Multiple-shot Person Re-identification by HPE signature

Loris Bazzani*,Marco Cristani*†,Alessandro Perina*,Michela Farenzena*,Vittorio Murino*†
*Computer Science Department, University of Verona, Italy
†Istituto Italiano di Tecnologia (IIT), Genova, Italy
Emails: {loris.bazzani,marco.cristani,alessandro.perina,michela.farenzena,vittorio.murino}@univr.it

***Abstract*—In this paper, we propose a novel appearance-based method for person re-identification, that condenses a set of frames of the same individual into a highly informative signature, called Histogram Plus Epitome, HPE. It incorporates complementary global and local statistical descriptions of the human appearance, focusing on the overall chromatic content, via histograms representation, and on the presence of recurrent local patches, via epitome estimation. The matching of HPEs provides optimal performances against low resolution, occlusions, pose and illumination variations, defining novel state-of-the-art results on all the datasets considered.**

***Keywords*-Pedestrian Re-identification, Pedestrian Description, Video-surveillance.**

I. Introduction

Person re-identification is a crucial issue in multi-camera tracking scenarios, where cameras with non-overlapping views are employed. Considering a single camera, the tracking captures several instances of the same individual, providing a *volume* of frames. The re-identification consists in matching different volumes of the same individual, coming from different cameras.

In the literature, the re-identification methods that focus solely on the appearance of the body are dubbed *appearance-based* methods, and can be grouped in two sets. The first group is composed by the *single-shot* methods, that model a person analyzing a single image [1], [2], [3]. They are applied when tracking information is absent. The second group encloses the *multiple-shot* approaches; they employ multiple images of a person (usually obtained via tracking) to build a signature [4], [5], [6], [7]. In [4], each person is subdivided into a set of horizontal stripes. The signature is built by the median color value of each stripe accumulated over different frames. A matching between decomposable triangulated graphs, capturing the spatial distribution of local temporal descriptions, is presented in [5]. In [6], a signature composed by a set of SURF interest points, collected over short video sequences, is employed. In [7], each person is described by local and global features, which are fed into a multi-class SVM for recognition. Other approaches 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 [8], but these cannot be considered purely appearance-based approaches.

In this paper, we present a novel multiple-shot appearance-based re-identification method, based on the extraction and matching of an ID signature that embeds global and local appearance features, called *Histogram Plus Epitome*, HPE. A pre-processing step extracts the body silhouettes from a set of images, reasonably, but not necessarily, acquired from a single-camera tracking phase. Afterwards, multiple *informative* images for each individual are selected by a clustering method that rejects the redundant information (similar images) and outliers (images containing occlusions). Then, complementary aspects of the human body appearance are extracted from the set of images highlighting: 1) the global chromatic content via a mean HSV histogram; 2) the presence of recurrent local patterns, through epitomic analysis [9]. The first aspect captures all the chromatic information of an individual's appearance, condensing it in a widely accepted descriptor for re-identification. The second aspect is supported by the paradigm of object recognition by local features, and lies on a model, the epitome [9], that encodes the pixels' local spatial layout with a set of frequently visible patches.

Our approach differs from the state of the art: unlike [4], [5], we do not rigidly link features to parts of the human structure, which is not reliable at low resolutions. We do not simply accumulate local features, as in [6], but we keep recurrent local aspects, that may reappear with higher chance in novel instances of the person. Finally, we do not employ as in [7] discriminative learning techniques, that have to be re-trained each time that a novel subject occurs.

We prove the reliability of our technique using the most recent databases with multiple images available for re-identification, *i.e.*, iLIDS for re-identification [3], and ETHZ [2]. For comparison, we consider the best results on these datasets: they are produced by single-image methods, that however exploit contextual information or discriminative strategies that makes the confrontation worthy. The rest of the paper is organized as follows. Sec. II details our approach. Several results are reported in Sec. III, and, finally, conclusions are drawn in Sec. IV.

II. The proposed method

A. Pre-processing: Foreground Extraction

In this step, the pixels depicting the person (the foreground, FG) are separated from the rest of the image (the background, BG), obtaining a set of *instances*; in this way,

our descriptor focuses on the sole person. This can be done by a BG subtraction strategy, or, more in general, using the STEL generative model [10]. STEL model captures the structure of an image class as a mixture of *component* segmentations, and isolates meaningful *parts* that exhibit tight feature distributions. We set 2 components and 2 parts (i.e., the FG and the BG). STEL has been learnt beforehand on a person database (not considering the experimental data), and the segmentation over new samples consists in a fast inference (see [10] for further details). For the sake of generality, we use STEL here.

B. Images Selection

The (single-camera) tracking output usually consists in a sequence of consecutive images of each individual in the scene. In order to discard redundant information, for each individual, we apply a completely unsupervised Gaussian clustering [11] to the HSV histograms of the instances. Clusters with a low number of elements ($= 3$ in our experiments) are discarded. For each cluster, an instance is randomly chosen, building the set $\mathbf{X}^k = \{X_n^k\}_{n=1}^{N_k}$ for the person k, with N_k instances. Each instance is scaled to $I \times J$ image size.

C. Histogram Plus Epitome Descriptor

The HPE descriptor is formed by three features, extracted from $\mathbf{X}^k$: the first captures chromatic global information; the last two analyze the presence of recurrent local patterns, i.e., the epitome.

The Global Feature, HSV histogram: The global appearance of each person is initially encoded by HSV histograms, in a 36-dimensional feature space $[H = 16, S = 16, V = 4]$, one for each instance. Then, the global feature $\mathrm{H}(\cdot)$ is built by averaging the histograms of the multiple instances of $\mathbf{X}^k$. This makes the feature robust to illumination and pose variations, keeping the predominant chromatic information only.

The Local Features, Epitomic Analysis: A image epitome e [9] is the result of collapsing an image, through a generative model, into a small collage of overlapped patches containing the essence of the textural, shape and appearance properties of the image. In this paper, we generalize the epitome by employing all the instances of an individual.

A set of P *ingredient* patches of fixed dimensionality $I_e \times J_e$ are uniformly sampled from each $X_n^k \in \mathbf{X}^k$, building a multi-shot set of patches $\{z_m\}_{m=1}^{N_k \times P}$. For each patch z_m, the generative model infers a hidden mapping variable $\tau_m(i,j)$ that maps (through translations) z_m into a equally sized portion of the epitome, having (i,j) as left-upper corner. The inference is possible by evaluating the variational distribution $q(\tau_m(i,j))$, that represents the probability of that mapping (see [9] for details).

By mapping all patches in the epitome space and averaging them, we extract the epitome's parameters $e = \{\mu, \phi\}$. μ is the epitome mean, i.e., an image that contains similar, recurrent patches present in several instances, while ϕ represents the standard deviation map associated to each pixel of e.

In this paper, we customize the use of the epitome for the task-at-hand, extracting two different features. The first is the *generic* epitome $\mathrm{Ge}(\cdot)$, that is the epitome's mean μ. Considering just μ is equivalent to disregard (i.e., being invariant to) small variations among the different instances' patches, usually due to small scale/pose discrepancies and illumination variations among the patches. μ is described by a HSV histogram, in order to permit an easy comparison between generic epitomes[1].

The second feature, the *local* epitome $\mathrm{Le}(\cdot)$, focalizes on individuating in the epitome local regions that portray highly informative recurrent ingredient patches. First, we estimate the prior probability on the transformation $P(\tau) = \frac{\sum_m q(\tau_m)}{N_k \cdot P}$, that for each pixel (i,j) of the epitome, gives the probability that the patch in the epitome having (i,j) as left-upper corner represents several ingredient patches $\{z_m\}$. Second, we rank in descending order of $P(\tau)$ all the patches in the epitome, retaining only the first $M = 40$, i.e., the most recurrent ones. We re-rank these patches in descending order by evaluating their entropy, retaining the first $F = 10$, i.e., the most informative ones. M and F's values are set after cross-validation on a small experimental data subset. As for the generic epitome, we use a HSV histogram for the patches' description.

D. Feature Matching

In the re-identification problem, we have two sets of volumes of instances, each one addressing a single person: a gallery set A and a probe set B. Re-identification consists in looking for matching between each volume in B with a volume in A. Our HPE descriptor is matched by combining three similarities scores (one for each feature), for each pair of volumes. In details, the matching between volumes $\mathbf{X}^A$ and $\mathbf{X}^B$ is carried out by minimizing the *matching distance* d:

$$d(\mathbf{X}^A, \mathbf{X}^B) = \beta_1 \cdot \log\left(d_c(\mathrm{H}(\mathbf{X}^A), \mathrm{H}(\mathbf{X}^B))\right) + \quad (1)$$

$$\beta_2 \cdot \log\left(d_c(\mathrm{Ge}(\mathbf{X}^A), \mathrm{Ge}(\mathbf{X}^B))\right) + \quad (2)$$

$$\beta_3 \cdot \log\left(d_e(\mathrm{Le}(\mathbf{X}^A), \mathrm{Le}(\mathbf{X}^B))\right) \quad (3)$$

where the $\mathrm{H}(\cdot)$, $\mathrm{Ge}(\cdot)$, and $\mathrm{Le}(\cdot)$ are the HSV histogram, the generic and the local epitome, respectively, and βs are normalized weights. d_c in Eqs. (1) and (2) is the Bhattacharyya distance, while d_e in Eq. (3) is estimated as the minimum distance of each patch b in $\mathrm{Le}(\mathbf{X}^B)$ to each

[1]Given a set of ingredient patches, learning two times an epitome results in two similar models with a possibly different spatial displacement. Adopting the histogram cancels out such discrepancy.

patch a in $\mathrm{Le}(\mathbf{X}^A)$ of the local epitome, i.e.:

$$d_e = \frac{1}{K} \sum_{b \in \mathrm{Le}(\mathbf{X}^B)} \min_{a \in \mathrm{Le}(\mathbf{X}^A)} d_c(\mathrm{H}(a), \mathrm{H}(b)), \qquad (4)$$

where K is a normalization constant.

In our experiments, we fix $\beta_{\{1,2,3\}} = [1/33, 1/27, 1/30]$. These values are estimated using the first 100 image pairs of the iLIDS dataset, and left unchanged for all the experiments. These values for βs underline the fact that the features are all important. Unbalancing the weights deteriorates the performances of the approach.

III. Experimental Results

In order to provide quantitative results for our approach, we consider the iLIDS for re-identification [12] and ETHZ [13] datasets. Both datasets cover challenging aspects of the person re-identification problem: shape deformation, illumination changes, occlusions, image blurring, low resolution images, *etc*. For comparison, we consider the best performances obtained so far on these datasets. The evaluation employs widely used state-of-the-art measurements [2], [3]: the Cumulative Matching Characteristic (CMC) curve, that represents the expectation of finding the correct match in the top n matches, and the Synthetic Recognition Rate (SRR) curve, that represents the probability that any of the n best matches is correct.

1) iLIDS for reindentification: The iLIDS dataset for re-identification (iLIDSfr) is composed by 479 images of 119 people, normalized to size 64×128. It is built from the iLIDS surveillance dataset [12]. It considers an airport arrival hall in the busy times under a multi-camera CCTV network. We reproduce the same experimental settings of [3], that get the best performances on iLIDSfr.

We randomly select a subset of N images for each person to build the gallery set, while the others form the probe set. Then, both gallery and probe sets are made up of the proposed HPE. Then, the matching between probe and gallery set is estimated. This whole procedure is repeated 20 times, and the average of the CMC and SRR curves over the trials is estimated. We test our algorithm using $N_k = N = \{2, 5\}\ \forall k$, i.e., all the HPEs are built with 2 or 5 instances, without employing the clustering method (Sec. II-B), simply because the number of images for person is low. The results, depicted in Fig. 1, show that 2 images are enough to outperform [3]. In that approach, the scene and the people around an individual are considered as characterizing the individual itself. The underlying intuition is that a person is better recognizable while considering the surrounding in which he is. Here we show that adding a second instance of an individual carries more information. Adding more images ($N = 5$) induces a further improvement[2]. Moreover, our method proves to be robust to occlusions and quite crowded situations (the images often contain more than a person).

[2]We do not report the results with $N = 1$ for every dataset. Though comparable with the best approaches, they do not mirror the nature of the proposed approach that is based on multiple images.

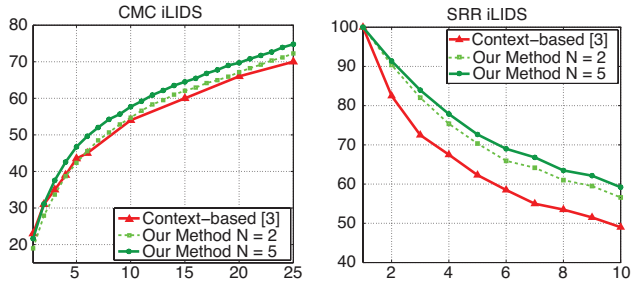

Figure 1. A comparison in term of CMC and SRR on iLIDSfr between our method and context-based method [3].

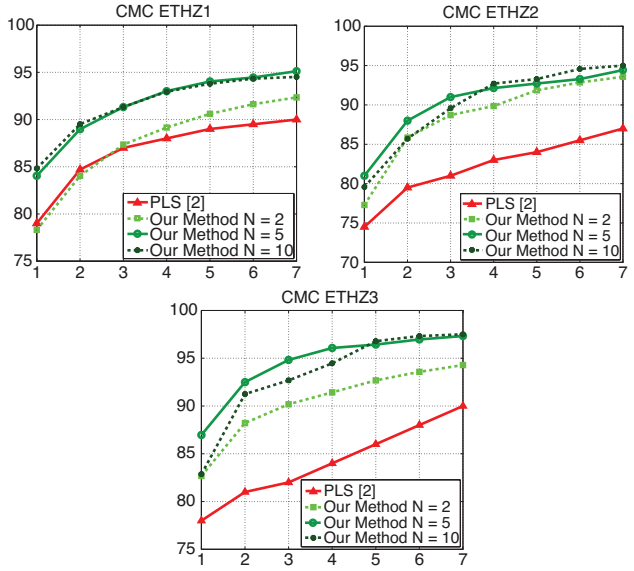

Figure 2. A comparison in terms of CMC on ETHZ between our method and PLS method [2].

2) ETHZ: This dataset is built from [13], and is captured from moving cameras in a crowded street. The best performances in this dataset are obtained by the Partial Least Squares (PLS)-based method [2]. The most challenging aspects of ETHZ are illumination changes, occlusions and low resolution (all images are 32×64 pixels). The dataset is structured as follows: Seq. #1 contains 83 people (4.857 images); Seq. #2 contains 35 people (1.936 images); Seq. #3 contains 28 people (1.762 images).

The experiments are carried out exactly as for iLIDSfr, choosing randomly the elements for the gallery set and the probe set. Repeating the same operation 20 times provides a reliable statistics. The experiments consider $N = \{2, 5, 10\}$, i.e., selecting randomly N instances from the N_k obtained by the clustering (Sec. II-B). The results of each sequence in ETHZ are reported on Fig. 2. In all the sequences, we obtain the best results compared to the best performances

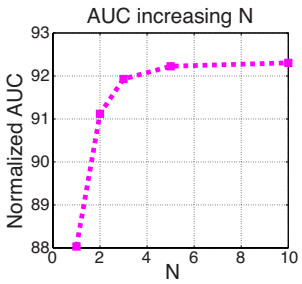

Figure 3. Normalized area under the CMC, increasing the number of images per person N.

on this dataset reported in [2]. Unlike our method, PLS uses all foreground and background information. In this case, background information makes the re-identification task simpler because the images of a person in both the probe set and the gallery set can have the same background. This assumption is not valid in a general multi-camera setting. In addition, PLS requires to have all the gallery images beforehand, in order to learn the weights of their descriptor. If one person is added the weights must be recomputed.

Like in iLIDS, the performances increase adding more images to the descriptor. However, there is a point after that adding more information does not enrich consistently the descriptive power of the descriptor any more, while increases the computational load, significantly slowing down the method. In order to choose the best value for N, we compute the Area Under the Curve CMC, normalized to the total area of the graph (nAUC). A high value of nAUC means high performances of the method. We average the nAUC for all iLIDS and ETHZ results, including the experiments with $N = \{1, 2, 3, 5, 10\}$ (see Fig. 3). A qualitative analysis of the profiles points outs as 5 the best choice for N.

IV. Conclusions

In this paper, we address the person re-identification problem proposing a novel descriptor, HPE, that is based on a collection of global and local features. The descriptor embeds information from multiple images per person, showing that the presence of several occurrences of an individual is very informative for re-identification. Our descriptor operates independently on each individual, not embracing discriminative philosophies that imply strong operating requirements. Employing HPE, we set novel best performances on all the available re-identification databases. The approach focuses on accuracy rather then efficiency, so we plan to customize it for an on-line processing.

Acknowledgements

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

References

[1] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensamble of localized features," in *Proceedings of the European Conference on Computer Vision*, Marseille, France, 2008, pp. 262–275.

[2] W. Schwartz and L. Davis, "Learning discriminative appearance-based models using partial least squares," in *XXII SIBGRAPI 2009*, 2009.

[3] W. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC 2009*, 2009.

[4] N. Bird, O. Masoud, N. Papanikolopoulos, and A. Isaacs, "Detection of loitering individuals in public transportation areas," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 2, pp. 167–177, June 2005.

[5] N. Gheissari, T. B. Sebastian, P. H. Tu, , J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1528–1535.

[6] 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*, 2008, pp. 1–6.

[7] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio, "Full-body person recognition system," vol. 36, no. 9, 2003.

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

[9] N. Jojic, B. J. Frey, and A. Kannan, "Epitomic analysis of appearance and shape," in *IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 34.

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

[11] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. PAMI*, vol. 24, no. 3, pp. 381–396, 2002.

[12] U. H. Office, "i-LIDS multiple camera tracking scenario definition," 2008.

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