Applied Computing and Informatics (2016) **12**, 142–153

Saudi Computer Society, King Saud University

Applied Computing and Informatics

(http://computer.org.sa)
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Ensemble of different approaches for a reliable person re-identification system

Loris Nanni [a], **Matteo Munaro** [a], **Stefano Ghidoni** [a], **Emanuele Menegatti** [a], **Sheryl Brahnam** [b,*]

[a] *Department of Information Engineering at the University of Padua, Via Gradenigo, 6-35131 Padova, Italy*
[b] *Computer Information Systems Department at Missouri State University, Springfield, MO 65804, USA*

Received 16 December 2014; revised 2 February 2015; accepted 17 February 2015
Available online 7 March 2015

KEYWORDS

Person re-identification;
Texture descriptors;
Ensemble;
Color space;
Depth map

Abstract An ensemble of approaches for reliable person re-identification is proposed in this paper. The proposed ensemble is built combining widely used person re-identification systems using different color spaces and some variants of state-of-the-art approaches that are proposed in this paper. Different descriptors are tested, and both texture and color features are extracted from the images; then the different descriptors are compared using different distance measures (e.g., the Euclidean distance, angle, and the Jeffrey distance). To improve performance, a method based on skeleton detection, extracted from the depth map, is also applied when the depth map is available. The proposed ensemble is validated on three widely used datasets (CAVIAR4REID, IAS, and VIPeR), keeping the same parameter set of each approach constant across all tests to avoid overfitting and to demonstrate that the proposed system can be considered a general-purpose person re-identification system. Our experimental results show that the proposed system offers significant improvements over baseline approaches. The source code used for the approaches tested in this paper will be available at https://www.dei.unipd.it/node/2357 and http://robotics.dei.unipd.it/reid/.

1. Introduction

Person re-identification is the task of recognizing a given individual when he or she is viewed across any number of non-overlapping views in a distributed network of cameras or at different time instants when captured by a single camera. Research in person re-identification is motivated by the need of automating many surveillance activities in airports, metro stations, etc. This task requires the creation of a model recording macroscopic characteristics, as many of the classic biometric cues (facial appearance and gait characteristics) are often not

* Corresponding author. Tel.: +1 417 8739979.
E-mail addresses: nanniemg@dei.unipd.it (L. Nanni), munaroemg@dei.unipd.it (M. Munaro), ghidoniemg@dei.unipd.it (S. Ghidoni), sbrahnam@missouristate.edu (S. Brahnam).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2015.02.002
2210-8327

available due to the low frame-rates and resolutions of many surveillance cameras. Appearance-based, non-collaborative scenarios are challenging because the system must measure the similarity between two person-centered bounding boxes and correctly identify the same person despite changes in illumination, pose, background, occlusions, and the variability in camera resolutions and viewpoints. New advances, however, such as using 3D sensors [16], are making it possible to extract some soft-biometric features such as a person's 3D shape, height, and the lengths of limbs.

This paper targets short-term re-identification, which aims at recognizing people within relatively short time frames, thus relying on the assumption that the person is wearing the same clothing during the training and testing phases. Unlike tracking, we assume that no motion information is available for this task.

In the literature on this topic, the features that are most commonly exploited are color, texture and shape. For instance, in [7,6,8], the body of each target is divided into smaller parts and evaluated with multiple color histograms, one for each part. Even though this method is simple and effective, it fails in the case of strong illumination changes. Texture-based and shape-based approaches, such as [1,11,24], usually make use of local features, which provide a detailed description of targets. These approaches exploit descriptors evaluated on a set of keypoints to generate the signature of a target. The performance of this method is thus dependent on the capability of the keypoint detector to select stable features. In [17] a texture-based signature is proposed that consists of local descriptors computed around the principal joints of the human body. To detect the body joints, 3D data from consumer depth sensors and state-of-the-art skeletal tracking algorithms are exploited. The resulting Skeleton-based Person Signature (SPS) has proved to be very robust in the presence of strong illumination changes. The main drawback of this approach, however, is its dependency on the skeletal tracker; when this fails to recognize the body pose, the provided signature is meaningless. For a recent survey on person re-identification, see [23].

In this paper we improve the performance of state-of-the-art person re-identification systems using an ensemble of methods combined by weighted sum rule. The different systems utilize different color spaces and several texture and color features for describing the images. To the best of our knowledge, this is the first work in which several different state-of-the-art person re-identification systems, and their variants, are combined to obtain a more robust approach.

To demonstrate the generality of our system, we validate our approach on the following well-known datasets: CAVIAR4REID, IAS, and VIPeR. Moreover, we test our system on a dataset derived from VIPeR, which we call VIPeR45 because it contains 45 image pairs from VIPeR that focus on some of the most difficult samples to re-identify images of persons containing strong pose changes, for instance, or wearing very similar clothing. VIPeR45 was created because person re-identification performance was tested in [7] using a dataset that was built in a similar fashion (i.e., using 45 difficult image pairs extracted from VIPeR); the human subjects obtained a Rank(1) of 75% and a Rank(10) of ~100% [7]. Thus, it is possible for other researchers in person re-identification to use VIPeR45 for approximately comparing the performance of their computer vision systems with the performance of human beings at this same task. The VIPeR45 dataset will be available at http://robotics.dei.unipd.it /reid/.

The remainder of this paper is organized as follows. In Section 2 we describe the base approaches used in our system and provide details of our weighted ensemble. In Section 3, we describe the datasets used in our experiments, and in Section 4 we provide the experimental results. Finally, in Section 5 we summarize the significance of our work and highlight some future directions of exploration.

2. Methods

In this work we compare and combine different recent state-of-the-art person re-identification systems, viz. a representation that combines biologically inspired features and covariance descriptors, called gBiCov [15], Symmetry-Driven Accumulation of Local Features (SDALF) [8], Custom Pictorial Structures (CPS) [7] based on chromatic content and color displacement (CCD), Color Invariants (CI) [12], and the Skeleton-based Person Signature (SPS) technique [17]. Moreover, we propose variants of such approaches, obtained by varying the features used for describing the images and by using different distance measures. Each of these state-of-the-art systems, our variants, and the different color spaces, distance measures (specifically, the Jeffery Divergence measure, which obtains the best performance), and the color and texture descriptors used in our approaches are described in this section.

The following descriptors (detailed in Section 2.8) are tested:

- Color: Color descriptor proposed in [3].
- WLD: Weber's Law Descriptor proposed in [5].
- LPQT: Local Phase Quantization from Three Orthogonal Plane proposed in [18].
- VLPQ: Volume Local Phase Quantization proposed in [19].

The best approach (see Fig. 1) is obtained by combining several methods (detailed in this section) that utilize different characteristics and can be described as follows:

- Convert the RGB image to XYZ.
- Extract the pictorial structures (PS) from both the RGB and XYZ image.
- Find skeleton joints from the RGB image in the 3D domain using the tracker.
- Extract gBiCov and SDALF from the RGB image: several descriptors are used to describe the region found by PS and the area around the skeleton joints.
- Match the two images using an appropriate distance measure: different Matching Functions (MF) are used in the different methods.
- Combine the set of matching scores by sum rule.

It is important to note the methods composing the ensemble schematized in Fig. 1 all work in parallel, i.e., each method is performed independently of the others. The scores of each method are simply summed (after normalization to mean 0 and std = 1). Moreover, the proposed ensemble uses no optimization algorithm: we simply combine the best methods for optimizing the average performance on the tested datasets.

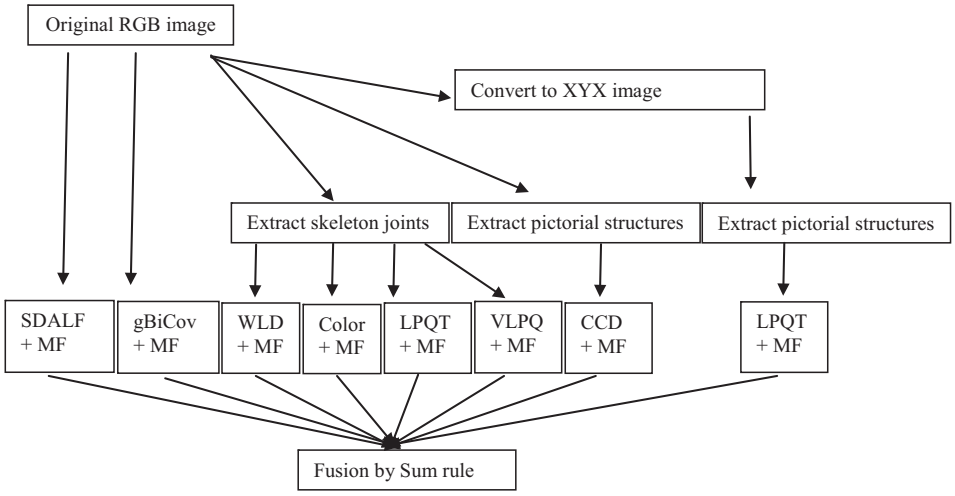

Figure 1 Flowchart of the proposed ensemble.

2.1. Color spaces

To improve the performance of each system, we utilize not only the RGB color space but also several other color spaces. A color space is an abstract mathematical model describing the way colors can be represented [4]. The input images in the tested databases are given in the RGB color space. To explore other spaces, the original images are transformed into the following codings: YUV, HSV, HSL, and XYZ.

YUV defines color in terms of one luma/brightness (Y) component and two chrominance (UV) components, taking into account human perception by reducing bandwidth for the two chrominance components. HSV (hue-saturation-value) and HSL (hue-saturation-lightness) are common cylindrical-coordinate representations of points in the RGB color space. The XYZ color space defines three primaries that are not tied to any particular physical device but rather are points that lie outside the visible gamut, thereby completely encoding all color perceptions possible in the real world.

2.2. Jeffery Divergence measure

Different distance measures were explored for comparing descriptors. Those that performed best are the angle distance, the Euclidean distance, and the Jeffrey Divergence measure [13], the last being numerically stable and symmetric. Jeffrey Divergence is an information-theoretic measure derived from Shannon's entropy theory that treats objects as probabilistic distributions. Thus, it is not applicable to features with negative values.

Given two objects $A, B \in \Re^N$ their Jeffrey Divergence is defined as

$$JD(A,B) = \sum_{i=1}^{n}\left(a_i \log\frac{2a_i}{a_i+b_i} + b_i \log\frac{2b_i}{a_i+b_i}\right) \quad (1)$$

2.3. gBiCov

Proposed in [15], gBiCov is a state-of-the-art person re-identification method that combines biologically inspired features (BIF) [20] and covariance descriptors [22], specifically by encoding the difference between BIF features at different scales. This image representation efficiently measures the similarity between two persons without needing a preprocessing step (e.g., to extract the background) since it is robust to illumination, scale, and background changes.

The extraction of the gBiCov descriptors is a three step process:

In Step 1 BIF features are extracted using Gabor filters and the max operator. Color images are split into the three HSV color channels and convolved with Gabor filters at 24 different scales, with neighboring scales grouped into 12 different bands. The BIF magnitude images ($B_i i \in [1, \ldots, 12]$) are obtained using the max operator within the same band of Gabor features.

In Step 2 similarity of BIF features is computed at neighboring scales using a covariance descriptor. The BIF magnitude images are divided into small overlapping regions to retain the spatial information, and the difference between the corresponding regions of the different bands and the covariance descriptors is computed, i.e., for each region the difference of covariance descriptors between two consecutive bands is computed as

$$d_{i,r} = d(C_{2i-1,r}, C_{2i,r}) = \sqrt{\sum_{p=1}^{P} In^2 \lambda_P(C_{2i-1,r}, C_{2i,r})}, \quad (2)$$

where $C_{i,r}$ is the covariance descriptor (see [15]), $i = [1, \ldots, 6]$, r is the region, and $\lambda_P(C_{2i-1,r}, C_{2i,r})$ is the p-th generalized eigenvalues of $C_{2i-1,r}, C_{2i,r}$.

In step 3 the BIF and covariance descriptors are combined into a single representation. Although $d_{i,r}$ can be taken as a direct gBiCov descriptor, they are nonetheless combined with the BIF magnitude features. The BIF and covariance descriptors are two different levels of the entire representation: BIF includes the appearance-based features while the covariance matrices are a description of the feature properties. Since color images in step 1 are split into three HSV color channels, the three separately extracted gBiCov descriptors are finally concatenated into a single signature that is then reduced using a dimensionality reduction method such as Principal Component Analysis (PCA). In this paper PCA is not used since it needs a training set.

2.4. SDALF

Proposed in [8] SDALF is a method that models three aspects of human appearance: (i) the overall chromatic content, (ii) the spatial arrangement of colors in specific regions, and (iii) the presence of recurrent local motifs with high entropy. This information is derived from different body parts and weighted by exploiting symmetry and asymmetry perceptual principles. This combination makes SDALF robust against very low resolution, occlusions, pose, viewpoint, and illumination. SDALF exploits both single-shot and multiple-shot approaches; in other words, the larger the number of images of a given person, the greater the expressivity of SDALF. In the description of SDALF that follows, the harder case of a single-shot approach will be described.

SDALF is a three-phase process. In phase 1, the background (BG) is extracted and a silhouette mask Z (bounded by a box of size $(I \times J)$) containing only foreground pixel values (FG) is obtained. Axes of asymmetry and symmetry are found for each pedestrian image using two operators: the *chromatic bilateral operator* and the special *covering operator*.

The chromatic bilateral operator is defined as

$$CH(i,\delta) = \sum_{B_{[i-\delta,i+\delta]}} d^2(p_i, \hat{p}_i), \tag{3}$$

where $d(\bullet, \bullet)$ is the Euclidean distance evaluated between the HSV pixel values $p_i, \hat{p}_i$ located symmetrically with respect to the horizontal height i. The Euclidean distance is summed over $B_{[i-\delta,i+\delta]}$, where δ is 1/4 the height of the image. In other words, B is the FG region lying in the box of J width and vertical extension $[i-\delta, i+\delta]$.

The *covering operator* calculates the difference for two regions of a FG area and is defined as

$$S(i,\delta) = 1/J\delta |A(B_{[i-\delta,i]}) - A(B_{[i,i+\delta]})|, \tag{4}$$

where $A(B_{[i-\delta,i]})$ is the FG area in the box of width J and vertical extension $[i-\delta, i]$.

CH and S are combined to give the axes of symmetry and asymmetry. The main x-axis of asymmetry Ax_{TL} is located at height i_{TL} and is obtained as $i_{TL} = argmin_i(1 - CH(i,\delta)) + S(i,\delta)$, with values of CH normalized. Ax_{TL} usually separates the two biggest body portions defined by different colors (e.g., shirt and pants). The other x-axis of asymmetry Ax_{HT} is located at height i_{HT} and is obtained as $i_{HT} = argmin_i(-S(i,\delta))$. Ax_{HT} separates regions that greatly differ in area (e.g., between head and shoulders).

The values of i_{HT} and i_{TL} isolate three regions R_k, $k = \{0,1,2\}$ that roughly correspond to the head, body, and legs, respectively. R_0 (the head) is discarded because its size is small and contains little information. Given R_1 and R_2, the y-axis of symmetry is located in $j_{LRk} = argmin_j CH(j,\delta) + S(j,\delta)$, where $k = (1,2)$ and δ is fixed to $J/4$.

In Phase 2 features are extracted from each part and accumulated into a single signature. The following methods for extracting features are used: Weighted Color Histograms (WH), Maximally Stable Color Regions (MSCR) [10], and Recurrent High-Structured Patches (RHSP). For all features, their distance from the j_{LRk} is considered to minimize effects of pose.

In WH, one histogram is made for each part and each pixel is weighted by a one-dimensional Gaussian Kernel $\varkappa(\mu,\varsigma)$, where μ is the y-coordinate of j_{LRk}, and ς is set to $J/4$. In this way, pixels near j_{LRk} are given more weight.

The MSCR operator detects blobs by iteratively clustering neighboring pixels with similar color, considering some threshold of maximal chromatic distance between colors. MSCR is extracted for each FG part and only within the Gaussian kernel used in WH.

In RHSP texture characteristics that are highly recurrent are highlighted. First, patches p of size $[I/6 \times J/6]$ are randomly extracted on each FG part, many around j_{LRk} to focus on symmetries, again taking into consideration the Gaussian Kernel used in WH. Entropy of the patches is used to select those patches with the most information and is computed as the sum H_p of each RGB channel. Only those patches whose H_p values are higher than a fixed threshold are selected. A set of transforms T_i, $i = 1,2,\ldots,N_T$ are then applied on p.

In phase 3 the matching of two signatures I_A and I_B is performed by estimating the SDALF matching distance as follows:

$$\begin{aligned} d_{\mathrm{SDALF}}(I_A, I_B) = {} & \beta_{\mathrm{WH}} \cdot d_{\mathrm{WH}}(\mathrm{WH}(I_A), \mathrm{WH}(I_B)) \\ & + \beta_{\mathrm{MSCR}} \cdot d_{\mathrm{MSCR}}(\mathrm{MSCR}(I_A), \mathrm{MSCR}(I_B)) \\ & + \beta_{\mathrm{RHSP}} \cdot d_{\mathrm{RHSP}}(\mathrm{RHSP}(I_A), \mathrm{RHSP}(I_B)) \end{aligned} \tag{5}$$

where β are normalized weights.

2.5. CPS

Proposed in [7] CPS is inspired by studying how human beings perform re-identification (examining eye-tracker information) and focuses on body parts, looking for pictorial structures (PS) and then comparing them part-to-part.

In PS the body model is decomposed into a set of parts $L = \{\boldsymbol{I}_p\}_{p=1}^{N}$, where $\boldsymbol{I}_p = (x_p, y_p, o_p, s_p)$ encodes the position, orientation, and scale of part p in image $\boldsymbol{I}$, respectively. Given image evidence D, the posterior of L is modeled as $p(L|D) \propto p(D|L)p(L)$, where $p(D|L)$ is the image likelihood and $p(L)$ is a prior modeling of the part's connectivity. The kinematic dependencies between body parts are mapped onto a directed acyclic graph with edges E. Image evidence D is obtained with discriminatively trained part models, each providing an evidence map $\boldsymbol{b}_p$. PS factorizes the likelihood in $p(D|L) = \prod_{p=1}^{N} p(\boldsymbol{d}_p \boldsymbol{I}_p)$, thereby making the posterior over the configuration L as follows:

$$p(L|D) \propto p(\boldsymbol{I}_1) \prod_{p=1}^{N} p(\boldsymbol{d}_p \boldsymbol{I}_p) \prod_{(i,j)\in E}^{N} p(\boldsymbol{I}_i \boldsymbol{I}_j), \tag{6}$$

where $\boldsymbol{I}_1$ is the root node (the torso) and $p(\boldsymbol{I}_i \boldsymbol{I}_j)$ models the joint between two connected parts.

PS is trained on a dataset of annotated images. For person re-identification, $N = 6$ parts are selected that completely describe the chest, head, torso, thighs, and legs.

After fitting PS the chromatic content and color displacement (CCD) in each of the six parts is considered. Chromatic content is computed using HSV color histograms, where hue and saturation are jointly taken by a two-dimensional histogram, along with a distinct count of full black to take into account areas of low brightness. Since different parts have different sizes (e.g., the torso is roughly three times larger than the head), part histograms are multiplied by a set of N weights.

The histograms are then normalized and concatenated to form a single feature vector. Color displacement is considered by extracting MSCR blobs (see Section 2.4 above) from the PS body mask.

2.6. CI

Proposed in [12], CI uses shape context descriptors to represent the intra-distribution of structure and is based on the intuition that colors composing a person (say wearing a red shirt and blue jeans) form invariant color clusters, i.e., "color cloud" shapes, that refer to specific parts of a person (torso/upper and limbs/lower). The claim of color invariance is based on a number of intuitive arguments: e.g., that the different colors observed in an object are strongly invariant. Another argument is that relative positions of colors remain invariant (if one part has a stronger red component than another, that difference will hold for a wide range of illuminants and imaging devices). Color uniformity is also taken into account.

CI adds an invariant signature that exploits the distribution of color in different parts of an object. This signature is composed of three descriptors: (i) *Cov*, a variant of the covariance descriptor defined in (2), (ii) a color histogram over a log color space using histogram intersection [21] as the similarity measure, and (iii) what is called in [12] *PartsSC*, which uses spatial information regarding the observed colors. The color histogram is straightforward; *Cov* and *PartsSC*, however, require further discussion.

Cov uses only the dominant color in the original RGB color space of each part of the object when computing the covariance descriptor. This variant of the covariance descriptor captures the texture not accounted for by signatures describing absolute colors or relations between colors.

PartsSC uses the Shape Context (SC) descriptor introduced in [2]. SC is a 2D log-polar histogram counting the number of points falling in radius $\log(r)$ and orientation θ from the reference point. Two cases are possible given a set of N color observations $O = \{x_1, \ldots, x_N\}$ within some color space: in the first case observations are given without spatial information. In the second case, observations are labeled $l_i = 1$, if they come from the upper part of the object, or $l_i = 0$ if they refer to the lower part. Two different signatures are extracted that correspond to these two cases.

For case one, if we let $O_L = \{x_i | l_i = 0\}$ denote the observations generated from the lower part of an object and $O_U = \{x_i | l_i = 1\}$ denote the observations generated from the upper part of an object, then $PartsSC(O_L, O_U)$ can be defined as

$$PartsSC(O_L, O_U) = \{sc(x, O_U) | x \in O_L\}, \tag{7}$$

where $\mathrm{sc}(x, O)$ is the shape context descriptor of the points in set O with respect to reference point x.

Note that the colors in the upper part of an object are encoded with respect to the colors in the lower part. This signature captures the upper part color cloud and the shape of the lower part color cloud, along with the relative positions of the two color clouds.

For case two, which encodes no spatial information, the signature is defined as

$$\mathrm{SC}(O) = \{sc(x, O) | x \in O\}, \tag{8}$$

where O is the set of available observations for a given object.

2.7. SPS

The Skeleton-based Person Signature (SPS) technique evaluates a signature vector for a given target based on the body pose. It takes as input the result of a skeletal tracker, namely a set of body joints, and evaluates a set of local descriptors on the image patches around each joint. It should be noted that the best-performing skeletal tracker algorithms work on 3D data, while when it comes to features the 2D approach performs better than the 3D counterpart. To improve the overall performance, the SPS is evaluated exploiting not only the 3D point cloud provided by the 3D sensor, but also the 2D image of the scene, which is usually also provided by 3D sensors. The skeleton joints are found in the 3D domain by the tracker; they are then projected onto the image plane (thanks to the calibration between 3D sensor and 2D camera). Once the joints are available in the image domain, they are exploited as keypoints for evaluating the local features. Each feature evaluated around a keypoint provides a feature vector (also called descriptor): the complete signature, describing the whole target, is obtained by concatenating all the feature vectors of the different body joints, following a pre-determined order.

The SPS for a given target T_k is given by

$$\mathrm{SPS}_k^J = \bigcup_{i=0}^{N-1} \{D(J_i, T_k)\}, \tag{9}$$

where $D(J_i, T_k)$ is the descriptor evaluated using the chosen feature on the i-th joint (J_i) for target k (T_k).

2.8. Texture/color descriptors

The following methods are coupled with state-of-the-art approaches (CPS and SPS) for improving their performance:

- Color, where the following features are concatenated for describing a patch [13]: mean and homogeneity of the three channels; mean, standard deviation and moments (3rd to 5th) of the three channels; and marginal histograms (8 bins per channel). Marginal histograms estimate the color content of an image through the probability distribution of colors as a function of each channel separately, thus discarding any information about the other channels [3].
- WLD [5], based on Weber's law which states that a change of stimulus that is just noticeable for human beings is a constant ratio of the original stimulus; if the change is less than this constant ratio then it is considered background noise. For each pixel of the input image, the differential excitation component is computed based on the ratio between: (i) the relative intensity differences of a current pixel against its neighbors, and (ii) the intensity of the current pixel. From the differential excitation component both the local salient patterns in the input image and the gradient orientation of the current pixel are computed. By combining the WLD feature per pixel, an image (or image region) is represented with a histogram.
- LPQT, or LPQ-TOP [18], is the application of Local Phase Quantization from Three Orthogonal Planes [25]. LPQ uses local phase information extracted using the two-dimensional short-term Fourier transform (STFT) computed over a rectangular $M \times M$ neighborhood centered at each pixel position x of an image. LPQT calculates LPQ histograms from three orthogonal planes (i.e., the xy, xt, and yt planes).

- VLPQ is an extension of LPQ where the quantized phase information of the Discrete Fourier Transform (DFT) is computed in pixel volume neighborhoods. The local Fourier transform is computed efficiently using 1-D convolutions for each dimension in a 3-D volume (see [19], for details).

SIFT is also tested for SPS since it is the best descriptor among those that were tested in [17], where SPS was first proposed. SIFT [14] is widely used in robotics. It is a keypoint detector and a descriptor invariant to image scale and rotation; it is also robust to changes in illumination, noise, and minor affine transformations. SIFT is computed as an 8-binned histogram of gradient distribution within the region around each keypoint. The descriptor is normalized to unit length to obtain illumination invariance.

3. Datasets

To verify our approach and to build a general person re-identification system, we exploited several datasets that are widely used to test intelligent video surveillance systems: VIPeR, CAVIAR4REID, and IAS.

VIPeR (Viewpoint Invariant Pedestrian Recognition) is a dataset composed of a large number of people (632) that are seen at different viewpoints and is available at http://vision.soe.ucsc.edu/node/178. Only one image pair for each person is available, and people are framed at a distance. This dataset is a widely used benchmark. As reported in the literature, results on VIPeR are produced by ten runs, each consisting of a partition of 316 randomly selected image pairs. Since this dataset is composed of 2D images, the skeletal tracker is unable to provide body joints; however, for a small subset of images (45 image pairs), keypoints were manually added by a human operator (we call this subset of images VIPeR45).

CAVIAR is a dataset where 72 different people were collected for the EC funded CAVIAR project/IST 2001 37540 and is available at http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Caviar is a dataset in which multiple test cases are considered; the CAVIAR4REID (Caviar for Re-identification) test case was used in our experiments. CAVIAR4REID is characterized by high level of occlusions, pose changes, and low-resolution images. As in the case of VIPeR, this dataset is composed by 2D images; keypoints were manually added by a human operator. We performed the same single-frame test described in [7], the frame selection is performed five times and the average results are reported.

[1] It is not clear how the images are selected, for future fair comparison we have selected the frames in the following way:

```
for ff = 1:5 %for five times
TR = []; TE = [];
  for person = 1:max(label) %for each person
    a = find(label = = person); %find the frames of that person
    TR = [TR a(ff)]; %id frame to insert in TR
    TE = [TE a(NUM(person)/2 + ff)]; %id frame to insert in
TE, NUM contains the number of frames of each person
  end
  %the images of TR and TE will be compared
...
```

IAS is a dataset that was originally acquired for testing the first version of the SPS algorithm. It includes 33 sequences and involves 11 people. For every subject, the training and testing sequences were collected in different rooms. The entire training set is composed by 2146 images, with 999 images belonging to the testing set. A 3D sensor was placed on a robot so image sequences are seen from a robot's perspective rather than from the perspective of a surveillance camera. The IAS dataset includes sequences of the same target seen under very different lighting conditions. IAS was used as a stress test for the SPS approach. IAS is available at http://robotics.dei.unipd.it/reid/index.php/8-dataset/5-overview-iaslabone.

4. Results

To verify our approach and to build a general person re-identification system, we use the well-known datasets described in Section 3. Across all databases the same parameters are maintained for each approach since the aim is not to optimize the performance of the proposed system for each dataset but rather to show that this generalized method works well across all datasets without ad-hoc tuning.

Rank(1) and Rank(10) are used as the performance indicators. Rank(k) is the average person recognition rate computed when considering a classification to be correct if the ground truth person appears among the subjects who obtained the k best classification scores.

In the first experiment the aim was to test the different color spaces applied to the different person re-identification systems. Results are presented in Tables 1–8. Each cell contains the Rank(1) and Rank(10) values, except in IAS where only Rank(1) is reported since IAS has only 11 individuals. In the approaches where different distances are applied, all results are reported, with the first row inside a cell reporting the Jeffrey distance, the second row the angle distance, the third row the Euclidean distance. Due to the computational issue of choosing the best methods to combine for fusion exploration, we ran experiments only on the VIPeR, VIPeR45, and CAVIAR4REID datasets. For IAS only the performance of the approaches used to build the ensembles and the baseline methods is reported. In the last column, the best average ensemble of the different color spaces is reported ($a \times X + b \times Y$ is the fusion by weighted sum rule between the color space X, with weight a, and Y, with weight b). In the last row, we report the computation time (CT) in seconds for extracting the descriptors from an image of size 128×64 using MATLAB R2013a (without the parallel toolbox) on an i5-3470 3.2 GHz processor with 8 GB of Ram.

It is interesting to note that the performance of gBiCov (see Table 1) is clearly related to the color space; in the original paper [15], gBiCov is calculated on the HSV space only. The fusion between HSV and YUV leads to a good performance in all the tested datasets. The performance obtained by the three distances is very similar. SDALF (see Table 2) obtains the best performance using the RGB color space, but its fusions are not useful.

CI (see Table 3) obtains the best improvement due to the fusion of the different color spaces. Unfortunately, CI works poorly in the IAS dataset because of the strong illumination change between training and testing sets caused by the

Table 1 gBiCov performance across different colorimetric spaces.

gBiCov	RGB		HSV		HSL		XYZ		YUV		RGB + HSV + YUV		HSV + YUV	
IAS	86.7		57.4		–		–		–		–		–	
	85.9		58.3											
	85.9		58.3											
VIPeR 45	15.6	44.4	22.2	46.7	15.6	46.7	8.9	37.8	**24.4**	**57.8**	15.6	44.4	24.4	48.9
	17.8	46.7	22.2	46.7	13.3	40.0	11.1	31.1	24.4	55.6	17.8	46.7	24.4	51.1
	17.8	46.7	22.2	46.7	13.3	40.0	11.1	31.1	24.4	55.6	17.8	46.7	24.4	51.1
VIPeR	3.2	14.2	12.9	37.3	8.6	31.7	1.6	8.3	10.7	31.8	12.8	33.9	**13.5**	**39.7**
	3.2	14.2	12.9	38.0	6.7	26.3	1.5	7.8	8.8	25.6	11.2	30.0	13.3	35.8
	3.2	14.2	12.9	38.0	6.7	26.3	1.5	7.8	8.8	25.6	11.2	30.0	13.3	35.8
CAVIAR4REID	**9.7**	34.4	7.8	31.4	6.1	23.3	8.6	29.4	6.7	31.7	8.6	35.3	8.1	31.7
	9.4	33.9	7.2	31.7	5.3	22.2	7.8	29.7	8.3	33.6	8.9	37.8	9.2	35.6
	9.4	33.9	7.2	31.7	5.3	22.2	7.8	29.7	8.3	33.6	8.9	37.8	9.2	35.6
CT	7.87													

Note: gBiCov was always performed without a mask.

Table 2 SDALF performance across different colorimetric spaces.

SDALF	RGB		HSV		HSL	XYZ		YUV		RGB + XYZ		RGB + XYZ + YUV	
IAS	86.0		–		–	–		–		–		–	
VIPeR 45	24.4	53.3	17.8	44.4	–	24.4	51.1	**26.7**	**53.3**	26.7	51.1	24.4	48.9
VIPeR	18.8	47.9	10.0	35.8	–	17.1	43.4	10.9	34.5	19.7	48.0	18.7	48.6
CAVIAR4REID	9.4	**39.4**	4.7	30.0	–	11.4	37.2	4.2	28.3	**11.9**	38.9	10.3	38.9
CT	2.70												

Table 3 CI performance across different colorimetric spaces.

CI	RGB		HSV		HSL		XYZ		YUV		RGB + YUV	
IAS	43.5		–		–		–					
VIPeR45	11.1	44.4	8.9	40.0	6.7	35.6	20.0	42.2	24.4	**46.7**	**28.9**	**46.7**
VIPeR	12.7	39.8	5.2	20.3	2.8	17.0	4.2	13.0	9.6	28.4	**15.5**	**42.8**
CAVIAR4REID	8.1	31.7	6.7	31.1	3.6	24.7	6.4	32.2	8.3	31.7	**9.7**	**33.9**
CT	0.33											

Table 4 CPS based on different colorimetric spaces.

CPS	RGB		HSV		HSL		XYZ		YUV	RGB + HSL	
IAS	96.7		–		–		–		–	–	
VIPeR45	24.4	55.6	15.6	40.0	26.7	40.0	**33.3**	**60.0**	–	22.2	51.1
VIPeR	14.4	43.8	12.0	39.9	10.4	38.3	11.8	39.1	–	**19.1**	**53.5**
CAVIAR4REID	**18.1**	**50.3**	7.8	35.0	6.4	26.9	16.7	48.3	–	13.3	44.4
CT	1.19										

different auto-exposure levels of the Kinect sensor. Moreover, CI has a low computational time. Our recommendation is to use the CI ensemble with RGB and YUV only when low computational power is available and in cases where there are no pronounced illumination changes.

CPS experiments span Tables 4–8 where results are reported for each texture/color descriptor (see Section 2.5). In general, CPS obtains very good results. It should be noted that unlike the results reported in the original paper [7] we did not change the parameters of this approach for the different datasets. Moreover, examining the results of CPS, it is clear that the proposed approach for extracting features from the mask obtained by CPS works quite well. The different distances, however, are quite similar in performance, except that the Jeffrey distance outperforms the Euclidean and angle distance measures.

Table 5 CPS-color performance across different colorimetric spaces.

CPS-color	RGB		HSV		HSL		XYZ		YUV	RGB + HSV	
VIPeR45	17.8	**57.8**	33.3	55.6	13.3	44.4	15.6	48.9	–	**33.3**	53.3
	24.4	55.6	17.8	40.0	13.3	48.9	17.8	46.7		24.4	55.6
	24.4	53.3	15.6	40.0	13.3	51.1	17.8	46.7		24.4	55.6
VIPeR	6.1	19.2	12.4	39.0	10.5	34.9	0.3	12.2	–	12.9	32.0
	6.7	21.8	5.8	14.2	8.3	36.1	0.2	8.5		7.5	21.7
	6.3	21.2	4.6	14.1	7.4	35.3	0.2	8.5		7.1	21.2
CAVIAR4REID	11.7	45.6	8.3	31.4	3.3	23.1	4.7	33.6	–	**13.3**	42.8
	11.9	**46.7**	7.2	29.2	3.9	23.6	4.4	31.9		12.5	45.8
	12.2	**46.7**	7.2	28.6	3.6	23.9	4.4	31.9		12.2	45.8
CT	1.22										

Table 6 CPS-WLD performance across different colorimetric spaces.

CPS-WLD	RGB		HSV		HSL		XYZ		YUV	HSV + HSL		RGB + HSV + HSL	
VIPeR45	13.3	42.2	33.3	46.7	37.8	46.7	13.3	51.1	–	35.6	46.7	31.1	48.9
	11.1	40.0	20.0	46.7	26.7	48.9	11.1	46.7		24.4	48.9	28.9	48.9
	13.3	35.6	20.0	44.4	20.0	51.1	6.7	48.9		24.4	51.1	22.2	48.9
VIPeR	3.0	12.6	14.2	40.2	12.1	35.2	3.9	16.2	–	16.1	39.5	13.5	36.1
	4.0	12.1	8.2	25.4	6.4	26.3	3.1	10.9		8.6	28.7	9.0	26.5
	2.7	10.7	5.7	21.5	5.7	21.4	1.5	10.4		6.5	23.5	6.1	22.8
CAVIAR4REID	10.3	37.5	10.3	32.5	6.9	28.1	8.9	38.1	–	8.1	34.2	9.7	**38.3**
	9.2	32.8	8.6	30.6	6.4	26.1	7.2	34.2		6.4	26.4	**10.8**	32.5
	8.9	32.8	6.7	28.6	5.8	24.4	6.1	33.9		6.9	26.9	9.7	33.9
CT	2.59												

Table 7 CPS-VLPQ performance across different colorimetric spaces.

CPS-VLPQ	RGB		HSV		HSL		XYZ		YUV	RGB + XYZ	
VIPeR45	17.8	42.2	22.2	46.7	15.6	44.4	20.0	48.9	–	22.2	46.7
	11.1	46.7	11.1	44.4	11.1	40.0	24.4	48.9		20.0	46.7
	11.1	42.2	11.1	33.3	6.7	31.1	22.2	42.2		13.3	35.6
VIPeR	8.3	37.4	7.6	31.3	2.8	15.9	9.5	33.0	–	12.8	42.8
	6.7	31.5	4.9	21.3	2.2	13.1	8.9	31.3		9.5	36.5
	3.7	19.8	2.3	16.2	1.5	9.4	4.9	21.3		5.1	23.4
CAVIAR4REID	7.8	32.2	6.7	33.6	4.2	20.3	8.9	34.4	–	9.4	35.0
	5.6	30.3	5.3	27.2	3.6	18.9	6.1	33.3		8.3	30.6
	5.3	31.4	5.0	26.7	2.5	18.3	7.5	31.7		8.1	30.3
CT	1.25										

Table 8 CPS-LPQT performance across different colorimetric spaces.

CPS-LPQT	RGB		HSV		HSL		XYZ		YUV	RGB + XYZ	
VIPeR45	17.8	44.4	22.2	**48.9**	20.0	46.7	**26.7**	**48.9**	–	24.4	46.7
	17.8	**48.9**	20.0	46.7	6.7	42.2	22.2	46.7		22.2	**48.9**
	15.6	**48.9**	17.8	46.7	13.3	42.2	**26.7**	46.7		**26.7**	46.7
VIPeR	12.2	44.2	12.7	37.6	4.0	27.0	13.0	36.8	–	**14.4**	**46.0**
	10.9	38.3	7.6	26.9	3.6	18.1	9.6	34.2		12.1	41.6
	10.1	36.6	7.0	25.6	3.3	17.8	9.3	31.9		10.4	41.3
CAVIAR4REID	11.1	32.8	8.1	33.6	4.7	21.7	8.9	35.6	–	**10.3**	**36.7**
	8.9	35.0	8.9	33.1	5.0	20.0	8.3	33.3		9.2	36.4
	7.2	35.8	8.3	33.1	5.0	22.8	7.5	33.3		8.6	36.4
CT	1.72										

The SPS experiments span Tables 9–13 where results are reported for each texture/color descriptor (see Section 2.5). In general, SPS performs very well. This approach is very fast since the features are extracted only from a small set of patches in the image. However, it does need to compute a depth map and a skeleton point detection step.

Table 9 SPS-SIFT performance across different colorimetric spaces.

SP-SIFT	RGB		HSV		HSL		XYZ		YUV		RGB + XYZ	
IAS	92.6		67.4		71.8		91.7		92.1		92.2	
VIPeR45	6.7	37.8	13.3	53.3	15.6	51.1	11.1	42.2	4.4	37.8	8.9	37.8
CAVIAR4REID	18.3	41.7	10.3	30.3	12.2	32.8	17.8	42.5	15.0	30.6	17.8	41.4
CT	0.4											

Table 10 SPS-color performance across different colorimetric spaces.

SPS-color	RGB		HSV		HSL		XYZ		YUV		RGB + XYZ		RGB + HSV + XYZ	
VIPeR45	22.2	57.8	28.9	51.1	15.6	42.2	17.8	46.7	20.0	33.3	17.7	57.8	**31.1**	57.8
	20.0	57.8	20.0	37.8	13.3	42.2	6.7	33.3	4.4	24.4	28.9	**60.0**	26.7	53.3
	26.7	57.8	20.0	37.8	13.3	44.4	8.9	33.3	4.4	24.4	28.9	53.3	24.4	53.3
CAVIAR4REID	18.3	57.2	12.8	41.7	4.4	26.1	15.3	51.4	5.8	31.9	19.2	58.1	**19.7**	58.9
	17.2	**61.1**	12.5	36.9	4.4	23.6	17.2	49.7	2.5	17.5	17.5	60.3	17.2	57.8
	16.7	60.3	12.5	36.9	4.2	24.4	17.2	49.7	2.5	17.5	16.7	59.4	17.2	56.4
CT	0.028													

Table 11 SPS-WLD performance across different colorimetric spaces.

SPS-WLD	RGB		HSV		HSL		XYZ		YUV		RGB + XYZ		RGB + HSV + XYZ	
VIPeR45	2.2	44.4	13.3	42.2	**17.8**	48.9	15.6	48.9	15.6	48.9	6.7	53.3	**17.8**	51.1
	6.7	44.4	13.3	42.2	13.3	33.3	17.8	57.8	**17.8**	**57.8**	11.1	48.9	11.1	51.1
	8.9	33.3	11.1	46.7	8.9	46.7	11.1	40.0	11.1	40.0	13.3	37.8	22.2	48.9
CAVIAR4REID	15.8	44.7	14.2	34.4	13.1	31.1	11.1	39.7	7.8	31.9	15.3	**44.7**	**16.7**	43.3
	8.6	37.2	13.6	43.6	11.4	34.7	9.7	36.7	8.6	32.5	8.9	39.4	11.9	41.7
	10.8	36.7	10.6	40.3	9.2	37.2	10.6	40.3	7.8	31.9	11.4	40.0	11.7	42.2
CT	0.042													

Table 12 SPS-VLPQ performance across different colorimetric spaces.

SPS-VLPQ	RGB		HSV		HSL		XYZ		YUV		RGB + XYZ		RGB + HSV + XYZ	
VIPeR45	20.0	44.4	22.2	44.4	6.7	35.6	22.2	44.4	11.1	35.6	33.3	46.7	**35.6**	**48.9**
	20.0	55.6	20.0	42.2	4.4	40.0	24.4	51.1	11.1	37.8	28.9	46.7	31.1	46.7
	13.3	46.7	13.3	42.2	8.9	31.1	22.2	42.2	8.9	35.6	22.2	40.0	22.2	37.8
CAVIAR4REID	13.9	42.8	12.8	40.6	5.3	18.6	15.0	38.9	14.2	43.1	**16.7**	43.3	**16.7**	45.8
	13.9	42.8	9.7	36.9	4.4	21.4	14.2	38.9	12.5	41.9	**16.7**	43.6	**16.7**	**47.5**
	3.6	21.9	6.1	25.6	1.7	18.1	5.3	28.3	8.9	33.3	6.1	24.4	4.4	24.7
CT	0.33													

Table 13 SPS-LPQT performance across different colorimetric spaces.

SPS-LPQT	RGB		HSV		HSL		XYZ		YUV		RGB + XYZ		RGB + HSV + XYZ	
VIPeR45	24.4	44.4	24.4	44.4	17.8	42.2	26.7	44.4	15.6	**48.9**	26.7	42.2	**35.6**	44.4
	20.0	**48.9**	26.7	44.4	8.9	40.0	22.2	40.0	8.9	46.7	22.2	42.2	35.6	42.2
	17.8	46.7	20.0	46.7	11.1	37.8	17.8	40.0	11.1	42.2	22.2	40.0	26.7	42.2
CAVIAR4REID	17.8	47.5	12.2	45.3	7.5	26.7	16.1	41.4	15.6	37.2	**17.2**	46.7	17.2	**48.3**
	13.1	44.4	10.6	41.1	5.6	22.5	13.6	38.9	13.1	38.1	13.9	41.7	15.0	45.3
	8.9	35.8	8.3	38.3	5.3	20.8	9.7 3	1.9	11.9	38.9	10.8	36.9	9.7	38.9
CT	0.12													

Table 14 Proposed ensemble approaches – rank as performance indicator.

	VIPeR		VIPeR 45		CAVIAR4REID		IAS
BS	14.4	43.8	24.4	55.6	18.1	50.3	96.7
BI	19.7	48.0	37.8	46.7	18.1	50.3	96.7
FUS1	**24.5**	**61.7**	40.0	53.3	16.1	49.4	83.5
FUS2(2)	22.9	56.2	33.3	53.3	18.3	55.3	95.1
FUS2(6)	21.2	53.8	28.9	55.6	20.3	54.4	95.4
FUS2(9)	20.3	51.6	26.7	57.8	20.3	52.8	95.7
FUS2(12)	18.9	50.7	26.7	57.8	20.3	53.3	96.4
FUS2(18)	17.9	49.2	28.9	60.0	20.3	51.9	96.7
FUS3(2)	–		**44.4**	57.8	25.3	65.0	97.6
FUS3(6)	–		42.2	60.0	**26.3**	63.9	**98.0**
FUS3(9)	–		42.2	60.0	26.1	63.3	97.8
FUS3(12)	–		37.8	64.4	25.6	61.9	97.6
FUS3(18)	–		33.3	**66.7**	24.2	59.4	97.4
NogBiCov	–		37.8	64.4	25.0	61.4	97.5

In Table 14 we compared the performance of some ensembles (before the fusion the scores of the approaches are normalized to mean 0 and standard deviation 1) with the best stand-alone methods:

- BS is the best stand-alone approach considering both Rank(1) and Rank(10) on average in the tested datasets.
- BI is the best method considering the different datasets separately.
- FUS1 is the sum rule ensemble of CPS(RGB) + CPS_LPQT(XYZ).
- FUS2(K) is the weighted sum rule ensemble of $K \times$ CPS(RGB) + gBiCov(RGB) + SDALF(RGB) + CPS_LPQT(XYZ).
- FUS3(K) is the weighted sum rule of $K \times$ CPS(RGB) + gBiCov(RGB) + SDALF(RGB) + CPS_LPQT(XYZ) + $2 \times$ SPS_COLOR(RGB) + $2 \times$ SPS_LPQT(RGB) + $2 \times$ SPS_WLD(RGB) + SPS_VLPQ(RGB).
- NogBiCov, is the method FUS3(12) without the expensive (from the computation time view) gBiCov(RGB).

NogBiCov obtains a performance that is similar to FUS3(12) (it is slightly lower in CAVIAR4REID) without using gBiCov, which takes several seconds to describe a given image.

For a deeper evaluation of the performance, Area Under ROC curve [9] is reported in Table 15. The Area Under the Curve (AUC) can be interpreted as the probability that a lower similarity is assigned to a randomly chosen positive match (i.e., for the same person) rather than to a randomly chosen negative match (i.e., for different persons). We have adopted the extension of AUC for multi-class datasets what is called a "one versus all" approach, where, if you have three classes, you would calculate three AUCs. In the first, you would choose the first class as the positive class, and group the other two classes together as the negative class, and so on. The average result is reported.

Notice that in VIPeR45 the best method (considering the Rank) obtains a lower AUC in respect to BS. Using AUC as the performance indicator shows that the fusions clearly outperform the stand-alone methods that built them.

The proposed combination is straightforward, but it has a potential drawback in computational speed, especially when combining several methods that are already known to be expensive, e.g. SDALF (foreground extraction) with CPS (parts detection). In Table 16, we report the computation time of the ensemble methods using an i5-3470 – 3.2 GHz processor with 8 GB of Ram, running MATLAB code with the parallel toolbox for exploiting the four cores. Descriptors are extracted from an image of size 128×64. However, as noted in the discussion of Fig. 1, the different approaches run independently of one another. Moreover, internally several approaches are highly parallelizable (e.g., the descriptor of each part of the image could be extracted in parallel). In our first tests using a better performing CPU (a Xeon E5 – 1620 v2.0) the computation time of NogBiCov is ~2.5 s.

As in several other machine learning problems, it is very difficult to find a stand-alone approach that works well on all the different datasets representing a specific problem. CPS works very well in IAS and CAVIAR4REID but does not obtain the best performance in VIPeR.

Table 15 Proposed ensemble approaches – AUC as performance indicator.

	VIPeR	VIPeR 45	CAVIAR4REID	IAS
BS	0.86	0.76	0.77	0.95
BI	0.91	0.60	0.77	0.95
FUS1	**0.92**	0.70	0.79	**0.96**
FUS2(2)	**0.92**	0.74	0.78	0.94
FUS2(6)	0.90	0.75	0.78	0.95
FUS2(9)	0.89	0.75	0.78	0.95
FUS2(12)	0.89	0.75	0.78	0.95
FUS2(18)	0.88	0.76	0.78	0.95
FUS3(2)	–	0.74	**0.85**	0.91
FUS3(6)	–	0.76	0.84	0.93
FUS3(9)	–	0.76	0.83	0.94
FUS3(12)	–	0.76	0.82	0.95
FUS3(18)	–	**0.77**	0.81	0.95
NogBiCov	–	**0.76**	0.82	0.95

Table 16 Computation time in seconds.

FUS1	FUS2	FUS3	NogBiCov
1.55	8.15	10.25	4.82

Table 17 Comparison with the Literature.

	VIPeR		CAVIAR4REID	
Here	22.9	56.2	25.3	65.0
OR_CPS	21.84	57.21	~9	~47
OR_SDALF	19.87	49.37	–	
eBicov	24.34	58.48	–	
OR_CI	24.00	58.00	~9	~45
kBiCov	31.11	70.71	–	
MCC [1]	15.19	57.59	–	
KISSME [11]	19.60	62.60	–	
PCCA-rbf [2]	19.27	64.91	–	

To obtain a more realistic validation of our approach, we used the same parameters in all the tested datasets. In other words, we did not optimize the performance of our systems for each dataset (to avoid overfitting). Nonetheless, our fusion method outperforms the average performance of all the stand-alone approaches. It should be noted that the results reported for SPS on the IAS dataset in [17] are not comparable with those reported in this paper. In [17] the extracted skeletons of low quality were removed; in the experiments reported here, the entire IAS dataset is used (i.e., no frames are pruned).

Finally, in Table 17 we compared our best approach with several state-of-the-art methods proposed in the literature. The methods named OR_*X* mean the performance as reported in the papers where method *X* is proposed. In several papers the parameters of the methods evaluated are fixed separately for each dataset; in contrast, our method, as mentioned above, always uses the same parameters across the datasets to avoid overfitting. With OR_CI we report the best approach reported in [12], whereas OR_CI is obtained using semi-automatically extracted masks; in our tests, we use automatically extracted silhouettes. EBICOV is an ensemble obtained combining SDALF and gBiCov (using the performance reported in [15]). We have also reported the performance of approaches based on the learnt metric (KBICOV, KISSME, and PCA-RBF), assuming a training set is available. Thus, the performance comparison with our approach is not fair. Yet it is interesting to note that our method obtains a performance that is very similar to methods based on learnt metrics also performed without skeleton detection.

An interesting example of the difference in performance when a method is optimized for a dataset can be observed in the performance of CPS. If CPS is optimized for VIPER, it obtains a Rank(10) of ~57%, while CPS optimized for CAVIAREID obtains a Rank(10) of ~53% (using our set of images[1]). In contrast, if we use a set of parameters that remain the same for both datasets, we obtain a rank(10) of 43.8% and 50.3%, respectively, clearly lower than those obtained when separately optimizing the parameters for each dataset.

For a more exhaustive comparison of methods with the literature, we suggest our results be compared to a recent survey [23]. Examining Table 4 in [23], it is clear that our proposed approach outperforms several other recent systems not compared in this paper.

5. Conclusion

In this paper we run experiments to develop an ensemble of person re-identification systems that works well on different datasets without any ad-hoc dataset tuning. Therefore, we are quite sure that our approach is stable and could be used in different image conditions.

For improving the state-of-the-art approaches, different color spaces, texture, and color features for describing the images were explored. We also considered different distances for comparing descriptors. Among the tested distances, the best performance was obtained with the Jeffrey Divergence measure.

The new methods proposed in this paper were tested across several benchmark databases: CAVIAR4REID; VIPeR; VIPeR45; IAS. The experimental results demonstrate that the proposed approach provides significant improvements over baseline algorithms. The VIPER45 is a new dataset of 45 image pairs taken from VIPeR that focus on difficult samples with strong pose changes and with subjects wearing similar clothing. It was created because human beings were tested in [7] in a dataset that was built in a similar fashion (i.e., using 45 difficult image pairs extracted from VIPeR). It is thus possible for other researchers in person re-identification to use VIPeR45 for approximately comparing the performance of their computer vision systems with the performance of human beings.

A drawback of our approach is computational time, which is not real-time, i.e., using MATLAB code. However, several methods used in our approach are internally highly parallelizable. The main focus of this paper was not on computational speed; our goal was to produce an approach that could match human performance. Unfortunately our results show that this goal has not been achieved (our ensemble obtains a Rank(10) of ~65%, while a human being obtains ~100%). Nonetheless, we have succeeded in producing a stable general-purpose

person re-identification system that offers significant improvements over baseline approaches.

The MATLAB code of the approach described in this paper will be freely available at https://www.dei.unipd.it/node/2357 as well as at http://robotics.dei.unipd.it/reid/.

References

[1] M.S. Bauml, R. Stiefelhagen, Evaluation of local features for person re-identification in image sequences, Paper presented at the 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Klagenfurt, Austria, 2011.

[2] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (24) (2002) 509–522.

[3] F. Bianconi, A. Fernández, E. González, S.A. Saetta, Performance analysis of colour descriptors for parquet sorting, Expert Syst. Appl. 40 (5) (2013) 1636–1644.

[4] L. Busin, N. Vandenbroucke, L. Macaire, Color spaces and image segmentation, Adv. Imaging Electron Phys. 51 (2008) 65–168.

[5] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, W. Gao, WLD: a robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1705–1720.

[6] D.C. Cheng, M. Cristani, Person re-identification by articulated appearance matching, in: S. Gong, M. Cristani, S. Yan, C.C. Loy (Eds.), Person re-identification, Springer, London, 2014, pp. 139–160.

[7] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: J. Hoey, S. McKenna, E. Trucco (Eds.), Proceedings of the British Machine Vision Conference, BMVA Press, University of Dundee, UK, 2011, pp. 1–11.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, Paper presented at the IEEE Computer Vision and Pattern Recognition, San Francisco, CA, 2010.

[9] T. Fawcett, ROC graphs: Notes and Practical Considerations for Researchers, HP Laboratories, Palo Alto, 2004.

[10] P.-E. Forssén, Maximally stable colour regions for recognition and matching, Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, 2007.

[11] K. Jungling, A. Michael, Feature based person detection beyond the visible spectrum, Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshop, Miami Beach, FL, 2009.

[12] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1622–1634.

[13] H. Liu, D. Song, S. Rüger, R. Hu, V. Uren, Comparing dissimilarity measures for content-based image retrieval, Paper presented at the Information Retrieval Technology: 4th Asia Information Retrieval Symposium, AIRS 2008, Harbin, China, 2008.

[14] D.G. Lowe, Object recognition from local scale-invariant features, Paper presented at the 7th IEEE International Conference on Computer Vision, Kerkyra, 1999.

[15] B. Ma, Y. Su, F. Jurie, Covariance descriptor based on bio-inspired feature for person re-identification and face verification, Image Vision Comput. 32 (2014) 379–390.

[16] M.B. Munaro, 3d reconstruction of freely moving persons for reidentification with a depth sensor, Paper presented at the IEEE International Conference on Robotics and Automation, Hong Kong, China, 2014.

[17] M.B. Munaro, S. Ghidoni, D.T. Tartaro, E. Menegatti, A feature-based approach to people re-identification using skeleton keypoints, Paper presented at the IEEE International Conference on Robotics and Automation, Hong Kong, China, 2014.

[18] V. Ojansivu, J. Heikkila, Blur insensitive texture classification using local phase quantization, Paper presented at the ICISP, 2008.

[19] J. Päivärinta, E. Rahtu, J. Heikkilä, Volume Local Phase Quantization for Blur-insensitive Dynamic Texture Classification Image Analysis, Springer, Berlin, Heidelberg, 2011, pp. 360–369.

[20] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nat. Neurosci. 2 (11) (1999) 1019–1025.

[21] M. Swain, D. Ballard, Indexing via color histograms, Paper presented at the IEEE International Conference on Computer Vision, Osaka, Japan, 1990.

[22] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on riemannian manifolds, IEEE Trans. Pattern Anal. Mach. Intell. 30 (10) (2008) 1713–1727.

[23] R. Vezzani, D. Baltieri, R. Cucchiara, People re-identification in surveillance and forensics: a survey, ACM Comput Surveys 46 (2) (2013) 29:21–29:23.

[24] K. Yoon, D. Harwood, L. Davis, Appearance-based person recognition using color/path-length profile, J. Visual Commun. Image Represent. 17 (3) (2006) 605–622.

[25] G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.