Multiple-shot Person Re-identification by Chromatic and Epitomic Analyses

L. Bazzani[a,*], M. Cristani[a,b], A. Perina[a], V. Murino[a,b]

[a]*Department of Computer Science, University of Verona, Verona, Italy*
[b]*Istituto Italiano di Tecnologia (IIT), Genova, Italy*

Abstract

We propose a novel appearance-based method for person re-identification, that condenses a set of frames of an individual into a highly informative signature, called the Histogram Plus Epitome, HPE. It incorporates complementary global and local statistical descriptions of the human appearance, focusing on the overall chromatic content via histogram representation, and on the presence of recurrent local patches via epitomic analysis. The re-identification performance of HPE is then augmented by applying it as human part descriptor, defining a structured feature called Asymmetry-based HPE (AHPE). The matching between (A)HPEs provides optimal performances against low resolution, occlusions, pose and illumination variations, defining state-of-the-art results on all the considered datasets.

Keywords: Person re-identification, Epitome, Asymmetry detection

1. Introduction

Person re-identification (*re-id*) aims at recognizing an individual captured in different times and/or locations, considering a large set of candidates. It has widespread applications such as tracking and person identification, re-acquisition and verification, as showed in Schwartz and Davis (2009). In literature, re-id methods that focus solely on the appearance of the body are dubbed *appearance-based* methods, and can be divided in two groups: *single-shot* and *multiple-shot* approaches. The former models a person analyzing a single image for each individual, not exploiting the temporal information provided by tracking, such as

*Corresponding author:
Email address: loris.bazzani@univr.it (L. Bazzani)
URL: www.lorisbazzani.info (L. Bazzani)

Prosser et al. (2010), Gray and Tao (2008), Schwartz and Davis (2009), Bak et al. (2010), and Zheng et al. (2009). The latter group, instead, employs multiple images of a person (obtained via tracking) to build the descriptor used for re-id, such as Bird et al. (2005), Gheissari et al. (2006), Hamdoun et al. (2008), Nakajima et al. (2003), and Farenzena et al. (2010). Bird et al. (2005) defines a descriptor built by subdividing the person in horizontal stripes, keeping the median color of each stripe accumulated over different frames. A matching between decomposable triangulated graphs, capturing the spatial distribution of local temporal descriptions, is presented by Gheissari et al. (2006). Hamdoun et al. (2008) uses SURF interest points, collected over short video sequences. Another supervised learning-based approach is proposed by Nakajima et al. (2003): local and global features accumulated over time are fed into a multi-class Support Vector Machine (SVM) for recognition. A recent method for re-id is proposed by Farenzena et al. (2010): it extracts features from different body parts, and weights them opportunely by exploiting symmetry and asymmetry perceptual principles. Other approaches, for example see Javed et al. (2007), add temporal reasoning on the spatial layout of the monitored environment, but these cannot be considered purely appearance-based approaches.

In this paper, we propose a novel multiple-shot appearance-based method for re-id, based on the extraction and matching of a signature that embeds global and local appearance features (see Fig. 1). Complementary aspects of the human appearance are extracted from the foreground region of the image (*i.e.*, the person) highlighting: 1) the global chromatic content via a mean color histogram, and 2) the presence of recurrent local patterns through epitomic analysis proposed by Jojic et al. (2003). The former captures all the chromatic information of an individual's appearance, condensing it in a widely accepted descriptor for re-id. The latter is supported by the paradigm of object recognition by local features, called epitome by Jojic et al. (2003), that encodes the pixels' local spatial layout with a set of frequently visible patches. Another advantage of the epitome is that we can properly accumulate images in a multi-shot descriptor. The descriptor is called *Histogram Plus Epitome*, HPE. We then exploit the asymmetry-based segmentation proposed by Farenzena et al. (2010) in order to apply our signature as human part descriptor, giving rise to the *Asymmetry-based* HPE (AHPE).

The proposed approach differs from the previous work: 1) Unlike Bird et al. (2005) and Gheissari et al. (2006), we do not rigidly link features to parts of the human structure, which is not reliable at low resolutions. 2) We do not simply accumulate local features with heuristics, as Hamdoun et al. (2008) and Farenzena et al. (2010), but we keep recurrent local aspects by analyzing the epitome result-

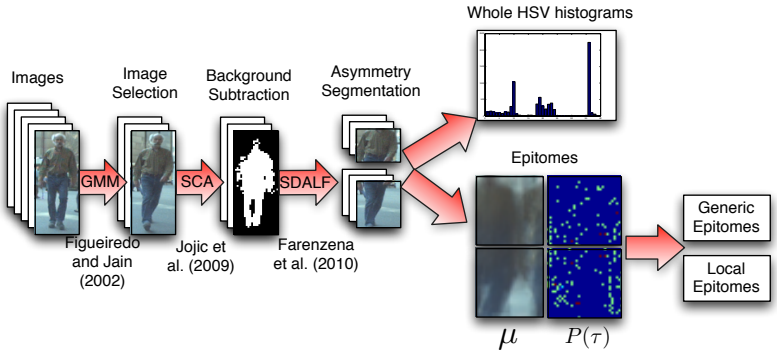

Figure 1: Overview of the proposed approach.

ing from the images of several person, that may reappear with higher probability in novel instances of the person. 3) We do not employ discriminative learning techniques as in Nakajima et al. (2003), that have to be re-trained each time a novel subject appears.

We test the proposed method on the most challenging public datasets with multiple images, *i.e.*, iLIDS for re-id of Zheng et al. (2009), and ETHZ of Schwartz and Davis (2009), comparing it with the best results on these datasets. As further analysis, we apply our approach on a dataset extracted from the CAVIAR repository [1]. This allows to clearly understand the benefit of having multiple instances per person in a re-id challenge. The rest of the paper is organized as follows. Sec. 2 details our approach, and related results are reported in Sec. 3. Finally, conclusive remarks are drawn in Sec. 4.

2. The proposed method

The overview of the proposed approach is shown in Fig. 1. Given a set of images for each pedestrian, the image selection phase discards the redundant temporal information that can be extracted by a tracking system, keeping only the essential data through a clustering method (Sec. 2.1). Then, the foreground extraction step splits the pixels that belong to the person (the foreground) from the rest of the image (the background), so that our descriptor will focus on the sole person (Sec. 2.2). A asymmetry-based segmentation of the images (Sec. 2.5) is done so as to extract interesting parts from the pedestrian. The descriptor is therefore extracted from each part of selected images (Sec. 2.3).

[1] `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`.

2.1. Image Selection

Since there is a temporal correlation between images of each tracked individual, redundancy is expected. It is discarded by applying the unsupervised Gaussian clustering method proposed by Figueiredo and Jain (2002) with automatic selection of the number of clusters N_k (k stays for the k-th person). HSV histogram is used as feature for clustering, in order to capture appearance similarities. Then, N_k images are randomly chosen from each cluster of each person, building the set $\mathbf{X}^k = \{X_n^k\}_{n=1}^{N_k}$. Experimentally, we found that clusters with low number of elements ($= 3$ in our experiments) contain outliers, such as occlusions, thus these cluster are discarded. It is worth noting that the clusters the method selects can still contain occlusions and bad images, hard for the re-id task.

2.2. Foreground Extraction

SCA model[2] by Jojic et al. (2009) captures the common structure of an image class: 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 L stel priors $p_l(s_i)$, that capture the common structure of the 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 image measurements. Instead of local appearance similarity, the model insists on consistent segmentation through the stel prior. Each component l 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.*, the foreground and the background) and $L = 2$, modeling the distribution over image measurements as a Mixture of Gaussians with 2 Gaussian components, as we want to capture segments with multiple color modes within them. SCA has been learnt beforehand on a person database different from those considered in the experiments, *i.e.*, VIPER dataset by Gray et al. (2007) and the segmentation over new samples consists in a inference on the model (it takes about 0.5 second per image using a non-optimized MATLAB version of the code).

2.3. Histogram Plus Epitome Descriptor

We define the HPE descriptor as composition of three features extracted from each $\mathbf{X}^k$: a chromatic global feature, that is, a color histogram; and two lo-

[2] In static-camera tracking, one could even exploit the temporal correlation for segmenting the moving objects. We use SCA, because it does not rely on the temporal correlation assumption.

cal epitome-based features, that capture the presence of recurrent local patterns. Moreover, those features can be extracted from parts of the person (Sec. 2.5).

Color histogram. As global appearance feature we use the Hue Saturation Value (HSV) histogram, proven to be very effective and largely adopted in several applications, *e.g.*, Gray and Tao (2008) and Sebastian et al. (2008). We encode it 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.

Epitomic Analysis. The main contribution of the work is that we employ the epitomic analysis by Jojic et al. (2003) to accumulate information/images over the time in order to build a multi-shot descriptor, without any assumption or heuristics. An image epitome is the result of collapsing an image or a set of images, through a generative model, into a small collage of overlapped patches embedding the essence of the textural, shape and appearance properties of the data.

A set of P *ingredient* patches of fixed size[3] $I_e \times J_e$ are uniformly sampled from each image $X_n^k \in \mathbf{X}^k$, 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 Jojic et al. (2003) for details). By mapping all patches in the epitome space and averaging them, we extract the epitome's parameters $e = \{\mu, \phi\}$, where 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 the epitome.

We customize the use of the epitome for the task at hand, extracting two different features from it: the *generic* epitome and the *local* epitome. The generic epitome $\mathrm{Ge}(\cdot)$ extracts information directly from the mean μ. Considering just μ is equivalent to disregarding (*i.e.*, being invariant to) small variations among the different instances' patches, usually due to small scale/pose discrepancies and illumination variations. A single HSV histogram is obtained from μ in order to

[3]To set the patch sizes we should fulfill the trade-off between too small patches, where the epitome converges to an histogram, and too big patches where the epitome loose its generalization properties. Experimentally, we found out that the patch area has to be $1/3$ of the area of the image.

have a robust appearance-based feature. Moreover, learning an epitome twice on the same data gives two similar models with a different spatial displacement. Adopting histograms cancels out such discrepancy.

On the other hand, the local epitome $\text{Le}(\cdot)$ is focused on detecting local regions in the epitome that portray highly informative recurrent ingredient patches. To this end, first, we estimate the prior probability on the transformation $P(\tau) = \frac{\sum_m q(\tau_m)}{N_k \cdot P}$ (see Fig. 1), that 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. Then, we rank again these M patches in descending order by evaluating their entropy, retaining the first $F = 10$, *i.e.*, the most informative ones[4]. We describe each *survived* patch with an HSV histogram (*i.e.*, F histograms in total).

2.4. Histogram Plus Epitome Matching

The re-id problem consists in matching elements B of a probe set to elements A of the gallery set. In general, we can define the re-id problem as a maximum log-likelihood estimation problem. In more detail:

$$A^* = \arg\max_A \left(\log P(\mathbf{X}^A|\mathbf{X}^B)\right) = \arg\min_A \left(d(\mathbf{X}^A, \mathbf{X}^B)\right) \tag{1}$$

where the equality is valid because we define $P(\mathbf{X}^A|\mathbf{X}^B)$ in Gibbs form $P(\mathbf{X}^A|\mathbf{X}^B) = e^{-d(\mathbf{X}^A,\mathbf{X}^B)}$ and $d(\mathbf{X}^A, \mathbf{X}^B)$ measures the distance between two descriptors. The HPE matching distance is defined by combining three similarities scores (one for each feature), for each pair of volumes:

$$\begin{aligned} d(\mathbf{X}^A, \mathbf{X}^B) = \beta_1 \cdot (d_c(\text{H}(\mathbf{X}^A), \text{H}(\mathbf{X}^B))) + \beta_2 \cdot (d_c(\text{Ge}(\mathbf{X}^A), \text{Ge}(\mathbf{X}^B))) + \\ + \beta_3 \cdot (d_e(\text{Le}(\mathbf{X}^A), \text{Le}(\mathbf{X}^B))) \end{aligned} \tag{2}$$

where the $\text{H}(\cdot)$, $\text{Ge}(\cdot)$, and $\text{Le}(\cdot)$ are the HSV histogram, the generic and the local epitome, respectively, and βs are normalized weights[5]. d_c is the Bhattacharyya distance, while d_e is estimated as the minimum distance of each patch b in $\text{Le}(\mathbf{X}^B)$ to each patch a in $\text{Le}(\mathbf{X}^A)$ of the local epitome, *i.e.*:

$$d_e = \frac{1}{C} \sum_{b \in \text{Le}(\mathbf{X}^B)} \min_{a \in \text{Le}(\mathbf{X}^A)} d_c(\text{H}(a), \text{H}(b)), \tag{3}$$

[4] M and F's values are set after cross-validation on a small experimental data subset.

[5] See Sec. 3 for a quantitative analysis of the performances when these weights vary.

where C is a normalization constant.

In terms of computational complexity, Eq. 1 is bounded by $\mathcal{O}(K \cdot (N^2 + F^2))$, because we have K pedestrians with N images each, which means $K \cdot N$ HSV histograms, K Ge$(\cdot)$ histograms, and $K \cdot F$ Le$(\cdot)$ histograms.

2.5. Asymmetry-based HPE

Semantic segmentation of objects has been largely exploited for characterizing salient parts of a structured object in object recognition tasks (Jojic and Caspi (2004)). We exploit the segmentation technique proposed by Farenzena et al. (2010) that uses Gestalt theory considerations on symmetry and asymmetry to segment the human body into horizontal stripes corresponding to head, torso and legs. The main idea is that horizontal parts are asymmetric in size and in appearance (for details, see Farenzena et al. (2010)). The main advantage of this strategy is that 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). The AHPE matching is defined by averaging the values of Eq. 2 for each part.

3. Experimental Results

The quantitative evaluation considers several public multi-shot datasets: ETHZ datasets by Ess et al. (2007), iLIDS for re-id by Zheng et al. (2009). We also consider a variant of the iLIDS, dubbed here iLIDS$_{\geq 4}$, and a dataset extracted from the CAVIAR repository for re-identification purposes, called CAVIAR4REID. These datasets cover challenging aspects of the person re-id problem, such as shape deformation, illumination changes, occlusions, image blurring, very low resolution images, *etc*. We compare HPE and AHPE with the best performances obtained so far on these datasets: PLS by Schwartz and Davis (2009), context-based re-id by Zheng et al. (2009), Spatial Covariance Regions (SCR) by Bak et al. (2010) and SDALF by Farenzena et al. (2010). State-of-the-art measurements are used, that is, the Cumulative Matching Characteristic (CMC) curve, that represents the expectation of finding the correct match in the top n matches and the normalized Area Under the Curve (nAUC), which is the area under the entire CMC curve normalized over the total area of the graph. nAUC gives an overall score of how well the re-identification methods do perform.

ETHZ dataset. The data are captured from moving cameras in a crowded street by Ess et al. (2007). The challenges covered by this dataset are illumination changes, occlusions and low resolution (32×64 pixels). This dataset contains

three sub-datasets: ETHZ1 with 83 people (4.857 images), ETHZ2 with 35 people (1.936 images), and ETHZ3 contains 28 with (1.762 images). Even if this dataset does not mirrors a genuine re-identification scenario but instead a person re-acquisition scenario (no different non-overlapping cameras are employed), it still carries important challenges not exhibited by other public dataset, as the high number of images per person.

iLIDS dataset for re-id. The dataset is composed by 479 images of 119 people, normalized to size 64×128. Zheng et al. (2009) built it from the iLIDS surveillance dataset. It considers an airport arrival hall in the busy times under a multi-camera network. This dataset does not fit well in a multi-shot scenario because the number of images per person is very low (4 in average). In tracking applications, it is usually possible to accumulate a higher number of instances per person (one for each frame). For this reason, we also created a modified version of the dataset, named iLIDS$_{\geq 4}$, where we selected only the individuals with at least 4 images. In total, iLIDS$_{\geq 4}$ contains 69 individuals.

CAVIAR for re-id Dataset. CAVIAR4REID[6] has been extracted from the CAVIAR database, in particular the recordings acquired from two different cameras in an indoor shopping center in Lisbon. The pedestrians images have been cropped and isolated in proper bounding boxes (whose sizes vary from 17×39 to 72×144) using the provided ground truth. 50 different individuals are captured under both the views, where each view has 10 images for each pedestrian. Such images have been selected by maximizing the variance with respect to resolution changes, light conditions, occlusions, and pose changes.

Results. We reproduce the same multi-shot experimental settings of Farenzena et al. (2010). We randomly select a subset of N images for each person to build the gallery set, and N for each person for the probe set. Whereas a pedestrian has less than $2N$ images in total (*e.g.*, in iLIDS the individuals with just 2 images are 18.5%), we build the signatures splitting in equal proportions the images for the probe and the gallery. When just 2 images are available, the descriptor becomes single-shot. Then, the matching between (A)HPEs of the probe set and the ones of the gallery set is estimated. To have a robust statistics, this whole procedure is repeated 20 times, and the CMC curves are averaged over the trials.

There are three aspects that we investigate with our experiments: i) we test the HPE descriptor varying the number of images N for each individual, to see how important is to have multiple instances per person. ii) We compare (A)HPE

[6]Available at `http://www.re-identification.net/`

with the state-of-the-art methods for better understanding pros and cons of our proposal. iii) We perform an analysis of the weights βs in order to find out which feature is more discriminant.

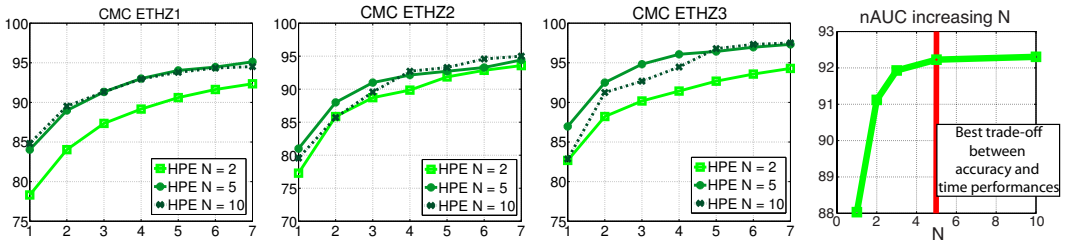

Figure 2: Evaluation on ETHZ 1,2,3 of HPE varying the number of images (first three columns). Normalized AUC averaged on ETHZ and iLIDS at increasing the number of images (last column).

First, we analyze HPE varying N. We focus on the ETHZ dataset (Fig. 2, first three columns) which has enough samples, setting $N = \{2, 5, 10\}$. Due to the nature of the datasets, the results prove that our method is robust to occlusions and quite crowded scenarios (*e.g.*, the images often contain more than a person). Moreover, the analysis of the nAUC (Fig. 2, right) shows that the accuracy increases sub-linearly with the number of images N. The trade-off between accuracy and time performances is provided by $N = 5$. We could use more images, but the computational time of the matching would increase significantly (it is quadratic in N), with a small gain in accuracy. Only ETHZ has such a number of images per person, while for iLIDS we have to set $N = 2$ as in Fig. 3. In other words, for iLIDS we cannot exploit our method at its best. In fact, learning epitomes with $N < 5$ is quite tricky because the model overfits the data and it is not able to generalize a common structure between the views. This effects is even more dramatic with $N = 1$, where performances are very low. In other words, our approach has to be intended solely as multi-shot approach for re-identification.

A comparison between different state-of-the-art methods, HPE and AHPE descriptor is shown in Fig. 3. On ETHZ, AHPE gives the best results, showing consistent improvements on ETHZ1 and ETHZ3. On ETHZ2, AHPE gives comparable results with SDALF, since the nAUC is 98.93% and 98.95% for AHPE and SDALF, respectively. Note that if we remove the image selection step (used for ETHZ), the performances decreases of 5% in terms of CMC, because the intra-variance between images of the same individual is low, and thus the multi-shot mode does not gain new discriminative information.

On iLIDS (Fig. 4, left), AHPE is outperformed only by SDALF. This witnesses again the fact, explained in the previous experiment, that the epitomic analysis

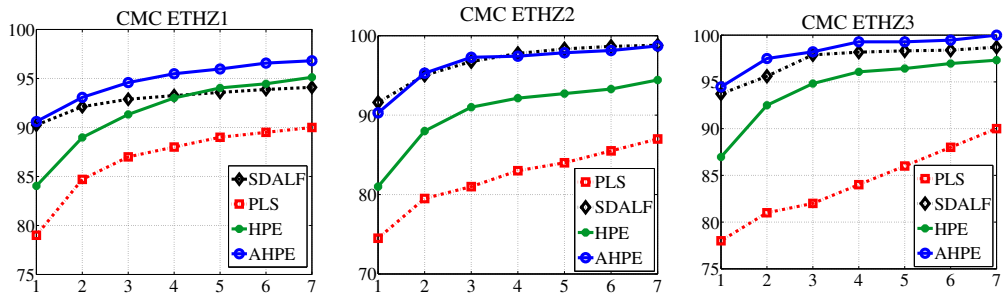

Figure 3: Comparisons on ETHZ 1,2,3 between AHPE (blue), HPE (green), SDALF by Farenzena et al. (2010) (black), PLS by Schwartz and Davis (2009) (red). For the multi-shot case we set $N = 5$.

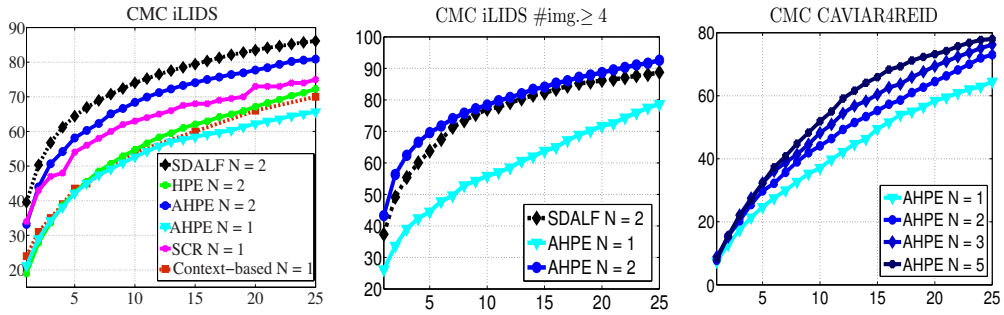

Figure 4: Comparisons on iLIDS (first column), iLIDS$_{\geq 4}$ (second column) and CAVIAR4REID (third column) between AHPE (blue), HPE (green, only iLIDS), SDALF by Farenzena et al. (2010) (black), SCR by Bak et al. (2010) (magenta, only iLIDS), and context-based by Zheng et al. (2009) (red, only iLIDS). For iLIDS and iLIDS$_{\geq 4}$ we set $N = 2$. For CAVIAR4REID, we analyze different values for N. Best viewed in colors.

works very well when the number of instances is appropriate (say, at least $N = 5$). This is also highlighted by the experiments on iLIDS$_{\geq 4}$ and CAVIAR4REID (Fig. 4, last two columns). Especially, if we remove from iLIDS the instances with less than 4 images, then AHPE outperforms SDALF (Fig. 4, center). The evaluation on CAVIAR4REID (Fig. 4, right) shows that: 1) as HPE in Fig. 2 the accuracy increases with N, and 2) the real, worst-case scenario of re-identification is still very challenging and an open problem.

The last analysis (Fig. 5) concerns the evaluation of the performances varying the weight βs of Eq. 2. This time, the quantitative analysis has been performed using the values of CMC at first position (CMC(1)). Fig. 5 shows the results for ETHZ1, ETHZ2 and iLIDS varying β_2 and β_3 (the value of β_1 can be derived by $\sum_i \beta_i = 1$). First of all, it is worth noting that if we use just the local epitome

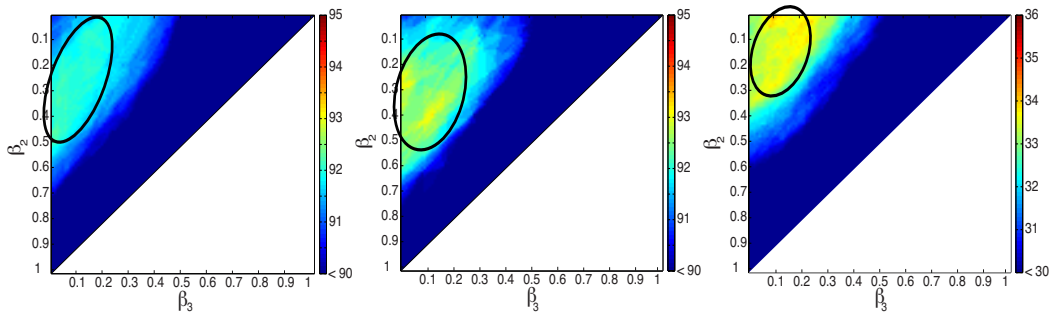

Figure 5: Analysis of the parameters $\beta_{1,2,3}$ in terms of the first rank of the CMC curve (CMC(1)). The black ellipses highlight the optimal parameters for each dataset. The main work is done by the wHSV descriptor, but using the epitome-based features in combination increase the accuracy. Note that values below 90 (for ETHZ) and 30 (for iLIDS) are all dark blue for better visualization and the right-bottom corner is white because the parameters sum to 1.

or the generic epitome the performances are not the best. Using only the color histogram (the upper-left corner) gives good performances, but again not the best. The best performance are highlighted for each dataset (with ellipses) in Fig. 5. This parameters optimization shows that there does not exist a unique set of parameters for all the dataset. Instead, we need to find a trade-off, for example, by intersecting the regions where the accuracy is good. In fact, we can notice that a good choice of the parameters is: $\beta_1 = 0.6$, $\beta_2 = 0.25$ and $\beta_3 = 0.15$. We used this parameters setting in our experiments.

4. Conclusions

In this paper, we address the person re-identification problem by proposing a novel descriptor, (A)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 (A)HPE, we set novel best performance scores on the available re-identification databases. The approach focuses on accuracy rather then efficiency, so the future work will focus on customizing it for on-line processing.

Acknowledgements. This research is funded by the European Project FP7 SAMURAI, grant FP7-SEC-2007-01 No. 217899.

References

Bak, S., Corvee, E., Bremond, F., Thonnat, M., 2010. Person re-identification using spatial covariance regions of human body parts. In: IEEE International Conference on Advances on Video and Signal Based Systems. pp. 435–440.

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

Ess, A., Leibe, B., Gool, L. V., 2007. Depth and appearance for mobile scene snalysis. In: IEEE International Conference on Computer Vision.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010. Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition.

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

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

Gray, D., Brennan, S., Tao, H., 2007. Evaluating appearance models for recongnition, reacquisition and tracking. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance.

Gray, D., Tao, H., 2008. Viewpoint invariant pedestrian recognition with an ensamble of localized features. In: European Conference on Computer Vision. Marseille, France, pp. 262–275.

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

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

Jojic, N., Caspi, Y., 2004. Capturing image structure with probabilistic index maps. In: IEEE Conference on Computer Vision and Pattern Recognition.

Jojic, N., Frey, B. J., Kannan, A., 2003. Epitomic analysis of appearance and shape. In: IEEE International Conference on Computer Vision.

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

Nakajima, C., Pontil, M., Heisele, B., Poggio, T., 2003. Full-body person recognition system. Pattern Recognition 36 (9).

Prosser, B., Zheng, W. S., Gong, S., Xiang, T., 2010. Person re-identification by support vector ranking. In: British Machine Vision Conference.

Schwartz, W., Davis, L. S., 2009. Learning discriminative appearance-based models using partial least squares. In: Brazilian Symposium on Computer Graphics and Image Processing.

Sebastian, P., Voon, Y. V., Comley, R., 2008. The effect of colour space on tracking robustness. In: Conference on Industrial Electronics and Applications. pp. 2512–2516.

Zheng, W., Gong, S., Xiang, T., 2009. Associating groups of people. In: British Machine Vision Conference.