparse Representation

for each query Cuboide **do**

Find the matched reference identity ref_id via Sparse Representation

Increment the number of votes associated with ref_id

end for

$W \leftarrow$ each component $W(i)$ corresponds to the sum of votes associated with the reference identity i

****

// Fusion

$\alpha \leftarrow$ weight parameter

$Z \leftarrow$ each component $Z(i)$ corresponds to the weighted sum of votes associated with the reference identity i

$Z = V + \alpha W$

****

// Re-identification

The query sequence is re-identified as the reference sequence having the maximum of votes in Z

Identity of the query video sequence $= \arg\max_i(Z(i))$

4. Feature extraction

Given an input video sequence, two IPs-based features, namely, SURF and Cuboids, are extracted over two stages: detection and description (Figure 3).

Table 2 summarizes the two features characteristics. In this section, we present an overview of SURF and Cuboids.

Figure 3: Diagram of features detection/description based on interest points

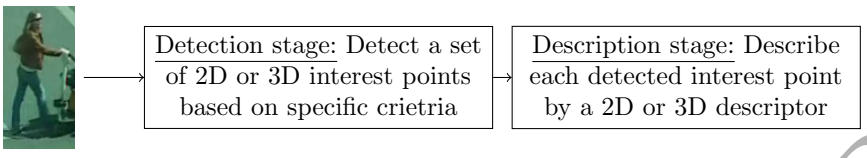


Table 2: Extracted features descriptions

Features	SURF	Cuboids
Encoding information	Appearance	Motion
Detection	Hessian Matrix Based Detector	Gabor filter
Description	Haar-Wavelet (64 components)	

4.1. Overview of the SURF method

SURF [21] operates in two main stages: detection and description. The detection stage analyzes an image and returns a list of interest points. Around each point, a descriptor vector is computed. The detection stage is performed through the Hessian Matrix, while the description stage is based on Haar wavelets.

The detection stage

The SURF detector is based on the approximation of the Hessian matrix determinant and the use of the integral image. It presents a good compromise between robustness to geometric transformations and computation time. Given a pixel $p = (x, y)$ of the image intensity I and integrating the scale information, the Hessian matrix $H_\sigma(p)$ at position p and scale σ, is defined by Eq.1.

$$H_\sigma(p) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{yx}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix} \quad (1)$$

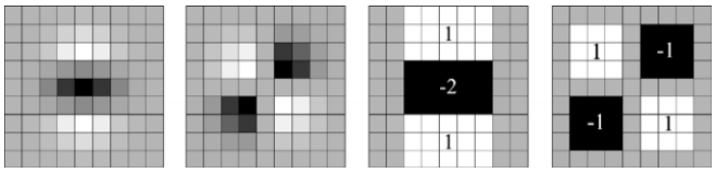
where the Laplacian $L_{xx}(p,\sigma)$ (respectively, $L_{yy}(p,\sigma)$, $L_{xy}(p,\sigma)$ and $L_{yx}(p,\sigma)$), refers to the convolution of the second order Gaussian derivative $\frac{\partial^2 g_\sigma}{\partial x^2}$ (respec-

tively, $\frac{\partial^2 g_\sigma}{\partial y^2}$, $\frac{\partial^2 g_\sigma}{\partial xy}$) and $\frac{\partial^2 g_\sigma}{\partial yx}$) with the intensity image I at position p (Eq.2).

$$L_{xx} = I * \frac{\partial^2 g_\sigma}{\partial x^2}, \quad L_{yy} = I * \frac{\partial^2 g_\sigma}{\partial y^2}, \quad L_{xy} = I * \frac{\partial^2 g_\sigma}{\partial xy}, \quad L_{yx} = I * \frac{\partial^2 g_\sigma}{\partial yx} \tag{2}$$

where * is the convolution product and $g_\sigma(p) = \frac{1}{2\Pi\sigma^2} e^{-\frac{\|p\|_2^2}{2\sigma^2}}$. SURF proposes to estimate the Laplacians by estimating the second order Gaussian derivative using a set of "box" filters. Figure 4 shows the appropriate filters used to estimate $\frac{\partial^2 g_\sigma}{\partial y^2}$ and $\frac{\partial^2 g_\sigma}{\partial xy}$. In the following, the estimation of L_{xx}, L_{yy} and L_{xy} are respectively named D_{xx}, D_{yy} and D_{xy} (Eq.3).

Figure 4: Left to right: the (discretized and cropped) Gaussian second order partial derivatives in y-direction and xy-direction, and approximations thereof using box filters: $filter_{yy}$ and $filter_{xy}$. The grey regions are equal to zero [21]



$$L_{yy} \approx D_{yy} = I * filter_{yy} \tag{3}$$

where the convolution between two functions A and B of two discrete variables i and j is defined as follows:

$$C(i,j) = \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} [A(k,n)B(i-k, j-l)] \tag{4}$$

The determinant of the Hessian $H_\sigma(p)$ is finally approximated as follows (Eq.5), where w is a constant empirically selected [21]:

$$\det(H_{approx}) = D_{xx}D_{yy} - (wD_{xy})^2 \tag{5}$$

The SURF detector uses the integral image to accelerate the convolution calculation. Once the determinant at each pixel is estimated, the maxima are searched in small neighborhoods (typically volumes of 3x3x3 pixels). These maxima correspond to SURF IPs. Then, each IP is described as follows.

The description stage

The description stage consists of two steps: orientation assignment and descriptor extraction. First, a characteristic orientation is estimated to ensure invariance to image rotation. For each pixel in a circular region around the IP, Haar wavelet responses are calculated and then weighted with a Gaussian centered at the IP. Each weighted response is interpreted as a 2D vector (x-response, y-response). Using the sliding window technique, for each window, all x-responses (d_x) and y-responses (d_y) are summed to one vector originating at the IP. The maximum resulting vector over all sliding windows determines the orientation assignment.

In the descriptor extraction step, we consider a square region around the IP, oriented according to the first step. This region is divided into 4x4 grids to form 16 sub-regions. Within each sub-region, Haar wavelet x-responses and y-responses are calculated at 5x5 equally spaced points. For these 25 sample points, we collect four components (Eq.6).

$$v_1 = \sum d_x,\ v_2 = \sum d_y,\ v_3 = \sum |d_x|,\ v_4 = \sum |d_y| \tag{6}$$

The collected components from each of the 16 sub-regions form the 64-dimensional SURF descriptor.

The Haar-like features have shown high discriminative power in computer vision tasks such as face detection [22] and person detection [23]. Besides, thanks to the precomputing of the integral image, the features can be computed efficiently in any part of the image.

4.2. Overview of the Cuboids method

The Cuboids [2] method is proposed originally to characterize periodical motions in video sequences. It also operates in two stages: detection and description.

The detection stage

The Cuboids detector is based on Gaussian and Gabor filters. The selection criterion R of Cuboids is the following (Eq.7):

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{7}$$

where h_{ev} and h_{od} define the Gabor filter:
$h_{od}(t, \tau, \omega) = -\sin(2\pi t\omega)e^{-\frac{t^2}{\tau^2}}$ and $h_{ev}(t, \tau, \omega) = -cos(2\pi t\omega)e^{-\frac{t^2}{\tau^2}}$, τ is an independent temporal scale value, $w = \frac{4}{\tau}$, and $g(x, y, \sigma)$ is a Gaussian smoothing function (Eq.8):

$$g(x, y, \sigma) = \frac{1}{2\Pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{8}$$

σ being an independent spatial scale value.

For each pixel, a response R is measured. Then the locations of Cuboids correspond to the points giving high responses of R [24], i.e., to the local maxima of R with respect to the 3D coordinates (x,y,t).

The description stage

Cuboids are originally used for action recognition, for which each action is acquired several times under similar training and test view angles. In this case, standard Cuboids' descriptors can encode motion sufficiently well for action representation. In the re-identification setting where only one reference sequence and one test sequence are available, the two associated view angles might be significantly different, which leads to huge discrepancy in their motion descriptors, entailing, thereby, a strong mismatch between the two sequences. For this reason, we have adapted the Cuboids descriptors to our re-identification context, by investigating the description of Cuboids by SURF and HOG representations.

5. IPs Matching via sparse representation

IPs matching can be performed in two ways: via a bag of features or directly for each IP. The former consists of converting the query IPs into a bag of k

clusters (obtained from the reference dataset using a clustering algorithm (e.g. k-means)) and then representing the image (or video) as a histogram of the occurrence of each cluster [25]. The latter consists of determining for each query IP the closest IP from the reference set. The closest element (IP) is defined as follows: given a set of elements $\{x_i \in \mathbb{R}^D, i = 1 \dots M\}$ and $y \in \mathbb{R}^D$, x_j is the closest element to y if $\forall i, d(y, x_j) \leqslant d(y, x_i)$, where d is a predefined distance measure. In this paper, we adopt the second scheme and perform IP matching via sparse representation. We first present the principle of sparse representation and we detail how we harness it in the person re-identification context.

5.1. Sparse representation principle

Sparse representation [19, 26] consists of representing a signal as a linear combination of the smallest number of elements of a preselected dictionary. Given a signal $y \in \mathbb{R}^D$ and a dictionary $\Phi \in \mathbb{R}^{D \times K}$ and given that usually $D \ll K$, i.e., the size of the dictionary is larger than the dimension of the input, our linear system is underdetermined [27, 28]. Therefore, there is an infinite number of solutions α corresponding to a linear subspace verifying Eq.9. In some cases, when there are inconsistencies in the linear equations, no solution can be obtained, but these cases rarely appear in practical image processing applications (e.g. face recognition as in [8] or in our own experiments).

$$y = \Phi\alpha \tag{9}$$

From all solutions of the underdetermined linear system above (Eq.9), we are interested in the sparsest solution α_s, that can be sought by minimizing the l_0-norm of α, i.e. the number of its non zero-coefficients. Hence, the original formulation of the sparse representation problem is as follows (Eq.10):

$$\alpha_s = \min_{\alpha} \|\alpha\|_0 \quad \text{subject to} \quad y = \Phi\alpha \tag{10}$$

To solve Eq.10, numerous algorithms have been proposed in the state-of-the-art. In the literature, sparse representation algorithms are categorized differently (as reported, for instance, in the survey in [29]). In most cases, sparse

representation categories include: greedy pursuit approach category and convex relaxation approach category. A greedy pursuit approach iteratively refines the current estimated α_s by selecting one or several dictionary atoms sequentially chosen to approximate the signal y. Examples of "greedy pursuit methods" include the Matching Pursuit (MP) [20] and the Orthogonal Matching Pursuit (OMP) [20]. On the other side, a convex relaxation approaches suggests a convexification of the problem presented in Eq.11 by replacing the l_0-norm with l_1-norm. In fact, the l_0-norm optimization problem in Eq.10 is NP-Hard and recent research showed that finding the solution of l_0-norm optimization problem is, in most practical cases, equivalent to the l_1-norm optimization problem (Eq.11) [30] [31]:

$$\alpha_s = \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad y = \Phi\alpha \tag{11}$$

Generally, in real cases where data are noisy, the equality constraint in Eq.11 can be relaxed to allow some error tolerance ($\varepsilon > 0$) as follows (Eq.12):

$$\alpha_s = \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad \|\Phi\alpha - y\|_2^2 < \varepsilon \tag{12}$$

The elements of Eq.12 can be combined in several ways to obtain equivalent formulations. Eq.12 seeks the sparsest possible solution at a given error ($\varepsilon > 0$). We can also seek the minimal possible error given the sparsity level ($\delta > 0$) as follows (Eq.13) :

$$\alpha_s = \min_{\alpha} \frac{1}{2}\|\Phi\alpha - y\|_2^2 \quad \text{subject to} \quad \|\alpha\|_1 < \delta \tag{13}$$

Usually, a tuning parameter λ is used to adjust the tradeoff between sparsity and error reconstruction (Eq.14):

$$\alpha_s = \min_{\alpha}(\|\Phi\alpha - y\|_2^2 + \lambda\|\alpha\|_1) \tag{14}$$

The function $f : \alpha \mapsto \|\Phi\alpha - y\|_2^2 + \lambda\|\alpha\|_1$ is convex and in order to minimize f, we need specific algorithms such as Coordinate Descent [32] and Least Angle Regression (LARS) [26].

5.2. Sparse representation for Interest Point Matching

In the context of IP matching, let us consider a query IP y and a set of reference IPs associated with M identities (persons). First, the reference dataset is arranged in a matrix (called dictionary), which is built using reference IPs: $\{S_{i,j}\} \in \mathbb{R}^D, i = 1 \dots M, j = 1 \dots k_i$, where k_i denotes the number of reference IPs for the i-th identity, and $K = k_1 + k_2 + \cdots + k_M$ denotes the number of IPs in the reference dataset. The k_i reference IPs of the i-th identity candidate constitutes the columns of the matrix Φ_i (Eq.15):

$$\Phi_i = [S_{i,1}; S_{i,2}; \dots ; S_{i,k_i}] \tag{15}$$

The matrix Φ of all K reference IPs is obtained by concatenating the Φ_i matrices (Eq.16):

$$\Phi = [\Phi_1; \Phi_2; \dots ; \Phi_M] = [S_{1,1}; S_{1,2}; \dots ; S_{M,k_M}] \tag{16}$$

The Sparse Representation scheme represents y as a linear combination of all the reference IPs:

$$y = \Phi\alpha_s = [\Phi_1, \Phi_2, \dots \Phi_M]\, \alpha_s \tag{17}$$

At this step, Eq.14 is applied to find the sparsest coefficient vector α_s. In noiseless conditions and if there is no ambiguity between the samples pertaining to different classes (person identities), the nonzero entries α_s are associated with the columns of Φ from a single person identity class i (Eq.18), and the query sample y is assigned to the i-th person identity [8].

$$\alpha_s = [0; \dots ; 0; \alpha_{i,1}; \alpha_{i,2}; \dots ; \alpha_{i,k_i}; 0; \dots ; 0]^T \tag{18}$$

In this ideal case, α_s has nonzero entries associated only with the i-th reference identity corresponding to actual identity of y [33]. In noisy conditions, however, non null coefficients may be associated with multiple person identities.

After calculating the sparsest solution, the non-zero coefficients of α_s can be used to determine the identity of the query IP y.

5.3. Sparse representation for IPs matching

In our approach, each IP is matched independently. To match one IP, three steps are applied: 1) Dictionary construction, 2) Sparse representation and 3) Identity assignment.

5.3.1. Dictionary construction (Algorithm 2)

The dictionary consists of all references IPs. Thus, for typical re-identification datasets, we would have to consider dictionaries of millions of reference interest points. The computation time for finding a sparse representation for one IP will then be huge. In this work, for each test IP, we select a dynamic and reduced dictionary A, consisting of the N closest reference IPs. As there is no theoretical method to infer the optimal dictionary size N, we set the latter in an empirical way by choosing a value not too small which may lead to missing relevant interest points, but not a too large value, for which a lot of irrelevant IPs will be introduced. The dimension of A is DxN where each column represents an IP descriptor of dimension D. To accelerate the search of nearest neighbors, a KD-tree is used [4].

The concept of dynamic (adaptive) dictionary has also been proposed in [34]. The dictionary construction in [34] is different from ours. In [34], an "Extreme Learning Machine" (ELM) is applied on the query element and only the first k largest reference elements in the ELM output are taken into consideration. This is due to the fact that the uncorrelated classes to the query element tend to have a small response in an ELM output. Then, all the points in the reference set that share the same classes (identities) as the k-largest entries are added to the dictionary. In our work, the dictionary corresponds directly in a dynamic way to the k-closest points, and we do not look for any additional points in the reference set with the same classes (identities).

5.3.2. Sparse representation (Algorithm 3)

In real-life conditions, re-identification is complex and dictionary IPs are very noisy. Our sparse representation formulation is based on equation Eq.14. To solve the sparse representation task, we use the Coordinate Descent algorithm

Algorithm 2 Dictionary construction

Require: - ϕ: Matrix containing all reference IPs
- N: Dictionary size
- q: query IP

// Tree construction
ϕ_{Sorted} = Sort-KD-Tree$(tree, q)$
// Dictionary construction
$A = [A_i]_{1 \leqslant i \leqslant N} = N$ first (closest) IPs from ϕ_{Sorted}

Ensure: - A: dictionary

(CD algorithm) [32] . This algorithm uses a regularized technique that copes with the ill-posed problem [35]. The Coordinate Descent algorithm inherits its numerical stability from the descent property at each iteration update [32, 36]. Moreover, the CD algorithm, according to the literature, is efficient if the correlation between the dictionary elements is small [37]. The algorithm takes as input an IP and its corresponding dictionary, and outputs a sparse vector with most coefficients equal to zero.

Algorithm 3 Sparse Representation

Require: - q: query IP
- A: dictionary corresponds to q
- λ: tuning parameter

// Solve the following equation with the Coordinate Descent algorithm
$\alpha_s = \min_{\alpha}(\frac{1}{2}||q - A\alpha||_2^2 + \lambda||\alpha||_1)$

Ensure: - α_s: sparse vector (most coefficients set to 0)

5.3.3. Identity assignment (Algorithm 4)

The non-zero coefficients of the sparse representation are used to assign an identity to a query IP. To determine IP identity, a reconstruction residual is calculated for each identity i having at least one non-zero coefficient in the following manner: denote L the number of identities having at least one non-zero coefficient in α_s. Let's q be a query IP and A its corresponding dictionary, we compute first $x_i : 1 \leqslant i \leqslant L$ as following (Eq.19):

$$x_i = [0, \ldots 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,k_i}, 0, \ldots, 0] \tag{19}$$

x_i is a coefficient vector obtained from α_s with all elements set to 0 except those associated with identity i. The dominant identity j satisfies the following equality (Eq.20):

$$j = \arg\min_i \|Ax_i - q\|_2^2 \tag{20}$$

j corresponds to the identity minimizing the reconstruction residual of q. Based on reconstruction error minimization, the identity of query IP is identified as the person j satisfying Eq.20 .

Algorithm 4 Identity assignment

Require: - q: query IP
- α_s: SR corresponds to query IP q and dictionary A

for each identity i having k_i non-zero coefficients **do**
// Compute x_i
$x_i = [0, \ldots, 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,k_i}, 0, \ldots 0]$
// Compute the reconstruction residual r_i
$r_i = ||Ax_i - q||_2^2$
end for
$j = \arg\min_i(r_i)$

Ensure: - j: identity of q

6. Fusion scheme

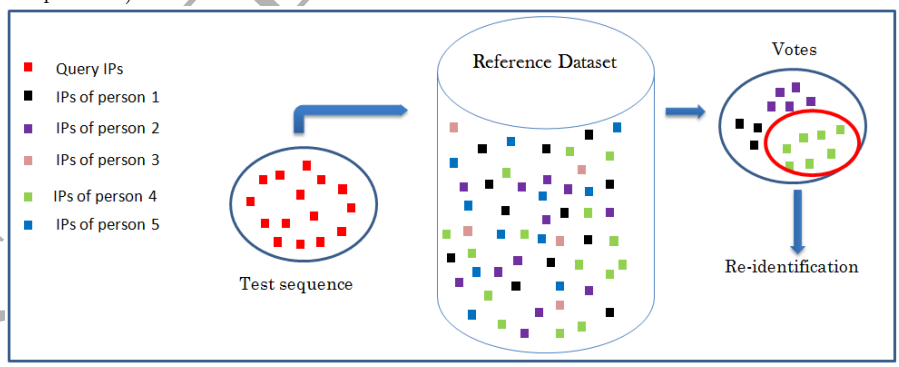
In the previous sections, we have proposed two independent person re-identification systems based on 2D static and 3D spatio-temporal IPs. Each of the two systems relies on SR for IP matching. As the 2D and 3D IPs emanate from two different sources of information (static and temporal respectively), they actually complement each other. 2D static points are usually detected on regions with rich texture structure and contrast, and can for instance detect textured person clothes that can help identification. They may however also detect texture background parts in the scene. Temporal IPs do not in general detect such regions but focus on those with high spatiotemporal variance associated with moving legs or arms. This complementarity motivates to seek fusion of the two systems for improved accuracy. Fusion can occur at any of the three

important biometric stages: 1) Feature extraction, 2) Matching and 3) Decision: 1) at the feature level, features are concatenated and fused into the same feature vector prior to matching; 2) at the matching level, the scores generated by different classifiers are combined into one score; 3) at the decision level, decisions of different classifiers are fused into a single decision, e.g. through a majority vote rule.

In our context, features' fusion is not possible because SURFs and Cuboids are detected at different interest points. Decision fusion is not possible either as it is not realistic to combine the decisions of only two classifiers. We use instead a score-level fusion method where the score is the number of votes for each person generated by each classifier (SURF-based and Cuboids-based).

Figure 5 illustrates the principle of majority vote rule. Given a query sequence and a reference dataset, each IP from the query sequence is classified into one identity from the reference dataset via SR as explained in section 5. Then, the found reference identities are submitted to the majority vote decision rule. For each query IP, a vote is added to the person associated with the reference selected identity. The person obtaining the majority of votes is claimed as the re-identified person.

Figure 5: Principle of majority vote rule (In this example, the query person is re-identified as the person 4).



After all query SURFs and Cuboids are matched, two votes' vectors are

generated. The dimension of each vector is the number of reference identities and each component reflects the number of times the associated reference has been matched to a query IP in the test sequence. The fusion strategy is based on a weighted sum of the two vote vectors. It then applies the majority vote rule: the query person is claimed as the person with the majority of votes.

Formally, given a query video sequence, described by n SURFs and m Cuboids, the goal is to determine the associated identity among M reference persons. Let V and W be the vote vectors corresponding to SURFs and Cuboids respectively. The vote vector Z obtained by fusion is written as equation Eq.21:

$$Z_i = \alpha V_i + (1 - \alpha) W_i \tag{21}$$

where α is a weighting parameter. As our approach is totally unsupervised, there is no way to learn the optimal value of α, so we derive it based on our a *priori* knowledge on the extracted 2D and 3D IPs. Assuming that 2D static IPs and 3D spatio-temporal IPs are equally discriminative, and giving that the number of detected 2D points is roughly twice the number of detected 3D points, we have set α to 0.3. With this fusion scheme, the re-identified query person is assigned to the reference Z_i with the maximum of votes.

7. Experiments

We evaluated our approach on two public multishot re-identification datasets: PRID-2011 and CAVIAR4REID. PRID-2011 [10] consists of hole video sequences and is the only available and large public video database that is adequate for the re-identification task. Although, CAVIAR4REID [38] does not consists of video sequences but only of a few unordered key frames, we used it as well in order to assess the effectiveness of the sparse representation in an additional experiment setting.

As SR parameters, the dimension of the dictionary A is DxN where each column represents an IP descriptor of dimension $D = 64$, and N is empirically set to 200. To compute the SR coefficients, the Coordinate Descent algorithm is used.

Results are shown in terms of the Cumulative Matching Characteristic (CMC) curve as commonly used in the literature. In the CMC curve, the identification rate at rank i gives the number of query video sequences where the actual reference is retrieved among the top i answers over the size of the query set. In the rest of the paper, by default, the re-identification rate will mean the re-identification rate at rank 1. It is also called the Correct Classification Rate (CCR).

Figure 6: Samples from different evaluated datasets: Left: CAVIAR4REID; Right: PRID-2011

7.1. Results on CAVIAR4REID

CAVIAR4REID [38] has been extracted from the CAVIAR database [39]. The recorded videos were captured from two different cameras in an indoor shopping center in Lisbon. The pedestrians' images have been cropped using the provided ground truth. From the 72 different individuals identified (with image sizes varying from 17x39 pixels to 72x144 pixels), 50 people are captured by both views and 22 from only one camera. For each pedestrian, 10 images from each camera view are selected, maximizing the variance with respect to resolution changes, light conditions, occlusions, and pose changes (see samples in Fig.6). These images or key frames are unordered and thus no spatio-temporal IP can be extracted. For CAVIAR4REID, therefore, only SURF-based SR is considered.

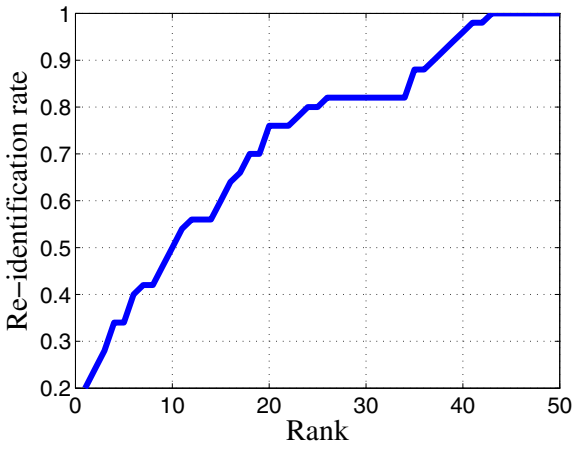
The CMC curve obtained by our approach is shown in Fig.7. Table 3 shows different state-of-the-art methods' performances (identification rate at rank 1). These methods share with us the same evaluation protocol.

Table 3: Results comparison on CAVIAR4REID

Approach	Re-identification rate (%)
Authors of [40] appear in [41]	10
[41]	10
SURF-1NN [42]	16
[38]	17
[43]	19
Our approach (only SURF)	20

The approaches based on appearance features (SDALF) [41] and MRGC [40] using essentially color descriptors achieve 10% of correct re-identification. The approach presented in [42], that is based on 1NN-based SURF matching using Euclidian distance and probabilistic filtering, achieves 16% of correct re-identification. Our approach achieves a re-identification rate of 20%, which is slightly higher than the one obtained in [43] and based essentially on spatiotemporal color features. CMC curves sources of [38, 41, 43] on CAVIAR4REID are not available to reproduce along with ours in the same figure. Overall, these results illustrate the power of SR for IPs matching compared to 1-NN.

Figure 7: CMC SURFs performance on CAVIAR4REID



In recent state of the art, the dataset CAVIAR4REID was evaluated with a supervised protocol. The dataset is divided into two parts : training and test.

The partition (36, 14), 36 people in training and 14 people in test, is widely used in the literature. Among these approaches that adopted this partition, [44] achieved a CCR = 36.19%, [45] achieved a CCR = 49.1% and a CCR = 32.86% is achieved in [46]. Recently, in other works like [47], all the 50 people are used in test, but only half of the available images are used in the test, while the remaining 5 frames are used as reference images. In evaluation, [47] achieved a CCR=35.2%. For this diversity of evaluation protocols on CAVIAR4REID, we compared our approach only with those using exactly the same protocol as ours. This comparison is presented in the table 3.

7.2. Results on PRID-2011

The dataset PRID-2011 (multi-shot version) [10] was created in 2011 by the Austrian Institute of Technology. The video sequences were obtained from two cameras (A and B) located on street (Fig.6). 385 people were filmed by camera-A and 749 people were filmed by Camera-B, 200 being common to the two cameras. The evaluation consists of searching the common 200 people filmed by Camera-A in the gallery set (Camera-B) of 749 people.

In the fusion stage, we selected a value of α proportional to the average number of IPs per person (Tab.4), roughly $\alpha = 0.3$, and compared the fusion result to that where the two systems are considered on equal footing, i.e., $\alpha = 0.5$.

Table 4: Average number of IPs per person

IPs	Average number of IPs per person
SURFs	1733
Cuboids	1030

Figure 8 shows the CMC (from rank 1 to rank 10) on PRID-2011 compared to the state-of-the art. Table 5 compares the re-identification rate at rank 1 of different methods. We see that the approach based on SURF matching via SR outperforms SURF matching via 1-NN, and achieves an improvement of 4.5 % in the re-identification rate at rank 1.

Figure 8: Comparison of CMC performances on PRID-2011

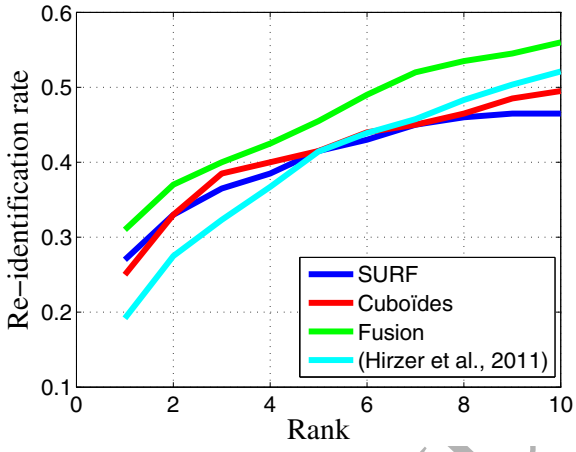


Table 5: Results comparison on PRID-2011

Approach	Re-identification rate (%)
Using all test dataset	
[10]	19.2
SURF-1NN [42]	22.5
Our approach (only SURF-SR)	27
Our approach (only Cuboids-SR)	25.5
Our approach (Fusion): $\alpha = 0.5$	28.5
Our approach (Fusion): $\alpha = 0.3$	31
Using only the 200 common people in evaluation	
Our approach	31
[48]	28.9

Our fusion approach compares favorably with the approach proposed in [10] that combines two appearances models. These results show that our fusion mechanism improves the system performance by 4% w.r.t using SURFs alone and 5.5% w.r.t using Cuboids alone.

Another result on the PRID-2011 database was published using a different protocol [48]: only the 200 common people in camera-A and camera-B are used in evaluation. Using this protocol, our approach outperforms that in [48] by 2.1% at rank 1 (Tab.5). Table 5 shows the comparison of our approach with the state of the art using only the 200 common people in evaluation.

7.3. Discussion

Two important conclusions can be drawn from the previous results: the first concerning the usefulness of SR and the second concerning the effectiveness of the fusion of appearance and motion features.

The approach based on SURF matching via SR outperforms that based on SURF matching via 1-NN. On PRID-2011, SR outperforms 1-NN at all ranks and achieves an improvement of 4.5% in the rate of re-identification at rank 1. This improvement is significant given the large size of the dataset and proves that the SR can provide richer information for decision making than [42] and other interest point matching methods like [5]. On CAVIAR4REID, SR is better again, but the improvement is small compared to the one obtained on PRID-2011. This may be explained by the small size of the database and the much fewer images available per person compared to PRID-2011.

The contribution of SR compared with the 1-NN is highlighted in cases where some reference SURFs are close to query SURF of the same person without being the closest though. Considering several neighbors to compute the SR is more effective than 1-NN matching.

Regarding fusion, we demonstrate the complementarity of the static aspect of person's appearance on one hand and the dynamic aspect related to movement on the other hand, for detecting discriminant interest points.

8. Supervised schemes for optimizing the re-identification task

Overall, the learning phase allows learning parameters to optimize the re-identification task. This optimization would be relevant if it is associated with increasing the re-identification rate or reducing the re-identification running time. The supervised parameters, as shown in current state-of-the-art, can be related to the metric used to compare images.

Recall that an unsupervised approach has been presented in this work under the same experimental protocol of the state-of-the-art, which allowed a fair comparison with the latter. Using this unsupervised protocol, all the dataset

PRID-2011 is considered as test dataset. In this section, we go beyond this experimental setting in order to show the usefulness of a learning phase: our approach (SURF only) is evaluated on the PRID-2011 using a supervised protocol. We divide the dataset PRID-2011 into two parts: training and test. The training set contains videos of the first 100 people shared by Camera-A and Camera-B. Each person has one video sequence per camera (A and B). The test set contains the remaining 649 people from Camera-B as reference and the remaining 100 common people in Camera-A as query. The training dataset is used to estimate a filtering model, added to our unsupervised approach, that aims at rejecting unreliable matched IPs pairs. Such matched pairs may result, for instance, from: 1) matching a background's IP to a foreground's IP (silhouette), 2) matching IPs associated with different parts of the silhouette and 3) matching IPs associated with different people.

The core of the filtering stage is based on the use of a Support Vector Machine (SVM) learned on the training dataset [49]. The SVM takes as input positive IP pairs (each pair {query IP, closest reference IP} is associated with the same person) and negative IP pairs (each pair {query IP, closest reference IP} is associated with different persons. For a query IP pair , SVM decide if it is associated with the same person or different persons. Then, only IP pairs associated each with a same person are used for re-identification.

In other words, to use SVM, two steps are needed: 1) model estimation and 2) class prediction. First, for model construction, SVM takes as input two vector sets: S_{Same} (positive vectors associated with class $+1$) and S_{Diff} (negative vectors associated with class -1). S_{Same} and S_{Diff} model respectively reliable and unreliable IPs. To construct S_{Diff} and S_{Same} from the training dataset, IPs corresponding to camera A are matched to those corresponding to camera B to form IP pairs; if the matched pair is associated with the same person, the difference vector between the pair descriptors is added to S_{Same}, else it is added to S_{Diff}. After sets S_{Same} and S_{Diff} are generated in this way, the associated SVM model is estimated by computing the hyperplane separating S_{Same} and S_{Diff} in the non-linear space defined by the RBF kernel (for "Radial Basis

Function kernel") [50]. Second, for class prediction, the filtering step consists of assigning a query IP to one of the two classes using the SVM-based filter model. The Acceptation/Rejection decision is performed as following: each query IP is matched to the closest reference IP. Then, the matched IP pair difference descriptor is input into the SVM model. If SVM's output decision is class +1, the query IP is retained, otherwise it is rejected. This supervised approach is evaluated and compared to the unsupervised protocol (Tab 6).

Table 6: Contribution of supervised schemes on PRID-2011

Approach	**Re-identification rate (%)**	**Running time (s/image)**
Our Approach (SURF only) evaluated on test set + No filtering is applied	36	1.8
Our Approach (SURF only) evaluated on test set + SURFs filtering	37	0.9

The experiment results proves the power of the proposed learning phase to reduce the re-identification running time. From the accuracy standpoint, the system above significantly outperforms the state of the art. In fact, our filtering scheme is able to maintain accuracy (even slightly outperforming the one obtained by the unsupervised one) while significantly increasing speed since it basically halves the processing time.

9. Conclusion

This paper proposed a new approach for human re-identification from video sequences, based on interest point matching, which exploits the complementary nature of appearance and temporal features. The proposed system mainly consists of three stages: person feature description, matching and fusion. For person representation, a robust description that takes into account both appearance and motion is proposed. We have described the appearance (respec-

tively movement) by spatial interest point (respectively spatiotemporal interest point). For matching, we proposed a new method of interest point matching via sparse representation that consists of representing each query interest point as the sparsest linear combination of reference interest points. For efficiency, a dynamic dictionary is selected based on a preset number of closest reference interest point obtained by KD-Tree neighborhood search. The experiments, performed on the large database of PRID-2011, showed that the fusion of the two descriptions allowed the re-identification system to achieve a Correct Classification Rate of 31% which outperforms the state of the art. On the other hand, the results obtained with only SURF matching via SR showed an improvement of 4% on CAVIAR4REID and 4.5% on PRID-2011 compared to 1-NN. These results demonstrate the relative power of sparse representation to match IPs in noisy and ambiguous conditions that are inherent to real video sequences.

Other future extensions of this work include the proposal of other filtering schemes for rejecting unreliable IPs, the study of other methods of dictionary construction by estimating for example the size of the dynamic dictionary (fixed in our experiences). Finally, for IPs matching, we propose to exploit the sparse representation in other ways than minimizing the reconstruction error.

References

[1] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, I. Rennes, I. I. Grenoble, L. Ljk, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[2] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[3] M. I. Khedher, M. A. El-Yacoubi, B. Dorizzi, Multi-shot surf-based person re-identification via sparse representation, in: International Conference on Advanced Video and Signal-Based Surveillance, 2013.

[4] J. H. Friedman, J. L. Bentley, R. A. Finkel, An algorithm for finding best matches in logarithmic expected time, ACM Transactions on Mathematical Software 3 (1977) 209–226.

[5] O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux, Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, in: ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2008, pp. 1–6.

[6] K. Jungling, M. Arens, View-invariant person re-identification with an implicit shape model, in: Proceedings of the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2011, pp. 197–202.

[7] N. Truong Cong, C. Achard, L. Khoudour, People re-identification by classification of silhouettes based on sparse representation, in: Proceedings of the International Confernce on Image Processing Theory, Tools and Applications, 2010, pp. 60–65.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 210–227.

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

[10] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Proceedings of the 17th Scandinavian conference on Image analysis, 2011, pp. 91–102.

[11] 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, volume Part II, 2006, pp. 589–600.

[12] G. Lisanti, I. Masi, A. Bagdanov, A. Del Bimbo, Person re-identification by iterative re-weighted sparse ranking, Pattern Analysis and Machine Intelligence, IEEE Transactions on 37 (2015) 1629–1642.

[13] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the 10th European Conference on Computer Vision, volume Part I, 2008, pp. 262–275.

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

[15] M. Dikmen, E. Akbas, T. Huang, N. Ahuja, Pedestrian recognition with a learned metric, in: Proceedings of the 10th Asian Conference on Computer Vision, volume Part IV, 2010, pp. 501–512.

[16] W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 649–656.

[17] L. An, M. Kafai, S. Yang, B. Bhanu, Reference-based person re-identification, in: Proceedings of the 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2013, pp. 244–249.

[18] S. Leurgans, R. Moyeed, B. Silverman, Canonical correlation analysis when the data are curves, J Roy Statistical Soc, Ser B 55 (1993) 725 – 740.

[19] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society (Series B) 58 (1996) 267–288.

[20] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Transactions on Signal Processing 41 (1993) 3397–3415.

[21] H. Bay, T. Tuytelaars, L. V. Gool, Surf: Speeded up robust features, in: Proceedings of 9th European Conference on Computer Vision, 2006, pp. 404–417.

[22] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, 2001, pp. 511–518.

[23] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, 2005, pp. 886–893.

[24] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzílez, Selective spatio-temporal interest points, Comput. Vis. Image Underst. 116 (2012) 396–410.

[25] J. Cao, T. Chen, J. Fan, Landmark recognition with compact bow histogram and ensemble elm, Multimedia Tools Appl. 75 (2016) 2839–2857.

[26] I. J. Bradley Efron, Trevor Hastie, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2004) 407–840.

[27] Z. Wang, J. Yang, H. Zhang, Z. Wang, Y. Yang, D. Liu, T. Huang, Sparse Coding and its Applications in Computer Vision, 2015.

[28] A. M. Bruckstein, D. L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (2009) 34–81.

[29] Z. Zhang, Y. Xu, J. Yang, X. Li, D. Zhang, A survey of sparse representation: Algorithms and applications, IEEE Access 3 (2015) 490–530.

[30] E. J. Candès, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and, Comm. Pure Appl. Math. (2006) 1207–1223.

[31] D. L. Donoho, For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution, Comm. Pure Appl. Math 59 (2004) 797–829.

[32] J. H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (2010) 1–22.

[33] K. Guo, P. Ishwar, J. Konrad, Action Recognition in Video by Sparse Representation on Covariance Manifolds of Silhouette Tunnels, 2010, pp. 294–305.

[34] J. Cao, K. Zhang, M. Luo, C. Yin, X. Lai, Extreme learning machine and adaptive sparse representation for image classification, Neural Networks 81 (2016) 91 – 102.

[35] J. H. Friedman, Regularized Discriminant Analysis, Journal of the American Statistical Association 84 (1989) 165–175.

[36] T. T. Wu, K. Lange, Coordinate Descent Algorithms for Lasso Penalized Regression, The Annals of Applied Statistics 2 (2008) 224–244.

[37] J. Mairal, Spams: a sparse modeling software, v2.4, 2009. URL: `http://spams-devel.gforge.inria.fr/`.

[38] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: Procedings of the British Machine Vision Conference, 2011, pp. 68.1–68.11.

[39] CAVIAR, http://homepages.inf.ed.ac.uk/rbf/caviar/, 2003.

[40] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Person re-identification using spatial covariance regions of human body parts, in: IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 435–440.

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

[42] M. Ibn Khedher, M. A. El-Yacoubi, B. Dorizzi, Probabilistic matching pair selection for surf-based person re-identification, in: Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG, 2012.

[43] D. Cheng, M. Cristani, Person re-identification by articulated appearance matching, in: Person Re-Identification, Advances in Computer Vision and Pattern Recognition, Springer London, 2014, pp. 139–160.

[44] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3318–3325.

[45] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, J. Bu, Semi-supervised coupled dictionary learning for person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014, pp. 3550–3557.

[46] W. Zuo, F. Wang, D. Zhang, L. Lin, Y. Huang, D. Meng, L. Zhang, Iterated support vector machines for distance metric learning, CoRR abs/1502.00363 (2015).

[47] M. Zeng, Z. Wu, C. Tian, L. Zhang, L. Hu, Efficient person re-identification by hybrid spatiogram and covariance descriptor, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 48–56.

[48] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Computer Vision, ECCV 2014, volume 8692, 2014, pp. 688–703.

[49] M. I. Khedher, M. A. El-Yacoubi, Local sparse representation based interest point matching for person re-identification, in: International Conference on Neural Information Processing (ICONIP), 2015, pp. 241–250.

[50] V. N. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.

Mohamed IBN KHEDHER obtained his engineering degree in Computer Science from the National School of Computer Sciences, Tunisia, in 2007. During his internship, he studied the video complexity estimation in the norm H.264. In 2009, he obtained his master degree from the National School of Computer Sciences, Tunisia where he implemented a biometric system for face and gait recognition. He obtained his PhD in computer science from Telecom SudParis with the collaboration of the University of Evry, France. During his PhD, he developed software for person re-identification from video sequence. Since 2015, he worked as a software Engineer in a research center specialized in Advanced Driver Assistance Systems. His main interests include Machine Learning, Video coding standards, Video Surveillance, Biometrics, Human gesture recognition and Handwriting Analysis.

Mounîm A. El Yacoubi obtained his PhD in Signal Processing and Telecommunications from the University of Rennes I, France, in 1996. During his PhD, he was with the Service de Recherche Technique de la Poste (SRTP) at Nantes, France where he developed software for Handwritten Address Recognition for Automatic French mail sorting. He was a visiting scientist for 18 months at the Centre for Pattern Recognition and Machine Intelligence (CENPARMI) in Montréal, Canada. He then became an associated professor (1998-2000) at the Catholic University of Parana (PUC-PR) in Curitiba, Brazil. From 2001 to 2008, he was a Senior Software Engineer at Parascript, Boulder (Colorado, USA), a world leader company in automatic processing of handwritten and printed documents (mail, checks, forms). Since June 2008, he is a Professor at Telecom SudParis, University of Paris Saclay. His main interests include Machine Learning, Human Gesture and Activity recognition, Human Robot Interaction, Video Surveillance and Biometrics, Information Retrieval, and Handwriting Analysis and Recognition for e-Health.

Bernadette Dorizzi got her PhD (Thèse d'état) in Theoretical Physics at the University of Orsay (Paris XI-France) in 1983, on the study of integrability of dynamical systems. She is Professor at Télécom SudParis (ex INT) since September 1989, and Dean of Research since September 2013. She has been leading the Electronics and PHysics department between 1995 and 2009. She is in charge of the Intermedia (Interaction for Multimedia) research team. Her research domain is related to pattern recognition and machine learning applied to activity detection, surveillance-video and biometrics. She has coordinated the BioSecure Network of Excellence and is now chairwoman of the BioSecure Foundation (http://biosecure.info). She is author of more than 300 research papers and has supervised more than 20 PhD thesis.