SDALF: Modeling Human Appearance with Symmetry-Driven Accumulation of Local Features

Loris Bazzani and Marco Cristani and Vittorio Murino

Abstract In video surveillance, person re-identification (re-id) is probably *the* open challenge, when dealing with a camera network with non-overlapped fields of view. Re-id allows the association of different instances of the same person across different locations and time. A large number of approaches have emerged in the last five years, often proposing novel visual features specifically designed to highlight the most discriminant aspects of people, which are invariant to pose, scale and illumination. In this chapter, we follow this line, presenting a strategy with three important key-characteristics that differentiate it with respect to the state of the art: 1) a symmetry-driven method to automatically segment salient body parts, 2) an accumulation of features making the descriptor more robust to appearance variations, and 3) a person re-identification procedure cast as an image retrieval problem, which can be easily embedded into a multi-person tracking scenario, as the observation model.

1 Introduction

Modeling the human appearance in surveillance scenarios is challenging because people are often monitored at low resolution, under occlusions, bad illumination conditions, and in different poses. Robust modeling of the body appearance of a person becomes mandatory for re-identification and tracking, especially when other

Loris Bazzani
Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy, e-mail: loris.bazzani@iit.it

Marco Cristani
University of Verona, Verona, Italy, e-mail: marco.cristani@univr.it

Vittorio Murino
Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy, and University of Verona, Verona, Italy e-mail: vittorio.murino@iit.it

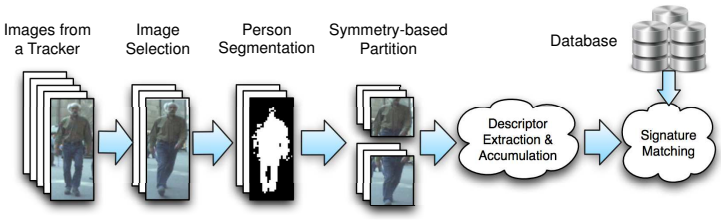


Fig. 1 Person re-id and re-acquisition pipeline. See the text for details.

classical biometric cues (*e.g.*, face, gait, or fingerprint) are not available or difficult to acquire.

Appearance-based re-id can be considered as a general *image retrieval* problem, where the goal is to find the images from a database that are more similar to the query. The only constraint is that there is an assumption of the presence of a person in the query image and the images in the database. On the other hand, person re-identification is also seen as a fundamental module for *cross-camera tracking* to keep unique identifiers in a camera network. In this setup, temporal and geometric constraints can be added to make easier re-id. In general, we define re-identification as matching the signature of each *probe* individual to a *gallery* database composed by hundreds or thousands of candidates which have been captured in various locations by different cameras and in different instants. Similarly to re-identification, multi-person tracking is another problem where the description of individuals plays an important role to ensure consistent tracks across time. In this context, the problem can be seen as matching across time the signature (also called template) of the tracked person with the set of detected individuals. Both re-identification and people tracking share the problem of modeling the human appearance in a way that is robust to occlusion, low resolution, illumination and other issues.

In this chapter, we describe the pipeline for re-identification that has become a standard in the last few years [14]. The pipeline and the descriptor used for characterizing the human appearance are called *Symmetry-Driven Accumulation of Local Features* (SDALF). The re-id pipeline is defined in six steps (Fig. 1): 1) *image gathering* collects images from a tracker, 2) *image selection* discards redundant information; 3) *person segmentation* discards the noisy background information; 4) *symmetry-based silhouette partition* discovers parts from the foreground exploiting symmetric and asymmetric principles; 5) *descriptor extraction and accumulation* over time, using different frames in a multi-shot modality; 6) *signature matching* between the probe signature and the gallery database.

SDALF is composed by a symmetry-based description of the human body, and it is inspired by the well-known principle that the natural objects reveal symmetry in some form. For this reason, detecting and characterizing symmetries is useful to understand the structure of objects. This claim is strongly supported by the Gestalt psychology school [30] that considers symmetry as a fundamental principle of perception: symmetrical elements are more likely integrated into one coherent

object than asymmetric regions. The principles of the Gestaltism have been largely exploited in computer vision for characterizing salient parts of structured objects [31, 10, 43, 44]. In SDALF, asymmetry principles allow to segregate meaningful body parts (head, upper body, lower body). Symmetries help to extract features from the actual human body, pruning out distracting background clutter. The idea is that features near the vertical symmetry axis are weighted more than those that are far from it, in order to obtain information from the internal part of the body, trusting less the peripheral portions more prone to noise.

Once parts have been localized, complementary aspects of the human body appearance are extracted in SDALF, highlighting: i) the global chromatic content, by the color histogram (see Fig. 4(c)); ii) the per-region color displacement, employing Maximally Stable Colour Regions (MSCR) [18] (see Fig. 4(d)); iii) the presence of *Recurrent Highly Structured Patches* (RHSP) [14].

Different feature accumulation strategies can be considered for re-id, and in this regard the literature is divided in single-shot and multi-shot modes, reflecting the way the descriptors are designed (see the next section for more details). In the former case, the signature is built using only one image for each individual, whereas in the latter multiple images are utilized. The multi-shot mode is strongly motivated by the fact that in several surveillance scenarios it is really easy to extract multiple images of the same individual from consecutive frames. For example, if an automatic tracking system is available, consecutive shots of a tracked individual can be used in refining the object model against appearance changes. SDALF takes into account these situations: it accumulates the descriptors from all the available images of an individual, increasing the robustness and the expressiveness of its description. After the signature is built, the matching phase consists of a distance minimization strategy to search for a probe signature across the gallery set, in a similar spirit of image retrieval algorithms.

In this chapter, we discuss how SDALF can be easily adapted to deal with the multi-person tracking problem, in the same spirit of [5]. The idea is to build a signature for each tracked target (the template). Then, the signature is matched against a gallery set: this set is composed by diverse hypotheses that come from a detection module or from the tracking dynamics. The idea is to employ the matching scores as probabilistic evaluations of the hypotheses. The template is then updated with SDALF, as multiple images are gathered over time in the multi-shot mode.

The proposed method is tested on challenging benchmarks: VIPeR [20], iLIDS for re-id [54], ETHZ [47], and CAVIAR4REID [9], giving convincing performance. These benchmarks represent different challenges for the re-id problem: pose, viewpoint and lighting variations, and occlusions. We test the limit of SDALF by subsampling these datasets up to dramatic resolutions (11×22 pixels). Moreover, the multi-person tracker based on SDALF was tested on CAVIAR, which represents a challenging real tracking scenario, due to pose, resolution and illumination changes, and severe occlusions.

The rest of the chapter is organized as follows. In Section 2, the state of the art of re-id is described, highlighting our peculiarities with respect to other approaches. Section 3 details the re-id pipeline and the SDALF descriptor. Section 4 describes

Table 1 Taxonomy of the existing appearance-based re-identification methods.

	Single-shot	Multiple-shot
Learning-based	[47, 41, 36, 32, 21, 46] [54, 1, 55, 23, 34, 16]	[48, 53]
Direct Methods	[2] **SDALF**	[7, 52, 19, 22, 45] **SDALF**

how the signature matching is performed. Section 5 describes how SDALF can be embedded into a particle filtering-based tracker. Several results and comparative analyses are reported in Section 6, and, finally, conclusions and future perspectives are discussed in Section 7.

2 Related Work

Re-id methods that rely only on visual information are addressed as *appearance-based* techniques. Other approaches assume less general operative conditions: *geometry-based techniques* exploit geometrical constrains in a scenario with overlapped camera views [50, 39]. *Temporal methods* deal with non-overlapped views adding a temporal reasoning on the spatial layout of the monitored environment, in order to prune the candidate set to be matched [25, 33, 42]. The assumption is that people usually enter in a few locations, spend a fixed period (learned beforehand) in the blind spots, and re-appear somewhere else in the field of view of a pre-selected set of cameras. *Depth-based approaches* consider other sensors (such as RGB-D cameras) to extract 3D soft-biometric cues from depth images in order to be robust to the change of clothes [3].

Appearance-based methods can be divided into two groups (see Table 1): the *learning-based* methods and the *direct* methods. Learning-based techniques are characterized by the use of a training dataset of *different individuals* where the features and/or the policy for combining them are utilized. The common assumption is that the knowledge extracted from the training set could be generalized to unseen examples. In [36], local and global features are accumulated over time for each subject, and fed into a multi-class SVM for recognition and pose estimation, employing different learning schemes. Viewpoint invariance is instead the main issue addressed by [21]: spatial and color information are here combined using an ensemble of discriminant localized features and classifiers selected by boosting. In [32], pairwise dissimilarity profiles between individuals are learned and adapted for a nearest neighbor classification. Similarly, in [47], a high-dimensional signature composed by texture, gradient and color information is projected into a low-dimensional discriminant latent space by Partial Least Squares (PLS) reduction. Multiple Component Learning is casted into the re-id scenario, dubbing it a Multiple Component Matching and exploiting SDALF as a descriptor, in [46]. The descriptor proposed in [54] uses contextual visual knowledge coming from the surrounding people that

form a group, assuming that groups can be detected. Re-id is casted as a binary classification problem (one vs. all) by [1] using Haar-like features and a part-based MPEG7 dominant color descriptor. In [41, 55, 53], the authors formulate re-id as a ranking problem and an informative subspace is learned where the potential true match corresponds to the highest ranking. Metric learning methods, which learn a distance metric from pairs of samples from different cameras, are becoming popular, see [23, 34] . In [16], re-id is defined as a semi-supervised single-shot recognition problem where multiple features are fused at the classification output level using the recent multi-view learning framework in [35].

The main disadvantage of learning-based methods is the need of retraining for environment covariates, *e.g.*, night-day, indoor-outdoor. In addition, some learning-based approaches also depend on the cardinality and the kind of training set: once a new individual is added to the gallery set, the classifier should be retrained from scratch.

The other class of approaches, the direct method class, does not consider training datasets of multiple people and works on each person independently, usually focusing on the design of features that capture the most distinguishing aspects of an individual. In [7], the bounding box of a pedestrian is equally subdivided into ten horizontal stripes, and the median HSL value is extracted in order to manage x-axis pose variations. These values, accumulated over different frames, generate a multiple signature. A spatio-temporal local feature grouping and matching is proposed by [19], considering ten consecutive frames for each person, and estimating a region-based segmented image. The same authors present a more expressive model, building a decomposable triangulated graph that captures the spatial distribution of the local descriptions over time, so as to allow a more accurate matching. In [52], the method consists in segmenting a pedestrian image into regions, and registering their color spatial relationship into a co-occurrence matrix. This technique proved to work well when pedestrians are seen under small variations of the point of view. In [22], the person re-id scheme is based on the matching of SURF interest points [4] collected in several images, during short video sequences. Covariance features, originally employed for pedestrian detection, are extracted from coarsely located body parts and tailored for re-id purposes [2].

Considering the features employed for re-id, in addition to color information, which is universally adopted, several other cues are textures [47, 41, 21], edges [47], Haar-like features [1], interest points [19], image patches [21], and segmented regions [52]. These features, when not collected densely, can be extracted from horizontal stripes [7], triangulated graphs [19], concentric rings [54], and localized patches [2].

Besides, the taxonomy (Table 1) for the re-identification algorithms distinguishes the class of the *single-shot* approaches, focusing on associating pairs of images, each containing one instance of an individual, from the class of *multiple-shot* methods. The latter employs multiple images of the same person as probe or gallery elements. The assumption of the multi-shot methods is that individuals are tracked, so that it is possible to gather lots of images. The hope is that the system will obtain a set of

images that vary in terms of resolution, partial occlusions, illumination, poses, etc. In this way, we can build a significant signature of each individual.

Looking at Table 1, which reports all these 4 paradigms of re-id, it is worth noting that direct single-shot approaches represent the case where the least information is employed. For each individual, we have a single image, whose features are independently extracted and matched against hundreds of candidates. The learning-based multi-shot approaches, instead, are in the opposite situation. The proposed method lies in the class of the direct strategies and works both in the single and in the multi-shot modality.

3 Symmetry-driven Accumulation of Local Features (SDALF)

As discussed in the previous section, we assume to have a set of trackers that estimate the trajectories of each person in the several (non-)overlapped camera views. For each individual, a set of bounding boxes can be obtained (from one or more consecutive frames), and SDALF analyzes these images to build a signature while performing matching for recognizing individuals in a database of pre-stored individuals. The proposed re-id pipeline of SDALF consists of six phases as depicted in Figure 1:

1. *Image Gathering* aggregates images given by the trajectories of the individuals and their bounding boxes.
2. *Image Selection* selects a small set of representative images, when the number of images is very big (*e.g.*, in tracking) in order to discard redundant information. [Section 3.1]
3. *Person Segmentation* separates the pixels of the individual (foreground) from the rest of the image (background) that usually "distracts" the re-id. [Section 3.2]
4. *Symmetry-based Silhouette Partition* detects perceptually salient body regions exploiting symmetry and asymmetry principles. [Section 3.3]
5. *Descriptor Extraction and Accumulation* composes the signature as an ensemble of global or local features extracted from each body part and from different frames. [Section 3.4]
6. *Signature Matching* minimizes a certain similarity score between the probe signature and a set of signatures collected in a database (gallery set). [Section 4]

The nature of this process is slightly different (steps 5 and 6) depending on if we have one or more images, that is, single- or multiple-shot case, respectively.

3.1 Image Gathering and Selection

The first step consists in gathering images of the tracked people. Since there is a temporal correlation between images of each tracked individual, redundancy is expected. Redundancy is therefore eliminated by applying the unsupervised Gaussian

clustering method [17] that is able to automatically select the number of clusters. Hue Saturation Value (HSV) histogram of the cropped image of the individual is used as the feature for clustering, in order to capture appearance similarities across different frames. HSV histograms are invariant to small changes in illumination, scale and pose, so different clusters will be obtained. The output of the algorithm is a set of N_k clusters for each person (k stays for the k-th person). Then, we build the set $\mathbf{X}^k = \{X_n^k\}_{n=1}^{N_k}$ by randomly selecting an image of the k-th person for each cluster. Experimentally, we found that clusters with a small number of elements ($= 3$ in our experiments) usually contain outliers, such as occlusions or partial views of the person, thus these clusters are discarded. It is worth noting that the selected clusters can still contain occlusions and bad images, hard for the re-id task.

3.2 Person Segmentation

Person segmentation allows the descriptor to focus on the individual foreground, avoiding being distracted from the noisy background. When videos are available (*e.g.*, a video-surveillance scenario), foreground extraction can be performed with standard motion-based background subtraction strategies such as [49, 13, 11, 40]. In this work, the standard re-id datasets, which contain only still images, constrained us to use the Stel Component Analysis (SCA) [27]. However, we claim that any other person segmentation method can be used as a component of SDALF.

SCA lies on the notion of "structure element" (stel), which can be intended as an image portion whose topology is consistent over an image class. In a set of given objects, a stel is able to localize common parts over all the instances (*e.g.*, the body in a set of images of pedestrians). SCA extends the stel concept as it captures the common structure of an image class by blending together multiple stels. SCA has been learned beforehand on a person database not considering the experimental data, and the segmentation over new samples consists in a fast inference (see [27, 5] for further details).

3.3 Symmetry-based Silhouette Partition

The goal of this phase is to partition the human body into salient parts, exploiting asymmetry and symmetry principles. Considering a pedestrian acquired at very low resolution (see some examples in decreasing resolutions in Fig. 2), it is easy to note that the most distinguishable parts are three: head, torso and legs. We present a method that is able to work at very low resolution, where more accurate part detectors, such as the pictorial structures [9], fail.

Let us first introduce the *chromatic bilateral operator* defined as:

$$C(i,\delta) \propto \sum_{B_{[i-\delta,i+\delta]}} d^2\left(p_i, \hat{p}_i\right) \tag{1}$$

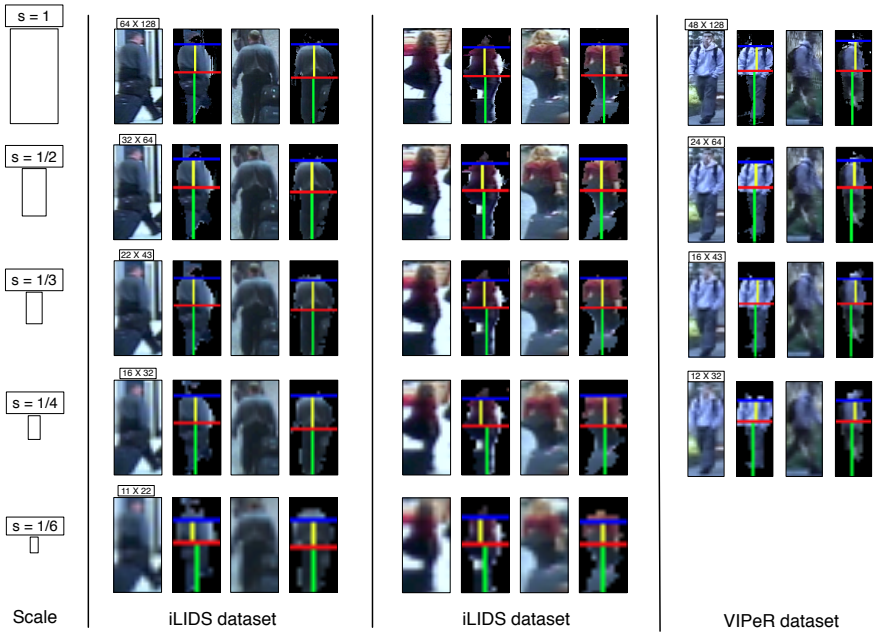


Fig. 2 Images of individuals at different resolutions (from 64×128 to 11×22) and examples of foreground segmentation and symmetry-based partitions.

where $d(\cdot,\cdot)$ is the Euclidean distance, evaluated between HSV pixel values $p_i, \hat{p}_i$, located symmetrically with respect to the horizontal axis at height i. This distance is summed up over $B_{[i-\delta,i+\delta]}$, *i.e.*, the foreground region (as estimated by the object segmentation phase) lying in the box of width J and vertical extension $2\delta+1$ around i (see Fig. 3). We fix $\delta = I/4$, proportional to the image height, so that scale independency can be achieved.

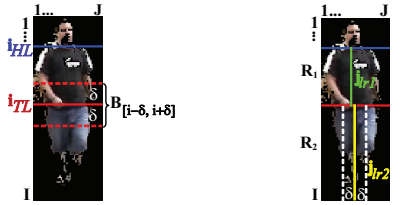


Fig. 3 Symmetry-based silhouette partition. First the asymmetrical axis i_{TL} is extracted, then i_{HT}; afterwards, for each R_k, $k = \{1,2\}$ region the symmetrical axes j_{LRk} are computed.

The second operator is the *spatial covering operator*, which calculates the difference of foreground areas for two regions:

$$S(i,\delta) = \frac{1}{J\delta}\left|A\left(B_{[i-\delta,i]}\right) - A\left(B_{[i,i+\delta]}\right)\right|, \tag{2}$$

where $A\left(B_{[i-\delta,i]}\right)$, similarly as above, is the foreground area in the box of width J and vertical extension $[i-\delta,i]$.

Combining opportunely C and S gives the axes of symmetry and asymmetry. The main x-axis of asymmetry is located at height i_{TL}:

$$i_{TL} = \underset{i}{\operatorname{argmin}}\,(1 - C(i,\delta)) + S(i,\delta), \tag{3}$$

i.e., we look for the x-axis that separates regions with strongly different appearance and similar area. The values of C are normalized by the numbers of pixels in the region $B_{[i-\delta,i+\delta]}$. The search for i_{TL} holds in the interval $[\delta,\ I-\delta]$: i_{TL} usually separates the two biggest body portions characterized by different colors (corresponding to t-shirt/pants or suit/legs, for example).

The other x-axis of asymmetry is positioned at height i_{HT}, obtained as:

$$i_{HT} = \underset{i}{\operatorname{argmin}}\,(-S(i,\delta)). \tag{4}$$

This asymmetry axis separates regions that strongly differ in area and places i_{HT} between head and shoulders. The search for i_{HT} is limited in the interval $[\delta, i_{TL}-\delta]$.

The values i_{HT} and i_{TL} isolate three regions R_k, $k = \{0,1,2\}$, approximately corresponding to head, body and legs, respectively (see Fig. 3). The head part R_0 is discarded, because it often consists in few pixels, carrying very low informative content.

At this point, for each part R_k, $k = \{1,2\}$, a (vertical) symmetry axis is estimated, in order to individuate the areas that most probably belong to the human body, *i.e.*, pixels near the symmetry axis. In this way, the risk of considering background clutter is minimized.

On both R_1 and R_2, the y-axis of symmetry is estimated in j_{LRk}, $(k = 1,2)$, obtained using the following operator:

$$j_{LRk} = \underset{j}{\operatorname{argmin}}\, C(j,\delta) + S(j,\delta). \tag{5}$$

This time, C is evaluated on the foreground region of the size of the height R_k timing the width δ (see Fig. 3). We look for regions with similar appearance and area. In this case, δ is proportional to the image width, and it is fixed to $J/4$.

In Fig. 2, different individuals are taken in different shots. As one can observe, our subdivision segregates correspondent portions independently on the assumed pose and the adopted resolution.

3.4 Accumulation of Local Features

Different features are extracted from the detected parts R_1 and R_2 (torso and legs, respectively). The goal is to extract as much complementary information as possible in order to encode heterogeneous information of the individuals. Each feature is extracted by considering its distance with respect to the vertical axes. The basic idea is that locations far from the symmetry axis belong to the background with higher probability. Therefore, features coming from those areas have to be either a) weighted accordingly or b) discarded.

Considering the literature in human appearance modeling, features may be grouped by considering the kind of information to focus on, that is, chromatic (histograms), region-based (blobs), and edge-based (contours, textures) information. Here, we consider a feature for each aspect, showing later their importance (see Fig. 4(c-e) for a qualitative analysis of the feature for the SDALF descriptor).

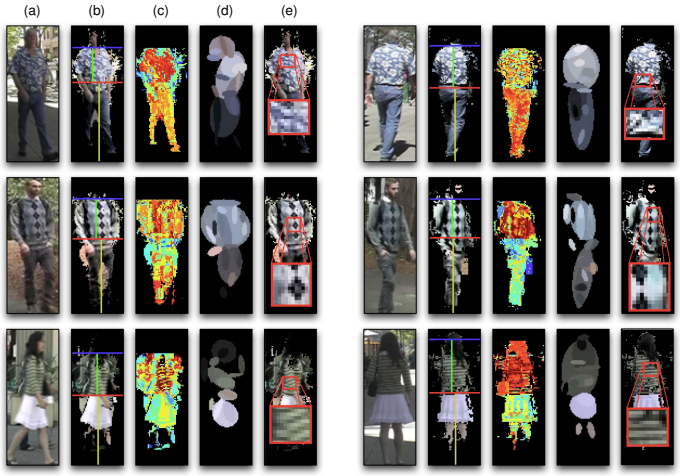


Fig. 4 Sketch of the SDALF descriptor for single-shot modality. (a) Given an image or a set of images, (b) SDALF localizes meaningful body parts. Then, complementary aspects of the human body appearance are extracted: (c) weighted color histogram, the values accumulated in the histogram are back-projected into the image to show which colors of the image are more important. (d) Maximally Stable Color Regions [18] and (e) Recurrent Highly Structured Patches. The objective is to correctly match SDALF descriptors of the same person (first column vs sixth column).

Weighted Color Histograms (WCH)

The chromatic content of each part of the pedestrian is encoded by color histograms. We evaluate different color spaces, namely, HSV, RGB, normalized RGB (where each channel is normalized by the sum of all the channels), per-channel normalized

RGB [2], and CIELAB. Among these, HSV has been shown to be superior and also allows an intuitive quantization against different environmental illumination conditions and camera acquisition settings.

We define *weighted color histograms* of the foreground regions that take into consideration the distance to the vertical axes. In particular, each pixel is weighted by a one-dimensional Gaussian kernel $\mathcal{N}(\mu,\sigma)$, where μ is the y-coordinate of j_{LRk}, and σ is a priori set to $J/4$. The nearer a pixel is to j_{LRk}, the more important it will be. In the single-shot case, a single histogram for each part is built. Instead, in the multiple-shot case, as M instances, all the M histograms for each part are considered during matching (see Section 4).

The advantage of using the weighted histogram is that in practice the person segmentation algorithm is prone to error especially in the contour of the silhouette. The weighted histogram is able to reduce the noise of the masks that contain background pixels wrongly detected as foreground.

Maximally Stable Color Regions (MSCR)

The MSCR operator[1] [18] detects a set of blob regions by looking at successive steps of an agglomerative clustering of image pixels. Each step clusters neighboring pixels with similar color, considering a threshold that represents the maximal chromatic distance between colors. These maximal regions that are stable over a range of steps represent the maximally stable color regions of the image. The detected regions are then described by their area, centroid, second moment matrix and average RGB color, forming 9-dimensional patterns. These features exhibit desirable properties for matching: covariance to adjacency, preserving transformations and invariance to scale changes, and affine transformations of image color intensities. Moreover, they show high repeatability, *i.e.*, given two views of an object, MSCRs are likely to occur in the same correspondent locations.

In the single-shot case, we extract MSCRs separately from each part of the pedestrian. In order to discard outliers, we select only MSCRs that lie inside the foreground regions. In the multiple-shot case, we opportunely accumulate the MSCRs coming from the different images by employing a Gaussian clustering procedure [17], which automatically selects the number of components. Clustering is carried out using the 5-dimensional MSCR sub-pattern composed by the centroid and the average RGB color of each blob. We cluster the blobs similar in appearance and position, since they yield redundant information. The contribution of the clustering is twofold: i) it captures only the relevant information, and ii) it keeps low the computational cost of the matching process, when the clustering results are used. The final descriptor is built by a set of 4-dimensional MSCR sub-pattern composed by the y coordinate and the average RGB color of each blob. Please note that x coordinates are discarded because they are strongly dependent on the pose and viewpoint variation.

[1] Code available at `http://www2.cvl.isy.liu.se/~perfo/software/`.

Recurrent High-Structured Patches (RHSP)

This feature was designed in [14], taking inspiration from the image epitome [26]. The idea is to extract image patches that are highly recurrent in the human body figure (see Fig. 5). Differently from the epitome, we want to take into account patches that are 1) informative, and 2) that can be affected by rigid transformations. The first constraint selects only those patches that are informative in an information theoretic sense. Inspired by [51], RHSP uses entropy to select textural patches with strong edges. The higher the entropy is, the more likely it is to have a strong texture. The second requirement takes into account that the human body is a 3D entity whose parts may be captured with distortions, depending on the pose. For simplicity, we modeled the human body as a vertical cylinder. In these conditions, the RHSP generation consists in three phases.

The first step consists in the random extraction of patches p of size $J/6 \times I/6$, independently of each foreground body part of the pedestrian. In order to take the vertical symmetry into consideration, we mainly sample the patches around the j_{LRk} axes, exploiting the Gaussian kernel used for the color histograms computation. In order to focus on informative patches, we operate a thresholding on the entropy values of the patches, pruning away patches with low structural information (e.g., uniformly colored). This entropy is computed as the sum H_p of the pixel entropy of each RGB channel. We choose those patches with H_p higher than a fixed threshold τ_H ($= 13$ in all our experiments). The second step applies a set of transforma-

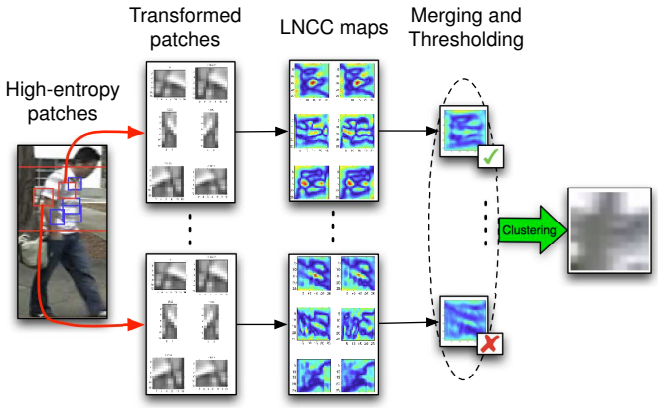


Fig. 5 Recurrent high-structured patches extraction.

tions T_i, $i = 1,2,\ldots,N_T$ on the generic patch p, for all the sampled p's in order to check their invariance to (small) body rotations, *i.e.*, considering that the camera may capture the one's front, back or side, and supposing the camera is at the face's

height. We thus generate a set of N_T *simulated* patches p_i, gathering an enlarged set $\hat{p} = \{p_1, \ldots, p_{N_T}, p\}$.

In the third and final phase, we investigate how much recurrent a patch is. We evaluate the Local Normalized Cross-Correlation (LNCC) of each patch in $\hat{p}$ with respect to the original image. All the $N_T + 1$ LNCC maps are then summed together forming an average map. Averaging again over the elements of the map indicates how much a patch, and its transformed versions, are present in the image. Thresholding this value ($\tau_\mu = 0.4$) generates a set of candidates RHSP patches. The set of RHSPs is generated through clustering [17] of the LBP description [37] in order to capture patches with similar textural content. For each cluster, the patch closer to each centroid composes the RHSP.

Given a set of RHSPs for each region R_1 and R_2, the descriptor consists of an HSV histogram of these patches. We have tested experimentally the LBP descriptor, but it turned out to be less robust than color histograms. The single-shot and the multiple-shot methods are similar, with the only difference that in the multi-shot case the candidate RHSP descriptors are accumulated over different frames.

Please note that, even if we have several thresholds that regulate the feature extraction, they have been fixed once, and left unchanged in all the experiments. The best values have been empirically selected using the first 100 image pairs of the VIPeR dataset.

4 Signature Matching

In a general re-id problem two sets of signatures are available: a gallery set A and a probe set B. Re-id consists in associating the signature $\mathbf{P}^B$ of each person in B to the corresponding signature $\mathbf{P}^A$ of each person in A. The matching mechanism depends on how the two sets are organized, more specifically, on how many pictures are present for each individual. This gives rise to three matching philosophies: 1) *single-shot vs. single-shot* (SvsS), if each image in a set represents a different individual; 2) *multiple-shot vs. single-shot* (MvsS), if each image in B represents a different individual, while in A each person is portrayed in different images, or *instances*; 3) *multiple-shot vs. multiple-shot* (MvsM), if both A and B contain multiple instances per individual.

In general, we can define re-id as a maximum log-likelihood estimation problem. More specifically, given a probe B matching is carried out by:

$$A^* = \arg\max_A \left(\log P(\mathbf{P}^A|\mathbf{P}^B)\right) = \arg\min_A \left(d(\mathbf{P}^A, \mathbf{P}^B)\right) \tag{6}$$

where the equality is valid because we define $P(\mathbf{P}^A|\mathbf{P}^B)$ in Gibbs form $P(\mathbf{P}^A|\mathbf{P}^B) = e^{-d(\mathbf{P}^A,\mathbf{P}^B)}$ and $d(\mathbf{P}^A, \mathbf{P}^B)$ measures the distance between two descriptors.

The *SDALF matching distance d* is defined as a convex combination of the local features:

$$d(\mathbf{P}^A, \mathbf{P}^B) = \sum_{f \in F} \beta_f \cdot d_f(f(\mathbf{P}^A), f(\mathbf{P}^B)) \tag{7}$$

where the $F = \{\text{WCH}, \text{MSCR}, \text{RHSP}\}$ is the set of the feature extractors, and βs are normalized weights.

The distance d_{WCH} considers the weighted color histograms. In the SvsS case, the HSV histograms of each part are concatenated channel by channel, then normalized, and finally compared via Bhattacharyya distance [28]. Under the MvsM and MvsS policies, we compare each possible pair of histograms contained in the different signatures, keeping the lowest distance.

For d_{MSCR}, in the SvsS case, we estimate the minimum distance of each MSCR element b in $\mathbf{P}^B$ to each element a in $\mathbf{P}^A$. This distance is defined by two components: d_y^{ab}, which compares the y component of the MSCR centroids; the x component is ignored, in order to be invariant with respect to body rotations. The second component is d_c^{ab}, which compares the MSCR color. In both cases, the comparison is carried out using the Euclidean distance.

The two components are combined as:

$$d_{\text{MSCR}} = \sum_{b \in \mathbf{P}^B} \min_{a \in \mathbf{P}^A} \gamma \cdot d_y^{ab} + (1-\gamma) \cdot d_c^{ab} \tag{8}$$

where γ takes values between 0 and 1. In the multi-shot cases, the set $\mathbf{P}^A$ becomes a subset of blobs contained in the most similar cluster to the MSCR element b.

The distance d_{RHSP} is obtained by selecting the best pair of RHSPs, one in $\mathbf{P}^A$ and one in $\mathbf{P}^B$, and evaluating the minimum Bhattacharyya distance among the RHSP's HSV histograms. This is done independently for each body part (excluding the head), summing up all the distances achieved, and then normalizing for the number of pairs.

In our experiments, we fix the values of the parameters as follows: $\beta_{\text{WCH}} = 0.4$, $\beta_{\text{MSCR}} = 0.4$, $\beta_{\text{RHSP}} = 0.2$ and $\gamma = 0.4$. These values are estimated by cross validating over the first 100 image pairs of the VIPeR dataset, and left unchanged for all the experiments.

4.1 Analysis

The signature of SDALF and its characteristics for both the single-shot and multi-shot descriptors are summarized in Table 2. The second column reports which feature the basic descriptor is constructed from. The third and forth columns show the encoding used as description and the distance used in the matching module, respectively, in the case of the single-shot version of SDALF. The last two columns report the same information for the multi-shot version.

Please note that even though the encoding of each descriptor is based on the color component, the way in which they are constructed is completely different. Therefore, the descriptors give a different mode/view of the same data. Color description

has revealed one of the most useful features in appearance-based person re-id that usually gives the main contribution in terms of accuracy.

		Single-shot		Multi-shot	
	Construction Cue	Encoding	Distance	Encoding	Distance
WCH	Color	HSV hist. per region	Bhattacharyya	Accumulate	Min over distance pairs
MSCR	Color	RGB color + y position per blob	Eq. 8	Clustering	Eq. 8 using clusters
RHSP	Texture	HSV hist. per recurrent patch	Bhattacharyya	Accumulate	Min over distance pairs

Table 2 Summary of the characteristics of SDALF.

In terms of computational speed[2], we evaluate how long the computation of the descriptor and the matching phase (Eq. 7) take in average on images of size 48×128. Partitioning of the silhouette in (a-)symmetric parts takes 56 milliseconds per image. SDALF is then composed by three descriptors WCH, MSCR and RHSP that take 6, 31 and 4843 milliseconds per image, respectively. It is easy to note that the actual bottleneck of the computation of SDALF is the RHSP. Matching is performed independently for each descriptor and it takes less that 1 millisecond per pair of images for WCH and RHSP and 4 milliseconds per image for MSCR. In terms of computational complexity, the computation of the SDALF descriptor is linear in the number of images, while the matching phase is quadratic.

5 SDALF for Tracking

In tracking, a set of hypotheses of the object position on the image are analyzed at each frame, in order to find the one which best fits the target appearance, *i.e.*, the *template*. The paradigm is different from the classical re-id: the gallery set is now the hypothesis set, which is different for each target. And the goal is to distinguish the target from the background and from the other visible targets. The problem of tracking shares some aspects with re-id: for example, the background can be hardly discernible from the background. Another example is when people are relatively close to each other in the video. In that case, hypotheses of a person position may go to the background or the wrong person. A descriptor specifically created for re-id better handles these situations.

The goal of tracking is thus to perform a soft matching, *i.e.*, compute the likelihood between the probe set (the target template) and the gallery set (the hypothesis set) without performing hard matching, like in re-id.

In this section, we briefly describe particle filtering for tracking (Sec. 5.1) and we exploit SDALF as appearance model (Sec. 5.2).

[2] The following values have been computed using our non-optimized MATLAB code on a quad-core Intel Xeon E5440, 2.83 GHz with 30 GB of RAM.

5.1 Particle Filter

Particle filter offers a probabilistic framework for recursive dynamic state estimation [12] that fits the tracking problem. The goal is to determine the posterior distribution $p(x_t|z_{1:t})$, where x_t is the current state, z_t is the current measurement, and $x_{1:t}$ and $z_{1:t}$ are the states and the measurements up to time t, respectively. The Bayesian formulation of $p(x_t|z_{1:t})$ enable us to rewrite the problem as:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int_{x_{t-1}} p(x_t|x_{t-1}) p(x_{t-1}|z_{1:t-1}) dx_{t-1}. \tag{9}$$

Particle Filter is fully specified by an initial distribution $p(x_0)$, a dynamical model $p(x_t|x_{t-1})$, and an observation model $p(z_t|x_t)$. The posterior distribution at previous time $p(x_{t-1}|z_{1:t-1})$ is approximated by a set of S weighted particles, *i.e.* $\{(x_{t-1}^{(s)}, w_{t-1}^{(s)})\}_{s=1}^{S}$, because the integral in Eq. 9 is often analytically intractable. Equation 9 can be rewritten by its Monte Carlo approximation:

$$p(x_t|z_{1:t}) \approx \sum_{s=1}^{S} w_t^{(s)} \, \delta(x_t - x_t^{(s)}). \tag{10}$$

where

$$w_t^{(s)} \propto w_{t-1}^{(s)} \, \frac{p(z_t|x_t^{(s)}) \; p(x_t^{(s)}|x_{t-1}^{(s)})}{q(x_t^{(s)}|x_{t-1}^{(s)}, z_t)} \tag{11}$$

where q is called *proposal distribution*. The design of an optimal proposal distribution is a critical task. A common choice is $q(x_t^{(n)}|x_{t-1}^{(n)}, z_t) = p(x_t^{(n)}|x_{t-1}^{(n)})$ because it simplifies equation (11) in $w_t^{(n)} \propto w_{t-1}^{(n)} \; p(z_t|x_t^{(n)})$. However, this is not an optimal choice. We can make use of the observation z_t in order to propose particles in more interesting regions of the state space. As in [38], detections are used in the proposal distribution to guide tracking and make it more robust.

Given this framework, tracking consists of observing the image z_t at each time t and updating the distribution over the state x_t by propagating particles as in Eq. 11.

5.2 SDALF as Observation Model

The basic idea is to propose a new observation model $p(z_t|x_t^{(s)})$ so that the object representation is made up by the SDALF descriptor. We define the observation model considering the distance defined in Eq. 7: $d(\mathbf{P}_A|\mathbf{P}_B) := d(x_t^{(s)}(z_t), \tau_t)$, where $\mathbf{P}_B$ becomes the object template τ_t made by SDALF descriptors, and $\mathbf{P}_A$ is the current hypothesis $x_t^{(s)}$. Minimization of Eq. 6 over the gallery set elements is not per-

formed for tracking. Instead, the probability distribution over the hypotheses is kept in order to approximate Eq. 9.

Some simplifications are required when embedding SDALF into the proposed tracking framework. First of all, since the descriptor has to be extracted for each hypothesis $x_t^{(s)}$, it should be reasonably efficient to compute. In our current implementation, the computation of RHSP for each particle is not feasible as the transformations T_i performed on the original patches to make the descriptor invariant to rigid transformations constitute a too high burden. Therefore, the RHSP is not used in the descriptor.

The observation model becomes:

$$p(z_t|x_t^{(s)}) = e^{-D(x_t^{(s)}(z_t),\tau_t)}, \qquad D(x_t^{(s)}(z_t),\tau_t) = \sum_{f\in F_R} \beta_f \cdot d_f(f(x_t^{(s)}), f(\tau_t)) \quad (12)$$

where $x_t^{(s)}$ is the hypothesis extracted from the image z_t, and τ_t is the template of the object and $F_R = \{\mathrm{WCH}, \mathrm{MSCR}\}$. During tracking, the object template has to be updated in order to model the different aspects of the captured object (for example, due to different poses). Therefore, τ_t is composed by a set of images accumulated over time (previous L frames). Then, in order to balance the number of images employed for building the model and the computational effort required, $N = 3$ images are randomly selected at each time step to form $\mathbf{P}_A$.

6 Experiments

In this section, an exhaustive analysis of SDALF for re-identification and tracking is presented. SDALF is evaluated on the re-id task against the state-of-the-art methods in Section 6.1. Then, it is evaluated on a tracking scenario in Section 6.2.

6.1 Results: Re-identification

In literature, several different datasets are available: VIPeR[3] [20], iLIDS for re-id [54], ETHZ[4] 1, 2, and 3 [47], and the more recent CAVIAR4REID[5] [9]. These datasets cover challenging aspects of the person re-id problem, such as shape deformation, illumination changes, occlusions, image blurring, very low resolution images, *etc*.

[3] Available at `http://users.soe.ucsc.edu/~dgray/VIPeR.v1.0.zip`

[4] Available at `http://www.liv.ic.unicamp.br/~wschwartz/datasets.html`

[5] Available at `http://www.lorisbazzani.info/code-datasets/caviar4reid/`

Datasets. The VIPeR dataset [20] contains image pairs of 632 pedestrians normalized to 48×128 pixels. It represents one of the most challenging single-shot datasets currently available for pedestrian re-id.

The ETHZ dataset [47] is captured from moving cameras in a crowded street and contains three sub-datasets: ETHZ1 with 83 people (4.857 images), ETHZ2 with 35 people (1.936 images), and ETHZ3 contains 28 with (1.762 images). ETHZ does not represent a genuine re-id scenario (no different cameras are employed), and it still carries important challenges not exhibited by other public datasets, as the big number of images per person.

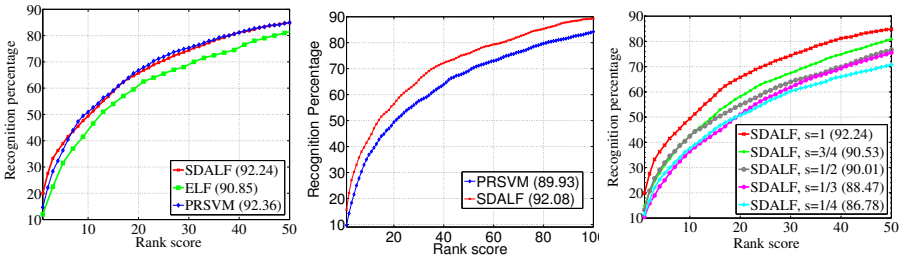
The iLIDS for re-id [54] dataset is composed by 479 images of 119 people acquired from non-overlapping cameras. However, iLIDS does not fit well in a multi-shot scenario because the average number of images per person is 4, and thus some individuals have only two images. For this reason, we also created a modified version of the dataset of 69 individuals, named $\text{iLIDS}_{\geq 4}$, where we selected the subset of individuals with at least 4 images.

The CAVIAR4REID dataset [9] contains images of pedestrians extracted from the shopping center scenario of the CAVIAR dataset[6]. The ground truth of the sequences was used to extract the bounding box of each pedestrian, resulting in a set of 10 images of 72 unique pedestrians: 50 with the two camera views and 22 with one camera view. The main differences of CAVIAR4REID with respect to the already-existing datasets for re-id are: 1) it has broad changes of resolution, and the minimum and maximum size of the images contained on CAVIAR4REID dataset are 17×39 and 72×144, respectively. 2) Unlike ETHZ, it is extracted from a real scenario where re-id is necessary due to the presence of multiple cameras and 3) pose variations are severe. 4) Unlike VIPeR, it contains more than one image for each view. 5) It contains all the images variations of the other datasets.

Evaluation Measures. State-of-the-art measurements are used in order to compare the proposed methods with the others: the Cumulative Matching Characteristic (CMC) curve represents the expectation of finding the correct match in the top n matches and the normalized Area Under the Curve (nAUC) is the area under the entire CMC curve normalized over the total area of the graph. We compare the proposed method with some of the best re-id methods on the available datasets: Ensemble of Localized Features (ELF) [21] and Primal-based Rank-SVM (PRSVM) [41] in VIPeR, Partial Least Squares (PLS) by [47] in ETHZ, Context-based re-id [54] and Spatial Covariance Region (SCR) [2] in iLIDS.

Results. Considering first the VIPeR dataset, we define CAM B as the gallery set, and CAM A as the probe set; each image of the probe set is matched with the images of the gallery. This provides a ranking for every image in the gallery with respect to the probe. We followed the same experimental protocol of [21]. In this work, the dataset is split evenly into a training and a test set, and matching is performed. In both algorithms a set of few random permutations are performed (5 runs for PRSVM, 10 runs for ELF), and the averaged score is kept. In order to fairly compare

[6] Available at `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`



(a) 316 ped. (b) 474 ped. (c) scale, 316 ped.

Fig. 6 Performances on the VIPeR dataset in terms of CMC and nAUC (within brackets). In (a) and (b), comparative profiles of SDALF against ELF [21] and PRSVM [41] on the 316-pedestrian dataset and the 474-pedestrian dataset, respectively. In (c), SDALF at different scales.

our results with theirs, we should know precisely the splitting assignment. Since this information is not provided we compare the existent results with the average of the results obtained by our method for 10 different random sets of 316 pedestrians and 474 pedestrians. In Fig. 6, we depict a comparison among ELF, PRSVM and SDALF in terms of CMC curves. We provided also the nAUC score for each method (within brackets in the legend of the plots of Fig. 6). Considering the experiment on 316 pedestrians (Fig. 6(a)), SDALF outperforms ELF in terms of nAUC, and we obtain comparable results with respect to PRSVM. Even if PRSVM is slightly superior to SDALF, one can note that the differences between it and SDALF are negligible (less than 0.12%). This is further corroborated looking at the different philosophy underlying the PRSVM and our approach. In the former case, PRSVM uses the 316 pairs as training set, whereas in our case we act directly on the test images, operating on each single image as an independent entity. Thus, no learning phase is needed for our descriptor. In addition, it is worth noting that SDALF slightly outperforms PRSVM in the first positions of the CMC curve (rank $1-6$). This means that in a real scenario where only the first ranks are considered, our method performs better.

Fig. 6(b) shows a comparison between PRSVM and SDALF when dealing with a larger test dataset where a set of 474 individuals has been extracted, as done in the PRSVM paper. This is further evidence about how the performance of PRSVM depends on the training set, which is now composed by 158 individuals. In this case, our approach outperforms PRSVM showing an advantage in terms of nAUC of about 2.15%.

The last analysis of this dataset is conducted by testing the robustness of SDALF when the image resolution decreases. We scaled the original images of the VIPeR dataset by factors $s = \{1, 3/4, 1/2, 1/3, 1/4\}$ reaching a minimum resolution of 12×32 pixels (Fig. 2 on the right). The results, depicted in Fig. 6, show that the performance decreases, as expected, but not drastically. nAUC slowly drops down from 92.24% at scale 1 to 86.78% at scale $1/4$.

Now let us analyze the results on iLIDS dataset. We reproduce the same experimental settings of [54] in order to make a fair comparison. We randomly select one

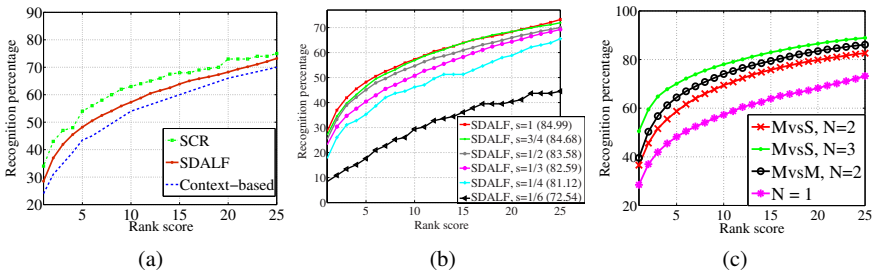


Fig. 7 Performances on iLIDS dataset. (a) CMC curves comparing Context-based re-id [54], SCR [2] and single-shot SDALF. (b) Analysis of SDALF performance at different resolutions. (c) CMC curves for MvsS and MvsM cases varying the average number of images N for each pedestrian. For reference, we put also the single-shot case ($N = 1$). In accordance with what reported by [54], only the first 25 ranking positions of the CMC curves are displayed.

image for each pedestrian to build the gallery set, while the others form the probe set. Then, the matching between probe and gallery set is estimated. For each image in the probe set the position of the correct match is obtained. The whole procedure is repeated 10 times, and the average CMC curves are displayed in Fig. 7.

SDALF outperforms the Context-based method [54] without using any additional information about the context (Fig. 7(a)) even using images at lower resolution (Fig. 7(b)). The experiments of Fig. 7(b) show SDALF when scaling factors are $s = \{1, 3/4, 1/2, 1/3, 1/4, 1/6\}$ with respect to the original size of the images, reaching a minimum resolution of 11×22 pixels. Fig. 7(a) shows that we get lower performance with respect to SCR [2]. Unfortunately, it has been impossible to test SCR on low resolution images (no public code available), but since it is based on covariance of features we expect that second order statistics on very few values may be uninformative and not significant.

Concerning the multiple-shot case, we run experiments on both MvsS and MvsM cases. In the former trial, we built a gallery set of multi-shot signatures and we matched it with a probe set of one-shot signatures. In the latter, both gallery and probe sets are made up of multi-shot signatures. In both cases, the multiple-shot signatures are built from N images of the same pedestrian randomly selected. Since the dataset contains an average of about 4 images per pedestrian, we tested our algorithm with $N = \{2, 3\}$ for MvsS, and just $N = 2$ for MvsM running 100 independent trials for each case. It is worth noting that some of the pedestrians have less than 4 images, and in this case, we simply build a multi-shot signature composed by less instances. In the MvsS strategy, this applies to the gallery signature only, and in the MvsM signature, we start by decreasing the number of instances that compose the probe signature, leaving unchanged the gallery signature; once we reach just one instance for the probe signature, we start decreasing the gallery signature too. The results, depicted in Fig. 7(c), show that, in the MvsS case, just 2 images are enough to increment the performance by about 10% and to outperform the Context-based method [54] and SCR [2]. Adding another image induces an increment of 20% with

respect to the single-shot case. It is interesting to note that the results for MvsM lie in between these two figures.

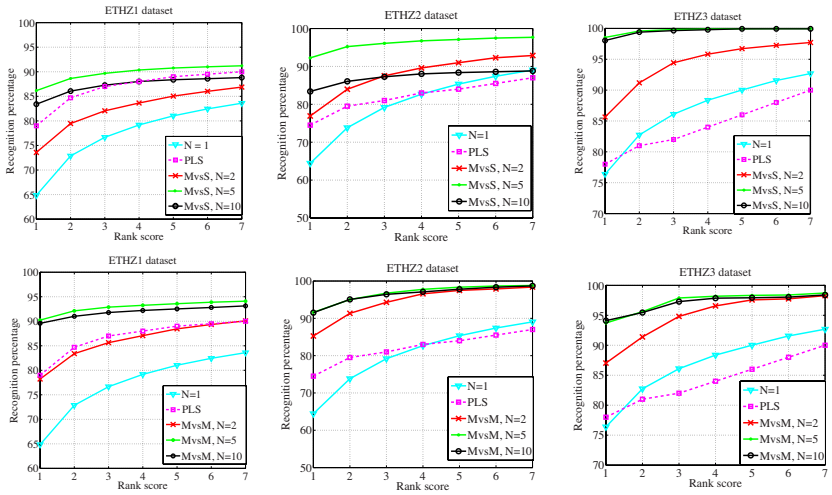


Fig. 8 Performances on ETHZ dataset. Left column, results on SEQ. #1; middle column, on SEQ. #2; right column, on SEQ. #3. We compare our method with the results of PLS [47]. On the top row, we report the results for single-shot SDALF ($N = 1$) and MvsS SDALF; on the bottom row, we report the results for MvsM SDALF. In accordance with [47], only the first 7 ranking positions are displayed.

In ETHZ dataset, PLS [47] produces the best performance. In the single-shot case, the experiments are carried out exactly as for iLIDS. The multiple-shot case is carried out considering $N = 2, 5, 10$ for MvsS and MvsM, with 100 independent trials for each case. Since the images of the same pedestrian come from video sequences, many are very similar and picking them for building the multi-shot signature would not provide new useful information about the subject. Therefore, we apply the clustering procedure discussed in Section 3.1.

The results for both single and multiple-shot cases for SEQ. #1 are reported on Fig. 8, and we compare the results with those reported by [47]. In SEQ. #1 we do not obtain the best results in the single-shot case, but adding more information to the signature we can get up to 86% rank 1 correct matches for MvsS and up to 90% for MvsM. We think that the difference with PLS is due to the fact that PLS uses all foreground and background information, while we use only the foreground. Background information helps here because each pedestrian is framed and tracked in the same location, but it is not valid in general in a multi-camera setting.

In SEQ. #2 (Fig. 8) we have a similar behavior: rank 1 correct matches can be obtained in 91% of the cases for MvsS, and in 92% of the cases for MvsM. The results for SEQ. #3 show instead that SDALF outperforms PLS even in the single-shot case. The best performance as to rank 1 correct matches is 98% for MvsS and

94% for MvsM. It is interesting to note that there is a point after that adding more information does not enrich the descriptive power of the signature any more. $N = 5$ seems to be the correct number of images to use.

Results: AHPE. To prove that the ideas introduced by SDALF should be used in combination with other descriptors, we modified the Histogram Plus Epitome (HPE) descriptor of [6]. HPE is made by two parts: color histograms accumulated over time in the same spirit of SDALF, and the epitome to describe local recurrent motifs. We extended HPE to Asymmetry HPE (AHPE) [6], where HPE is extracted from (a-)symmetric parts in the same partition method used by SDALF. The quantitative evaluation of HPE and AHPE considers the six multi-shot datasets: ETHZ 1, 2, and 3, iLIDS for re-id, iLIDS$_{\geq 4}$, and CAVIAR4REID.

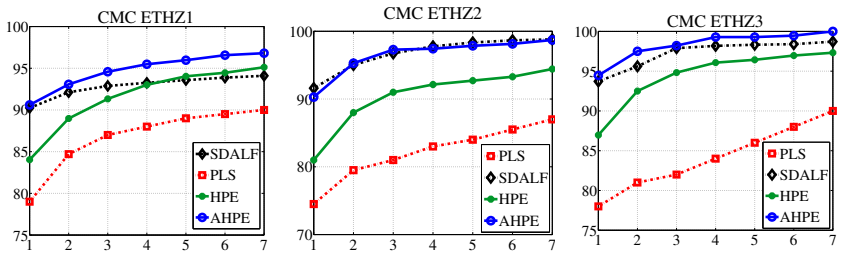


Fig. 9 Comparisons on ETHZ 1,2,3 between AHPE (blue), HPE (green), SDALF (black), PLS [47] (red). For the multi-shot case we set $N = 5$.

A comparison between different state-of-the-art methods in the multi-shot setup ($N = 5$), HPE and AHPE descriptor is shown in Fig. 9. On ETHZ, AHPE gives the best results, showing consistent improvements on ETHZ1 and ETHZ3. On ETHZ2, AHPE gives comparable results with SDALF, since the nAUC is 98.93% and 98.95% for AHPE and SDALF, respectively. Note that if we remove the image selection step (used for ETHZ), the performance decreases of 5% in terms of CMC, because the intra-variance between images of the same individual is low, and thus the multi-shot mode does not gain new discriminative information.

On iLIDS (Fig. 10, left), AHPE is outperformed only by SDALF. This witnesses again the fact, explained in the previous experiment, that the epitomic analysis works very well when the number of instances is appropriate (say, at least $N = 5$). This statement is clearer by the experiments on iLIDS$_{\geq 4}$ and CAVIAR4REID (Fig. 10, last two columns). Especially, if we remove from iLIDS the instances with less than 4 images, then AHPE outperforms SDALF (Fig. 10, center). The evaluation on CAVIAR4REID (Fig. 10, right) shows that: 1) the accuracy increases with N, and 2) the real, worst-case scenario of re-id is still a very challenging open problem.

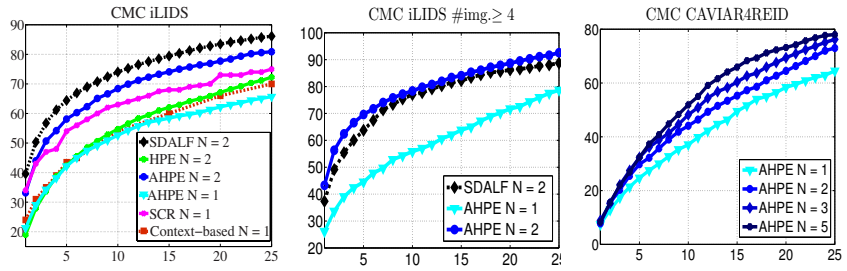


Fig. 10 Comparisons on iLIDS (first column), iLIDS$_{\geq 4}$ (second column) and CAVIAR4REID (third column) between AHPE (blue), HPE (green, only iLIDS), SDALF (black), SCR [2] (magenta, only iLIDS), and context-based [54] (red, only iLIDS). For iLIDS and iLIDS$_{\geq 4}$ we set $N = 2$. For CAVIAR4REID, we analyze different values for N. Best viewed in colors.

6.2 Results: Tracking

As benchmark, we adopt CAVIAR, as it represents a challenging real tracking scenario, due to pose, resolution and illumination changes, and severe occlusions. The dataset consists of several sequences along with the ground truth captured in the entrance lobby of the INRIA Labs and in a shopping center in Lisbon. We select the shopping center scenario, because it mirrors a real situation where people move in the scene. The shopping center dataset is composed by 26 sequences recorded from two different points of view, at the resolution of 384×288 pixels. It includes individuals walking alone, meeting with others, window shopping, entering and exiting shops.

We aim to show the capabilities of SDALF as appearance descriptor in a multi-person tracking case. We use the particle filtering approach described in Sec. 5, since it represents a general tracking engine employed by many algorithms. As proposal distribution, we use the already-trained person detector [15] in the same way exploited by the boosted particle filter [38]. For generating new tracks, weak tracks (tracks initialized for each not associated detection) are kept in memory, and it is checked whether they are supported continuously by a certain amount of detections. If this happens, the tracks are initialized [8].

The proposed SDALF-based observation model is compared against two classical appearance descriptors for tracking: joint HSV histogram and part-based HSV histogram (partHSV) [24] where each of three body parts (head, torso, legs) are described by a color histogram.

The quantitative evaluation of the method is provided by adopting the metrics presented in [29][7]:

- Average Tracking Accuracy (**ATA**): measures penalizing fragmentation phenomena in both the temporal and spatial dimensions, while accounting for the number of objects detected and tracked, missed objects, and false positives;

[7] For the sake of fairness, we use the code provided by the authors. For the metric ATA, we use the association threshold suggested by the authors (0.5).

Table 3 Quantitative comparison between object descriptors: SDALF, part-based HSV histogram and HSV histogram; the performance are given in terms of the number of tracks estimated (# Est.) vs. the number of tracks in the ground truth (# GT), Multi-Object Tracking Precision (MOTP) and Multi-Object Tracking Accuracy (MOTA).

	# Est.	# GT	ATA	MOTP	MOTA
SDALF	**300**	235	**0.4567**	**0.7182**	**0.6331**
partHSV	522	235	0.1812	0.5822	0.5585
HSV	462	235	0.1969	0.5862	0.5899

- Multi-Object Tracking Precision (**MOTP**): considers the spatio-temporal overlap between the reference tracks and the tracks produced by the test method.
- Multi-Object Tracking Accuracy (**MOTA**): considers missed detections, false positives, and ID switches by analyzing consecutive frames.

For more details, please refer to the original paper [29]. In addition, we provide also an evaluation in terms of:

- the number of tracks estimated by our method (# **Est.**) vs. the number of tracks in the ground truth (# **GT**): an estimate of how many tracks are wrongly generated (for example, because weak appearance models cause tracks drifting).

The overall tracking results averaged over all the sequences are reported in Table 3. The number of estimated tracks using SDALF is closer to the correct number than partHSV and HSV. Experimentally, we noted that HSV and partHSV fail very frequently in the case of illumination, pose, and resolution changes and partial occlusions. In addition, several tracks are frequently lost and then re-initialized.

Considering the temporal consistency of the tracks (ATA, MOTA, and MOTP), wa can notice that SDALF definitely outperforms HSV and partHSV. The values of ATA are not so high, because track fragmentation is frequent. This is due to the fact that the tracking algorithm does not explicitly cope with complete occlusions. ATA shows that SDALF gives the best results. This experiment promotes SDALF as an accurate person descriptor for tracking, able to manage the natural noisy evolution of the appearance of people.

7 Conclusions

In this chapter, we presented a pipeline for re-identification and a robust symmetry-based descriptor for modeling the human appearance. SDALF relies on perceptually relevant parts localization driven by asymmetry/symmetry principles. It consists of three features that encode different information, namely, chromatic and structural information, as well as recurrent high-entropy textural characteristics. In this way, robustness to low resolution, pose, viewpoint and illumination variations is achieved. SDALF was shown to be versatile, being able to work using a single image of a person (single-shot modality), or several frames (multiple-shot modality). Moreover,

SDALF was also showed to be robust to very low resolutions, maintaining high performance up to 11×22 windows size.

References

1. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person Re-identification Using Haar-based and DCD-based Signature. In: 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (2010)
2. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In: International Conference on Advanced Video and Signal-based Surveillance (2010)
3. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: European Conference on Computer Vision. Workshops and Demonstrations, *Lecture Notes in Computer Science*, vol. 7583, pp. 433–442 (2012)
4. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Proceedings of the European Conference on Computer Vision, pp. 404–417 (2006)
5. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding **117**(2), 130 – 144 (2013)
6. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters **33**(7), 898–903 (2012)
7. Bird, N., Masoud, O., Papanikolopoulos, N., Isaacs, A.: Detection of loitering individuals in public transportation areas. IEEE Transactions on Intelligent Transportation Systems **6**(2), 167 – 177 (2005)
8. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: IEEE International Conference on Computer Vision (2009)
9. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference (BMVC) (2011)
10. Cho, M., Lee, K.M.: Bilateral symmetry detection and segmentation via symmetry-growing. In: British Machine Vision Conference (2009)
11. Cristani, M., Bicego, M., Murino, V.: Multi-level background initialization using hidden markov models. In: First ACM SIGMM international workshop on Video surveillance, IWVS '03, pp. 11–20. ACM, New York, NY, USA (2003)
12. Doucet, A., Freitas, N.D., Gordon, N.: Sequential Monte Carlo methods in practice (2001)
13. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of the IEEE **90**(7), 1151 – 1163 (2002)
14. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360 –2367 (2010)
15. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. IEEE Conference on Computer Vision and Pattern Recognition (2010)
16. Figueira, D., Bazzani, L., Minh, H., Cristani, M., Bernardino, A., Murino, V.: Semi-supervised multi-feature learning for person re-identification. In: International Conference on Advanced Video and Signal-based Surveillance (2013)
17. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Transaction on Pattern Analysis and Machine Intelligence **24**(3), 381–396 (2002)
18. Forssén, P.E.: Maximally stable colour regions for recognition and matching. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)

19. Gheissari, N., Sebastian, T.B., Tu, P.H., Rittscher, J., Hartley, R.: Person reidentification using spatiotemporal appearance. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1528–1535 (2006)
20. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (2007)
21. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision (2008)
22. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: IEEE International Conference on Distribuited Smart Cameras, pp. 1–6 (2008)
23. Hirzer, M., Roth, P.M., Kostinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: European Conference on Computer Vision, *Lecture Notes in Computer Science*, vol. 7577, pp. 780–793 (2012)
24. Isard, M., MacCormick, J.: Bramble: a bayesian multiple-blob tracker. In: IEEE International Conference on Computer Vision, vol. 2, pp. 34–41 vol.2 (2001)
25. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking accross non-overlapping views. Computer Vision and Image Understanding **109**, 146–162 (2007)
26. Jojic, N., Frey, B., Kannan, A.: Epitomic analysis of appearance and shape. International Conference on Computer Vision **1**, 34–41 (2003)
27. Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B.: Stel component analysis: Modeling spatial correlations in image class structure. IEEE Conference on Computer Vision and Pattern Recognition pp. 2044–2051 (2009)
28. Kailath, T.: The divergence and bhattacharyya distance measures in signal selection. IEEE Transactions on Communications **15**(1), 52–60 (1967)
29. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 319–336 (2009)
30. Kohler, W.: The task of Gestalt psychology. Princeton NJ (1969)
31. Levinshtein, A., Dickinson, S., Sminchisescu, C.: Multiscale symmetric part detection and grouping. In: International Conference on Computer Vision (2009)
32. Lin, Z., Davis, L.S.: Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: International Symposium on Advances in Visual Computing, pp. 23–34 (2008)
33. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II–205–II–210 Vol.2 (2004)
34. Mignon, A., Jurie, F.: Pcca: A new approach for distance learning from sparse pairwise constraints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2666–2672 (2012)
35. Minh, H.Q., Bazzani, L., Murino, V.: A unifying framework for vector-valued manifold regularization and multi-view learning. In: Proceedings of the 30th International Conference on Machine Learning (2013)
36. Nakajima, C., Pontil, M., Heisele, B., Poggio, T.: Full-body person recognition system. Pattern Recognition Letters **36**(9), 1997–2006 (2003)
37. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51 – 59 (1996)
38. Okuma, K., Taleghani, A., de Freitas, N., Little, J.J., Lowe, D.G.: A boosted particle filter: Multitarget detection and tracking. In: European Conference on Computer Vision, pp. Vol I: 28–39 (2004)
39. Pham, N.T., Huang, W.M., Ong, S.H.: Probability hypothesis density approach for multi-camera multi-object tracking. In: Asian Conference on Computer Vision, pp. I: 875–884 (2007)

40. Pilet, J., Strecha, C., Fua, P.: Making background subtraction robust to sudden illumination changes. European Conference on Computer Vision pp. 567–580 (2008)
41. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: British Machine Vision Conference (2010)
42. Rahimi, A., Dunagan, B., Darrel, T.: Simultaneous calibration and tracking with a network of non-overlapping sensors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 187–194 (2004)
43. Reisfeld, D., Wolfson, H.J., Yeshurun, Y.: Context-free attentional operators: The generalized symmetry transform. International Journal of Computer Vision **14**(2), 119–130 (1995)
44. Riklin-Raviv, T., Sochen, N., Kiryati, N.: On symmetry, perspectivity, and level-set-based segmentation. IEEE Transactions on Pattern Recognition and Machine Intelligence **31**(8), 1458–1471 (2009)
45. Salvagnini, P., Bazzani, L., Cristani, M., Murino, V.: Person re-identification with a ptz camera: an introductory study. In: International Conference on Image Processing (2013)
46. Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V.: A multiple component matching framework for person re-identification. In: Proceedings of the 16th international conference on Image analysis and processing, pp. 140–149 (2011)
47. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing (2009)
48. Sivic, J., Zitnick, C.L., Szeliski, R.: Finding people in repeated shots of the same scene. In: Proceedings of the British Machine Vision Conference (2006)
49. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 2, pp. 252–259 (1999)
50. Taylor, G.W., Sigal, L., Fleet, D.J., Hinton, G.E.: Dynamical binary latent variable models for 3d human pose tracking. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **0**, 631–638 (2010)
51. Unal, G., Yezzi, A., Krim, H.: Information-theoretic active polygons for unsupervised texture segmentation. International Journal of Computer Vision **62**(3), 199–220 (2005)
52. Wang, X., Doretto, G., Sebastian, T.B., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. In: International Conference on Computer Vision, pp. 1–8 (2007)
53. Wu, Y., Minoh, M., Mukunoki, M., Lao, S.: Set based discriminative ranking for recognition. In: European Conference on Computer Vision, pp. 497–510. Springer (2012)
54. Zeng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: British Conference on Machine Vision (2009)
55. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 653–668 (2013)