Symmetry-Driven Accumulation of Local Features for Human Characterization and Re-identification

Loris Bazzani[b], Marco Cristani[a,b], Vittorio Murino[a,b]

[a]*Dept. of Computer Science, University of Verona, Strada le Grazie 15 - 37134 Verona, Italy*
[b]*Istituto Italiano di Tecnologia, Via Morego, 30 - 16163 Genova, Italy*

Abstract

This work proposes a method to characterize the appearance of individuals exploiting body visual cues. The method is based on a symmetry-driven appearance-based descriptor and a matching policy that allows to recognize an individual. The descriptor encodes three complementary visual characteristics of the human appearance: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy. The characteristics are extracted by following symmetry and asymmetry perceptual principles, that allow to segregate meaningful body parts and to focus on the human body only, pruning out the background clutter. The descriptor exploits the case where we have a single image of the individual, as so as the eventuality that multiple pictures of the same identity are available, as in a tracking scenario. The descriptor is dubbed Symmetry-Driven Accumulation of Local Features (SDALF). Our approach is applied to two different scenarios: re-identification and multi-target tracking. In the former, we show the capabilities of SDALF in encoding peculiar aspects of an individual, focusing on its robustness properties across dramatic low resolution images, in presence of occlusions and pose changes, and variations of viewpoints and scene illumination. SDALF has been tested on various benchmark datasets, obtaining in general convincing performances, and setting the state of the art in some cases. The latter scenario shows the benefits of using SDALF as observation model for different trackers, boosting their performances under different respects on the CAVIAR dataset.

Keywords: Human Characterization, Appearance Modeling, Re-identification, Multi-target tracking

1. Introduction

Characterizing humans in surveillance scenarios is a hard task: most of the time people are captured by different low resolution cameras, under occlusions conditions, badly illuminated, and in different poses. The modeling problem becomes even harder when human descriptions serve as ID signatures in a recognition scenario. In this context, a robust modeling of the entire body appearance of a person is mandatory, especially when other classical biometric cues (face, gait) are not available or difficult to catch, due to the sensors' scarce resolution or low frame-rate.

In this paper, we propose a method that exploits a novel descriptor for characterizing human beings. Such method may be cast naturally in the context of the re-identification, *i.e.*, the recognition of a "probe" individual in different locations over cameras with non-overlapping fields of view, employing an opportune matching policy, that considers a large "gallery" set of candidates. In addition, the descriptor can be exploited as object model for tracking, and the matching policy in this case serves to evaluate the model against a set of observations. The idea is that at each frame a soft or probabilistic matching between a probe set (the person template) and the gallery set (the tracking hypotheses or particles) is performed.

The descriptor is the core of our approach, and is dubbed *Symmetry-Driven Accumulation of Local Features* (SDALF). It works on a rectangular region in which a pedestrian has been detected, and supposes that the pedestrian is in upright pose (Fig. 1(a)). It represents a convenient trade-off between the more complex pictorial structures for humans [2], and the whole-body representation [3, 4] employed in many surveillance methods. SDALF is a symmetry-based description of the human body, and it was inspired by the fact that most natural objects and phenomena manifest symmetry in some form, so detecting and characterizing symmetry is a natural way to understand the structure of objects. To support this, the Gestalt psychology school [5] considers symmetry as a fundamental principle of perception: symmetrical elements are more likely integrated into one coherent object than asymmetric regions. This principle has been also largely exploited in Computer Vision for characterizing salient parts of a structured object [6, 7, 8, 9]. In SDALF, asymmetry principles allow to segregate meaningful body parts (head, upper body, lower body), whereas symmetry criteria help in extracting features from the actual human body, pruning out distracting background clutter (Fig. 1(b)). The idea is that features near the vertical axis of symmetry are weighted more than those that are far from it, ensuring to get information from the internal part of the body, trusting less the peripheral portions. This perceptual part localization is robust as it operates at dramatic low resolution (up to 11×22), under pose, viewpoint, and illumination changes. This promotes the use of SDALF for surveillance purposes.

Once body parts are localized, complementary aspects of their appearance are extracted, highlighting: i) the global chro-

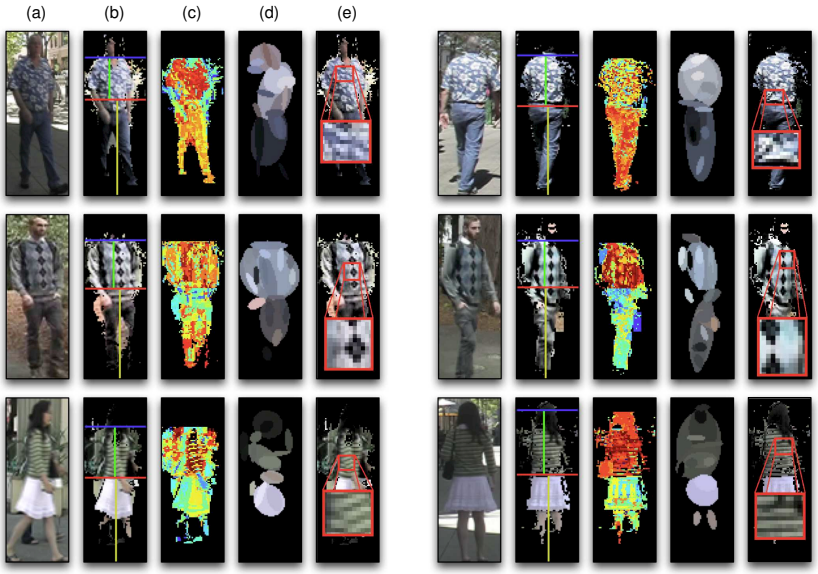


Figure 1: Sketch of the proposed descriptor. (a) Given an image or a set of images, (b) SDALF localizes meaningful body parts. Then, complementary aspects of the human body appearance are extracted: (c) weighted HSV histogram, represented here by its (weighted) back-projection (brighter pixels mean a more important color), (d) Maximally Stable Color Regions [1] and (e) Recurrent Highly Structured Patches. The objective is to correctly match SDALF descriptors of the same person (first column vs. sixth column).

matic content, by the HSV histogram (see Fig. 1(c)); ii) the per-region color displacement, employing Maximally Stable Colour Regions (MSCR) [1] (see Fig. 1(d)); iii) the presence of *Recurrent Highly Structured Patches* (RHSP), estimated by a novel per-patch similarity analysis (see Fig. 1(e)). Such cues have been selected after a comparative analysis of a large set of features, and each one is necessary to capture diverse aspects of each person. Please note that the feature set can be easily adapted to low-cost video surveillance systems, that are usually characterized by acquisitions of a gray level streaming. The HSV histogram can be replaced by the gray levels histograms, the MSCR by the maximally stable extremal regions [10] and the RHSP works also in the case of gray level images.

An important aspect of the proposed method is that it can exploit the presence of multiple instances of the same individual for reinforcing its characterization. This occurs in several surveillance scenarios: to quote a few, human operators may employ Pan-Tilt-Zoom (PTZ) cameras to grab as many images of a suspect as possible. Another example is when a tracker is exploited, consecutive shots of individuals can be used in revising thier object models against appearance changes. SDALF takes into account these situations, collecting features from all the available pictures of an individual, thus augmenting the robustness and the expressiveness of the description. After the descriptor is created, our method adopts a simple distance minimization strategy to match a probe individual with a gallery composed by multiple signatures.

We tested our method in two surveillance applications, i.e. re-identification and tracking. For re-identification we consider several standard public datasets: ViPER [11], iLIDS [12], ETHZ [13] and CAVIAR4REID [14], setting in most of the cases state-of-the-art performances. These datasets embed different challenges for the re-identification problem: pose, viewpoint and lighting variations, and occlusions. Moreover, we test the limit of SDALF by subsampling these dataset up to dramatic resolutions (11×22 pixels). Exploiting SDALF as an appearance model for the tracking, we consider the widely-known CAVIAR [15] sequence dataset and diverse trackers. We show that SDALF outperforms the classical object descriptors considered in the literature, increasing the performances of different trackers.

This paper extends the work of [16], which was focused on re-identification with a smaller experimental section. In this paper, we fully detail the SDALF description and its experimental validation introducing novel tests for the re-identification. In addition, we promote SDALF as object model for multi-person tracking.

The rest of the paper is organized as follows. In Sec. 2, the state of the art of re-identification and tracking is described, highlighting the differences of the existing methods with respect to our strategy. Sec. 3 details the proposed descriptor, and Sec. 4 and 5 report the use of SDALF for re-identification and tracking, respectively. Several comparative results are reported in Sec. 6, and, finally, conclusions and future perspectives are discussed in Sec. 7.

2. Related Work

In this section, we review the state of the art of the re-identification approaches, and we propose a short[1] essay on the different appearance model employed for tracking.

[1]Please note that a review of tracking methods is out of the scope of this paper. Interested readers may refer to [17] for a good review.

	Single-shot	Multiple-shot
Learning-based	[21, 22, 23, 24] [25, 26, 27]	[28]
Direct Methods	[14, 29] **Our Approach**	[14, 30, 31, 32] [33, 34, 35] **Our Approach**

Table 1: Taxonomy of the existing re-identification methods.

Person Re-identification. Re-identification methods that rely only on visual information are addressed here as *appearance-based* techniques. Our approach lies in this category, so the review will mainly focus on this class of methods. Other approaches assume easier and less general operative conditions, for instance, adding temporal reasoning on the spatial layout of the monitored environment, in order to prune away physically impossible matches [18, 19, 20].

Appearance-based methods can be divided into two groups: the *learning-based* methods and the *direct* methods (Table 1). The former group is characterized by the use of a training dataset of *different individuals* where the features and/or the policy for combining them that ensures high re-identification accuracies are analyzed [21, 22, 23, 24, 25, 26, 28, 27]. The underlying assumption is that the knowledge extracted from the training set could generalize to unseen examples. In [23], 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 [25]: spatial and color information are here combined using an ensemble of discriminant localized features and classifiers selected by boosting. In [24], pairwise dissimilarity profiles between individuals are learned and adapted for a nearest neighbor classification. Similarly, in [21], 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. An "unconventional" approach is proposed by [26], where the description of a person is enriched by contextual visual knowledge coming from the surrounding people that form a group. The method implies that a group association between two or more people holds in different locations of a given environment, and exploits novel visual group descriptors, embedding visual words into concentric spatial structures. Re-identification is cast as a binary classification problem (one vs. all) by [27] using Haar-like features and a part-based MPEG7 dominant color descriptor. In [22], the re-identification problem is reformulated as a ranking problem and an informative subspace is learned where the potential true match is given highest ranking. Ensemble RankSVM is proposed as ranking method, reducing significantly the memory requirements.

It is worth noting that the learning-based approaches are strongly dependent on the cardinality and the kind of training set. Such approaches may suffer of generalization problems so that they have to be frequently re-trained/updated, when facing real scenarios (*e.g.*, an airport), while the gallery set changes quickly and consistently (*e.g.*, new individuals entering into the monitored area).

The other class of approaches, the direct methods, does not consider training datasets but rather work on each person independently [14, 30, 31, 29, 32, 33, 34], Those works are usually focused on designing novel features for capturing the most distinguishing aspects of an individual. In [30], 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 [32], 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 [31], 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 from small variations of the point of view. In [33], the person re-identification scheme is based on the matching of SURF [36] interest points collected in several images during short video sequences. Covariance features, originally employed for pedestrian detection [37], are extracted from coarsely located body parts and tailored for re-identification purposes in [29]. The work has been extended in [35] by considering the case where multiple images of the same individual are available. The authors adopt the manifold mean as surrogate of the different covariances coming from the multiple images. Similar features (*i.e.*, MSCR and color histograms) to the ones proposed in this work have been employed also in [14]. The features are extracted from human parts estimated using the pictorial structure detector [2].

Considering the features employed for re-identification, in addition to color information which is universally adopted, several other features of interest are textures [21, 22, 25], edges [21], Haar-like features [27], interest points [32], image patches [25], and segmented regions [31]. These features, when not collected densely, can be extracted from horizontal stripes [30], triangulated graphs [32], concentric rings [26], and localized patches [29].

Another complementary taxonomy (Table 1) for the re-identification algorithms distinguishes the class of the *single-shot* approaches from the class of *multiple-shot* methods. The former focuses on associating pairs of images for each individual, while the latter employs multiple images of the same person as probe and gallery elements.

These four paradigms of re-identification give rise to the taxonomy reported in Table 1. Looking at the table, 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 matched against hundreds of candidates. The learning-based multi-shot approaches, instead, deal with the highest amount of information. Our approach lies in the class of the direct methods, and is versatile, working both in the single and in the multi-shot modality.

In general, learning-based approaches produce higher per-

formances than the direct approaches. However, how stated before, they are not truly suited for a practical usage in surveillance scenarios.

Person Tracking. Here, we focus the discussion on the features for object representation commonly exploited in tracking. Such representations should be robust to deal with the hard recording conditions (*e.g.*, low resolution and scarce illumination). In addition, they have to be computationally efficient in order to comply to the huge number of hypothesis a tracker should evaluates at each time step.

For this review, we follow the scheme proposed by [17], discussing first the data structures useful to represent objects, and then specifying the most common features employed. Points are the poorest object representation, which are suitable for modelling targets that occupy small regions in an image with little overlap. The object can be represented by a single point (the centroid) [38], or a set of sparse points [39]. Covariance matrices of elementary features have been recently adopted to deal with non-rigid objects under different scales [40]. Geometric shapes as rectangle or regular ellipses serve to primarily model simple rigid objects affected by translation, affine, or projective (homography) transformations [41]. Elementary shapes may be employed to encode different body parts, such as head, torso and legs [3, 42]. Patches may also be employed to track salient parts of a target [43]. Contours are suitable for tracking complex non-rigid shapes [44]. Articulated shapes are represented as rigid body parts held together with joints, such as the pictorial structures [45, 46, 2]. Such structures essentially rely on two components, one capturing the local appearance of body parts, and the other representing an articulated body structure. Inferring and detecting pictorial structures involve finding the maximum-a-posteriori spatial configuration of the parts that best fits with the image. Skeletal models [47, 48] can also be extracted by considering the object silhouette, and can be used to model both articulated and rigid objects.

There are many appearance features for objects and the most employed are represented under the form of probability densities. They can be either parametric, such as Gaussian distributions [49] or mixtures of Gaussians [50], or non-parametric, such as Parzen windows [51] and histograms [4, 41]. The probability densities of object appearance features (color, texture) can be computed from the image regions specified by the shape models, *i.e.*, the internal region of an ellipse, a box, or a contour. Templates are formed using simple geometric shapes or silhouettes that model the whole targets or a portion of them [3, 42, 52, 53]. Their advantage is primarily due to the fact that they carry both spatial and appearance information, however they can only encode the object appearance generated from a single view. Thus, they are only suitable for tracking objects whose poses do not vary considerably during the course of tracking. Active appearance models are generated by simultaneously modeling the object shape and appearance [54]. In general, the object shape is defined by a set of landmarks, and similar to the contour-based representation the landmarks can reside on the object boundaries or inside the object region. For each landmark, an appearance vector is stored which is in the form of color, texture, or gradient magnitude. Active appearance models require a training phase where both the shape and its associated appearance is learned from a set of samples using, for instance, principal component analysis [55]. The multi-view appearance models encode different views of an object. One approach to represent the different object views is to generate a subspace from the given views. Subspace approaches such as principal component analysis or independent component analysis have been used for both shape and appearance representations [56, 57].

It is worth noting that a weak appearance modeling of the target is not the only cause of tracking failure. Tracking may also fail when the object model is not properly updated [40, 55, 58, 59], or if a target becomes (even partially) occluded [42, 60, 61, 62].

In this scenario, SDALF is used as appearance representation for tracking. As data structure, we employ the bounding box containing the target, using the symmetry and asymmetry axes to obtain the segmentation of the body parts. As appearance features for objects, we use the weighted histogram (Fig. 1(c)) and the MSCR (Fig. 1(d)) accumulated over time. Please note that RHSP has not been considered for the tracking descriptor, because in practice a system can afford the computation of the RHSP only when dealing with single images (such as the person re-identification task). Instead, since we use particle filtering methods, RHSP has to be computed for each hypothesis, and therefore there is a computational issue in doing this.

3. The SDALF Descriptor

Once an individual has been detected and segregated within a bounding box in one or more frames, the SDALF descriptor can be assembled. The nature of this process is slightly different depending on the modality we are considering, *i.e.*, single- or multiple-shot. The building process of SDALF consists of three phases:

1. *Background subtraction* separates the pixels of the individual (foreground) from the rest of the image (background);
2. *Symmetry-based silhouette partition* individuates perceptually salient body portions;
3. *Symmetry-driven accumulation of local features* composes the signature as an ensemble of features extracted from the body parts.

In the following, each step is described and analyzed focusing on the differences between single-shot and multi-shot modality.

3.1. Background subtraction

The aim of this phase is to separate the genuine body appearance from the rest of the scene. This allows the descriptor to focus solely on the individual, disregarding the context in which it is immersed. We suppose that in a real scenario, a person can be captured at completely different locations, like the arrival hall of an airport, and in the parking lot. In the case of a sequence of consecutive images, the object/scene classification may be operated by a whatsoever background subtraction strategy. In the

case of a single image, the separation is performed by Stel Component Analysis (SCA) [63]. SCA lies on the notion of "structure element" (stel), which can be intended as an image portion (often discontinuous) whose topology is consistent over an image class. This means that in a set of given objects (faces or pedestrian images), a stel individuates the same part over all the instances (*e.g.*, the hair in a set of faces, the body in a set of images each one containing a single pedestrian). In other words, an image can be seen as a segmentation, where each segment is a stel. SCA enriches the stel concept as it captures the common structure of an image class by blending together multiple stels: it assumes that each pixel measurement x_i, with its 2D coordinate i, has an associated discrete variable s_i, which takes a label from the set $\{1, \ldots, S\}$. Such a labeling is generated from K stel priors $p_k(s_i)$, which capture the common structure of the set of images. The model detects the image self-similarity within a segment: the pixels with the same label s are expected to follow a tight distribution over the image measurements. Instead of the local appearance similarity, the model insists on consistent segmentation via the stel prior. Each component k represents a characteristic (pose or spatial configuration) of the object class at hand, and other poses are obtained through blending these components. We set $S = 2$ (i.e., foreground/background) and $K = 2$, modeling the distribution over the image measurements as a mixture of Gaussians as we want to capture segments with multiple color modes within them. SCA is learnt beforehand on a generic person database not considering the experimental data, and the segmentation over new samples consists in a fast inference. Each Expectation-Maximization iteration of the inference algorithm takes in average 18 milliseconds[2] when dealing with images of size 48×128. In our experiments, we set the number of iterations to 100: this for being sure that the learning process reached a local minima of the likelihood function. In practice, we saw that 10-20 iterations are enough in most of the cases.

3.2. Symmetry-based silhouette partition

Background subtraction is used to extract the foreground pixels and also to subdivide the human body into salient parts, exploiting asymmetry and symmetry principles. Considering a pedestrian acquired at very low resolution (see Fig. 3), it is easy to note that the most distinguishable parts are three: head, torso and legs. Focusing on such parts is thus reasonable, and their detection can be exploited observing natural asymmetry properties in the human appearance. In addition, the relevance of head, torso and legs as salient regions for human characterization also emerged from the boosting approach proposed by [25].

Let us define the *chromatic bilateral operator* as:

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

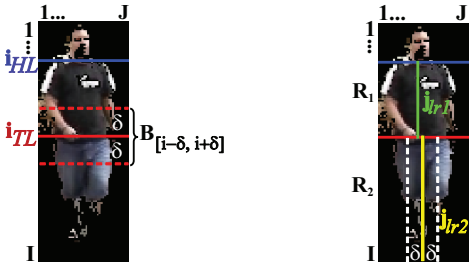


Figure 2: Silhouette Partition: first the asymmetrical axis i_{TL} is extracted, then i_{HT}; afterwards, for each region R_k, $k = \{1, 2\}$ region the symmetrical axis j_{LRk} are computed.

where $d(\cdot, \cdot)$ is the Euclidean distance, evaluated between HSV pixel values[3] $p_i, \hat{p}_i$, located symmetrically with respect to the horizontal axis at height i. The sum is over $B_{[i-\delta, i+\delta]}$, that is the foreground region lying in the box of width J and vertical extension $2\delta + 1$ around i as depicted in Fig. 2. The value of δ is experimetally set to $I/4$, where I is the image height.

Let us also introduce the *spatial covering operator*, that 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 $A(b)$ is a function that computes the foreground area in a given box b and J is the image width.

Combining opportunely the two operators C and S enables us to find the axes of symmetry and asymmetry. To locate the horizontal asymmetry axes, we want to maximize the difference in appearance and the similarity between foreground areas. Therefore, the main x-axis of asymmetry (usually the torso-legs axis) is located at height i_{TL} by solving the following problem:

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

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 (usually the shoulders-head axis) is positioned at height i_{HT}. The goal is to find a local grandient variation in the foreground area:

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

The search for i_{HT} is limited in the interval $[\delta, i_{TL} - \delta]$.

Once computed i_{HT} and i_{TL}, three regions of interest are isolated R_k, $k = \{0, 1, 2\}$, approximately corresponding to head, body and legs, respectively. For re-identification purposes, it is common to discard the information of the head/face region because standard biometric algorithms usually fail at low resolution. Therefore, here R_0 is discarded. For each part

[2] We used the authors' MATLAB code [63] on a quad-core Intel Xeon E5440, 2.83 GHz with 4 GB of RAM.

[3] In case of gray level videos, replacing HSV values with gray levels is straightforward.

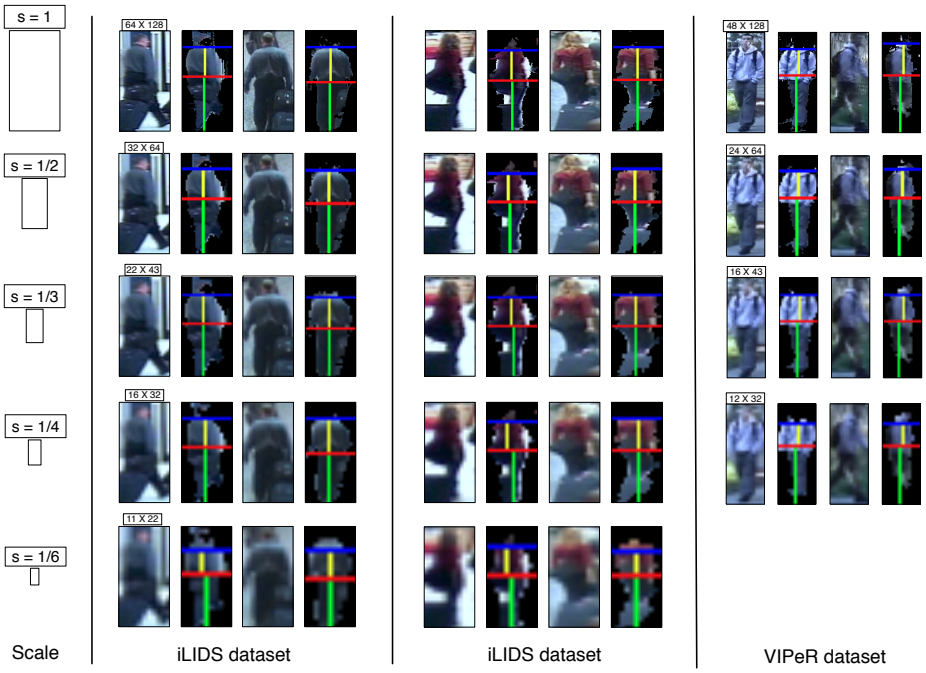


Figure 3: Images of individuals at different resolutions (from 64×128 to 11×22) and examples of foreground segmentation and symmetry-based partitions.

$R_k,\ k = \{1, 2\}$, a (vertical) symmetry axis is estimated, in order to localize 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.

To this end, both the chromatic and spatial covering operator are used on both R_1 and R_2. The y-axes of symmetry $j_{LRk},\ (k = 1, 2)$ are obtained as follows:

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

C is evaluated on the foreground region of size the height of R_k and width δ (see Fig. 2). The goal is to search for regions with similar appearance and area. In this case, δ is proportional to the image width, and it is fixed to $J/4$.

Some results of the optimization process applied to images at different resolutions are shown in Fig. 3. As one can observe, our subdivision segregates correspondent portions independently on the assumed pose and the adopted resolution.

3.3. Symmetry-driven Accumulation of Local Features

The SDALF descriptor is computed by extracting features from each part R_1 and R_2. The goal is to distill as much complementary aspects as possible in order to encode heterogeneous information, so capturing distinctive characteristics of the individuals. Each feature is extracted by taking into account its distance with respect to the j_{LRk} axes. The basic idea is that locations far from the symmetry axis belong to the background with higher probability. Therefore, features coming from that areas have to be a) weighted accordingly or b) discarded. Depending on the considered features, one of these two mechanisms will be applied.

There are many possible cues useful for a fine visual characterization. Considering the previous 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. SDALF considers a feature for each aspect.

Weighted Color Histograms. 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 [29], CIELAB. Among these, HSV has shown to be superior and also allows a intuitive quantization against different environmental illumination conditions and camera acquisition settings.

Therefore, we build *weighted histograms*, so taking into consideration the distance to j_{LRk} 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 to j_{LRk}, the more important. In the single-shot case, a single histogram for each part is built. Instead, in the multiple-shot case, N histograms for each part are considered, where N is the number of images for each pedestrian. Then, the matching policy will handle these multiple histograms properly (see Sec. 4).

Maximally Stable Color Regions (MSCR). The MSCR operator[4] [1] detects a set of blob regions by looking at successive steps of an agglomerative clustering of image pixels. At each

[4]We used the author's implementation, downloadable at `http://www2.cvl.isy.liu.se/~perfo/software/`.

step, neighboring pixels with similar color are clustered, considering a threshold that represents the maximal chromatic distance between colors. Those regions that are stable over a range of steps constitute the maximally stable color regions of the image. The descriptor of each region is a 9-dimentional vector containing area, centroid, second moment matrix and average RGB color. MSCR exhibits desirable properties for matching useful also in re-indeitification: 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, MSCRs are extracted separately from each part of the pedestrian. To discard outliers, MSCRs that do not lie inside the foreground regions are ruled out. In the multiple-shot case, MSCRs from multiple images have to be oppurtunely accumulated. To this end, a mixture of Gaussian clustering procedure [64] that automatically selects the number of components is utilized. 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. This phase helps in discarding redundant information, and keeping low the computational cost during matching because only the representants of each cluster are used. The descriptor is then a set of 4-dimensional MSCR sub-pattern: the y coordinate and the average RGB color of each blob. x coordinates are discarded because they are strongly dependent on the pose and viewpoint variation.

Recurrent High-Structured Patches (RHSP). We design this feature taking inspiration from the image epitome [65]. The idea is to extract image patches that are highly recurrent in the human body figure (see Fig. 4). Differently from the epitome, we want to take into account patches 1) that are informative (in an information theoretic sense, *i.e.*, carrying out high entropy values), and 2) that can be affected by rigid transformations. The first constraint selects only those patches with strong edgeness, such as textures. 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. Since the images have low resolution, we can approximate the human body with 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 on 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. Thus, a Gaussian kernel centered in j_{LRk} is used similarly to the color histograms computation. The patches that do not underline structure (e.g., uniformly colors) are removed by thresholding on the entropy values of the patches. The patch 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). In the second step, a set of transformations T_i, $i = 1, 2, \ldots, N_T$ are applyed on the generic patch p, for all the sampled p's in order to check

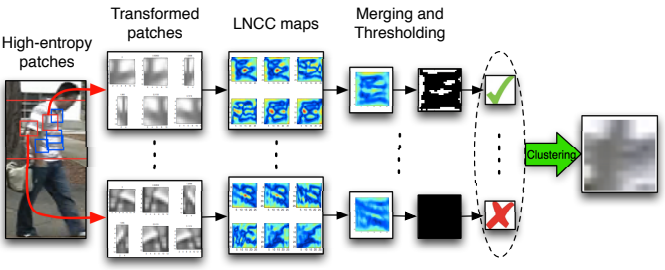


Figure 4: Recurrent high-structured patches extraction. The final result of this process is a set of patches (in this case only one) characterizing each body part of the pedestrian.

their invariance to (small) body rotations. 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, only the most recurrent patches are kept. 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, is present in the image. Thresholding this value does select the RHSP patches. As threshold, we fix $\tau_\mu = 0.4$.

The RHSPs is computed for each region R_1 and R_2, and the descriptor consists again of an HSV histogram of them. The multi-shot case differs from the single-shot case from the fact that 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 selected by qualitatively analyzing the results on the VIPeR dataset.

4. SDALF for Re-identification

In re-identification, two sets of pedestrian images are available: the gallery set A (the database of signatures whose label is known) and the probe set B (the set of tracked pedestrians without label). Re-identification consists in matching each signature in the set B, I_B to the corresponding signature of the set A, I_A. The association 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. The MvsM philosophy is preferred when trajectories of people are available, because one can exploit the redundancy and diversity of the data to make the signature more robust.

Re-identification can be seen as a maximum log-likelihood estimation problem [34]. More in details, given a probe B the

matching is carried out by:

$$A^* = \arg\max_A (\log P(I_A|I_B)) = \arg\min_A (d(I_A, I_B))$$

During testing, we want to match the given probe signature I_B against the gallery set signatures I_A. The goal is to optimize the likelihood of I_A given the probe I_B. The right-hand term of the formula is given by the fact that, in this work, we define the matching probability $P(I_A|I_B)$ in a Gibbs form $P(I_A|I_B) = e^{-d(I_A,I_B)}$ and $d(I_A, I_B)$ measures the distance between two descriptors. The *SDALF matching distance* d is defined as a convex combination of the local features:

$$d(I_A, I_B) = \beta_{\text{WH}} \cdot d_{\text{WH}}(\text{WH}(I_A), \text{WH}(I_B)) + \quad (6)$$

$$\beta_{\text{MSCR}} \cdot d_{\text{MSCR}}(\text{MSCR}(I_A), \text{MSCR}(I_B)) + \quad (7)$$

$$\beta_{\text{RHSP}} \cdot d_{\text{RHSP}}(\text{RHSP}(I_A), \text{RHSP}(I_B)) \quad (8)$$

where the WH$(\cdot)$, MSCR$(\cdot)$, and RHSP$(\cdot)$ are the weighted histograms, MSCR, and Recurrent High-Structured Patch descriptors, respectively, and βs are normalized weights.

The distance d_{WH} 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 [66]. 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 I_B to each element a in I_A. This distance is defined by two components: d_y^{ab}, that 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}, that compares the MSCR color. In both cases, the comparison is carried out using the Euclidean distance. The two components are combined as:

$$d_{\text{MSCR}} = \sum_{b \in I_B} \min_{a \in I_A} \gamma \cdot d_y^{ab} + (1 - \gamma) \cdot d_c^{ab} \quad (9)$$

where γ takes values between 0 and 1. In the multi-shot cases, the set I_A of Eq. 9 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 RHSP, one in I_A and one in I_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: $\beta_{\text{WH}} = 0.4$, $\beta_{\text{MSCR}} = 0.4$, $\beta_{\text{RHSP}} = 0.2$ and $\gamma = 0.4$. These values have been estimated once with cross-validation using a subset of 100 image pairs of the VIPeR dataset and left unchanged for all the experiments.

4.1. Detecting new instances

The literature of re-identification does not take into account the case where an individual I_B is not already in the gallery set.

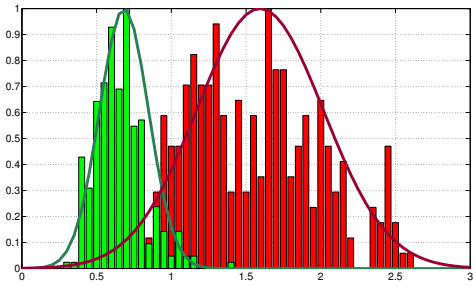


Figure 5: Bimodal distribution of distances. The correct matching distances and the wrong matching distances are depicted by the green and the red histogram, respectively. These curves have been computed for the ETHZ, MvsM N=2 experiment, discussed in Sec. 6.

In a real re-identification setting, this is a very frequent scenario where new people enters the scene for the first time ever.

We address this issue by observing the distances of correct matches and the distances of wrong matches. Experimentally, we have found out that these distances follow the bimodal distribution of Fig. 5, where the correct matching distances and the wrong matching distances are depicted by the green (on the left) and the red bars (on the right), respectively. By simply fitting two Gaussian distribution on the distances data, we are able to distinguish between correct matches and wrong matches. This means that given the minimum distance $d(I_{A^*}, I_B)$ between the best matching I_{A^*} and I_B, if $d(I_{A^*}, I_B)$ is associated to the mode of "wrong" distances (Fig. 5 on the right), the individual is not in the gallery set. If $d(I_{A^*}, I_B)$ is associated to the other mode (Fig. 5 on the left), re-identification is performed. Instead of manually choosing a threshold, that may has to be changed for different scenarios, the likelihood ratio of the two Gaussian can be exploited. We estimate the parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$ of the two Gaussians ($\mathcal{N}(\mu_1, \sigma_1)$ for the mode of "correct" distances and $\mathcal{N}(\mu_2, \sigma_2)$ for the mode of the "wrong" distances) in a training phase, shown in Fig. 5. At testing time, given a distance d if $\frac{\mathcal{N}(d;\mu_1,\sigma_1)}{\mathcal{N}(d;\mu_2,\sigma_2)} >= 1$ there is re-identification, otherwise we identify a new individual.

5. SDALF for Tracking

In tracking, a set of hypotheses of the object position on the image is analyzed at each frame, in order to find the one which best fits with the target appearance, usually called *the template*. The paradigm is different to the classical re-identification: the gallery set is now the hypothesis set, which is different for each target. In addition, it may contain background clutter (hypotheses that explore the scene) and it is not ensured that the exact correspondence will be present (for example, because of occlusions). The goal 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 any ranking.

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

5.1. Particle Filter

Particle Filter offers a probabilistic framework for recursive dynamic state estimation [67] that fits with 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:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int_{x_{t-1}} p(x_t|x_{t-1}) p(x_{t-1}|z_{1:t-1}) dx_{t-1}. \quad (10)$$

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. 10 is often analytically intractable. Equation 10 can be rewritten by its Monte Carlo approximation:

$$p(x_t|z_{1:t}) \approx \sum_{s=1}^{S} w_t^{(s)} \, \delta(x_t - x_t^{(s)}). \quad (11)$$

where

$$w_t^{(s)} \propto w_{t-1}^{(s)} \, \frac{p(z_t|x_t^{(s)}) \; p(x_t^{(s)}|x_{t-1}^{(s)})}{q(x_t^{(s)}|x_{t-1}^{(s)}, z_t)} \quad (12)$$

where q is called *proposal distribution*.

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 reconsidering Eq. 6: $p(z_t|x_t^{(s)}) = P(I_A|I_B)$, where this time I_B is the object template made by SDALF descriptors, and I_A is the current hypothesis $x_t^{(s)}$. In this case, we do not perform minimization like in Eq. 6, but the probability distribution over the hypotheses is kept in order to approximate Eq. 10.

Some semplifications are required when embedding SDALF into the proposed tracking framework. 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. We performed some preliminar experiments including the RHSP, but it turned out that the small improvement in accuracy that it gives is not worth the price in terms of computations that we have to pay. Further evidence is given by the weight we estimated for re-identification ($\beta_{\text{RHSP}} = 0.2$). It highlights that the RHSP gives a small contribution with respect to the other features. For these reasons, the RHSP is drop out from the descriptor for tracking.

The observation model becomes:

$$p(z_t|x_t^{(s)}) = e^{-D(x_t^{(s)}(z_t), \tau_t)} \quad (13)$$

$$D(x_t^{(s)}(z_t), \tau_t) = \beta_{\text{WH}} \cdot d_{\text{WH}}(\text{WH}(x_t^{(s)}), \text{WH}(\tau_t)) + \beta_{\text{MSCR}} \cdot d_{\text{MSCR}}(\text{MSCR}(x_t^{(s)}), \text{MSCR}(\tau_t))$$

where $x_t^{(s)}$ is the hypothesis extracted from the image z_t, and τ_t is the template of the object. 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 I_A.

The computation of the observation model of Eq. 13 consists in evaluating the distances of the hypotheses $\{x_t^{(s)}\}$ against I_B, as dictated by the MvsS strategy in the re-identification. In other words, we have a gallery set of S images (the hyphotheses), and a multi-shot signature as probe.

6. Experimental Results

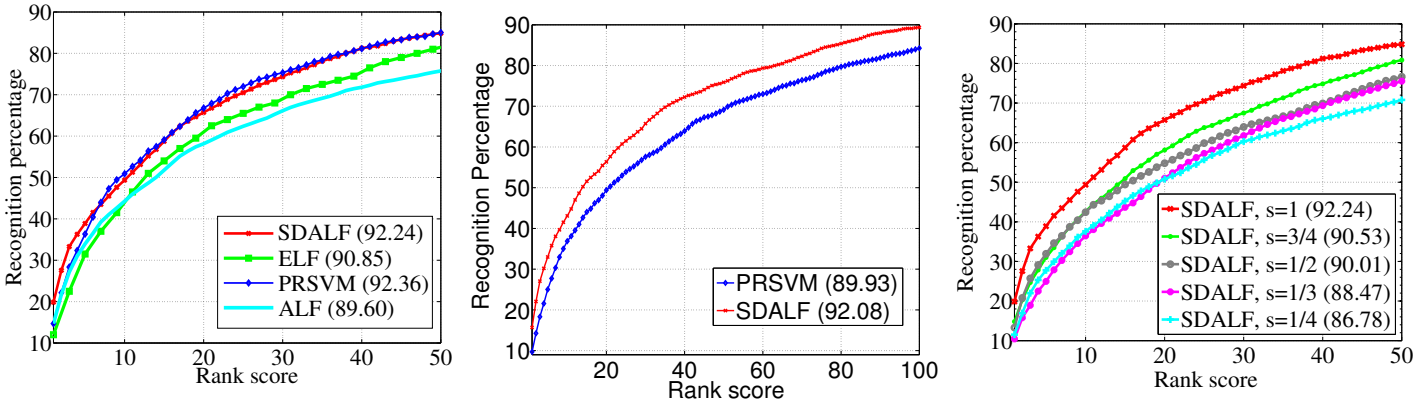
This section shows the evaluation of our approach in the two main applications discussed so far: re-identification and tracking. In the former case, an accurate comparative analysis is carried out considering six public datasets and the most widely-adopted re-identification protocols. In case of tracking, we compare different object models applied to a standard particle filtering approach, analyzing how SDALF performs in a general tracking framework. Moreover, we consider a state-of-the-art tracker, Predator [68], demonstrating how SDALF can be embedded to obtain new best results.

6.1. Person re-identification

We take into account six different public datasets, VIPeR [11], iLIDS for re-id [12], ETHZ 1, 2, and 3 [13], CAVIAR4REID [14]. Each one covers different aspects and challenges for the person re-identification problem[5]. We also create a scenario where the pedestrian images are tiny (up to 11×22).

All the results are shown in terms of recognition rate by the Cumulative Matching Characteristic (CMC) curve, as commonly performed in the literature, and the normalized Area Under Curve (nAUC) score for the CMC curve [34]. The CMC curve is a plot of the recognition performance vs the ranking score and represents the expectation of finding the correct match in the top n matches. On the other hand, nAUC gives an overall score of how well a re-identification method does perform. The parameters' values are fixed for all the experiments unless stated. The parameters are either estimated in a cross-validation phase (*e.g.*, βs) or by a qualitative analysis (*e.g.*, the RHSP thresholds) as described in Sec. 3. To obtain better results they could have been optimized for each dataset, but we preferred to fix them to have a more general re-identification setting.

[5] A video that shows examples of re-identification by SDALF can be found at `http://www.youtube.com/watch?v=3U5Aacyg-No`.



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

Figure 6: Performances on the VIPeR dataset in terms of CMC and nAUC (within brackets). In (a) and (b), comparative profiles of SDALF and other methods (ELF [25] and PRSVM [22]) on 316 pedestrian dataset and 474 pedestrian dataset, respectively. In (c), comparison of SDALF at different scales.

VIPeR Dataset [11, 69]. This dataset [69] contains two views of 632 pedestrians. Each pair is made up of images of the same pedestrian taken from different cameras, under different viewpoints, poses and lighting conditions. All images are normalized to 48×128 pixels. Most of the examples contains a viewpoint change of 90 degrees. Each pair is randomly split into two sets: Cam A and Cam B. Considering images from Cam B as the gallery set, and images from 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. Ideally rank 1 should be assigned only to the correct pair matches.

As competitors, we take into account the Primal-based RankSVM (PRSVM) [22], and the Ensemble of Localized Features (ELF) approach [25], following its commonly-used experimental protocol: the dataset is split randomly into a training and a test set, and the matching is performed. This procedure is repeated several times for the sake of crossvalidation (5 for PRSVM, 10 for ELF), averaging the obtained CMC curves. It is worth noting that in our case we discard the training data, since SDALF is a direct re-identification method and does not need any training stage. In particular, we report the performances using a test set of 316 and 474 pedestrians, respectively (see Fig. 6;). The nAUC score for each method is provided within brackets in the legend of the plots of Fig. 6. Considering the experiment on 316 test pedestrians (Fig. 6(a)), we also compute a modified version of SDALF, in which the features have been extracted directly from the whole human body without any partition driven by asymmetry/symmetry principia, and without any weighting. The method is named Accumulation of Local Features (ALF), and it is useful for highlighting the importance of focusing on separate parts of the human body and considering the internal regions of the parts more reliable.

Many considerations could be drawn. First of all, SDALF outperforms ELF in terms of nAUC, and we obtain comparable results with respect to PRSVM (less than 0.12%). Moreover, SDALF slightly outperforms PRSVM in the first positions of the CMC curve (rank $1-6$). This clearly shows the effectiveness of the SDALF descriptor: without knowing beforehand how the appearance information is transferred across different cameras (this is actually studied by the learning methods), it is still able to capture discriminant appearance traits. Finally, it is easy to notice that the use of the symmetry axes increase the results of $5-10\%$ of the CMC, considering the ALF curve.

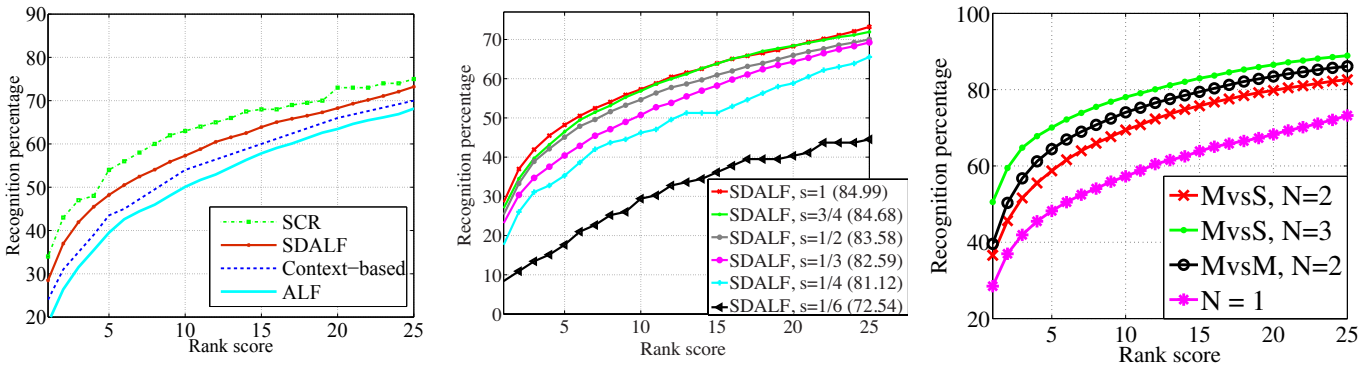
Fig. 6(b) shows a comparison between PRSVM and SDALF when dealing with a larger test dataset (474 individuals, as done in the PRSVM paper). In this case, our approach outperforms PRSVM of about 2.15%, in terms of nAUC. This reconfirms that the performances of PRSVM, now lower than the previous experiment, strictly depend on the training set (158 individuals in this case) while the performances of SDALF remain similar.

The most considerable source of error for SDALF derives from the illumination that in some cases tend to saturate the colors so that many individual look very similar, and to the severe lighting changes across the two views.

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

iLIDS Dataset [12]. The iLIDS Multiple-Camera Tracking Scenario repository is a videosequence dataset captured at an airport arrival hall at the rush hour, exploiting a multi-camera CCTV network. An excerpt of 479 images of 119 pedestrians was extracted from these videos for testing a context-based pedestrian re-identification method [26]. The images, normalized to 64×128 pixels, derive from non-overlapping cameras, under quite large illumination changes and subject to occlusions (not present in VIPeR). Since there are more than two examples for each pedestrian, we can evaluate both single- and multiple-shot cases.

Regarding the single-shot case, we take into account the Spatial Covariance Region (SCR) approach [29], and the context-based method of [26], adopting also its testing protocol. We



(a) (b) (c)

Figure 7: Performances on iLIDS dataset. (a) CMC curves comparing Context-based re-id [26], SCR [29], single-shot SDALF and ALF. (b) Analysis of SDALF performances at different resolution. (c) CMC curves for MvsS and MvsM cases varying the average number of images N for each pedestrian.

also report the performances of ALF. We randomly select one image for each pedestrian to build the gallery set, while the other images 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 is displayed in Fig. 7. SDALF outperforms the Context-based method [26] without using any additional information about the context (Fig. 7(a)) even using images at lower resolution (Fig. 7(b)). As expected, ALF is undoubtedly less performing than SDALF.

The experiments of Fig. 7(b) show the performances of our approach 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 performances with respect to SCR [29]. Unfortunately, it has been applied solely to the iLIDS datasets, so that its performances cannot be generalized in a consistent way. In particular, an interesting challenge would be that of working on extremely low resolutions, as in the CAVIAR4REID benchmark. The SCR approach uses covariances of features which are computed on localized patches. At a very low resolution this would mean computing second order statistics on very few values, that could be uninformative and subjected to dimensionality issues.

Concerning the multiple-shot case, we run experiments on both the Multiple vs Single (MvsS) and the Multiple vs Multiple (MvsM) paradigms. In the former situation, we build a gallery set of multi-shot signatures and we match it with a probe set composed by 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: therefore, in such a case, we simply build a multi-shot signature composed by less instances. Intuitively, in the MvsS situation, this policy applies to the gallery signature only; in the MvsM signature, we start by decreasing the number of instances that compose the probe signature, leaving unchanged the number of elements that build the gallery signature; once we reach just one instance for the probe signature, we start decreasing the elements of 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 performances of about 10% and to outperform the Context-based method [26]. Adding another image gives an increment of 20% with respect to the single-shot case. It is interesting to note that the results for MvsM lie below the MvsS (N=3) curve. This is probably due to the fact that a signature with 3 images captures more heterogeneous aspects, encoding the information exhibited by the fourth test image.

ETHZ Dataset [13, 70]. This repository [70] is formed by images captured from a moving camera, and it has been used originally for pedestrian detection. [21] extracted a set of samples for each different person in the videos, and use the resulting set of images to test their PLS method. The moving camera setup provides a range of variations in people's appearance. Though, variation in pose is relatively small in comparison with the other two datasets. The most challenging aspects of ETHZ are the illumination changes and the occlusions, other than the low resolution: all images are normalized to 32×64 pixels. The dataset is structured as follows: Seq. #1 contains 83 pedestrians, for a total of 4.857 images; Seq. #2 contains 35 pedestrians, for a total of 1.936 images; Seq. #3 contains 28 pedestrians, for a total of 1.762 images.

In the single-shot case, the experiments are carried out exactly as for iLIDS. The multiple-shot case is performed 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 beforehand a clustering algorithm [64] on the original frames, based on their HSV histograms: this way, consecutive similar frames would end up in the same cluster. At this point, we select randomly one frame from each cluster and use them as

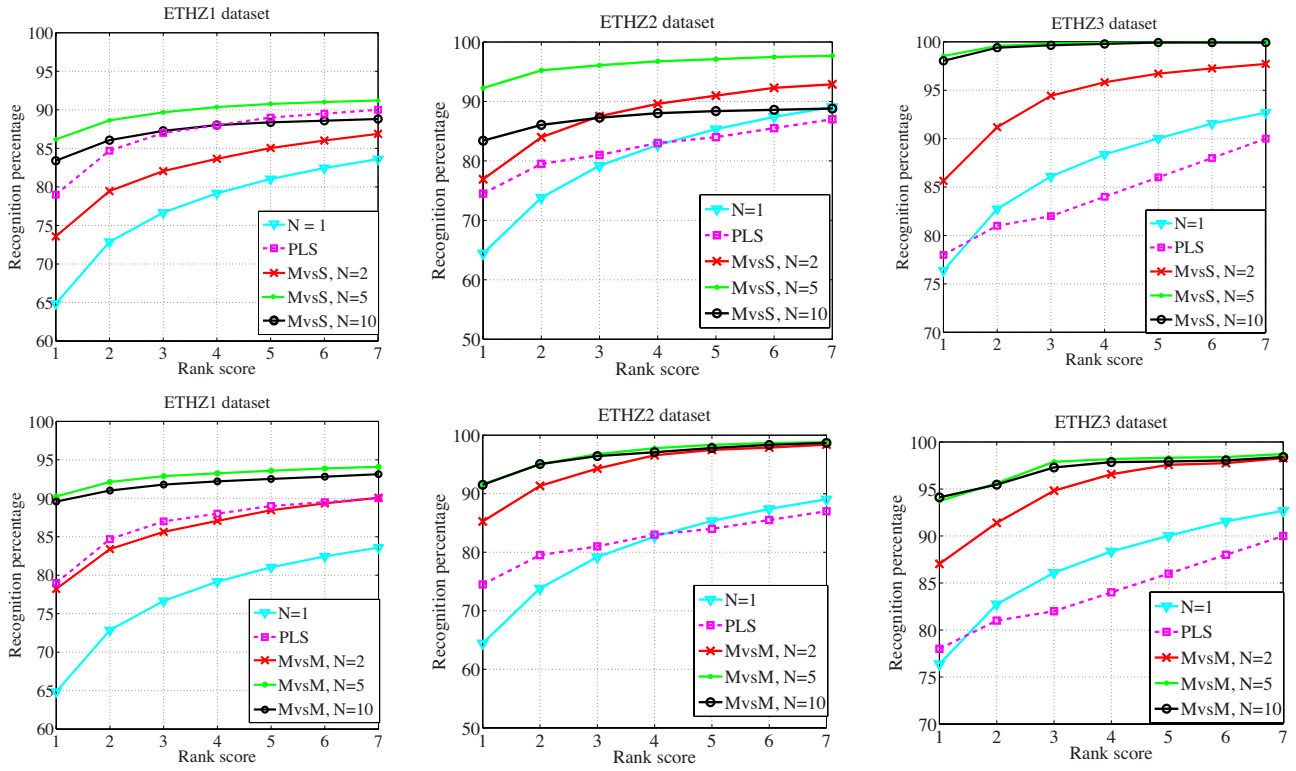


Figure 8: Performances on the ETHZ dataset. Left column, results on SEQ. #1; middle column, on SEQ. #2; right column, on SEQ. #3. We compare our method with the results of PLS [21]. On the top row, we report the results for single-shot and MvsS SDALF; on the bottom row, we report the results for MvsM SDALF.

keyframes for the multi-shot signature. Then, the gallery and probe sets are built like the MvsM in iLIDS considering the selected keyframes.

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 [21]. 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 multicamera setting. In addition, PLS requires to have all the gallery signatures beforehand in order to estimate the weights on the appearance model. So, if one pedestrian is added the weights must be recomputed , weakening its effectiveness in real scenarios.

In SEQ. #2 (Fig. 8) we note 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 performances as to rank 1 correct matches is 98% for MvsS and 94% for MvsM. It is worth noting 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 a good trade-off between accuracy and computational performance.

CAVIAR4REID Dataset [14]. The CAVIAR4REID dataset contains images of pedestrians extracted from CAVIAR repository [15]. It is composed by 72 pedestrians with 10 images for each of them, for two camera views. For this reason, it is more interesting than the ETHZ, where images are extracted from a single camera. The other challenging features of this dataset are: a broad change in the image resolution, with a minimum and maximum size of 17×39 and 72×144, respectively; pose variations are severe, as so as the illumination changes and the occlusions.

Fig. 9 reports the results of SDALF with a single image (left), and under the multi-shot policy (right). The results show that in a more realistic scenario the results are much worse - nAUC is around 70% in the single-shot case. (C)PS [14] outperforms SDALF in this dataset especially in the multi-shot setting. This is expected because 1) (C)PS uses features similar to SDALF and 2) it relies on a finer description of the human body using the pictorial structures instead of just segment the body in 3 parts. Those results prove again that having a good body segmentation is essential for re-identification. In the multi-shot case, SDALF performs comparably with MRCG [35][6].

6.2. Tracking

As benchmark, we adopt CAVIAR [15], as it represents a challenging real tracking scenario, due to pose, resolution and

[6]We thank the authors of [35] for providing us the results of their algorithm on CAVIAR4REID.

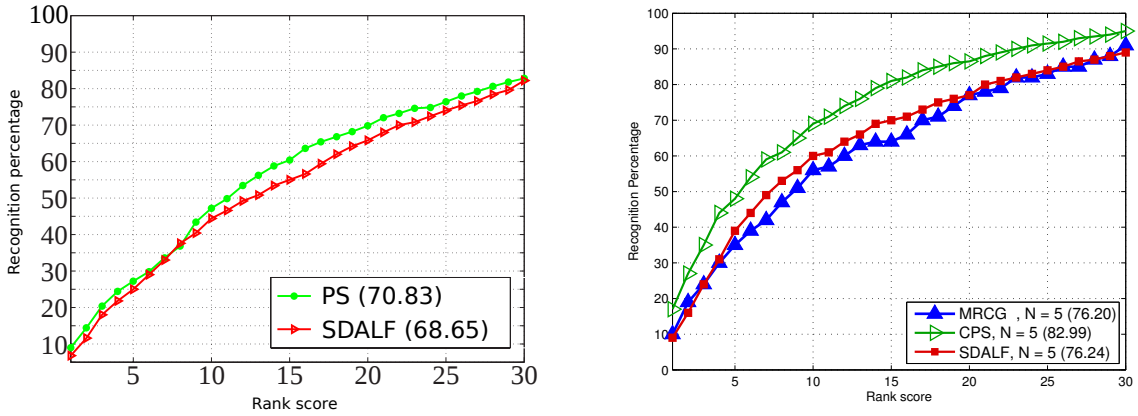


Figure 9: Performances of the single-shot descriptor (on the left) and the multi-shot descriptor (on the right) on the CAVIAR4REID dataset comparing SDALF, (C)PS [14] and MRCG [35].

illumination changes, and severe occlusions. The dataset consists of several ground-truthed sequences 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.

As first experiment, we want 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 offline-trained person detector [71] in the same way exploited by the Boosted Particle Filter [53]. 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 [72].

The proposed SDALF-based observation model is compared against two classical appearance descriptors for tracking: joint HSV histogram and part-based HSV histogram (partHSV) [3] where each of three body parts (head, torso, legs) are described by a color histogram.

The quantitative evaluation of the descriptors is provided by adopting the metrics proposed by [73], that consist of:

- False Positives (**FP**): An estimate exists that is not associated with a ground truth object;
- False Negatives (**FN**): A ground truth object exists that is not associated with an estimate;

The single-frame values are averaged over time for estimating an overall statistics. 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).

Finally, we adopt metrics that take in account also the temporal coherence of the estimated trajectories [74][7]:

- Average Tracking Accuracy (**ATA**): measure that penalizes fragmentation phenomena in both the temporal and spatial dimensions, while accounting for the number of objects detected and tracked, missed objects, and false positives;
- Multi-Object Tracking Precision (**MOTP**): considers the spatiotemporal 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 [74].

The overall tracking results averaged over all the sequences are reported in Table 2. Our approach is better in terms of FP, which means that tracking is performed with higher accuracy (*e.g.*, not too large bounding boxes), and in terms of FN, that is, it is less probable to lose targets. The number of estimated tracks using SDALF are 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 outperforms HSV and partHSV in all the metrics. 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

[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).

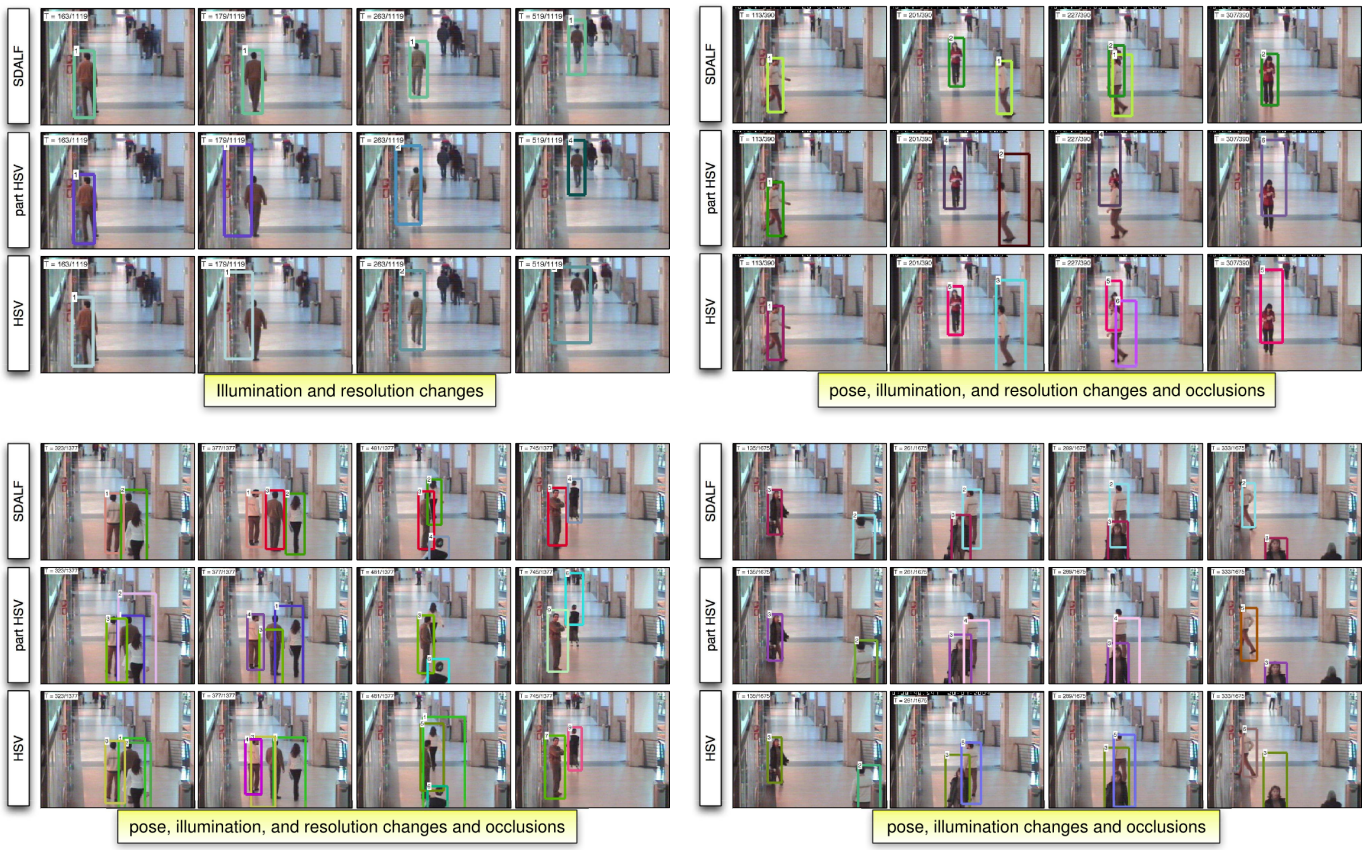


Figure 10: Qualitative comparison between the descriptors on the sequence OneLeaveShop2cor (top-left), OneLeaveShopReenter1cor (top-right), ShopAssistant1cor (bottom-left) and OneShopOneWait1cor (bottom-right). Those sequences poses the problem of multi-target tracking when dealing with resolution, pose, illumination and resolution changes and occlusions.

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.

	FP	FN	# Est.	# GT	ATA	MOTP	MOTA
SDALF	**0.0608**	**0.1852**	**300**	235	**0.4567**	**0.7182**	**0.6331**
partHSV	0.2094	0.4145	522	235	0.1812	0.5822	0.5585
HSV	0.2364	0.3858	462	235	0.1969	0.5862	0.5899

Table 2: Quantitative comparison between object descriptors: SDALF, part-based HSV histogram and HSV histogram; the performances are given in terms of False Positives (FP), False Negatives (FN), 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).

A qualitative analysis that highlights the performances discussed above is provided in Fig. 10 and the videos reported at `http://www.youtube.com/watch?v=JiW2unf5gwg`. In particular, the sequence of Fig. 10 (top-left) shows the problem of single-target tracking when dealing with illumination and resolution changes. HSV (third row) and partHSV histograms (second row) are not able to deal properly with these problems even if the sequence is quite simple (no occlusions, simple background, only one target) resulting in many target misses: three times for partHSV and two times for HSV (if the target is lost in a particular frame, it is reinitialized the next frames). Conversely, our approach follows the target for the whole sequence without any track hijacking. In Fig. 10 (top-right), tracking becomes more challenging, because the appearance model has to face pose changes and partial occlusions. As in the previous figure, HSV and partHSV lose the track several times. SDALF outperforms the competitors, and shows to be robust to partial occlusions.

A similar behavior is reported in the results of Fig. 10 (bottom). When dealing with pose, illumination and resolution changes and partial occlusions, SDALF outperforms the HSV and partHSV descriptors in terms of less misses and higher accuracy.

In terms of computational speed, we evaluate how long takes the computation of Eq. 13[8]. Two steps are required: first, the SDALF descriptor for the current hypothesis is extracted, second, the distances on Eq. 13 are computed. The first and second phase take in average 18 and 15 milliseconds, respectively, when the hypothesis has size 12×36. When the hypothesis increases his size to 40×46, these phases take in average 26 and 24 milliseconds, respectively. Let S be the number of particles, N the number of images and K the average number of targets

[8]The following values have been computed using our non-optimized MATLAB code on a quad-core Intel Xeon E5440, 2.83 GHz with 4 GB of RAM.

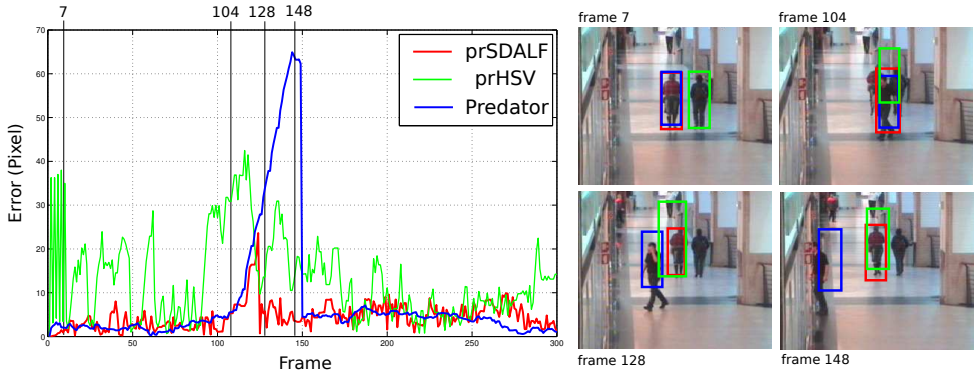


Figure 11: Comparison between Predator (blue), Predator with SDALF (red) and Predator with joint HSV histogram (green) on EnterExitCrossingPaths1cor. The goal is to track the person on the left.

Sequence.	Frames	prSDALF	prHSV	Predator
EnterExitCrossingPaths1cor	1-300	4.48 (3.10)	12.95 (9.39)	7.82 (13.59)
OneLeaveShopReenter2cor	1-500	8.00 (5.82)	14.37 (12.75)	26.73 (21.86)
OneStopMoveNoEnter2cor	401-1000	12.17 (8.89)	48.39 (66.18)	44.91 (53.98)
Average	–	**8.22 (5.94)**	25.24 (29.44)	26.49 (29.81)

Table 3: Error on three representative sequences comparing Predator with SDALF, Predator with joint HSV histogram and Predator.

($N = 3$ and $S = 100$ in our experiments), then the computational complexity of Eq. 13 is $O(K \cdot S \cdot N)$. HSV and partHSV have the only difference that they are computed only for a single instance, therefore N is dropped out. Moreover, it is worth noting that SDALF has an additional component to extract, that is the MSCR descriptor, and this brings in an additional cost.

The last test we carry out is to evaluate the SDALF observation model embedded into a state-of-the-art tracker, that is, the well-known Predator [68]. The general aim is to see if SDALF might be adopted by novel trackers, hoping for an amelioration of their performances due to our person modeling. We already showed that using a generic particle filter, so this test should be considered as an additional confirmation of our thesis.

Strictly speaking, Predator combines a non-parametric single-object tracker that estimates the target position using optical flow and a detector that is specifically trained to detect the particular target. We want to check how much SDALF can improve its performances, focusing on heterogeneous sequences where occlusions happen at some point. For this reason, three different samples (EnterExitCrossingPaths1cor, OneLeaveShopReenter2cor, OneStopMoveNoEnter2cor) have been selected from CAVIAR. The goal is to track a single person during its presence in the scene.

We thus modify the Predator prediction part as a cross-correlation phase. The algorithm generates different hypotheses around the target (in space and scale). Then the target template is compared with those hypotheses in the same way described by Eq. 13. The hypothesis that maximizes Eq. 13 is chosen as the target estimate. As in the original algorithm, a single track is manually initialized at the first frame.

In Table 3, we reported the error measures of the original Predator, Predator with SDALF (prSDALF) and also Predator with joint HSV histogram[9] (prHSV) that works in similar way of SDALF. The results are given in terms of mean distance between the ground truth and the tracker estimates, and standard deviation. The "frame" column in the table indicates the initial and final frame of the person that has been considered in the experiment. The mean error and its standard deviation of prSDALF are considerably lower than the results of the other competitors.

Moreover, we report a detailed analysis on a particular sequence (EnterExitCrossingPaths1cor): two people are walking together and an occlusion occurs. The distances between the estimate and the ground truth over the time are shown in Fig. 11. While comparing Predator with prSDALF, it is easy to notice that in case of occlusion (frames 95-125) Predator drifts away from the target and it takes more time to realize that the target is lost. Instead, prSDALF after the occlusion is immediately able to recover the target. The approach prHSV is not so good, because in the first frames it gets confused between the two people that are walking one next to each other and the error is also high after the occlusion. To summarize the results in the other sequences, when we do not have occlusions or drastic changes of illumination/pose, the performances of prSDALF and Predator are comparable. When the situation becomes hard, SDALF gives a substantial improvement.

7. Conclusions

In this paper, we introduced a novel robust descriptor characterizing the human appearance, SDALF. SDALF is able to capture discriminant appearance information of individuals independently from many factors, as pose, resolution, illumination changes, occlusions, and in general clutter. The descriptor has been tested extensively in two open issues for computer vision, re-identification and tracking. SDALF consists in the robust detection of human parts, driven by asymmetry/symmetry principles. After that, three complementary kinds of features are

[9] partHSV is not used because it has already been proved to perform comparably to HSV.

extracted, focusing on different perceptual aspect of the human appearance. In particular, chromatic and structural information, as well as recurrent high-entropy textural characteristics are distilled from the human body. When multiple images of the same person are available for the characterization, SDALF accumulate the features into a unique descriptor, encoding as much information as possible.

In the re-identification case, SDALF is used as a simple feature extractor, making it well-suited for the direct approaches, where minimization techniques are employed to rank the gallery candidates. In the tracking case, SDALF comes along with a matching strategy, providing thus an observation model that can be embedded into many tracking paradigms. In both the scenarios SDALF offers convincing results, that promote it a basic tool for researchers dealing with the human appearance.

As future works, for re-identification SDALF may be embedded into the learning strategies, for improving the already good performances such approaches do provide. For tracking, SDALF should be optimized and parallelized to reach real-time performances, in order to be embedded into industrial products and approach the market.

Acknowledgments

This research is founded by the EU-Project FP7 SAMURAI, grant FP7-SEC- 2007-01 No. 217899.

References

[1] P.-E. Forssén, Maximally stable colour regions for recognition and matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[2] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1014 –1021. doi:10.1109/CVPR.2009.5206754.

[3] M. Isard, J. MacCormick, BraMBLe: a bayesian multiple-blob tracker, in: International Conference on Computer Vision, Vol. 2, 2001, pp. 34–41.

[4] P. Pérez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic tracking, in: European Conference on Computer Vision, 2002, pp. 661–675.

[5] W. Kohler, The task of Gestalt psychology, Princeton NJ, 1969.

[6] A. Levinshtein, S. Dickinson, C. Sminchisescu, Multiscale Symmetric Part Detection and Grouping, in: International Conference on Computer Vision, 2009.

[7] K. L. M. Cho, Bilateral symmetry detection and segmentation via symmetry-growing, in: British Machine Vision Conference, 2009.

[8] D. Reisfeld, H. Wolfson, Y. Yeshurun, Context-free attentional operators: The generalized symmetry transform, International Journal of Computer Vision 14 (2) (1995) 119–130.

[9] T. Riklin-Raviv, N. Sochen, N. Kiryati, On symmetry, perspectivity, and level-set-based segmentation, IEEE Transactions on Pattern Recognition and Machine Intelligence 31 (8) (2009) 1458–1471.

[10] J. Matas, O. Chum, U. Martin, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proceedings of the British Machine Vision Conference, Vol. 1, 2002, pp. 384–393.

[11] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition and tracking., in: PETS, 2007.

[12] UK home office, i-LIDS multiple camera tracking scenario definition, `http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/`.

[13] A. Ess, B-Leibe, L. V. Gool, Depth and appearance for mobile scene analysis, in: International Conference on Computer Vision, 2007.

[14] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: British Machine Vision Conference (BMVC), 2011.

[15] Caviar dataset, `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/` (2004).

[16] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: CVPR, 2010.

[17] A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, ACM Computer Survey 38. doi:http://doi.acm.org/10.1145/1177352.1177355.

[18] O. Javed, K. Shafique, Z. Rasheed, M. Shah, Modeling inter-camera space-time and appearance relationships for tracking accross non-overlapping views, Computer Vision and Image Understanding 109 (2007) 146–162.

[19] D. Makris, T. J. Ellis, J. K. Black, Bridging the gaps between cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. 205–210.

[20] A. Rahimi, B. Dunagan, T. Darrel, Simultaneous calibration and tracking with a network of non-overlapping sensors, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2004, pp. 187–194.

[21] W. Schwartz, L. Davis, Learning discriminative appearance-based models using partial least squares, in: SIBGRAPI, 2009.

[22] B. Prosser, W. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: British Machine Vision Conference, 2010, pp. 1–11.

[23] C. Nakajima, M. Pontil, B. Heisele, T. Poggio, Full-body person recognition system, Pattern Recognition 36 (9).

[24] Z. Lin, L. Davis, Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance, in: International Symposium on Visual Computing, 2008, pp. 23–34.

[25] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensamble of localized features, in: European Conference on Computer Vision, 2008, pp. 262–275.

[26] W. Zheng, S. Gong, T. Xiang, Associating groups of people, in: British Machine Vision Conference, 2009.

[27] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Person Re-identification Using Haar-based and DCD-based Signature, in: Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, 2010.

[28] J. Sivic, C. L. Zitnick, R. Szeliski, Finding people in repeated shots of the same scene, in: Proceedings of the British Machine Vision Conference, 2006.

[29] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Person Re-identification Using Spatial Covariance Regions of Human Body Parts, in: AVSS, 2010.

[30] N. Bird, O. Masoud, N. Papanikolopoulos, A. Isaacs, Detection of loitering individuals in public transportation areas, IEEE Transactions on Intelligent Transportation Systems 6 (2) (2005) 167 – 177. doi:10.1109/TITS.2005.848370.

[31] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, P. H. Tu, Shape and appearance context modeling, in: International Conference on Computer Vision, 2007, pp. 1–8.

[32] N. Gheissari, T. B. Sebastian, P. H. Tu, , J. Rittscher, R. Hartley, Person reidentification using spatiotemporal appearance, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 1528–1535.

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

[34] L. Bazzani, M. Farenzena, A. Perina, V. Murino, M. Cristani, Multiple-shot person re-identification by HPE signature, in: IEEE International Conference on Pattern Recognition, 2010.

[35] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Multiple-shot human re-identification by mean riemannian covariance grid, in: Advanced Video and Signal-Based Surveillance, Klagenfurt, Autriche, 2011.

[36] H. Bay, T. Tuytelaars, L. V. Gool, SURF: Speeded Up Robust Features, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 404–417.

[37] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on riemannian manifolds, IEEE Trans. PAMI (2008) 1713–1727.

[38] C. J. Veenman, C. J. Veenman, M. J. T. Reinders, E. Backer, Resolving motion correspondence for densely moving points, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 54–72.

[39] D. Serby, E. Koller-Meier, L. V. Gool, Probabilistic object tracking using multiple features, in: International Conference on Pattern Recognition, 2004, pp. 184–187. doi:http://dx.doi.org/10.1109/ICPR.2004.716.

[40] F. Porikli, O. Tuzel, P. Meer, Covariance tracking using model update based on lie algebra, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 728–735.

[41] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transaction on Pattern Analysis and Machine Intelligence 25 (2003) 564–575. doi:10.1109/TPAMI.2003.1195991.

[42] O. Lanz, Approximate bayesian multibody tracking, IEEE Transaction on Pattern Analysis and Machine Intelligence.

[43] J. Kwon, K. Lee, Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1208–1215.

[44] A. Yilmaz, X. Li, M. Shah, Contour-based object tracking with occlusion handling in video acquired using mobile cameras, IEEE Transaction on Pattern Analysis and Machine Intelligence 26 (2004) 1531–1536. doi:http://dx.doi.org/10.1109/TPAMI.2004.96.

[45] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1) (2005) 55–79.

[46] M. Andriluka, S. Roth, B. Schiele, Monocular 3d pose estimation and tracking by detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 623–630.

[47] R. Urtasun, D. J. Fleet, P. Fua, 3d people tracking with gaussian process dynamical models, in: EEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 238–245. doi:10.1109/CVPR.2006.15.

[48] M. Brubaker, L. Sigal, D. Fleet, Video-based people tracking, Handbook of Ambient Intelligence and Smart Environments (2010) 57–87.

[49] S. C. Zhu, A. Yuille, Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1996) 884–900. doi:http://doi.ieeecomputersociety.org/10.1109/34.537343.

[50] N. Paragios, R. Deriche, Geodesic active regions and level set methods for supervised texture segmentation, International Journal of Computer Vision 46 (2002) 223–247. doi:http://dx.doi.org/10.1023/A:1014080923068.

[51] A. Elgammal, R. Duraiswami, D. Harwood, L. Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, Proceedings of the IEEE 90 (7) (2002) 1151 – 1163. doi:10.1109/JPROC.2002.801448.

[52] P. Fieguth, D. Terzopoulos, Color-based tracking of heads and other mobile objects at video frame rates, in: IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[53] K. Okuma, A. Taleghani, N. de Freitas, J. Little, D. Lowe, A boosted particle filter: Multitarget detection and tracking, in: European Conference on Computer Vision, 2004, pp. 28–39.

[54] G. J. Edwards, C. J. Taylor, T. F. Cootes, Interpreting face images using active appearance models, in: International Conference on Face & Gesture Recognition, 1998.

[55] J. Lim, D. Ross, R. sung Lin, M. hsuan Yang, Incremental learning for visual tracking, in: Advances in Neural Information Processing Systems, 2004, pp. 793–800.

[56] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Transaction on Pattern Analysis and Machine Intelligence 19 (1997) 696–710. doi:http://dx.doi.org/10.1109/34.598227.

[57] M. J. Black, A. D. Jepson, Eigentracking: Robust matching and tracking of articulated objects using a view-based representation, International Journal of Computer Vision 26 (1998) 63–84. doi:10.1023/A:1007939232436.

[58] I. Matthews, T. Ishikawa, S. Baker, The template update problem, IEEE Transactions on Pattern Analysis and Machine Intelligence (2004) 810–815.

[59] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: Computer Vision and Pattern Recognition, 2009.

[60] C.-H. Kuo, C. Huang, R. Nevatia, Multi-target tracking by online learned discriminative appearance models, IEEE Conference on Computer Vision and Pattern Recognition (2010) 685–692doi:http://doi.ieeecomputersociety.org/10.1109/CVPR.2010.5540148.

[61] L. Bazzani, M. Cristani, M. Bicego, V. Murino, Online subjective feature selection for occlusion management in tracking application, in: International Conference on Image Processing, 2009.

[62] J. Xing, H. Ai, S. Lao, Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1200–1207.

[63] N. Jojic, A. Perina, M. Cristani, V. Murino, B. Frey, Stel component analysis: Modeling spatial correlations in image class structure, IEEE Conference on Computer Vision and Pattern Recognition (2009) 2044–2051.

[64] M. Figueiredo, A. Jain, Unsupervised learning of finite mixture models, IEEE Transaction on Pattern Analysis and Machine Intelligence 24 (3) (2002) 381–396.

[65] N. Jojic, B. Frey, A.Kannan, Epitomic analysis of appearance and shape, in: Proc. of International Conference on Computer Vision, 2003.

[66] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, IEEE Transactions on Communications 15 (1) (1967) 52–60.

[67] A. Doucet, N. De Freitas, N. Gordon (Eds.), Sequential Monte Carlo methods in practice, 2001.

[68] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2012) 1409–1422. doi:http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.239.

[69] VIPeR dataset, `http://vision.soe.ucsc.edu/?q=node/178`.

[70] ETHZ dataset, `http://www.liv.ic.unicamp.br/~wschwartz/datasets.html`.

[71] P. Felzenszwalb, R. Girshick, D. McAllester, Cascade Object Detection with Deformable Part Models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[72] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. V. Gool, Robust tracking-by-detection using a detector confidence particle filter, in: IEEE International Conference on Computer Vision, 2009.

[73] K. Smith, D. Gatica-Perez, J. Odobez, S. Ba, Evaluating multi-object tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 36–43.

[74] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, J. Zhang, 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 (2009) 319–336.