Person Re-Identification by Symmetry-Driven Accumulation of Local Features

M. Farenzena[1], L. Bazzani[1], A. Perina[1], V. Murino[1,2], M. Cristani[1,2]
[1]*Dipartimento di Informatica, University of Verona,* Italy
[2]*Istituto Italiano di Tecnologia (IIT),* Genova, Italy

Abstract

In this paper, we present an appearance-based method for person re-identification. It consists in the extraction of features that model three complementary aspects 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. All this information is derived from different body parts, and weighted opportunely by exploiting symmetry and asymmetry perceptual principles. In this way, robustness against very low resolution, occlusions and pose, viewpoint and illumination changes is achieved. The approach applies to situations where the number of candidates varies continuously, considering single images or bunch of frames for each individual. It has been tested on several public benchmark datasets (ViPER, iLIDS, ETHZ), gaining new state-of-the-art performances.

1. Introduction

Person re-identification consists in recognizing an individual in diverse locations over different non-overlapping camera views, considering a large set of candidates. It represents a valuable task in video surveillance scenarios, where long-term activities have to be modeled within a large and structured environment (e.g., airport, metro station). In this context, a robust modeling of the entire body appearance of the individual is essential, because other classical biometric cues (face, gait) may not be available, due to sensors' scarce resolution or low frame-rate. Usually, it is assumed that individuals wear the same clothes between the different sightings. The model has to be invariant to pose, viewpoint, illumination changes, and occlusions: these challenges call for specific human-based solutions.

Re-identification methods that rely only on visual information are addressed here as *appearance-based* techniques [2, 3, 6, 8, 9, 12, 15, 19, 20, 21]. Other approaches assume easier operative conditions: they simplify the problem by adding temporal reasoning on the spatial layout of the monitored environment, in order to prune the candidate set to be matched [10, 14, 17].

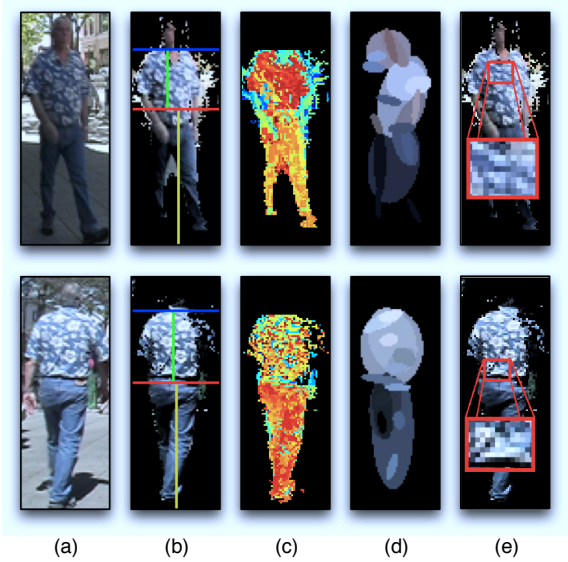

Figure 1. Sketch of the approach: a) two instances of the same person; b) x- and y-axes of asymmetry and symmetry, respectively; c) weighted histogram back-projection (brighter pixels mean a more important color), d) Maximally Stable Color Regions; e) Recurrent Highly Structured Patches.

In this paper, we present a novel and versatile appearance-based re-identification method, based on a pondered extraction of local features. After a pre-processing step, we select salient parts of the body figure by adopting perceptual principles of symmetry and asymmetry. First, we find two horizontal axes of asymmetry that isolate three main body regions, usually corresponding to head, torso and legs. On the last two, a vertical axis of appearance symmetry is estimated. Then, complementary aspects of the human body appearance are detected on each part (see Fig. 1), highlighting: i) the general chromatic content via HSV histogram; ii) the per-region color displacement, through Maximally Stable Colour Regions (MSCR) [5]; iii) the presence of *Recurrent Highly Structured Patches* (RHSP), estimated through a novel per-patch similarity analysis. The extracted features are weighted by the distance with respect to the vertical axis, so that the effects of pose variations are minimized. Matching these features gives the similarity measure

between the candidates.

We name our approach *Symmetry-Driven Accumulation of Local Features* (SDALF). It applies to both the case where a single image for each candidate is present, and the case where multiple images for each individual (not necessarily consecutive) are available. We properly accumulate the local features in a single signature: intuitively, the higher the number of images for each person, the higher the expressivity of the signature.

SDALF approach is simple and effective. Apart from RHSP, the other features are not novel in the literature. Novel is the way – derived from attentive considerations about the body structure and its appearance – they are assembled together for the task at hand. We tested SDALF on different compelling public databases: ViPER [7], iLIDS [16], and ETHZ [4], attaining in most cases novel state-of-the-art performances. In general, these datasets face all the different challenges of the re-identification problem: pose, viewpoint and lighting variations, and occlusions. Moreover, iLIDS is taken from a real surveillance scene.

Finally, SDALF is independent from the number of candidates we consider, so it can cope with a varying, arbitrary large number of elements.

The rest of the paper is organized as follows. In Sec. 2, a taxonomy of the present literature on the appearance-based re-identification methods is reported, highlighting the differences with respect to our strategy. Sec. 3 is the core of the paper, detailing our approach. Several comparative results are reported in Sec. 4, and, finally, conclusions are drawn and future perspectives are envisaged in Sec. 5.

2. State of the Art

Person appearance-based re-identification techniques can be organized in two main groups. The first focuses on the hardest situation, that is, associating pairs of images, each containing one instance of an individual. We name these methods as *single-shot* approaches. The second group employs multiple images of the same person as training data, considering still images or short sequences as testing observations. We name these methods as *multiple-shot* approaches.

As to *single-shot* approaches, in [20] the method consists in segmenting a pedestrian image into regions, and registering their color spatial relationship into a co-occurrence matrix. This technique works well when pedestrians are seen from small varying points of view. Viewpoint invariance is instead the main issue addressed in [8]: spatial and color information are combined using an ensemble of discriminant localized features and classifiers, selected by boosting. Other approaches focus on enhancing the discriminative power of each individual signature with respect to the others. In [12], pairwise dissimilarity profiles between individuals are learned and adapted for a nearest neighbor classification. Similarly, in [19], 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. In both methods, a learning phase based on the pedestrians to re-identify is required. If a novel person is added to the set, this phase has to be re-computed. Finally, an approach that potentially opens a new direction is [21], where the description of a person is enriched by contextual visual knowledge coming from the surrounding people. 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.

As to *multiple-shot* approaches, in [15], for each considered subject, a set of local and global features are extracted from a set of training images and fed into an SVM, employing different learning schemes. In [3], the bounding box of a pedestrian is equally divided into ten horizontal stripes, extracting the median HSL value 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 in [6], 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. This permits a more accurate matching. In [9], the person re-identification scheme is based on the matching of SURF [2] interest points, collected in several images during short video sequences.

Considering all these methods, our SDALF approach differs on the following aspects. i) Unlike [12, 19], no discriminative training is needed on the joint battery of candidates: our approach applies independently on each new individual at hand; this is advantageous when dealing with a large and variable number of persons. ii) It copes better with viewpoint, pose and illumination variations, as witnessed by the experimental results. iii) It is flexible, working in both the single- and the multi-shot case.

3. The Approach

SDALF is a three-phase process. The phases apply in a slightly different way whether we are in the single or in the multiple-shot case.

In the first phase, axes of asymmetry and symmetry are found for each pedestrian image. This phase assumes the presence of the silhouette of the individual, disregarding the background (BG). In the single-shot case, a silhouette mask Z containing only foreground (FG) pixel values is obtained for each person by inferring over the STEL generative model [11]. STEL model captures the general structure of an image class as a blending of several *component*

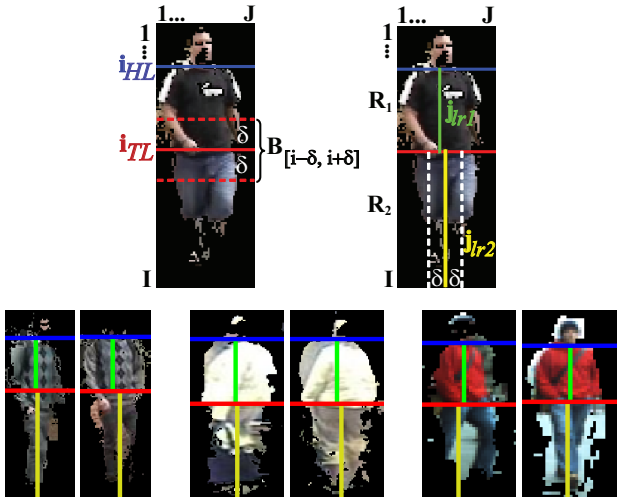

Figure 2. Symmetry-based Silhouette Partition. On the top row, overview of the method: first the asymmetrical axis Ax_{TL} is extracted, then Ax_{HT}; afterwards, for each R_k region the symmetrical axis j_{LRk} are computed. On the bottom row, examples of symmetry-based partitions on images from the datasets. As you can notice, they coherently follow the pose variation.

segmentations, isolating meaningful *parts* that exhibit tight feature distributions. The model has been customized here for the FG/BG separation (we set 2 components and 2 parts, corresponding to the FG and BG), and learned beforehand using a pedestrian database. The segmentation over new samples consists in a fast inference (see [11] for further details).

In the multiple-shot case, if a video sequence is processed, a BG subtraction strategy is enough to obtain the silhouette. For convenience, let us suppose that Z is bounded by a box of size $I \times J$.

In the second phase, features are extracted from the equalized FG images and they are accumulated in a single signature. In the third and final phase matching of the signatures is performed.

3.1. Symmetry-based Silhouette Partition

Gestalt theory considers symmetry as a fundamental principle of perception: symmetrical elements are more likely integrated into one coherent object than asymmetric regions. This finding has been largely exploited for characterizing salient parts of a structured object [13, 18].

Here, we apply this principle for individuating salient human parts that lend themselves to being robustly described. A straightforward way would be to simply use fixed partitions of the bounding box. However, 1) it is no guaranteed that a person's body is well centered in the bounding box, and 2) we experimentally found that a more principled search gives better results, since the segmentation is prone to errors.

Let us first define two basic operators. The first one is the *chromatic bilateral operator*:

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

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 FG region lying in the box of width J and vertical extension $[i-\delta, i+\delta]$ (see Fig. 2). We fix $\delta = I/4$, proportional to the image height, so that scale independency can be achieved.

The second one is the *spatial covering operator*, that calculates the difference of FG 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 FG 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 Ax_{TL} is located at height i_{TL}, obtained as:

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

i.e., we look for the x-axis that separates regions with strongly different appearance and similar area. The values of C are normalized. The search for i_{TL} holds in the interval $[\delta,\ I-\delta]$: Ax_{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 (area) asymmetry Ax_{HT} is positioned at height i_{HT}, obtained as:

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

This separates regions that strongly differ in area and places Ax_{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. 2). The head part R_0 is discarded, because it often consists in few pixels, carrying very low informative content.

On R_1 and R_2, a y-axis of symmetry is estimated. This is located in $j_{LRk},\ (k = 1,2)$, obtained from:

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

This time, C is evaluated on the FG region of size the height of R_k and width δ (see Fig. 2). We look for regions with

similar appearance and area. In this case, δ is proportional to the image width, and it is fixed to $J/4$.

This simple perceptually-driven strategy individuates body parts which are dependent on the visual and positional information of the clothes, robust to pose, viewpoint variations, and low resolution (where pose estimation techniques usually fail or cannot be satisfactorily applied).

3.2. Symmetry-driven Accumulation of Local Features

Once the asymmetry/symmetry axes have been set, different features are extracted from each part, in order to encode their visual appearance. For all features, their distance with respect to the j_{LRk}-axes is taken into account in order to minimize the effects of pose variations.

Weighted Color Histograms. In order to encode all the chromatic content of each part of the pedestrian, HSV histograms are employed; in fact, local histograms have proven to be very effective and largely adopted [1, 8]. More specifically, we build *weighted histograms*, thus taking into consideration the distance to the j_{LRk}-axes: each pixel is weighted by a one-dimensional Gaussian kernel $\mathcal{N}(\mu, \sigma)$, where μ is the y-coordinate of the j_{LRk}, and σ is a priori set to $J/4$. In this way, pixel values near j_{LRk} count more in the final histogram. In the single-shot case, one has a single histogram for each part. In the multiple shot case, one has multiple histograms for each part, and we leave to matching method to choose the most representative histogram (see Sec. 3.3).

Maximally Stable Color Regions (MSCR). The MSCR operator[1] [5] 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. Those maximal regions that are stable over a range of steps constitute the maximally stable color regions of the image. The detected regions are then described by their area, centroid, second moment matrix and average color, forming 9-dimensional patterns. These features exhibit desirable properties for feature 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 location.

In the single-shot case, we extract MSCRs separately from each (FG) part of the pedestrian. In order to discard outliers, we select only MSCRs that lay inside the Gaussian kernel used for color histograms. In the multiple-shot case, we opportunely accumulate the MSCRs coming from the different images by employing a Gaussian clustering procedure [22], which automatically selects the number of components. The clustering is carried out using the 5-dimensional MSCR sub-pattern composed by the centroid and the average color of the blob. We cluster the blobs similar in appearance and position, since they yield redundant information. The contribution of this clustering operation is two-fold: i) it captures only the relevant information, and ii) it keeps low the computational cost of the matching process, where the clustering results are used.

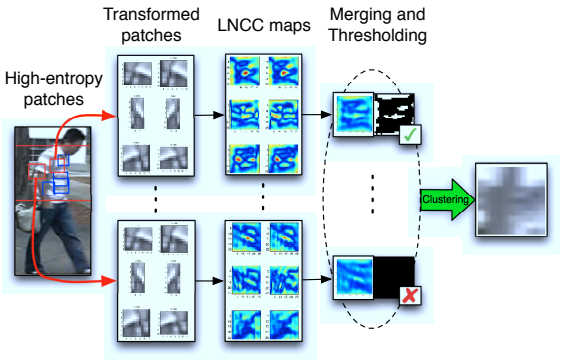

Figure 3. Recurrent high-structured patches extraction. The final result of this process is a set of patches (in this case only one) characterizing each body part of the pedestrian.

Recurrent High-Structured Patches (RHSP). The novel feature we propose aims at highlighting those image patches with texture characteristics that are highly recurrent in the pedestrian appearance (see Fig. 3). The first step consists in the random extraction of patches p of size $[I/6 \times J/6]$, independently on each (FG) part of the pedestrian. In order to take symmetries into consideration, we mainly sample these 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 values of entropy of the patches, thus pruning patches with low structural information (e.g., uniform color). This entropy is computed as the sum H_p of the 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 next step is to apply a set of transformations T_i, $i = 1, 2, \ldots, N_T$ on p, in order to check its invariance to geometric variations of the object. By these transformations we generate a set of N_T patches p_i, and obtain an enlarged set $\hat{p} = \{p_1, \ldots, p_{N_T}, p\}$. As T_i we consider rotations along the y central axis of the patch.

Subsequently, we investigate how recurrent a patch is. We evaluate the Local Normalized Cross-Correlation (LNCC) for each patch in $\hat{p}$. We do not consider the LNCC values of the whole image part, but only a local region containing p. All the $N_T + 1$ LNCC maps are then merged

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

together into the average map. Patches containing small values in this map are discarded. Finally, given all remaining p, we cluster them together, in order to avoid patches with similar content. To this end, we employ the Gaussian clustering [22] on the HSV histogram of the patches, maintaining for each final cluster the patch nearest to the cluster's centroid.

The single-shot and the multiple-shot methods are similar, with the only difference that in the multi-shot case the candidate RHSPs are accumulated over different frames.

3.3. Feature Matching

In this section, we illustrate how the different features are jointly used as a single signature for matching. In general, we have two sets of pedestrian images: a gallery set A and a probe set B. Re-identification consists in associating each person of B to the corresponding person of A. This association depends on the content of two sets: 1) *single-shot vs single-shot* (SvsS), if each image represents a different individual; 2) *multiple-shot vs single-shot* (MvsS), if each image in B represents a different individual, and in A a single person is described by a multiple images signature; 3) *multiple-shot vs multiple-shot* (MvsM), if both A and B contain signatures from multiple images. Groups of images of the same individual can be obtained from tracking information, if available.

In general, the matching of two signatures I_A and I_B is carried out by estimating the *SDALF matching distance* d:

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

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

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

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

The distance d_{WH} evaluates the weighted color histograms. The HSV histograms of each part are concatenated, channel by channel, and compared via Bhattacharyya distance. In the MvsM and MvsS association, we compare each possible pair of histograms contained in the different signatures, selecting the obtained 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, and d_c^{ab}, that compares their mean color. In both cases, the comparison is carried out using the Euclidean distance. This results:

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

where γ takes values between 0 and 1.

In the MvsM and MvsS association, in order to speed up the computation, we first calculate d_{MSCR} on each MSCR b of I_B and each cluster representative of I_A. The representative that gives the lowest distance indicates the cluster, i.e. the set of MSCRs, with which b must be compared with.

d_{RHSP} is obtained by selecting the best pair of RHSP, one in I_A and one in I_B. We evaluate the minimum Bhattacharyya distance among the RHSP's HSV histograms. This is done independently for each body part, summing all the distances achieved and then normalizing for the number of pairs. We used HSV histograms, instead of a feature more specific for describing textures, because the RHSP's content is not necessarily a texture, since it exhibits less regularity.

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 are estimated using the first 100 image pairs of the VIPeR dataset, and left unchanged for all the experiments.

4. Experimental Results

In this section we show extensive experiments to evaluate our approach, providing comparisons with other methods in the state of the art on benchmark datasets. We consider three different datasets. Each one covers different aspects and challenges for the person re-identification problem. The results are shown in terms of recognition rate, by the Cumulative Matching Characteristic (CMC) curve, and the re-identification rate, by the Synthetic Recognition Rate (SRR) curve. The CMC curve represents the expectation of finding the correct match in the top n matches. The SRR curve represents the probability that any of the m best matches is correct. This follows the validation method suggested in [7] for the person re-identification problem. All the following results are obtained using the parameters' values as detailed in Sec. 3.

VIPeR Dataset [7]. This dataset[2] contains two views of 632 pedestrians. Each pair is made up of images of the same pedestrian taken from different cameras, under different viewpoint, pose and light conditions. All images are normalized to 128×48 pixels. Most of the examples contain a viewpoint change of 90 degrees. Each pair is randomly split into two sets: CAM A and CAM B. It is the most challenging dataset currently available for pedestrian re-identification.

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

[2]The dataset is available to download at the web address `http://vision.soe.ucsc.edu/?q=node/178`

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.

The best performance on this dataset is obtained in [8]. In their experimental section, the authors split the dataset evenly into a training and a test set, and run their ELF algorithm. In order to fairly compare our results with theirs, we should know precisely the splitting assignment. Since this information is not provided we compare the results published in [8] with the average of the results obtained by our SDALF method for 10 different random sets of 316 pedestrians[3]. In Fig. 4, we depict the comparative graphs for the CMC and SRR curves. It can be seen that SDALF outperforms ELF. In particular, rank 1 matching rate is around 20% for SDALF, versus around 12% in ELF, while the correct match can be found in the top 10% (rank 31) around 75% of the times for SDALF, versus around 70% for ELF. The most erroneous matchings are due to severe lighting changes, and to the fact that many people tend to dress in very similar ways. In these cases more cues may be necessary, like higher resolution images, in order to grab finer image details, or to consider spatio-temporal information.

We analyze also the robustness of the proposed method when the image resolution decreases. We scale the original images of the VIPeR dataset by factors 0.75, 0.5 and 0.33. The results, depicted in Fig. 4, show that the performances decrease only slightly.

iLIDS Dataset [16]. The iLIDS MCTS dataset is a publicly available video dataset captured at an airport arrival hall in the busy times under a multi-camera CCTV network. It is a real scenario. From these videos a dataset of 479 images of 119 pedestrians was extracted by Zheng et al for testing their Context-based pedestrian re-identification method in [21]. The images of this dataset, normalized to 128×64 pixels, derive from non-overlapping cameras, under quite large illumination changes and subject to occlusions (not present in VIPeR). [21] produces the best performances on this dataset. Since there are more than two examples for each pedestrian, we can evaluate both single and multiple-shot cases.

As to single shot case, we reproduce the same settings of the experiments in [21] in order to make a fair comparison. We randomly select one 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. This whole procedure is repeated 10 times, and the average CMC and SRR curves are displayed in Fig. 5. We outperform [21], without using any additional information about the context, and even using images at lower resolution. SDALF proves to be robust enough to deal with occlusions and quite crowded situations. Indeed, some images of the dataset contain more than one person.

As to the multiple-shot case, we run experiments both on MvsS and MvsM cases. For the former, we build a gallery set of multi-shot signatures and we match it with a probe set of one-shot signatures. For 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 a mean of about 4 images for pedestrian, we test our algorithm with $N = \{2, 3\}$ for MvsS and just $N = 2$ for MvsM. For each case, we run 100 independent trials. The results, depicted in Fig. 5, show that, in the MvsS case, just 2 images are enough to increment the performances of about 10%. 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 lay between these two performances.

ETZH Dataset [4]. This dataset is captured from moving cameras, and it has been used originally for pedestrian detection. Schwartz and Davis in [19] extract a set of samples for each different person in the videos, and use the resulting set of images to test their PLS method[4]. The moving camera setup provides a range of variations in people's appearance. Variation in pose is relatively small, though, in comparison with the other two datasets. The most challenging aspects of ETHZ are illumination changes and occlusions. All images are normalized to 64×32 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. [19] produces the best performances on this dataset.

In the single-shot case, the experiments are carried out exactly as for iLIDS. We repeat the same operation 10 times, in order to provide a robust statistics. 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 information about the subject. Therefore, we apply beforehand a clustering algorithm [22] on the original frames, based on their HSV histograms. Consecutive similar frames would end up in the same cluster. At this point, we select randomly one frame for each cluster: these are the keyframes to use for the multi-shot signature. The results for both single and multiple-shot case for SEQ. #1 are reported on Fig. 6. We

[3]To receive the exact partitions used, in order to facilitate future comparative experiments, please contact us.

[4]The dataset is available to download at the web address `http://www.umiacs.umd.edu/~schwartz/datasets.html`

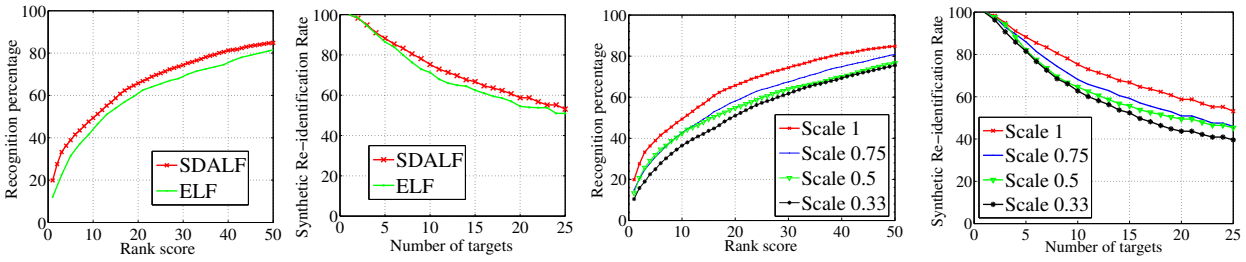

Figure 4. Performances on VIPeR dataset. In (a) and (b), comparison with ELF method. In (c) and (d), comparison of SDALF at different scales. In (a) and (c) the CMC curve. Only the first 50 ranking positions are displayed. In (b) and (d) the SRR curve.

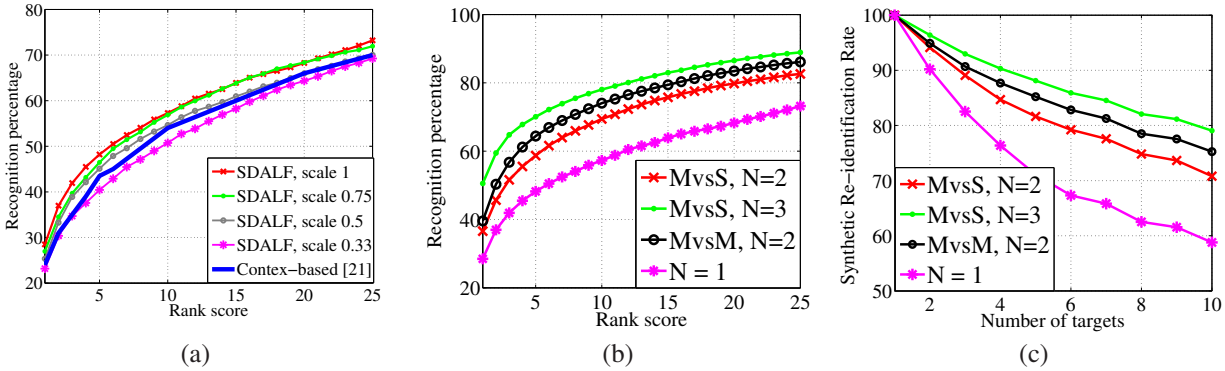

Figure 5. Performances on iLIDS dataset. (a): CMC curves, comparing the results reported in [21] and single-shot SDALF at different resolutions. (c) and (d): CMC and SRR curves for MvsS and MvsM cases. For reference, we put also the single-shot case ($N = 1$). In the CMC curves, in accordance with what reported in [21], only the first 25 ranking positions are displayed.

compare the results with what reported in [19]. 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. Additionally, PLS requires to have all the gallery signatures beforehand, in order to estimate the weights on the appearance model. If one pedestrian is added the weights must be recomputed.

In SEQ. #2 (Fig. 6) we have a similar behavior: rank 1 correct matches can be obtained in 91% of the cases for MvsS, 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.

Acknowledgments

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

5. Conclusions

In this paper, we addressed the appearance-based re-identification problem proposing a feature extraction and matching strategy. This strategy is based on the localization of perceptual relevant human parts, driven by asymmetry/symmetry principles, followed by the extraction of three complementary kinds of features. Each type of feature encodes different information, namely, chromatic information, structural information through uniformly colored regions, and the nature of recurrent informative (in an entropy sense) patches. In this way, robustness to pose, viewpoint and illumination variations is achieved.

The method works by using a single image of a person (single-shot modality), or several frames (multiple-shot modality). We tested our approach (in single-shot modality) on three challenging public databases (VIPeR, iLIDS, and ETHZ), outperforming the highest performances on all but one dataset. In the multiple-shot modality, our performances strongly increase, setting new state-of-the-art re-

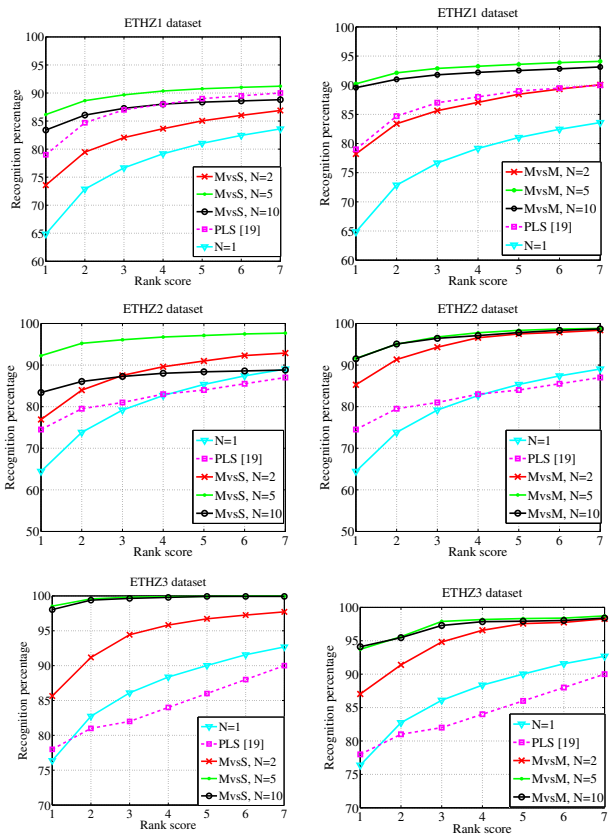

Figure 6. Performances ETHZ dataset. Top row, results on SEQ. #1; middle row, on SEQ. #2; bottom row, on SEQ. #3. We compare our method with the results of PLS method in [19]. On the left column, we report the results for single-shot SDALF ($N = 1$) and MvsS SDALF; on the right column the results for MvsM SDALF. In accordance with what reported in [19], only the first 7 ranking positions are displayed.

sults. Finally, we would like to outline the fact that our technique operates independently on each individual, not embracing discriminative philosophies which need the knowledge of the entire dataset: we simply extract a novel set of reliable, robust, and descriptive localized features. This opens up to a wide range of future developments and customizations, including feature boosting, multi-scale reasoning, to be possibly used in a multi-object tracking technique.

References

[1] U. Park, A.K. Jain Kitahara, K. Kogure, N. Hagita. ViSE: Visual Search Engine Using Multiple Networked Cameras. In *ICPR*, pages 1204-1207. 4

[2] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded Up Robust Features. In *ECCV*, pages 404–417, 2006. 1, 2

[3] N. Bird, O. Masoud, N. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):167–177, June 2005. 1, 2

[4] A. Ess, B-Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007. 2, 6

[5] P.-E. Forssén. Maximally stable colour regions for recognition and matching. In *CVPR*, Minneapolis, USA, 2007. IEEE Computer Society. 1, 4

[6] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, vol. 2, pages 1528–1535, 2006. 1, 2

[7] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recongnition, reacquisition and tracking. In *PETS*, 2007. 2, 5

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensamble of localized features. In *ECCV*, pages 262–275,2008. 1, 2, 4, 6

[9] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proceedings of the IEEE Conference on Distributed Smart Cameras*, pages 1–6, 2008. 1, 2

[10] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking accross non-overlapping views. *In CVIU*, 109:146–162, 2007. 1

[11] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey. Stel component analysis: Modeling spatial correlations in image class structure. In *CVPR*, pages 2044–2051, 2009. 2, 3

[12] Z. Lin and L. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing*, pages 23–34, 2008. 1, 2

[13] K. L. M. Cho. Bilateral symmetry detection and segmentation via symmetry-growing. In *BMVC*, 2009. 3

[14] D. Makris, T. J. Ellis, and J. K. Black. Bridging the gaps between cameras. In *CVPR*, vol. 2, pages 205–210, 2004. 1

[15] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. In *Pattern Recognition*,36(9), 2003. 1, 2

[16] U. H. Office. i-LIDS multiple camera tracking scenario definition, 2008. 2, 6

[17] A. Rahimi, B. Dunagan, and T. Darrel. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *CVPR*, volume 1, pages 187–194, 2004. 1

[18] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: The generalized symmetry transform. *IJCV*, 14(2):119–130, 1995. 3

[19] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least squares. In *XXII SIBGRAPI*, 2009. 1, 2, 6, 7, 8

[20] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV* pages 1–8, 2007. 1, 2

[21] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 1, 2, 6, 7

[22] M. T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002. 4, 5, 6